GENOMIC VARIATIONS, PATHOGENICITY AND EVOLUTIONARY GENETICS OF *Salmonella enterica* SEROVAR TYPHI

YAP KIEN PONG

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

INSTITUTE OF BIOLOGICAL SCIENCES

FACULTY OF SCIENCE

UNIVERSITY OF MALAYA

KUALA LUMPUR

2017

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: YAP KIEN PONG

Registration /Matric No: SGR110122

Name of Degree: Master of Science (M.Sc.) (by full research)

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

Genomic Variations, Pathogenicity and Evolutionary Genetics of *Salmonella enterica* serovar Typhi

Field of Study: Molecular Microbiology

I do solemnly and sincerely declare that:

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

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

ABSTRACT

Typhoid fever is an infectious disease of global importance and is caused by the etiologic agent, *Salmonella enterica* serovar Typhi (*S.* Typhi). This disease causes an estimated of 21.7 million infections and 217,000 deaths annually. *S*. Typhi is a human-restricted pathogen, and human-carriers are the main reservoirs. Although the completed *S*. Typhi genomes are available in the public domain, the true extent of genetic variations remains obscure. Thus, this study could potentially shed light on the genetic basis of *S.* Typhi. In the first part of this study, extensive comparative genomics analyses of *S.* Typhi strains isolated from different backgrounds were carried out. A comparative genomics analysis combined with phylogenomic analyses revealed that an outbreak strain (BL196) was closely related to a human carrier strain (CR0044) and possibly derived from a common ancestor. Further comparison with other completely sequenced *S*. Typhi genomes showed that the strains BL196 and CR0044 exhibited unusual genomic variations despite *S*. Typhi being generally regarded as highly clonal. The two genomes shared distinct chromosomal architectures and uncommon genome features. Mutations in the two highly related virulence-determinant genes, *rpoS* and *tviE* were detected in strains BL196 and CR0044. These proteins were further studied using protein modelling and molecular dynamics simulations, which revealed that the mutation in *rpoS* is stabilising, while that of *tviE* is destabilising, potentially leading to altered protein functions. These microvariations between the two highly related strains with distinct epidemiological characteristics provide novel insight into the genes optimisation of the pathogen for its successful adaptation and persistence in the host. On the contrary, the sporadic strain, ST0208 was found to be far more conserved in comparison with other strains. In the second part of the study, the global MLST distribution of *S*. Typhi was described by utilizing the genome sequences from the first part of the study together with the recently released, publicly available 1,826 *S*. Typhi draft genome sequences. The global MLST analysis confirms the predominance of two sequence types (ST1 and ST2) co-existing in the endemic regions. Interestingly, *S*. Typhi strains with ST8 are currently confined to the African continent. Comparative genomic analyses of ST8 revealed unique mutations in important virulence genes such as *flhB*, *sipC*, and *tviD* that may explain the genetic variations that differentiate between seemingly successful (widespread) and unsuccessful (poor dissemination) *S*. Typhi populations. Large scale whole-genome phylogeny demonstrated evidence of phylogeographical structuring and showed that ST8 may have diverged from the earlier ancestral populations of ST1 and ST2, in which later lost some of its fitness advantages, leading to poor worldwide dissemination. This study demonstrated for the first time the utility of large-scale genome-based MLST as a quick and effective approach to narrow the scope of in-depth comparative genomic analysis and consequently provided new insights into the fine scale of pathogen evolution and population structure. In summary, this study has contributed to the understanding of genetic blueprints of *S*. Typhi associated with different clinical outcomes, particularly the less studied, yet important human carrier strain. Besides, this study has discovered a number of important key genes, novel genes, and mutations related to virulence and pathogenesis of *S*. Typhi, which could have profound implications in disease control of *S*. Typhi and other species.

ABSTRAK

Tifoid merupakan penyakit berjangkit yang penting di dunia yang disebabkan oleh agen etiogi, *Salmonella enterica* serovar Typhi (*S*. Typhi). Penyakit ini dijangka menyebabkan 21.7 juta jangkitan and 217,000 kematian setiap tahun. *S*. Typhi ialah patogen terhad manusia dan takungan utamanya ialah pembawa penyakit. Walaupun genom penuh *S*. Typhi sedia ada di domain awam, tetapi sejauh mana variasi genetik ini masih kurang jelas. Dengan itu, kajian ini berpontensi untuk memberikan pemahaman lebih mendalam mengenai genetik asas *S*. Typhi. Di bahagian pertama kajian ini, analisis perbandingan genomik yang luas bagi strain *S*. Typhi yang berkaitan dengan keadaan epidemiologi berlainan dijalankan. Analisis perbandingan genomik bersama dengan analisis filogenomik membuktikan bahawa strain berasal dari wabak demam tifoid besar (BL196) dan pembawa penyakit manusia (CR0044) adalah saling berkait rapat dengan mikrovariasi dan kemungkinannya berasal dari moyang yang sama. Selepas perbandingan dengan genom-genom penuh *S*. Typhi yang lain, kajian mendapati bahawa strain BL196 dan CR0044 mempamerkan variasi genomik yang luar biasa walaupun *S*. Typhi secara umumnya dikatakan sangat klonal. Dua genom tersebut berkongsi struktur kromosom yang berbeza dan ciri-ciri genom luar biasa. Mutasi pada dua gen-gen ketentuan kebisaan, *rpoS* and *tviE* dikesan pada strain-strain BL196 dan CR004. Protein-protein ini kemudiannya dikaji dengan menggunakan permodelan protein dan simulasi dinamik molekul dan mendedahkan mutasi di *rpoS* adalah menstabilkan, manakala di *tviE* adalah menyahstabilkan, dimana ia berpotensi membawa kepada pengubahan fungsi protein. Mikrovariasi ini di antara dua strain yang berkait rapat dengan ciri-ciri epidemiologi yang berbeza memberikan pandangan baru dalam optimisasi gen dari patogen untuk adaptasi dan kegigihan yang berjaya di dalam perumah manusia. Tetapi, strain sporadik, ST0208 didapati jauh lebih dipulihara berbanding dengan strain yang lain. Dalam bahagian kedua kajian ini, pengedaran

MLST global *S*. Typhi telah digambarkan dengan menggunakan jujukan genom dari bahagian pertama dengan 1,826 draf genom *S*. Typhi yang baru dikongsikan di domain awam. MLST global analisis mengesahkan predominan dua bentuk jujukan (ST1 and ST2) yang wujud bersama-sama di rantau endemik. Menarikya, *S*. Typhi dengan ST8 kininya terhad dalam benua Afrika. Analisis perbandingan genomik bagi ST8 mendedahkan mutasi unik di gen-gen kebisaan yang penting seperti *flhB*, *sipC* dan *tviD* yang mungkin menjelaskan variasi genetik yang membezakan populasi yang nampaknya berjaya (penyebaran luas) dengan yang tidak berjaya (penyebaran terhad). Filogenomik keseluruhan genom berskala besar memberikan bukti-bukti penstrukturan filogeografik dan menunjukkan bahawa ST8 mungkin disimpang dari populasi moyang awal ST1 dan ST2 di mana kemudiannya hilang beberapa kecergasan yang penting, membawa kepada penyebaran dunia yang terhad. Kajian ini menunjukkan penggunaan skala besar MLST sebagai pendekatan yang pantas dan berkesan untuk menyempitkan skop analisis perbandingan genomik dan seterusnya memberikan pandangan baru kepada evolusi patogen dan struktur populasi yang terperinci. Kesimpulannya, kajian ini menyumbangkan pemahaman kepada pelan genetik *S*. Typhi yang berkaitan dengan pelbagai hasil klinikal, terutamanya strain pembawa penyakit yang kurang dikaji tetapi penting. Selain itu, kajian ini menemukan beberapa gen-gen, gen baru dan mutasi penting yang berkaitan dengan kebisaan dan patogenesis *S*. Typhi yang mempunyai implikasi yang mendalam kepada kawalan penyakit *S*. Typhi dan lain-lain spesies.

ACKNOWLEDGEMENT

This thesis would not be successful without the support of my supervisors, lab mates, as well as my collaborators, colleagues, friends and family, who provided intellectual stimulation, thought-provoking conversation, ideas, motivation, and assistance at all levels. Although I mentioned a few, the contributions of many are anonymously appreciated.

Foremost, I would like to express my deepest appreciation to my thesis and research supervisor, Professor Dr. Thong Kwai Lin, for the continuous support and guidance. Her vast research experience, achievements, accolades and aptitude inspired a young scientist like me to work meticulously with unconditional efforts for the benefits of human mankind. Dr. Chai Lay Ching has been my invaluable co-supervisor, who I owed my biggest thanks for her extraordinary talents, support and dedication in my research, not to mention her immense contribution in statistical analyses. My sincere thanks must also go to my great collaborator, Dr. Gan Han Ming, a bioinformatics wizard who contributed enormously in my learning and research. He also demonstrated what a brilliant and hard-working scientist can contemplate and accomplish. The whole team which I mentioned was fantastic, and as always, they pressed me, in a positive ways, to abide by Albert Einstein’s dictum that “the whole of science is nothing more than a refinement of everyday thinking”— my manuscripts were revised and refined countless times.

I am most grateful to my collaborators from University of Hyderabad, India for lending me their expertise and intuition in the initial phase of my research: Dr. Niyaz Ahmed, Dr. Tiruvayipati Suma Avasthi, Dr. Narender Kumar, and Dr. Ramani Baddam. Several field-experts, including editors, reviewers, and thesis advisory and examination panel who read my manuscripts and offered their valuable comments toward improving my work are anonymously thanked here. Thanks to my brilliant lab mates, seniors and juniors for the stimulating discussions and technical assistance, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last few years (in the alphabetical order): Boon Hong, Hannah, Hossein, Hui Key, Jesse, King Ting (Ph.D.), Lai Kuan, Natelie, Nurul, Sarah, Shiang Chiet, Shu Yong, Solomon, Soo Ling, Soo Tein, Wing Sze, Xiu Pei and many others. Not to forget my gratitude to all the other current and former Prof. Thong's undergraduates, postgraduates students and visitors whom I know of (not mentioned). A special thanks to my senior and also my collaborator, Dr. Cindy Teh for her prompt support, care and specialty assistance that lead to the completion of my work. The administrative staff of the IPS, ISB and Genetic Department is memorable which I have especially benefited from their truly professional assistance.

As always my friends, especially the close one (you know who you are) who have made my journey and the work-life balance infinitely exciting and challenging. To this end I owe a huge debt of gratitude to the incomparable best friend and confidant of mine, Chen Young for the unconditional supports, unbounded curiosity and illumination. Leaving the best for last—my family is the greatest source of my happiness and total well-being. I thank my Papa and Mama, the greatest parents of all time for their unconditional affection and support which help me get through agonizing periods in the most positive ways. To my younger brother, who I indebted a lot for taking care of our parents while I was away and struggling. Finally, not to forget our greatest companion and the human's best friend, Wai Wai, our pet dog for keeping my parents forever young and cheerful despite difficulty and hardship in life. *Merci beaucoup*!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

BLASTn	:	Basic Local Alignment Search Tools Nucleotide
BLASTp	:	Basic Local Alignment Search Tools Protein
S. Typhi	:	*Salmonella enterica* serovar Typhi
tRNA	:	Transfer Ribonucleic Acid
tmRNA	:	Transfer-Messenger Ribonucleic Acid
rRNA	:	Ribosomal Ribonucleic Acid
DNA	:	Deoxyribonucleic Acid
PCR	:	Polymerase Chain Reaction
PFGE	:	Pulsed-field Gel Electrophoresis
MLST	:	Multilocus Sequence Typing
SNP	:	Single Nucleotide Polymorphism
nsSNP	:	Non-Synonymous Single Nucleotide Polymorphism
STs	:	Sequence Types
SPIs	:	Salmonella Pathogenicity Islands
MLNs	:	Mesenteric Lymph Nodes
TNF-α	:	Tumor Necrosis Factor-α
ERIC-PCR	:	Enterobacterial Repetitive Intergenic Consensus PCR
RAPD	:	Random Amplification of Polymorphic DNA
AFLP	:	Amplified Fragment Length Polymorphism
MLVA	:	Multiple-Locus Variable Number of Tandem Repeat Analysis
WGS	:	Whole Genome Sequencing
CRISPR	:	Clustered Regularly Interspaced Short Palindromic Repeats

LIST OF SYMBOLS AND ABBREVIATIONS

DRs	:	Direct Repeats
IS	:	Insertion Sequences
(REP)-PCR	:	Repetitive Extragenic Palindromic PCR
bp	:	Bases Pair
Mbp	:	Mega bases Pair
CDS	:	Coding Sequences
ORFs	:	Open Reading Frames
HRM	:	High-Resolution Melting
MD	:	Molecular Dynamics
HGT	:	Horizontal Gene Transfer
RM System	:	Restriction-Modification System
MRCA	:	Most Recent Common Ancestor

LIST OF APPENDICES

CHAPTER 1

GENERAL INTRODUCTION

The research work presented in this thesis primarily focuses on the comparative genomics analyses of *Salmonella enterica* serovar Typhi (*S*. Typhi) (sequenced through Next-Generation Sequencing Technology) using bioinformatics analyses coupled with wet-lab experiments. *S*. Typhi, the etiologic agent of typhoid, causes 21.7 million infections and 200,000 deaths annually throughout the world. In the recent years, the study and understanding of the pathogen's whole genome have become the hallmark in deciphering the genetic blueprints of the pathogen that have broad applications such as in disease control, vaccine development, epidemiological investigation, bacterial DNA fingerprinting, etc. Prior to this work, there were only two complete genomes of *S*. Typhi (CT18, Ty2) available in the public database. The first comparative genomic analyses were performed since the completion of Ty2 genome. However, considering the potential vast genomic variations carried by the pathogens, particularly from various endemic zones, it is crucial to conduct genomic interrogation/dissection of strains isolated from different endemic regions and clinical backgrounds. This is to understand better the contribution of the genetic variations in contributing to differential virulence capability, pathogenicity and the evolutionary relationships of strains understudied.

In the first part of this thesis, *S*. Typhi strains isolated from different clinical outcomes (outbreak, sporadic, carrier) were hypothesized to carry genetic variations that may contribute to various clinical manifestations. Further, two strains which are associated with two distinct clinical outcomes (from a massive outbreak of typhoid fever and human carrier) that shared a very similar genetic fingerprints inferred from Pulsed-field Gel Electrophoresis (PFGE) may harbor micro-evolutionary related variations that define the associated clinical outcomes. To test this hypothesis, in-depth comparative genomic analyses were performed on strains associated with different clinical outcomes and to identify genetic variations that distinguished them. The details of the analyses and findings are in Chapter 3

In the second part of the study, it was hypothesized that the identification of the uncommon sequenced types from the predominant STs globally may present an important genome model to study the genetic variations that differentiate the seemingly successful *S.* Typhi from the unsuccessful *S.* Typhi population. Comparative genomics analyses were performed on *S.* Typhi genome associated with the uncommon sequences types identified from the analyses of both experimentally and *in silico* derived genome data. The details and findings of the analyses will be discussed in Chapter 4.

With respect to the hypotheses made, the objectives of this study were

1. To determine the genome features, genomic variations, virulence, pathogenicity and evolution of *S.* Typhi associated with various clinical outcomes and uncommon sequence types via extensive comparative genomics analyses.
2. To determine the phylogenomic relationship among Malaysian *S.* Typhi strains and uncommon sequence typed-*S.* Typhi strain in relative to global *S.* Typhi strains
3. To elucidate the potential protein alteration of nsSNPs from microevolution via *in silico* protein modelling
4. To determine the sequence types of global *S.* Typhi strains by performing both experimental MLST and *in silico* MLST on locally acquired as well as database derived global strains.

With the objectives above, it serves to answer the following research questions and the gap of knowledge;

1. Which of the genes in the genomes are conserved or shared (core genes) and which accessory genetic signatures that may give rise to differential phenotypes as observed?
2. The whole-genome based phylogenetic relationship of the studied genomes in relative to all other available genomes globally
3. The elucidation of the unique and shared genomic features carried by both highly related outbreak associated strain and human carrier strain.
4. The discovery of unique mutations (nsSNPs) and the prediction of the effects on protein alteration together with the protein modelling carried either by the outbreak strain or human carrier in relative to other genomes.
5. The most updated MLST prevalence of global *S*. Typhi and the genetic variations associated with uncommon STs.
6. The discovery of genes linked to genomes associated with the "less successful trait"

CHAPTER 2

LITERATURE REVIEW

2.1 *Salmonella enterica*

Salmonella serovars are predominately pathogenic Enterobacteriaceae which are thought to have evolved from a common ancestor with *Escherichia coli* (*E.coli*) ~100 million years ago (Doolittle, Feng, Tsang, Cho, & Little, 1996). The genus *Salmonella* consists of *Salmonella enterica* (*S. enterica*) and *Salmonella bongori* (*S. bongori*). Unlike *S. bongori* which has been found predominantly associated with cold-blooded animals (i.e., reptiles), *Salmonella enterica,* which comprised of six subspecies; *enterica*, *salamae, arizonae*, *diarizonae*, *houtenae,* and *indica*, infect mainly warm-blooded animals. *S. enterica* subspecies are further subdivided into more than 2,500 serovars (Crosa, Brenner, Ewing, & Falkow, 1973). The characterization of these serovars is based on their surface antigens, the O (somatic) the two H (flagellar) (Le Minor, Veron, & Popoff, 1982). The O antigens are part of the variable long-chain lipopolysaccharide located on the outer part when the two flagellar structures are expressed. Among *S. enterica*, some serovars such as *S*. Typhimurium, are adapted to a broad host range while *S*. Typhi and *S*. Paratyphi are human-restricted without other known natural hosts (Kauffmann, 1966).

2.2 Typhoid and *Salmonella* Typhi

Typhoid is caused by *S*. Typhi, a highly adapted human-specific pathogen. *S*. Typhi is clinically important and responsible in causing febrile illnesses and death in human population living in poorly sanitized and crowded environment. The risk of contracting the disease is higher in a population exposed to contaminated food and water. It also poses a high risk to travelers visiting endemic countries (Crump & Mintz, 2010). *S*. Typhi is a rod-shaped (bacilli) Gram-negative, flagellated, non-sporulating, encapsulated, and facultative anaerobic bacterium with diameter and length of 0.4-0.6 µm and 2-3 µm, respectively. *S*. Typhi is serologically positive for lipopolysaccharide antigens O9 and O12, protein flagellar antigen type H;*d*, and polysaccharides capsular antigen Vi (Parry *et al*., 2002). The Vi capsular antigen is almost unique to *S*. Typhi, although this antigen has also been detected in some strains of *S. enterica* Paratyphi C and Dublin (Chowdhury, Shumy, & Anam, 2014). *Salmonella* spp. nomenclature is complex, and scientists use different systems to refer and communicate about this genus. However, it was later referred to as *Salmonella enterica* subsp. enterica serotype (serovar) Typhi (Brenner, Villar, Angulo, Tauxe, & Swaminathan, 2000) and this standard convention has been used since then.

2.3 Global Epidemiology of Typhoid

The morbidity and mortality of typhoid are high, particularly in poor and resourced-limited endemic zones. According to the latest global estimates, there are approximately 20-30 million cases and 217, 000 deaths resulted from typhoid fever globally each year, predominantly in children and young adults (Crump & Mintz, 2010). The global burden of typhoid may be underestimated particularly in rural areas due to unreported cases and lacks of resources for proper diagnosis and clinical differentiation. Unlike the developed countries where typhoid is controlled by good sanitation practice, developing countries are usually suffering from poor sanitation, poverty, and contaminated food and water. The regions with the highest incidence (>100 cases /100,000 population per year) of typhoid include South Asia and Southeast Asia. Regions of medium incidence (10-100 cases/100,000 population per year) include mainly the rest of Asia, Africa, Latin America and the Caribbean, and Oceania, except for Australia and New Zealand. On the other hand, the rest of the developed world have low incidence of typhoid (<10 cases/100,000 population per year), which due in large part to travelers returning from endemic countries (Buckle, Walker, & Black, 2012; Crump & Mintz, 2010; Crump *et al*., 2004; Ochiai *et al*., 2008). The current trends of urbanization, migrations, travels and trades have tremendously increased the risk of transmission from endemic areas. More alarming, the increasing resistance to major antibiotics has severely impacted the infection control (Dougan & Baker, 2014). At present, the World Health Organization (WHO) (WHO, 2012) urges typhoid vaccination for persons living in or travelling to endemic areas. However, the current typhoid vaccine is of limited use as the existing typhoid vaccine lacks effectiveness in children (Rahman, Hosen, & Chakraborty, 2013).

2.4 Virulence and Pathogenesis

The infectious dose required for infection in patients could vary between 1000 to 1 million organisms; these numbers may, in fact, lower than expected as pointed out in recent studies (Glynn, Hornick, Levine, & Bradley, 1995; Waddington *et al*., 2014). It has been shown that Vi-negative strains of *S*. Typhi are less virulent and less infectious compared to Vi-positive strains (WHO, 2010; Parry *et al*., 2002). For *S*. Typhi to reach the small intestine, the pathogen must survive the gastric acid barrier, which is the first line of defense mechanism in humans. It has been reported that individuals with previous gastrectomy, or treatment with proton-pump inhibitors or achlorhydria from aging may lower the infective doses (Chowdhury *et al*., 2014; Hornick *et al*., 1970). Upon reaching the small intestine, the bacteria attach themselves to mucosal cells followed by the invasion of the mucosa layers. The pathogens are capable of penetrating the mucosal epithelium via microfold cell (M cells) or enterocytes to arrive the lamina propria, where they trigger the influx of macrophages to ingest them but do not generally kill them (Dougan & Baker, 2014; Hornick *et al*., 1970; Rahman *et al*., 2013) (Figure 2.1). Although *S*. Typhi remains within the macrophage of the intestinal lymphoid tissues, some are translocated to intestinal lymphoid follicles and the mesenteric lymph nodes in which they enter the thoracic duct and the general blood circulation (Jones & Falkow, 1996). Hence, *S*. Typhi is highly invasive but not necessarily trigger rapid immune and diarrheal responses.

The pathogen reaches the intracellular spaces within 24 hours after ingestion through the aforementioned silent primary bacteraemia. *S*. Typhi can survive and multiply within the mononuclear phagocytic cells of the lymphoid follicles, liver, spleen and bone marrow. The incubation period is expected to be around 7 to 14 days (Parry *et al*., 2002). However, the incubation period varies depending on the infective dosage, their virulence, and the host responses. The clinical manifestation is usually accompanied by a fairly sustained low level of secondary bacteraemia (One bacterium per millilitre of blood and approximately ten bacteria per millilitre of bone marrow). Although typhoid induces systemic and local humoral and cellular responses, but it usually confers incomplete protection against relapse or/and reinfection (Rahman *et al*., 2013). Notably, the interaction between the host immunologic mediators and bacterial factors in infected tissues may contribute to the necrosis of Peyer’s patches in severe disease stages (Gilman, Terminel, Levine, Hernandez-Mendoza, & Hornick, 1975; Marks *et al*., 2010).

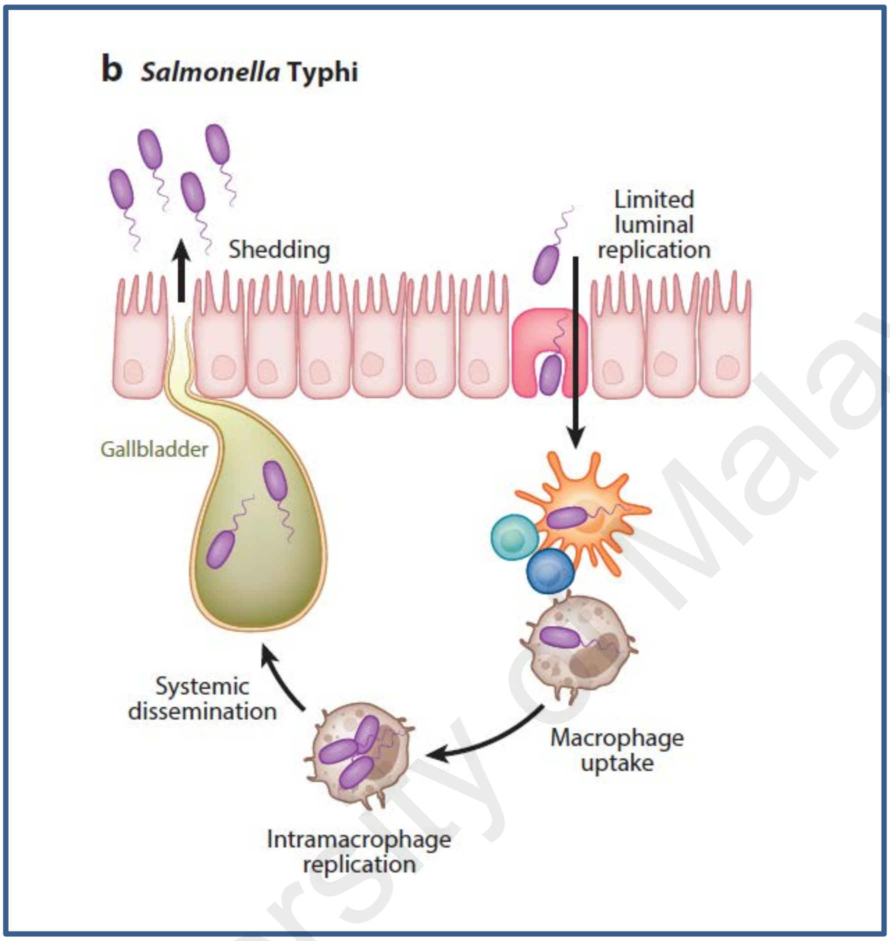

Figure 2.1: A generalized infection route diagram of typhoid (*S*. Typhi) (b). Vi expression and other factors such as effector repertoires may influence inflammatory responses to *S*. Typhi (Dougan & Baker, 2014)

.

2.5 Molecular Pathogenicity

S. Typhi harbors many virulence factors, such as genes encoding adhesion, invasion and toxin, which are clustered together in the chromosome and hence, termed "Salmonella Pathogenicity Islands" (SPI). These arrays of genes are located either on the chromosome or a plasmid and are usually flanked by repeat sequences. This region is also known to exhibit deviated GC contents in relative to its neighboring regions

(Pickard *et al*., 2003; Sabbagh, Forest, Lepage, Leclerc, & Daigle, 2010). SPIs have varied base composition from the core genome and are often associated with tRNA genes and mobile genetic elements such as *IS* elements, transposons and phage genes (Siriken, 2013). SPIs are known to encode for genes that produce and excrete invasion protein that facilitates uptake of the bacterium by non-phagocytic cells, and in turns, the pathogens are able to live intracellularly. In addition, genes in SPIs can inhibit oxidative burst of leucocyte, renders the inmate immune response ineffective (den Bakker *et al*., 2011; Dougan & Baker, 2014; Sabbagh *et al*., 2010). For example, *S*. Typhi harbors a large SPI-7 that encodes the Via locus, a repertoire of genes that involve in Vi capsule expression and host cell attachment (Sabbagh *et al*., 2010). *S*. Typhi also carries a collection of conserved phages and SNPs/Indels that are distinct from other *Salmonella* serovars (Jacobsen, Hendriksen, Aaresturp, Ussery, & Friis, 2011; Thomson *et al*., 2004).

2.6 Host specificity and Adaptation

During the persistence typhoid infection, the entry of bacteria into neutrophils and macrophages trigger inflammation and recruitment of T and B cells (Wyant, Tanner, & Sztein, 1999). In most cases, dendritic cells and/or macrophages are targeted as these cells favor dissemination through the blood circulatory system to the mesenteric lymph nodes (MLNs) and deeper tissues (Rescigno *et al*., 2001; Swart & Hensel, 2012). This leads to the following transport of the pathogens to the spleen, bone marrow, liver and gall bladder, in which the pathogen can persist for life, and periodic reseeding of the mucosal surface via the bile ducts and MLNs. It has been documented that the interferon (IFN), which is secreted by the T cells, play a key role in maintaining cell persistence by controlling the intracellular replication (Jones & Falkow, 1996). Also, the interleukin which can increase the IFN productions and *Salmonella* pro-inflammatory cytokine tumor necrosis factor-α (TNF-α) may also contribute to the persistence of pathogens (Raffatellu *et al*., 2006; Wyant *et al*., 1999).

2.7 Human carriers of typhoid

S. Typhi has no known environmental reservoir. Hence, the asymptomatic carrier state is thought to be the key reservoir of continued maintenance of the bacterium within human populations and serves as the link of the unbroken chain of transmission (Gonzalez-escobedo, Marshall, & Gunn, 2011). Despite the important role of the human carriage in maintaining typhoid in the human population, the understanding of the molecular mechanisms that supports the carriage remains unclear. It is thought that about 3-5% of recovered individuals, which termed as human carriers will develop a chronic infection of the gall bladder and may shed the bacteria through the faeces for decades (Gonzalez-escobedo *et al*., 2011; Gopinath, Carden, & Monack, 2004). They are highly contagious and are typically asymptomatic, making the identification of carriers highly challenging. The identification is further exacerbated by the fact that about 25% of carriers show no clinical manifestation even during the acute phase of the disease (Gunn *et al*., 2014). Various epidemiological studies from the endemic regions established some strong evidence of association between the chronic state of the human carrier and the presence of gallstones (Gonzalez-escobedo *et al*., 2011). Hence, the *S*. Typhi associated with typhoid carrier is an elusive population to investigate, complicated by the unknown factors contributing to its carriage-state. More startling, human carriers with or without gallstones were reported to carry predisposing factor for the development of gallbladder cancer (Dutta, Garg, Kumar, & Tandon, 2000; Shukla, Singh, Pandey, Upadhyay, & Nath, 2000). Recently, a systematic review with meta-analysis unequivocally indicated that the chronic *S*. Typhi carrier state is an important risk factor among patients with carcinoma of the gallbladder (Nagaraja & Eslick, 2014a; Nagaraja & Eslick, 2014b). It is estimated that chronic carriers may have 3-200 times higher risk of developing gallbladder cancer than non-carriers.

2.8 Molecular Subtyping of *S*. Typhi

The epidemiological studies of pathogens are of great importance for disease control such as during a disease outbreak to identify individual cases from related ones. Hence, the capability to subtype strains is paramount in epidemiological investigations. In the past few decades, various molecular typing methods have been developed for *S*. Typhi. These include the classical methods of phage typing and isoenzyme analyses (Olsen, Skov, Angen, Threlfall, & Bisgaard, 1997). Most of these conventional methods have been gradually replaced by newer molecular methods such as pulsed-field gel electrophoresis (PFGE) (Thong, Cordano, & Yassin, 1996a; Thong, Goh, Yasin, & Lau, 2002; Thong *et al*., 1995; Thong, Passey, Clegg, & Combs, 1996b), ribotyping (Navarro *et al*., 1996), multiplex-PCR-based VNTR profiling (Liu *et al*., 2003), Eric-PCR (Nath, Maurya, & Gulati, 2010), random amplification of polymorphic DNA (RAPD) (Nath *et al*., 2010), mobile genetic element *IS200* DNA fingerprinting (Rowe & Gibert, 1994), amplified fragment length polymorphism (AFLP) (Lin, Usera, Barrett, & Goldsby, 1996), and Multiple-Locus Variable number of tandem repeat Analysis (MLVA) (Chiou *et al*., 2013; Tien, Ushijima, Mizuguchi, Liang, & Chiou, 2012). Despite the availability of the aforementioned methods, there is still a lack of “gold standard” methods with the perfect combination of rapidity, ease of use, reproducibility, inter- and intra-lab comparability and discriminatory power for typing *S*. Typhi. The most apparent drawback of these methods is the subjective interpretation of banding patterns and repeat numbers. Due to the recent expansion of DNA sequencing technology, multilocus sequence typing (MLST), a sequenced based approach has become a widely accepted method for molecular subtyping and population study (Urwin & Maiden, 2003). This method allows discrete characterization of isolates by using the internal fragments of the housekeeping genes sequences. Recently, in the advent of robust high-throughput whole genome sequencing technology (WGS), bacterial subtyping has been rapidly shifted from traditional subtyping into whole genome-scale typing such as using the single nucleotide polymorphism (SNP) of the genomes *(*Chen *et al*., 2013; Didelot, Bowden, Wilson, Peto, & Crook, 2012; Pallen, Loman, & Penn, 2010; Sherry *et al*., 2013; Underwood *et al*., 2013). Although WGS typing is the best available method to date but the cost of sequencing for a large number of strains and the ease of use are still less practical for routine subtyping and surveillance purposes, particularly in the developing countries where resources remains limited and cost of WGS is still prohibitive.

2.9 Whole Genome Sequencing of *S*. Typhi

The availability of genome sequences provides a genetic blueprint in which scientists could understand better the genetic basis of the pathogen and their hosts. The WGS era of *S*. Typhi was jumpstarted with the release of the first complete genome of *S*. Typhi CT18, a multidrug resistant strain which was isolated from a 9-year-old girl in Vietnam (Parkhill *et al*., 2001). This was followed by the publication of a second complete genome, Ty2 isolated in the early 1970s from Russia, which marked the first comparative genomics analysis between two *S*. Typhi genomes (Deng *et al*., 2003). With increasing affordability to perform WGS, more genomes have been sequenced and studied, including those strains with clinical values. In 2008, a joint research group led by The Wellcome Trust Sanger Institute (UK), University of Cambridge and University of Oxford have deposited 19 *S*. Typhi draft genomes in the database, all sequenced using two different platforms (Roche and Illumina) (Holt *et al*., 2008).

However, there was a lack of representative *S*. Typhi genomes originated from Southeast Asian endemic zone despite the fact that prevalence of typhoid in this region remains alarmingly high. Therefore, *S*. Typhi genomes derived from this region are anticipated (considering the endemicity of this part of the world) to better understand the true genomic variations of *S*. Typhi obtained from various geographical regions. As so much of data can be derived from sequencing a bacterial isolate, it is becoming ever important to obtain accurate strain provenance, particularly with respects to regions and clinical outcomes for meaningful biological interpretations.

2.10 Comparative genomics of *S*. Typhi

2.10.1 Core and dispensable genomes

Comparison of the genomes of several strains identifies the common set of genes called "core genome" that are shared and regarded as genes that perform "housekeeping" functions such as of those involved in central metabolism, structural biosynthesis, intestinal colonization, and transmission. Core genes are commonly conserved in the same order along the genome, a feature which referred as "synteny". Accompanying the core genome, blocks of genes, single gene or gene remnants are scattered throughout the genome (Anjum, Marooney, Fookes, & Baker, 2005; Welch *et al*., 2002). These genes usually have limited or no homology with the core genome and exhibit significant divergence among the member of the species. However, these genes could be novel and share some related function (Baker & Dougan, 2007). For example, they might work together to enhance the virulence or involve in the pathogenesis of the pathogen. Such array of genes are commonly found in *Salmonella* spp. and referred to as SPIs. For instance, SPI-1 and SPI-2 are two common SPIs found in *Salmonella* spp., which involve invasion and the ability to survive in the host (Jôrg Hacker & Kaper, 2000). SPIs are often have deviated GC contents that differ from that of the core genome, suggesting their recent integration, most likely acquired through horizontal gene transfer events (Hacker, Blum-Oehler, Mühldorfer, & Tschäpe, 1997). One of the most remarkable characteristics of *S*. Typhi is the ability to synthesize the Vi-polysaccharides capsule, in which its production is linked to a set of genes located in SPI-7. This island carries large numbers of virulence and virulence-like genes clusters, including genes encoding the Vi locus, *sopE* effector protein, type IV pilus and secretion systems with typical characteristics of horizontally acquired DNA (Nair, Alokam, Kothapalli, & Porwollik, 2004; Pickard *et al*., 2003). The previous comparison between the CT18 and Ty2 has shown a remarkable degree of conservation with minor differences particularly the additional cluster of genes (Baker & Dougan, 2007; Deng *et al*., 2003). However, this comparison yielded more questions than answers as to what extend the genome diversity of *S*. Typhi around the world are with respect to its genetic variations.

2.10.2 Evolution and selection pressures of *S*. Typhi

Holt *et al.* (2008) initiated the evolutionary study of *S*. Typhi with the use of high-throughput whole genome sequencing (Illumina and Roche) of 19 strains, isolated from 10 different countries. The study demonstrated the first use of the WGS in a large-scale study, replacing older methods such as the high-throughput genotyping (Baker *et al*., 2008) and PFGE (Thong, Cheong, & Puthucheary, 1994). Most of the enteric pathogens are under intense selection pressure due to factors including normal flora, nutrient sources, host immune system, etc. The use of WGS successfully captured the fine variations exist among strains of *S*. Typhi, which is generally thought as highly clonal with limited genetic variations. The estimation of evolutionary mean dN/dS of *S*. Typhi in comparison with the last common ancestor has indicated a weak trend towards stabilizing selection since the common ancestor (Holt *et al*., 2008). The SNPs analyses of the study showed little evidence of recombination, diversifying selection and antigenic variations among strains, suggested that the *S*. Typhi is under a relatively little selective pressure from its human host, in agreement with the existence of long-term carriage in the human carrier.

2.10.3 Genomic organization and arrangement

Typhoid fever is sometimes associated with unusual clinical manifestation, such as neuropsychiatric complications and varying disease severity, and such differences may be attributed to its varied genetic contents (Ali, Kamili, Shah, Koul, & Aziz, 1992; Thong *et al*., 1996). Many earlier studies, which tapped into larger sets of strains, demonstrated that *S*. Typhi is more diverse, evidenced by the variations in genome sizes (Thong, Puthucheary, & Pang, 1997), genomic rearrangement (Liu, Sanderson, & Anderson, 1996) and genetic contents (Boyd, Porwollik, Blackmer, Mcclelland, & Icrobiol, 2003). However, it remains obscure in terms of the extent of these variations in the gene pool, or perhaps the variants are gradually displaced by the seemingly more successful and dominants strains as a result of selective pressures. These notion were previously demonstrated by the observation of genetic recombination/rearrangement in between CT18 and Ty2. The half-genome inter-replicore inversion that spans the origin of replications has been implicated in the Ty2 genome. The inversion that lies within rRNA operons (located at rrnG and rrnH) is apparently mediated by the homologous recombination and may result in distinguishable ribotypes that are closely linked to host adaptation. Besides, minor inversion was also detected in Ty2 in which the small inverted region of the genome was translocated, yet these effects of the inversion remain speculative (Deng *et al*., 2003).

2.10.4 Pseudogenization and horizontal gene transfer

Since the pathogen is host-specific, genome optimization for host adaptation is not surprising. This optimization has been actively driven by genome degradation through the process of pseudogenization of metabolic genes (Holt *et al*., 2009; Parkhill *et al*., 2001). The annotation of CT18 yielded an approximate 200 pseudogenes in which these genes are either disrupted or inactivated. Interestingly, comparison with *S*. Typhimurium LT2 genome showed that most of the inactivated genes in *S*. Typhi remain intact and functional in LT2 (generalist), suggesting that *S*. Typhi may have lost some of the genes required for the infection of wide host range *(*McClelland *et al*., 2001; Sabbagh *et al*., 2010).

On the contrary, *S*. Typhi and *S*. Paratyphi shared a number of pseudogenes and some of it was inactivated by identical mutations, suggesting a common evolutionary origin (Holt *et al*., 2009). Pseudogenes formation seems to vary within *S*. Typhi; comparison of Ty2 and CT18 yielded considerable numbers of variations, ranging from single point mutations to extensive deletions/insertions (Deng *et al*., 2003). However, it is difficult to determine as to how the differences in pseudogene contents affect the functions of the protein, attributable to the incomplete knowledge of the structural-functional relationship, particularly in the field of pseudogenization in the current bodies of knowledge.

2.10.5 Insertion sequences, phages and CRISPRs

Bacteriophages are highly abundant and demonstrated a high degree of host specificity. Phages play important roles in bacterial evolution, such as being the vehicles of horizontal gene transfer (Moreno *et al*., 2013; Sabbagh *et al*., 2010). Several *Salmonella* phages and prophages have been reported (e.g., Fels-1, Gifsy-2, P22, FelixO1), and some of these phages have been sequenced and characterized (Switt *et al*., 2013; Pang *et al*., 2013). While the phage diversity of *Salmonella* is generally well studied, our knowledge on *S*. Typhi phage genomic diversity is still rather limited especially with respect to *S*. Typhi that was isolated from various endemic zones. The earlier comparison of CT18 and Ty2 showed similar but non-identical phages region, in which each strain carried a unique set of phages (mainly associated with prophages). But the Gifsy-1 phage which was identified in *S*. Typhimurium LT2 were absent in both CT18 and Ty2, and only partial region of the Gifsy-2 was identified (McClelland *et al*., 2001). Similarly, the previously identified LT2 phages Fels-1 and Fels-2 are neither present in Ty2 nor CT18 (Deng *et al*., 2003; Parkhill *et al*., 2001). It was suggested that the modular nature of prophage genomes may have important implication to strain variations. Repeated sequences such as transposable elements (insertion sequences) are common in prokaryotes. In the earlier classical studies, these sequences may be considered selfish in nature, having no function other than self-perpetuation (Brunet & Doolittle, 2015; Doolittle & Sapienza, 1980; Orgel & Crick, 1980). In spite of that, there is increasing evidence that they may involve in conferring an evolutionary advantage to the carrier. The *IS*200, for instance, is relatively common in *Salmonella* spp. and have been used as a target for bacterial subtyping. Similarly, *S*. Typhi also carries a relatively abundance of *IS* (Gibert, Barbé, & Casadesús, 1990; Lam & Roth, 1983; Rowe & Gibert, 1994). This has been evidenced by a large number of *IS* carried by Ty2 and CT18. Ty2 has relatively fewer insertion sequences (IS) compared to CT18 (Deng *et al.*, 2003; Parkhill *et al.*, 2001). Notwithstanding, currently, *IS* of *S*. Typhi cannot be reliably detected using short reads genome sequencing approach (Holt *et al*., 2008).

Jansen *et al*. (2002) identified a new family of repeated DNA sequences, namely CRISPRs (clustered regularly interspaced short palindromic repeats) in many prokaryotes. The CRISPRs is characterized by DNA direct repeats (DRs) of ~24-47 bp, separated by 21-72 bp varied sequences called "spacers." A "leader sequence" and Cas (CRISPRs-associated sequence) genes are often identified adjacent to the CRISPRs locus (Fabre, Zhang, Guigon, Hello, & Guibert, 2012; Jansen, Embden, Gaastra, & Schouls, 2002; Sorek, Kunin, & Hugenholtz, 2008; Touchon & Rocha, 2010). The presence of these short sequences in the CRISPRs cluster highlights the evolutionary history of multiple viral infections in the host, providing the evidence of rapid evolution of this dynamic program (Fabre *et al*., 2012; Fricke *et al*., 2011). Several studies have reported the presence of two CRISPR loci in *Salmonella* spp. (Fabre *et al*., 2012; Fricke *et al*., 2011; Liu *et al*., 2011). Despite that discovery, the genetic program of CRISPRs is not fully understood in *S*. Typhi, especially in respect to its genetic organization and functions.

2.10.6 Genome-based phylogenetic analysis

To gauge the diversity of bacterial species such as *Salmonella enterica*, a simpler method such as MLST can be utilized to subtype the strains into relevant sequence types. Some serovars are polyphyletic, following the distinct evolutionary lineages (Achtman, 2012; Pérez-Losada, Cabezas, Castro-Nallar, & Crandall, 2013). However, *S*. Typhi was shown to exhibit a high degree of conservation with limited MLST sequence types. Various studies suggest that *S*. Typhi strains are clonally related and have evolved from a relatively recent common ancestor, estimated around 50,000 years ago (Kidgell *et al*., 2002). Since the *S*. Typhi is highly related and accrues very limited variations, it is difficult to discern genetic differences among strains and it is even more challenging to determine the evolutionary lineages within the population (Kidgell *et al*., 2002; Parkhill & Wren, 2011). Classically, there are few relatively simple methods, such as PFGE and MLVA, which are useful in differentiating strains but these methods have limitations such as poor inter- and intra-laboratory comparison and have limited potential in defining evolutionary lineages, much less if the purpose is to understand the base-to-base differences at the molecular level (Hopkins, Maguire, Best, Liebana, & Threlfall, 2007; Hopkins, Peters, de Pinna, & Wain, 2011; van Belkum *et al*., 2007; Werner, Klare, & Witte, 2007). In recent decades, genome-based phylogenetic analysis has been widely used to study phylogenetic relationships of large numbers of strains (Didelot *et al*., 2012; Holt *et al*., 2008; Wong *et al*., 2015). In *S*. Typhi, the phylogenomic analysis was proven valuable and superior to conventional methods for unravelling the genetic variations among strains (Holt *et al*., 2008; Parkhill & Wren, 2011). Besides, phylogenomic analysis has been employed effectively to determine the genetic relatedness of epidemiologically related and clinically relevant strains (Chen *et al*., 2013; Didelot *et al*., 2012; Sherry *et al*., 2013).

CHAPTER 3

COMPARATIVE GENOMICS OF *Salmonella enterica* SEROVAR TYPHI STRAINS ASSOCIATED WITH DISTINCT CLINICAL OUTCOMES

3.0 Introduction

Typhoid fever is a human systemic infection that is caused by *Salmonella enterica* serovar Typhi (*S*. Typhi). This human-restricted and highly adapted pathogen is transmitted via the oral-faecal route. *S*. Typhi is responsible for 21.7 million infections and results in approximately 217,000 deaths worldwide annually (Crump & Mintz, 2010). The disease primarily causes acute systemic infection with life-threatening complications, and the recovering patient may develop into a chronic carrier state (Gonzalez-escobedo *et al*., 2011). The risk of developing gallbladder diseases, including carcinoma, is also higher among typhoid carriers (Crawford *et al*., 2010; Gonzalez-escobedo *et al*., 2011; Shukla *et al*., 2000).

Typhoid is endemic, with periodic outbreaks and sporadic cases occur in developing countries, particularly in Southeast Asia, South-central Asia, Latin America and Southern Africa where sanitary conditions are poor (Crump & Mintz, 2010). Among the 13 states of Malaysia, Kelantan has a significantly higher incidence of typhoid fever (MOH, 2006). A large typhoid outbreak occurred in Kelantan state, which resulted in 735 cases of infection and two deaths in a short period of three months (from April to June of 2005) (MOH, 2006). The representative clonal strain of this outbreak, *S*. Typhi BL196 was isolated from the blood samples of a severe typhoid case in Kelantan during the year 2005 notorious outbreak. In the year 2007, an *S*. Typhi strain CR0044 was isolated from a stool sample of an asymptomatic carrier in

Kelantan, Malaysia. This strain was highly similar to the outbreak strain BL196 by pulsed-field gel electrophoresis (PFGE) (Appendix B).

Human carriers are the main reservoirs of *S*. Typhi transmission, but the genetic basis, and the underlying mechanisms, in particular, are unclear (Holt *et al*., 2008). It has been suggested that a carrier strain will likely lack gene acquisition capabilities and have little fitness advantages compared with those strains causing symptomatic infections because the human reservoir is small and physiologically isolated (Holt *et al*., 2008; Vaishnavi *et al*., 2005). Although typhoid fever is endemic in many countries, including Malaysia, little is known about the mechanism of survival and persistence of *S*. Typhi in the host. Therefore, it is of great interest to know as to what extent these closely related strains differed or shared in its genomic contents despite being isolated from these two distinct epidemiological settings. It was hypothesized that the genome sequences of the underlying strains would provide more insights to enhance understanding of endemicity or persistence of typhoid in Kelantan. On the contrary, an *S*. Typhi strain, ST0208 was isolated from the stool sample of a typhoid fever patient admitted to University Malaya Medical Centre (UMMC), Kuala Lumpur, Malaysia, in 2008. The strain was characterized by pulsed-field gel electrophoresis (PFGE), repetitive extragenic palindromic (REP)-PCR, and antimicrobial susceptibility profiling (Tiong *et al*., 2010). PFGE analyses showed that this strain is more distantly related to the outbreak and carrier strains, but the epidemiological link is unknown. The genome sequence of this sporadically associated strain will provide insights into possible genetic events that would confer a fitness advantage.

Recent whole-genome sequencing of *S*. Typhi has demonstrated that the pathogen shows limited genetic variation with little evidence of purifying selection, antigenic variation or recombination between isolates (Achtman, 2008; Holt *et al*., 2008). This clonal pathogen, however, is associated with varying degrees of disease severity in different regions (Parry *et al*., 2002). Previous PFGE studies have also demonstrated genome size variations and distinct PFGE patterns in relation to fatal and non-fatal typhoid cases (Thong, *et al*., 1996; Thong *et al*., 1997). Although the health conditions of the host cannot be completely ruled out, various reports have suggested that the gain and loss of genes through mutations and gene transfers that have occurred independently in different lineages have markedly contributed to the varying pathogenic potentials (Fraser-liggett, 2005; Holt *et al*., 2008). However, these important factors are poorly understood because there is limited genomic information for *S*. Typhi, mainly involving strains that are associated with diverse epidemiological settings. Its genomic heterogeneity is likely due to the adaptation of the pathogen to the host and its exposure to mobile elements, such as bacteriophages (Dobrindt & Hacker, 2001). Organisms having common core genomes could differ in their dispensable (strain-specific) genes, reflecting their unique physiological and virulence properties (Ahmed, Dobrindt, Mayne, Wright, & Cartmel, 2008; Wren, 2000). Although not all genetic variations are essential for adaptation, some dispensable genes are believed to be responsible for conferring fitness advantages to the pathogen to thrive in its host. Horizontal gene transfer is also thought to be the predominant force in bacterial evolution, which contributes to novel gene acquisition. The acquired genes provide new characteristics, which either aid in host adaptation and persistence or enhance virulence capabilities (Dobrindt & Hacker, 2001; Lawrence, Ochman, & Ragan, 2002; Ahmed *et al*., 2008; Wren, 2000). A previous study on the pan-genome of *Salmonella enterica* revealed that the pan-genome (total known genomic content) of all strains will continue to increase as new genomes are sequenced (Medini, Donati, Tettelin, Masignani, & Rappuoli, 2005). With the availability of robust next-generation sequencing technologies, high-quality whole genome sequences can be generated and analysed, which will be especially useful for capturing fine variations among highly conserved *S*.

Typhi strains. Multiple whole-genome sequence comparisons of closely related strains will not only lead to the better understanding of their relationships but also provide novel insights into the functional roles of strain-specific genes.

In this study, detailed and comprehensive comparative functional analyses were performed on three sequenced genomes of *S*. Typhi strains that were isolated from typhoid patients during a massive outbreak in 2005, a sporadic case in 2008 and an asymptomatic carrier from Malaysia, where typhoid is endemic. These Malaysian *S*. Typhi strains were compared with previously published *S*. Typhi genomes with the following aims: 1) to determine and describe the genomes signatures and conserved and unique regions of the strains that were being studied; 2) to elucidate the phylogeny and genetic relatedness of these Malaysian strains compared with 17 other published strains using phylogenomic analysis; 3) to compare those strains that have been associated with various epidemiological settings (outbreak, carrier, and sporadic cases), and particularly the regions of plasticity that may contribute to the varying pathogenic potentials; 4) to identify potential novel pathogenic factors that are harboured by the analysed strains; 5) to provide insight into the possible differential functionalities of the genes, focusing mainly on virulence- and persistence-related genes based on non-synonymous SNPs of closely related strains and particularly on carrier strains to gain insight into the persistence of the carrier state; and 6) to understand how the potential nsSNPs affect protein structures and functions. The data that are generated will be useful for the profiling of strains, marker development and the increased understanding of outbreak, sporadic and less studied asymptomatic typhoid carriage infection.

3.1 Material and Methods

3.1.1 Choice of strains

Three Malaysian *S*. Typhi strains (BL196, ST0208, and CR0044) were selected for the comparative genomic analyses that were based on previous PFGE data. These strains are associated with diverse epidemiological settings. Strain BL196 was isolated from a typhoid patient with diarrhoea during a large outbreak in Kelantan, Malaysia that resulted in 735 cases and two deaths in the year 2005. Strain ST0208 was isolated from a typhoid patient, who was a sporadic case, at a local tertiary hospital in Kuala Lumpur, Malaysia. Strain CR0044 was isolated in 2007 from a carrier (food handler) following the large 2005 outbreak in Kelantan, Malaysia (Table 3.1). The initial molecular analysis showed that both the CR0044 and BL196 strains were highly similar with only one band difference as revealed by PFGE (Appendix B). These three new genomes were compared with all three of the available *S*. Typhi full genomes at the time of analysis (CT18, Ty2, and P-stx-12). CT18 and Ty2, the former is a fairly recent and geographically related multidrug-resistant strain that was isolated from Vietnam, the latter was isolated from Russia in the early 1970s, a geographically more distant strain known to be used for oral typhoid vaccine development. A comparison was also made using a carrier associated strain, P-stx-12, which was isolated from a carrier in India recently. All of these strains represent *S*. Typhi from diverse temporal and spatial backgrounds in association with variable epidemiological settings. The details of the bacterial strains are provided in Table 1 [GenBank accession number: BL196 (AJGK00000000.1), ST0208 (AJXA00000000.1), CR0044 (AKZO00000000.1), CT18 (AL513382), Ty2 (AE014613) and P-stx-12 (CP003278).

Strain name[a]	**Year of isolation**	**Location of isolation**	**Specimen**[b]	**Epidemiological information**[b] (if available)
BL196	2005	Malaysia	Blood	Outbreak
CR0044	2007	Malaysia	Stool	Carrier
ST0208	2008	Malaysia	Stool	Sporadic
CT18 (8)	1993	Vietnam	Blood	NA
Ty2 (9)	1916	Russia	NA	NA
P-stx-12 (10)	2009	India	Stool	Carrier

[a]Names in bold font refer to genome-sequenced strains in this study. Numbers in parentheses refer to bibliographic references of previously published sequenced strains. [b]NA: not available

Table 3.1: Details of bacterial strains used in this study

3.1.2 DNA sequencing, assembly, and annotation

Whole genome sequencing was carried out on three *S*. Typhi strains using the Illumina Genome Analyser (GA2X, pipeline version 1.6, insert size 300), generating >10 total gigabytes of data. Genome assembly was constructed with Velvet (Zerbino, Daniel, & Birney, 2008) using the *de novo* approach. The open reading frames (ORFs) of the resultant contigs were predicted using RAST (Aziz *et al*., 2008). In brief, the predicted ORFs were annotated by searching against clusters of orthologous group (Disz *et al*., 2010) and SEED databases (Overbeek *et al*., 2014; Tatusov, Koonin, & Lipman, 2012) of the RAST (Aziz *et al*., 2008). The annotation results were validated with Prodigal (Hyatt *et al*., 2010) and Blast2GO (Conesa *et al*., 2005), whereas tRNA and rRNA genes were identified with tRNAscan-SE (Lowe & Eddy, 1997) and RNAmmer (Lagesen *et al*., 2007), respectively.

3.1.3 Multilocus sequence typing

Multilocus sequence typing (MLST) housekeeping gene sequences (*thrA* (aspartokinase + homoserine dehydrogenase), *purE* (phosphoribosylaminoimidazole carboxylase), *sucA* (alpha ketoglutarate dehydrogenase), *hisD* (histidinol dehydrogenase), *aroC* (chorismate synthase), hemD (uroporphyrinogen III cosynthase) and *dnaN* (DNA polymerase III beta subunit) according to PubMLST were extracted from the genome sequences (Kidgell *et al*., 2002). The alignments for each of these genomic regions were bioinformatically extracted, trimmed and concatenated into final sequence lengths of 3,336 bp using MEGA 5 (Kumar, Nei, Dudley, & Tamura, 2008). The sequences were subsequently submitted to the MLST database (http://mlst. warwick.ac.uk) and assigned existing or novel allele type numbers. The composite sequence types (STs) were defined by the database based on the allelic profile that was derived from each of the seven loci. The STs from the fragmented, incomplete genomes were derived by comparing the three less conserved alleles *hemD*, *hisD*, and *thrA* while assuming that the other four alleles, *aroC*, *dnaN*, *purE* and *sucA*, were conserved. The results with positive BLASTn hits of 100% query sequence coverage (E < 1 × 10−6) were only considered in the analysis.

3.1.4 Comparative genomic analysis

Protein coding gene predictions were performed using Prodigal (Hyatt *et al*., 2010). The predicted genes were then subjected to annotations using Blast2GO (Conesa *et al*., 2005) (E < 1 × 10−30). The genomic sequences and functional annotations of the CDSs were validated based on the results of homology searches against the public non-redundant nucleotide and protein databases (http://www.ncbi.nlm.nih.gov/) using BLASTn and BLASTp (Lipman, Altschul, Gish, Miller, Myers, 2014). The genes were selected based on the top BLAST hits (E < 1 × 10−30, ≥60% query coverage and ≥60% protein identity). The open reading frames (ORFs) of the genomes were reciprocally compared (ORF-dependent comparisons) using RAST (Aziz *et al*., 2008). The subsystem category distributions were compared among the genomes. The circular map of genes that was based on the similarities of the amino acid sequences of the BL196, CR0044, ST0208, TY2 and P-stx-12 genomes against that of CT18 was generated using the BLAST Ring Image Generator (BRIG) (Alikhan, Petty, Ben Zakour, & Beatson, 2011). A synteny-based analysis was performed by aligning the genomes with CT18 as a reference, and the contigs were reordered with iterative refinements using progressiveMauve (Darling, Mau, & Perna, 2010) and Nucmer (Delcher, Phillippy, Carlton, & Salzberg, 2002). The best alignments were chosen for the multiple genome alignments. The reference-ordered and oriented genomic scaffold that was used for the subsequent analysis was generated by concatenating reordered contigs by inserting 5Ns between the contigs using an in-house Python script. A bioinformatic pipeline, Pan-Genomes Analysis Pipeline (PGAP) (Zhao *et al*., 2012) was utilised to identify the homologous regions of the compared ORFs at an E-value cut-off of 1×10^{-10}. The nucleotide and amino acid sequences of the query ORF and selected target homologous regions were then aligned and validated using BLAST against the NCBI redundant database. The resulting matched and validated homologues, paralogues, and orthologues were used for the multiple alignment comparisons. A genomic island analysis and prediction were performed using IslandViewer (Langille & Brinkman, 2009) based on three methods (Island Pick, IslandPath-DIMOB and SGI-HMM). The IS elements were analysed using IS Finder (http://www-is.biotoul.fr). The phages were analysed with PHAST (Zhou, Liang, Lynch, Dennis, & Wishart, 2011). The regions that were algorithmically identified as intact and those sharing high similarities were compared and analysed. The sequence content comparison was performed using ACT

(Carver *et al*., 2005) and MEGA 5 (Kumar *et al*., 2008). The regions of interest were then manually curated to improve the annotations and gene predictions. The nsSNP analysis was carried out using PGAP (Zhao *et al*., 2012) by sorting them from synonymous SNPs, deletions, and insertions, and the results were validated using the CLC Genomic Workbench version 5.1 (CLC Bio, Aarhus, Denmark).

3.1.5 Phylogenomic analysis

The genome sequences of 20 global [*S*. Typhi strains and their Genbank accession numbers were as follows: BL196 (AJGK00000000.1), CR0044 (AKZO00000000.1), ST0208 (AJXA00000000.1), UJ308A (AJTD00000000.1), UJ816A (AJTE00000000.1), CR0063 (AKIC00000000.1), 404ty (CAAQ00000000.1), E00-7866 (CAAR00000000.1), E01-6750 (CAAS00000000.1), E02-1180 (CAAT00000000.1), E98-0664 (CAAU00000000.1), E98-2068 (CAAV00000000.1), J185 (CAAW00000000.1), M223 (CAAX00000000.1), AG3 (CAAY00000000.1), E98-3139 (CAAZ00000000.1), STH2370 (JABZ00000000.1), CT18 (AL513382), Ty2 (AE014613) and P-stx-12 (CP003278) (Appendix K). These sequences were submitted to the Reference Sequence Alignment-based Phylogeny Builder (RealPhy) (Bertels, Silander, Pachkov, Rainey, & Van Nimwegen, 2014) for the identification of sites that were relevant for the phylogenomic analysis using the default parameters. The complete genome of *S*. Typhi CT18 was chosen as the reference genome, and all of the query genomic sequences were divided into possible sequences of 50 bp (default) and subsequently mapped to the reference genome via Bowtie2 with a default k-mer length of 22, allowing for one mismatch within the k-mers to maximise sensitivity. The generated multiple genome sequence alignments were subsequently used to construct an unrooted phylogenetic tree that was inferred via the approximate maximum likelihood method using FastTreeMP (Price, Dehal, & Arkin, 2010).

3.1.6 Phylogenetic analysis of Zonula occluden toxin (Zot)

The *zot* amino acid sequence data (the use of amino acid sequence is more informative and appropriate for distantly related homologues) from 40 closely and distantly related bacterial strains were downloaded from the Genbank. The sequences from both BL196 and CR0044 were aligned with those of the 40 bacterial strains that were selected using the MAFFT (Katoh & Daron M, 2013) with E-INS-I strategy. Phylogenetic analysis was subsequently performed using maximum likelihood phylogenetic algorithms with the PhyML module of SeaView V4.5 (Gouy, Guindon, & Gascuel, 2010) which was supported by 1000 bootstrap replicates. The Maximum Likelihood tree was constructed using the best substitution model (Blosum62 algorithm) after being tested and optimised by ProtTest 2.4 (Darriba, Taboada, & Posada, 2011).

3.1.7 PCR validation of selected genes and SNPs

PCR was carried out to validate the identified high-quality nsSNPs. Genomic DNA for the sequencing reactions was extracted using the Wizard® Genomic DNA Purification Kit (Promega, Madison, WI, USA). The amplification of the selected genes was performed using a standard PCR protocol. Each 25 µl PCR reaction contained 150 µM (each) deoxynucleoside triphosphates, 1× PCR colourless buffer, 1.2 mM MgCl2, 0.2 µM of primer and 0.5 U of Go Taq Flexi DNA Polymerase (Promega, Madison, WI, USA). The PCR was performed under the following conditions: initial denaturation at 95°C for 30 s, 30 cycles of denaturation at 95°C for 30 s, 30 s at the respective annealing temperature (Appendix G; Appendix E) and an extension step at 72°C for 40 s; a final extension was performed at 72°C for 1 min. The reactions were carried out using a PCR Master Cycler (Eppendorf AG, Hamburg, Germany). The primer sets that were used for the target genes are shown (Appendix G; Appendix E).

3.1.8 Sequencing and high-resolution melting (HRM) analysis

The PCR products were purified using the PCR Clean-up Kit (Qiagen, Valencia, CA) according to the manufacturer's instructions. The PCR products were then sent to a commercial sequencing facility (First BASE Laboratory Sdn Bhd, Selangor, Malaysia) for direct sequencing. The nsSNP variations were validated with a pair of primers as described (Appendix G) and subsequently used for a high-resolution melting (HRM) analysis using the Kapa HRM Fast PCR Kit (Kapa Biosystems, Boston, Massachusetts, USA) and Eco Real-Time qPCR System (Illumina, San Diego, California, USA) according to the manufacturer's instructions. The melting curve profiles that were generated were analysed with the Eco-qPCR software using both homozygous and heterozygous controls.

3.1.9 Analysis of molecular effects of nsSNPs and protein structure modelling

The nsSNPs of *tviE* and *rpoS* were selected for the predictions and analyses of the further molecular effects. The Poly-Phen2 (Maathuis, Colombo, Kalisch, & Bühlmann, 2000), SIFT (Ng & Henikoff, 2003), Provean (Choi, Sims, Murphy, Miller, & Chan, 2012), SNAP (Bromberg & Rost, 2007), I-Mutant 3.0 (Capriotti, Fariselli, Rossi, & Casadio, 2008) and PredictSNP 1.0 (PredictSNP, MAPP, PhD-SNP and Panther) tools

(Bendl, Stourac, Salanda, & Pavelka, 2014) were used to examine the functional modifications and predictions of the tolerated and deleterious nsSNPs. The details of the methods and scores that were used for each tool are included in (Appendix J). A combination of different prediction methods was used to increase the prediction accuracy and confidence. The protein structures of TviE and RpoS were modelled using the I-TASSER server (Zhang, 2008). The best model was selected based on the optimal C-score. Further, the native structure was mutated by introducing a point mutation in the native RpoS protein at P193L (proline to leucine) and native TviE protein at H53Y (histidine to tyrosine) using FixPDB with the NOMAD-Ref server (Lindahl, Azuara, Koehl, & Delarue, 2006) and validated with the SPDB viewer (Kaplan & Littlejohn, 2001). The native and mutant structures were checked, fixed, refined and energetically optimised by MDWeb (Andrio, Fenollosa, Cicin-Sain, Orozco, & Gelpí, 2012), ModRefiner (Xu & Zhang, 2011), FG-MD (Andrio *et al.*, 2012) and the SPDB viewer (Kaplan & Littlejohn, 2001). The qualities of the model structures were independently verified with the PROCHECK (Laskowski, Rullmannn, MacArthur, Kaptein, & Thornton, 1996), WHATCHECK (Hooft, Vriend, Sander, & Abola, 1996) and PROSA programs (Sippl, 2007).

3.1.10 Molecular dynamics (MD) simulation and energy minimisation

The molecular dynamics (MD) simulations were carried out using MDWeb (Andrio *et al.*, 2012). The optimised structures of the native and mutant RpoS and TviE proteins were used as input data for the MD simulations. GROMACS topologies were first generated by removing the crystallographic water molecules and adding missing side chain and hydrogen atoms. Histidine residues were protonated according to the protpKa program algorithm with the GROMACS package. Water molecules were added at energetically favourable positions of the structure surfaces. Hydrogen atoms were energetically minimised for 500 steps of hydrogen conjugate gradients, while the remainder of the structures were fixed and followed by energy minimisations for 500 steps of conjugate gradients, restraining heavy atoms with a force constant of 500 KJ/mol.nm2 to their initial positions. The system was solvated with simple point charge (SPC) water molecules at spacing distances of 15 Å around the molecules. Chloride (Cl-) and/or sodium (Na+) ions were added until the system was neutralised at a concentration of 50 mM. The step involving the minimisations of the structures for 500 steps of conjugate gradients to restrain the heavy atoms with a force constant of 500 KJ/mol.nm2 to their initial positions was repeated. The whole molecular system was subjected to energy minimisations of 500 iterations by the steepest descent algorithm implementing a GROMOS96 43a1 force field. The comparative analysis of structural deviations between the native and mutant proteins of RpoS and TviE was assessed by their respective RMSD values.

3.2 Results and Discussion

3.2.1 General genome signatures of *S*. Typhi in association with outbreak, sporadic case, and carrier

The genome sequencing of *S*. Typhi strain CR0044 has generated 1.0 gigabyte of data with 90× depth coverage and a 73-bp read length. The genome assembly produced 201 contigs with a minimum contig length of more than 200 bp and an average size of 23,367 bp. The predicted genome size is approximately 4,769,054 bp, with a mean GC content of 52.1% and coding percentage of 85.8. The genome revealed approximately 4,884 coding sequences (CDS) with an average length of 825 bp. The genome also contains predicted 69 tRNA and 22 rRNA genes. For *S*. Typhi ST0208, an approximate

1.83 gigabytes of data was generated with an average coverage of 165× and yielded 1,499,986 paired-ends reads with a 100-bp read length. The genome assembly generated 222 contigs. The draft genome size is approximately 4,798,272 bp in length, with an average GC content of 52.0%, and is composed of 4,890 predicted coding sequences with an average length of 810 bp. A mean percentage of 83.7% of nucleotides of the genome are predicted to encode proteins. The genome reveals 71 tRNA and 22 rRNA predicted genes. On the other hand, the size of the single chromosome genome sequence of *S*. Typhi BL196 was approximately 4,744,056 bp, with a G+C content of 53.21% and a coding percentage of 87.1. There were 4,875 protein coding sequences found, with an average length of 875 bp. The genome revealed 76 tRNA and 22 rRNA genes (Baddam *et al*., 2012).

The genomes of the three sequenced Malaysian *S*. Typhi strains from different sources were compared to identify potential genomic features that may help to elucidate the various disease outcomes. However, because of the limited numbers of strains studied and information regarding the pan-genome of the *S*. Typhi population, the genetic differences observed should be taken as *ad hoc* basis. The aim of the detailed comparative analysis was to provide a better understanding and insights into the unusual genome dynamics of *S*. Typhi, a highly clonal organism.

In general, genome sequencing analyses have generated high-quality assemblies with an average genome coverage of 100× for the three Malaysian *S*. Typhi genomes, including BL196 (an outbreak strain that was isolated from a blood sample; Genbank accession number AJGK00000000.1), CR0044 (a strain that was isolated from a stool sample of a carrier; Genbank accession number AKZO00000000.1) and ST0208 (a sporadic strain that was isolated from a stool sample of a typhoid case; Genbank accession number AJXA00000000.1). The approximate predicted genome sizes and average guanine-plus-cytosine (G + C) contents of all three genomes ranged from 4.7

Mb to 4.8 Mb and 52.0% to 53.2%, respectively (Table 1). These genomes form a single and circular chromosome with no plasmids detected. An *in silico* multi-locus sequence typing (MLST) analysis subtyped both BL196 and CR0044 as ST1 and ST0208 as ST2, which are the main sequence types that have been associated with the worldwide distribution of *S*. Typhi out of the four STs that have been identified to date (Kidgell *et al*., 2002). The predicted coding sequences (CDSs) of the genomes based on RAST subsystem-based annotations varied from 4,875 to 4,890 with an average coding percentage of 86.0%. The average sizes of the CDSs were similar (ranging from 810 bp to 875 bp), indicating that the size differences among the genomes are largely attributable to a number of CDSs and intergenic regions. Approximately 12% of the CDSs were annotated as uncharacterised proteins (Table 3.2). Some of these genes (4.2%) were observed to vary from one strain to another. These "dispensable" genomes carried genes that were present in one or more strains and could even be unique to a single strain (Parkhill *et al*., 2001), indicating a possible open pan-genome of *S*. Typhi. The chromosomes of the three assembled genomes exhibited overall structural conservation and co-linearity with each other as evidenced by the homologous and conserved regions that were shared and the very small strain-specific regions (Figure 3.1), which may harbour genes that are relevant to the specific adaptations and fitness advantages of each of the strains. These regions most likely represent DNA that was acquired during events of HGT that may provide the strains with greater metabolic versatility or even virulence capabilities as will be further discussed in the plasticity section.

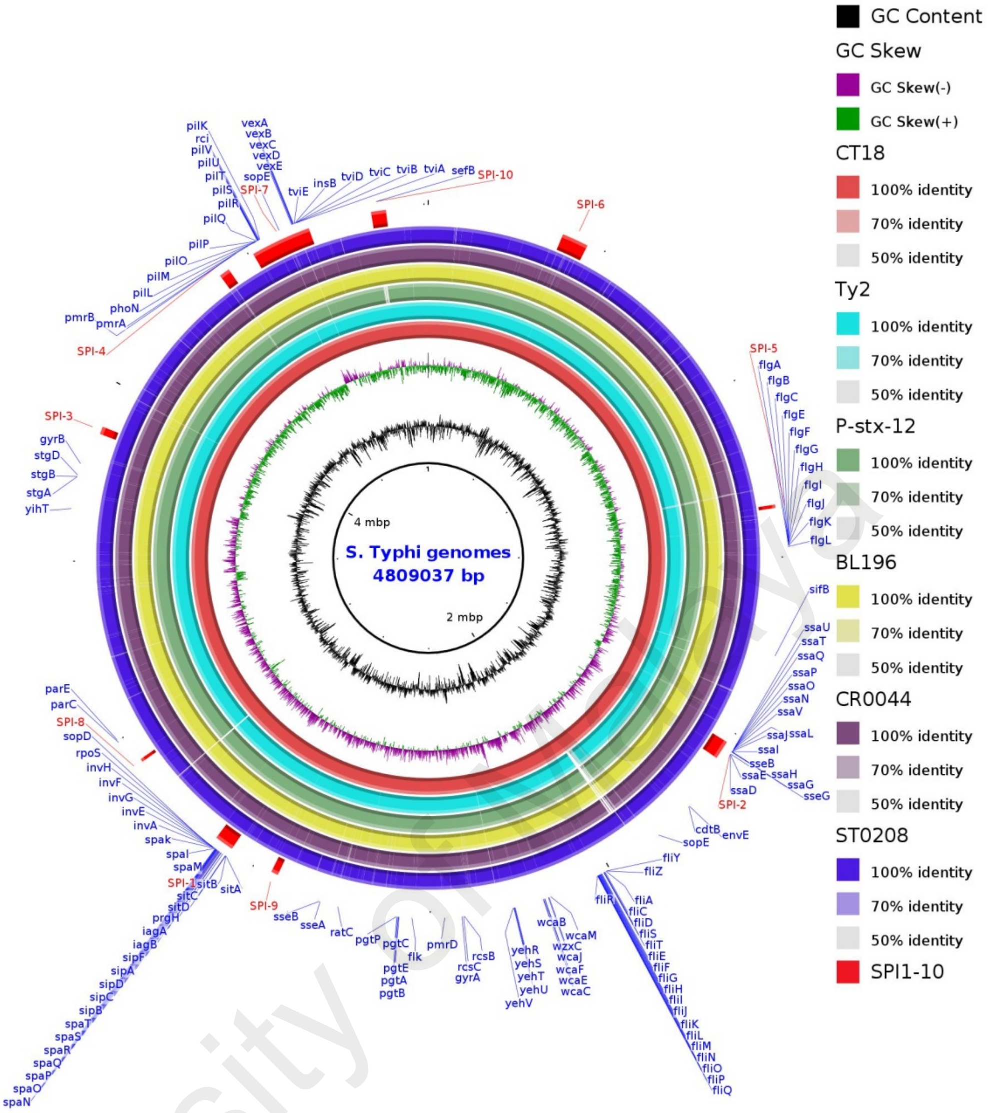

Figure 3.1: Circular genomes representation map and genome comparison of *S*. Typhi (CT18, Ty2, P-stx-12, BL196, CR0044, and ST0208). The circle is divided into arcs representing the sequence of the chromosomes of all six genomes as labeled. The inner ring shows the coordinates in scale and total genomes size of the reference sequence, CT18 in scale (Mbps), with black histogram bar representing GC content and purple-green histogram bar representing GC deviation. The orthologs for each genome with respect to CT18 (innermost red arch) are showed in order (inside-out) with the percentage of similarity based on nucleotide sequences (100%, 70% and 50%, colour tone from darkest to lightest) as indicated on the legend located at the right of the figure. The outermost arch in red represents the location of SPI1 to SPI10 in relative to CT18 and labeled in red. The position of the genes related to pathogenicity and host persistent are shown and labeled in blue at the edge of the rings.

Features	BL196	CR0044	ST0208
Approximate genome size (Mbp)	4.74	4.76	4.8
G+C content (mol %)	53.21	52.1	52
Protein coding sequences	4875	4884	4890
Percentage of coding region	87.1	85.8	83.7
Average CDS length	875	825	810
No. of tRNAs	76	69	71
No. of rRNAs	22	22	22
Percentage of hypothetical proteins	4.2	4.2	4.3
Percentage of uncharacterized protein	11.5	11.5	11.6

Table 3.2: The general genetic features of BL196, CR0044, ST0208 genomes

3.2.2 Comparative genomics of *S*. Typhi

Genomic comparisons were performed on the three Malaysian *S*. Typhi strains and the three completed *S*. Typhi genomes (the only full genomes available from the NCBI database at the point of study), CT18 (Genbank accession number AL513382) *(*Parkhill *et al*., 2001), Ty2 (Genbank accession number AE014613) (Deng *et al*., 2003) and P-stx-12 (Genbank accession number CP003278) (Ong *et al*., 2012). The comparisons of the studied strains with the reference genomes allowed for the elucidation of the novel and additional genes that are carried by the Malaysian *S*. Typhi genomes that may be of significance. The shared and unique genes of all six genomes have been analysed to determine their distinct virulence and pathogenic features. As expected, the six genomes exhibited high similarities and syntenies with each other with limited evidence of genomic rearrangements, which collectively indicate stable genomic structures. The majority of the ORFs from the compared genomes were part of a conserved genomic core, in which 4532 ORFs were shared among the genomes. These shared ORFs provide clear evidence of conservation among the genomes of the *S*. Typhi strains. The rest of the unshared ORFs or accessory genes are present in one or more strains, which represent the salient differences in the genomes. Most of these

ORFs (4.2% to 4.9%) that were harboured by each genome were annotated as hypothetical proteins (50% to 75%). The extended homology analysis has shown that the remaining unshared ORFs (25% to 50%) were likely to encode for functional proteins from diverse categories, including virulence-related proteins, secretory proteins, conserved domain proteins, transporter proteins and phage proteins among others. These important regions of the genomes could provide important functional clues for understanding the virulence and persistence of the pathogen more clearly, anticipating the need for extensive future studies focusing on their possible roles in bacterial pathogenesis. However, the numbers of shared and unshared ORFs may have been underestimated because the genomes were incomplete. Among the shared genes that were found in strains BL196 and CR0044, uncommon ORFs that encoded for the VI Icm-F secretion protein, Icm-F-related protein and type VI secretion protein EvpB were identified whose products are related to the type VI secretion system. The genes shared 99% similarity with the type VI secretion protein of *Salmonella* Typhimurium strain D23580 (Kingsley *et al*., 2009) and were only found in BL196 and CR0044. This protein was recently recognised as one of the main virulence determinants in *Burkholderia pseudomallei*, *Legionella pneumophila* and *Vibrio cholerae,* but its function in *S*. Typhi remains to be elucidated (Pukatzki, Mcauley, & Miyata, 2009; Schell *et al*., 2007). T6SS genes are believed to be involved in either structural components of the secretory apparatus, secretory products or assisting with protein translocation; for example, providing the energy to push substrates through the channel of the apparatus (Bingle, Bailey, & Pallen, 2008). These genes are also proposed to be involved in the surface reorganisation, enhancing adherence to epithelial cells, intracellular multiplication, and human macrophage killing (Bingle *et al*., 2008; Parsons & Heffron, 2005; Schreiber *et al*., 2015). Other T6SS clusters were found intact as in reference genomes. The high similarities of the genetic contents of BL196 and CR0044 with minor variations, and particularly the presence of unique accessory genes (in addition to SNPs, which are discussed in another section), are in agreement with the PFGE pulsotype data, which revealed that both strains are genetically similar, showing a difference of only one band (Appendix B).

The chromosome of *Salmonella enterica* is commonly integrated with a large portion of horizontally acquired DNA apart from its core, which is termed the Salmonella pathogenicity islands (SPIs) (Ochman & Groisman, 1996). These acquired SPIs have led to divergence and host restriction similar to those in *S*. Typhi. The identification of conserved SPIs and their variations have important implications for a wide range of microbiological applications, such as antigen and marker discovery and the identification of essential genes and their respective traits. In this study, the SPIs (SPI1-10) and the genetic variant of *S*. Typhi in the genome were annotated, which is characterised by its deviated GC content, flanking by tRNA genes and the presence of phages, integrases, recombinases and genes that are related to DNA integration. All of the SPIs (1-10) (Figure 3.1) were detected in the genomes, but marked variations were observed, notably in SPI-10. The presence of a large number of transposition-related genes in this SPI suggests that the site may be actively involved in the integration and transposition of genetic elements, which drive genetic variation. Interestingly, the comparative analysis revealed that the carrier strain P-stx-12 lacks a ~10 kb *prpZ* cluster and adjacent gene clusters harbouring 14 ORF with a deviated GC skew of 49.2 % at SPI-10 but remains fully intact in the carrier strain (Figure 3.1). Previously, a *prpZ* cluster deletion study showed that the mutant has a significantly lower survival rate compared with the parental strain, which may be due to a signalling pathway that controls the long-term survival of *S*. Typhi in host cells, and particularly, the survival in human macrophages (Faucher, Viau, Gros, Daigle, & Le Moual, 2008). In fact, the findings support this study with the fine-tuned postulation that the deletion led to reduced virulence that enabled the carrier strain to coexist with the host; for example, in the tissue of gall bladder. This possibly explains as to why the long-term survival in the macrophage is no longer necessary, which is presumed that the pathogens have colonised and persisted in other cells of the host during adaptation. However, the deletion was not detected in the human carrier strain CR0044, suggesting that the genes may not be the only factors that could lead to a human carrier state. Furthermore, the region is flanked by multiple transposases, integrases, ligases and uncharacterised proteins, which are known to be involved in transposition. Additionally, genes coding for the DNA mismatch repair protein mutC and transposase were identified both upstream and downstream of the cluster, suggesting that the region could have possibly been acquired earlier during horizontal gene transfer. The presence of a DNA mismatch protein gene has been previously implicated to be involved in modulating recombination events by incorporating or inhibiting the transfer of mobile genetic elements (Schofield & Hsieh, 2003). The deletion of the gene cluster and the presence of a vast number of genes that are related to transposition indicate that SPI-10 may be unstable and prone to excision similar to the precise excision of the crucial SPI-7, which has been recently reported in *S*. Typhi (Bueno *et al*., 2004), suggesting that gene deletion may be relevant to the host adaptation of this organism although independent acquisition or gene gain by other strains cannot be completely ruled out. These findings expand upon previous studies, which have reported that other SPIs are relatively stable in the genome (Holt *et al*., 2008), highlighting the importance of a future extensive evaluation of the stability of the other SPIs. Apart from these remarkable differences, the comparison of the two carrier strains, CR0044, and P-stx-12, revealed that CR0044 carries several additional genes that were not identified in P-stx-12 that encode for unknown functions and phages that are present in the phage region. Almost all of the major potential virulence- and persistence-associated genes have homologues in P-stx-

12, suggesting that they are not specifically associated with the unique persistence of the carrier strains but are common to *S*. Typhi. The genomic structure of the sporadic strain ST0208 is more conserved in comparison and has relatively fewer dispensable ORFs, which are mainly genes that code for hypothetical proteins and phages, indicating the conservation of large numbers of genes, which is essential for strict host adaptation and virulence optimisation.

3.2.3 Phylogenomics of *S*. Typhi revealed shared common ancestry

A core genome-based phylogeny by mapping 20 query genomic sequences against CT18 into a single non-redundant alignment was determined (Figure 3.2) (see Materials and methods). The phylogenomic tree showed that the outbreak strain BL196 and carrier strain CR0044 were closely related and could be differentiated by only 50 SNPs. These data are in agreement with the PFGE results that showed that both strains are highly related (Appendix B). The observed close phylogenetic relationship of these two strains is consistent with the earlier speculation that the large outbreak that occurred in 2005 could share a common ancestor with the human carrier that may have been circulating for a long period in the country. It is challenging to determine how these two strains are related, considering their short-term evolutionary relationship. However, three evolutionary postulations are possible. First, the human carrier strain may have been derived from the outbreak strain after the large outbreak. Second, the human carrier strain may have long existed in a human carrier who served as a reservoir and the source of the outbreak. Finally, both of these strains may have diverged independently from the common ancestor, which was possibly harboured in a carrier to give rise to two independent cases, considering the geographical proximities. Notably, these two highly related strains with unique gene repertoires clustered with four epidemiologically and geographically unrelated strains from India (P-stx-12), Russia (TY2), Vietnam (AG3) and Senegal in West Africa (E01-6750). Interestingly, all of these strains, including BL196, CR0044, P-stx-12, Ty2 and E01-6750, were subtyped as ST1 (the ST from the genomic sequence of AG3 could not be established due to missing determining allelic sites). Conversely, the Malaysian sporadic *S*. Typhi strain ST0208 clustered closely with the multidrug-resistant strain CT18 from Vietnam together with other geographically related strains from Indonesia (404ty and J185) and Bangladesh (E98-2068), which were all subtyped as ST2, suggesting the possible movement of clonally related strains among the Southeast Asian countries. Such close genetic relatedness that is based on macrorestriction and SNP typing analyses has been previously reported (Roumagnac *et al*., 2006; Tiong *et al*., 2010). The rest of the genomes were clustered as ST2. From the analysis, no temporal or geographical signals were found, but ST was highly correlated with the phylogenomic clustering. Apparently, ST2 could be further differentiated into two clusters according to the phylogenomic analysis but was limited in the current MLST scheme for *S*. Typhi. These results indicate that the phylogenomic analysis has much better resolution power compared with MLST in separating highly clonal strains as in *S*. Typhi. This information is essential for devising a new set of MLST alleles. These data further support the widespread distribution of ST1 and ST2 as the major genotypes that occur worldwide, although ST3 and ST8 have also been isolated in a previous study (Kidgell *et al*., 2002). The two human carrier strains, CR0044, and P-stx-12 belonged to different subtypes, indicating the non-universality of the genotypes in association with the human carrier strains.

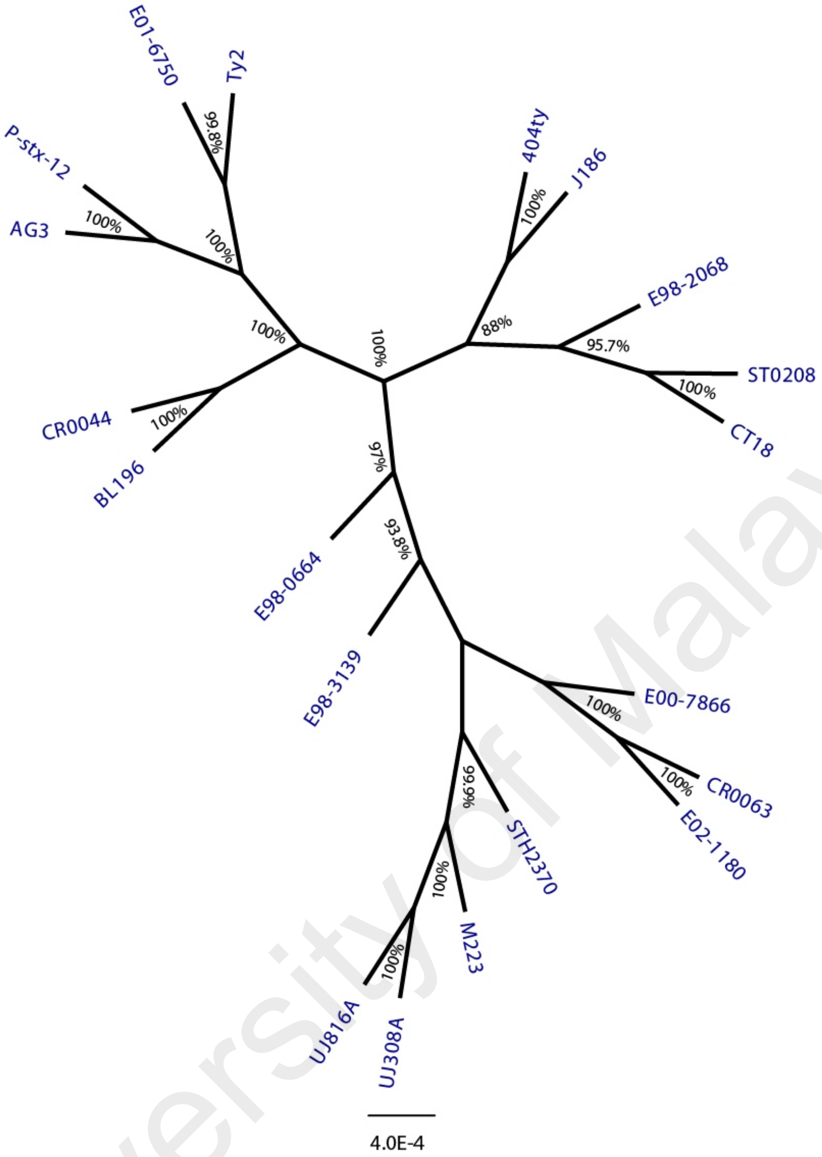

Figure 3.2: Phylogenomic tree inferred by approximately-maximum-likelihood method from the aligned core genomes. Multiple genomes alignments were generated by mapping genome sequences of the 20 global *S*. Typhi strains against CT18 at all sites relevant for phylogenomic analysis using RealPhy (78). Phylogenetic tree (unrooted) was inferred via approximately-maximum-likelihood method using FastTreeMP (79). Branch length scale as indicated and bootstrapping values are shown on the node of the tree.

3.2.4 Genome plasticity (insertion sequences, phages and CRISPRs)

The analyses of the relationships among the strains and their genomic variations are further supported and extended to the study of the genome plasticity of *S*. Typhi. In many organisms, genomic plasticity is commonly observed, but *S*. Typhi has generally demonstrated few variations compared with other *Salmonella* spp. (Edwards, Olsen, & Maloy, 2002). However, in this study, considerable genetic variations were detected, predominantly involving IS elements, phages, and CRISPRs. In fact, marked variations in the numbers and types of IS elements were detected. IS200 (200, 200F, 200C, 200G, and 200H) and IS1541 (1541A, 1541B, 1541C and 1541D) were both abundant in the three Malaysian *S*. Typhi genomes. The IS elements, such as IS200, have been widely used as molecular markers for subtyping due to their genome-wide distributions and high levels of diversity, but their roles in modulating the gene expression in *S*. Typhi have yet to be clarified (Rowe & Gibert, 1994). Recent studies on enterohaemorrhagic *E. coli* O157 have shown that the presence of IS could play a role in the gene inactivation and immobilisation of incoming phages and plasmids, leading to the diversification and evolution of the bacterial genome (Ooka *et al*., 2009). IS elements have also been shown to affect the expression of neighbouring genes and induce genomic rearrangements (deletions, inversion, and duplications) (Kusumoto *et al*., 2011). However, little is understood about their roles in modulating virulence and gene expression in *S*. Typhi. The variations that were detected may provide clues on how these differences affect the virulence and fitness strategies of the pathogens.

The *S*. Typhi strains BL196, CR0044 and ST0208 carry eight, seven and eight phages, respectively. Substantial phage variations among the *S*. Typhi genomes were identified. One of the differentiating features was a distinct set of prophages that were harboured by both BL196 and CR0044, which rendered them less unique compared with ST0208 except a few ORFs that encoded for phages and hypothetical proteins. Interestingly, the phages that were carried by ST0208 had relatively shorter lengths (in bp) compared with the phages that were identified in the other strains. As expected, both of the closely related strains (BL196 and CR0044) had highly similar phage contents and carried an additional intact *Salmonella* phage RE-2010 with uncommon ORFs, which mainly contained genes coding for hypothetical proteins, phage proteins, prophage-like proteins, repressor proteins, excisionases, terminases and integrases. The variations that were detected in the numbers of predicted prophages and prophage-like regions illustrated the dynamics of phage gain and loss that distinguished one strain from the others, indicating that phages may play important roles in genomic diversity. Based on the phylogenomic alignments, this region appears to show the typical gain and loss of sequences during the course of genomic evolution. This evolutionary relationship is consistent with the phage variations, providing a useful framework for investigating the relationships of strains and their respective phenotypes. The phage was likely acquired prior to the divergence of the common ancestor of both BL196 and CR0044 through horizontal gene transfer rather than phage loss. Alternatively, due to the advantageous roles of the HGT events, the phage proteins that were acquired by the strains could promote theirs *in vivo* survival and pathogenesis (Casjens, 2003). Among the detected phages, the two typical *S*. Typhi phages, Gifsy-2, and Fels-2 were both found to be intact and conserved in all six of the *S*. Typhi genomes, suggesting that these regions may play essential roles and provide fitness advantages to the pathogen. Apart from the intact phages, three incomplete phages, including Burkholderia_phage_BcepMu, Enterobacteria_phage_cdtI and Lactococcus_phage_bIL312, were also observed in the six genomes, suggesting that these common phages were likely present in their common ancestors and may possess important functions for host adaptation and survival.

Clustered regularly interspaced short palindromic repeats (CRISPRs together with the *cas* gene) were recently found to be important in bacteria as a primary defence strategy against foreign nucleic acids, including phages and conjugative plasmids (Barrangou *et al*., 2007). In fact, CRISPR regions have been found to be integrated in response to infecting phages. These regions are known to have hypervariable genetic loci due to the high diversities of the interspaced regions between the palindromic repeats and frequently match to phage and other extrachromosomal elements. The presence of CRISPRs in genomes may affect short-term phenotypic changes and mediate long-term sublineage divergences (Barrangou *et al*., 2007; Horvath & Barrangou, 2010). CRISPR regions were identified in all three of the Malaysian *S*. Typhi genomes using CRISPRs Finder online (crispr.u-psud.fr/Server/). The identified regions were found to be located around the CRISPR-associated protein cas1 and flanked by genes coding for alkaline phosphatase isoenzyme conversion aminopeptidase and clusters of CRISPR-associated genes, including *cas* and *cse*. CRISPR_1 was identified with the palindromic repeats, showing strikingly high similarities among all of the analysed *S*. Typhi genomes with minor variations in spacers, indicating the important evolutionary conservation of the *S*. Typhi strains. Apparently, the gene orders of CRISPRs are also conserved among the genomes. The sporadic *S*. Typhi strain ST0208 harbours a shorter CRISPR region (designated as confirmed CRISPRs by CRISPRs Finder) of 333 bp in length compared with CT18 (385 bp), Ty2 (394 bp) and P-stx-12 (394 bp). The CRISPR region of ST0208 has identical palindromic repeats, and a repeat length compared with the other genomes but lacks one spacer. Similar observations were also observed in CT18, which had shorter spacers compared with the rest, which is in agreement with phylogenomic analysis to be genetically related to ST0208 and cluster together with Ty2 and P-stx-12, which were found to possess CRISPR regions of identical lengths. A previous study showed that the addition or deletion of spacers can modify the phenotype of phage resistance (Horvath & Barrangou, 2010). Alternatively, non-identity spacers have been suggested to mediate the interactions between CRISPRs and phages (Cady & O'Toole, 2011). The diversity of spacers in CRISPRs of *S*. Typhi may be relevant to other interesting roles that are yet to be understood. Interestingly, additional CRISPR-like regions (designated as possible CRISPRs by CRISPRs Finder) were observed to be identical in the two closely related strains, BL196, and CR0044. All of the *S*. Typhi spacers that were analysed showed homology to many eukaryotic sequences, extrachromosomal sequences, and phages, supporting the immunity roles of CRISPRs against phages and other incoming DNA. Variations in CRISPR regions were identified, including those in BL196 and CR0044, providing evidence of evolutionary relevance that is useful for distinguishing between closely related strains and inferring ancestral relationships.

3.2.5 Putative pathogenomic island harbouring *zot,* a potential novel virulence-determining factor

A total of eight and 10 ORFs were identified in BL196 and CR0044, respectively, with eight commonly shared (100% sequence similarities) ORFs being identified (Figure 3.3) compared with the other genomes. These additional genomic clusters were absent in the other *S*. Typhi genomes and the *S*. Typhi sequences in NCBI database. Most of these regions contained ORFs that were related to transposable elements and were flanked by genes coding for phage proteins, which may be regarded as evidence of their horizontal acquisitions. The examination of the GIs that were specific to BL196 and CR0044 also supported their common evolutionary lineages, suggesting that they were acquired fairly recently. The sizes of the clusters were approximately 7.7 kb (BL196) and 9.6 kb (CR0044), and they contained deviated GC contents of 42.3% (BL196) and

40.51% (CR0044) compared with the rest of the genomes. These clusters were predicted to be genomic islands (GIs) using IslandViewer (predicted by at least one method) (Langille & Brinkman, 2009). GIs are non-self-mobilising elements that code for proteins with diverse functions that may be integrated or excised, thus playing important roles in bacterial diversification and adaptation. Their features resembled those of the previously reported pathogenicity islands, such as the presence of ISs, integrases, transposases and deviated GC contents (Hacker & Kaper, 2000). The exact sizes of these GIs are likely to be larger than 10 kb, and they may have contained many elements of GIs. The GIs were found to carry the IS element ISBf10 (IS66 family) at its 5'-end and notably, it harbours a novel gene coding for the zonular occluden toxin family protein (*zot)*. Apparently, the *Salmonella* spp. including *S*. Typhi, that carry this gene are very limited (positive BLASTp hit only to *Salmonella enterica* subsp. salamae and *Salmonella enterica* subsp. houtenae in NCBI nr database as of 10 May 2014), but it was found commonly in more distantly related ancestors, such as diarrheagenic *E. coli* DEC13A, which shared 100% amino acid sequence identity but was less similar (<30%) compared with *zot* in *Vibrio cholerae* (*V. cholerae*) and *Neisseria meningitidis* (*N. meningitidis*) (Radnedge, Agron, Worsham, & Andersen, 2002) (Figure 3.4). Zot has been well characterised in *V. cholerae* as a toxin that disrupts the integrity of the intestinal barrier by targeting the tight junction to increase tissue permeability (Hazen *et al*., 2012; Rosso *et al*., 2008). However, its functions in other pathogens, particularly the *Salmonella* spp., are poorly understood. *Zo*t was previously found to be an uropathogenic-specific protein in uropathogenic *E. coli* and a potentially important toxin in many pathogens that have received minimal attention (Kurazono *et al*., 2000). Recently, *Campylobacter concisus* was the only species among the Campylobacteriales to harbour *zot*, suggesting its importance in gastrointestinal pathogenesis (Mahendran, Tan, Riordan, & Grimm, 2013). The extended homology analysis has further revealed that this gene homologue is surprisingly diverse, being present in the genomes of plant pathogens, such as *Ralstonia solanacearum* and *Xanthomonas campestris,* and opportunistic pathogens that are found in soil and water, such as *Providencia stuartii* and *Pandoraea* spp. Interestingly, this homologue was also found in lithoautotrophic bacteria, such as *Caminibacter mediatlanticus*, *Sulphurimonas* spp., *Pseudogulbenkiana ferrooxidans* and *Beggiatoa alba,* which were isolated from the extreme environments, such as the deep hydrothermal vents and sulphur springs (Figure 3.4). The importance of the necessity of gene targeting tight junctions in these extremophiles is yet to be understood but may be due to the role of the toxins in facilitating the penetration of the tissue layers that cover the organisms that they colonise. The peptide that is encoded by the *zot* was predicted to be 38 KDa with an average isoelectric point of 8.3. The analysis of the region immediately upstream revealed putative prokaryotic consensus promoters at -10 and -35 containing TTTTATTAT and TTATCT motifs. The promoter site is situated 21 bp upstream from the start of the coding region. This site harboured matching transcriptional-binding motifs of cAMP receptor proteins (CRP), which are required for efficient transcription through direct protein-protein interactions with RNA polymerase. A Shine-Dalgarno (SD) consensus motif (AGGGTG) that is required for the initiation of translation in prokaryotes was also detected. The gene contains four highly conserved domains when aligned with the *zot* genes of *V. cholerae* and *N. meningitidis*. These conserved domains are likely to be important to toxin activity. The P-loop region of the gene between amino acids 3-134 contains a nucleoside triphosphate hydrolase (NTP)-binding motif. The transmembrane region at amino acids 187-209 and the two low complexity regions at amino acids 346-351 and 387-396 were also found. However, the previously identified active domain (FCIGRL) of *V. cholerae* was absent in the Zot of *S.* Typhi but retained a putative binding motif adjacent to the active domain within the

C-terminus. A previous study showed that the partial refolding of the denatured binding peptide of this domain did not prevent its specific binding to the Zot receptor on Caco-2 cells, demonstrating that the conformationally varied domain was still able to induce activity (Di Pierro *et al*., 2001). The flexible conformation of Zot is in agreement with its identity as a membrane-spanning protein. The secondary structures that were predicted from these species (*S*. Typhi, *V. cholerae* and *N. meningitidis*) were conserved with some minor variations. However, their predicted tertiary structures were highly variable, suggesting that the toxin might have different mechanisms of action despite maintaining some core binding activities. The possible mechanism of Zot in *S*. Typhi may be similar to that of *V. cholerae*, in which tight junction disassembly is induced through the activation of the proteinase-activated receptor 2 (Groschwitz & Hogan, 2009). The presence of *zot* as a potentially new enterotoxin in *S*. Typhi may be relevant to the pathogenesis of these strains. The prevalence of this gene was further screened in 41 *S*. Typhi strains (from diverse locations that were collected over a span of 25 years from 1983-2008; Appendix C), and remarkably, none of the strains carried it except the two highly related strains, BL196 and CR0044 (Appendix D). The primers that were used are listed in Appendix E. This gene could act as a marker to distinguish between the closely related outbreak and carrier strains and the sporadic strain and other tested strains. *Zot* is flanked by genes encoding a conserved domain protein and bacterial Type II and III secretion proteins. Given that the secretion protein is often associated with an extrabacterial target, these genes may be related to the transportation of Zot into the host intestine. Apart from the secretion system, the GIs also contained a set of unique genes, including membrane proteins, phage hypothetical proteins, hypothetical proteins and phage plasmid replication proteins, suggesting the presence of phage-mediated integration as occurs in most genomic elements. The genomic elements together with the hypothetical proteins shared remarkable sequence similarities of

>90% with *Yersinia pestis* A1122 and *Yersinia pestis* CO92 (Goldblum *et al*., 2011; Lee, White, & Van Der Walle, 2003), which are the causative agents of the Black Death (Figure 3.3).

The high similarities between the GIs of the BL196 and CR0044 strains further provide strong evidence of the common ancestry between the strains. Alarmingly, these strains are more diverse than was previously thought, emphasising the concern that such strains that have acquired new genes through HGT are circulating in the country or elsewhere. Additionally, the HGT events that occurred in these *S*. Typhi strains are relevant to the speculations that both the clonally related strains are responsible for the outbreak and carrier cases. Future studies could be carried out to fully characterise the GIs and their relevant roles in the pathogenesis of *S*. Typhi. *zot* and other genes in the cluster merit further extensive studies due to their potential contributions to the virulence of *S*. Typhi and other pathogens.

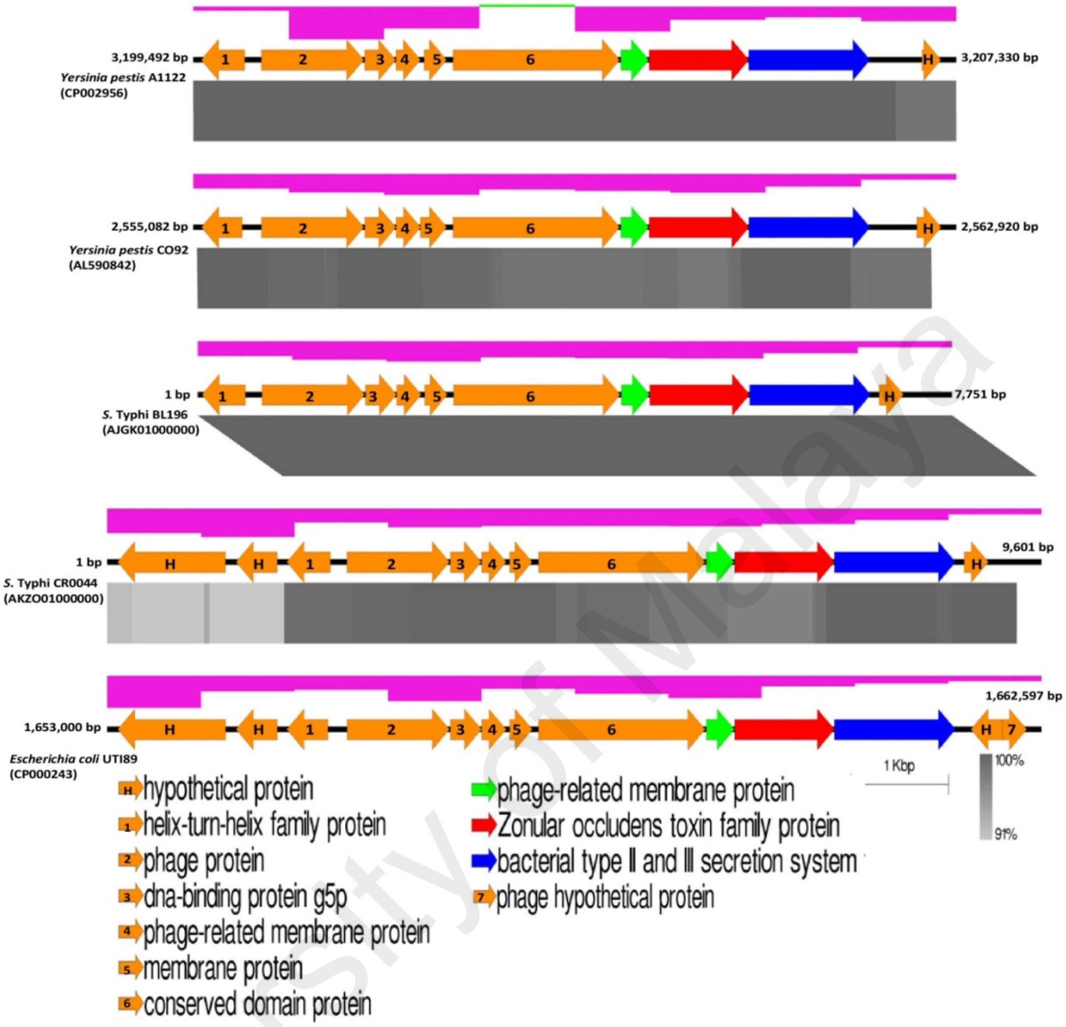
3,199,492 bp
3,207,330 bp
Yersinia pestis A1122 (CP002956)
2,555,082 bp
2,562,920 bp
Yersinia pestis CO92 (AL590842)
1 bp
7,751 bp
S. Typhi BL196 (AJGK01000000)
1 bp
9,601 bp
S. Typhi CR0044 (AKZO01000000)
1,653,000 bp
1,662,597 bp
Escherichia coli UTI89 (CP000243)
H
1
2
3
4
5
6
7
1 Kbp
100%
91%
H hypothetical protein
1 helix-turn-helix family protein
2 phage protein
3 dna-binding protein g5p
4 phage-related membrane protein
5 membrane protein
6 conserved domain protein
phage-related membrane protein
Zonular occludens toxin family protein
bacterial type II and III secretion system
7 phage hypothetical protein

Figure 3.3: Schematic representative of novel putative pathogenomic island harboring *zot* of CR0044 and BL196. 7.7-kb and 9.6-kb genomic island fragment of strain BL196 and CR0044, respectively. The schematic diagram shows the presence of various virulence-associated genes detected, particularly *zot* and tBLASTx comparison with *Yersinia pestis* A1122 and *Yersinia pestis* CO92. The position of the regions in the genomes is labeled (bp). The arrow bars denote annotated genes as in the legend based on BLAST classification (the BLASTP analysis was carried out across a non-redundant protein database in GenBank). The arrow direction showed transcription direction of the gene. The green and pink blocks above the arrow bars denote GC deviation

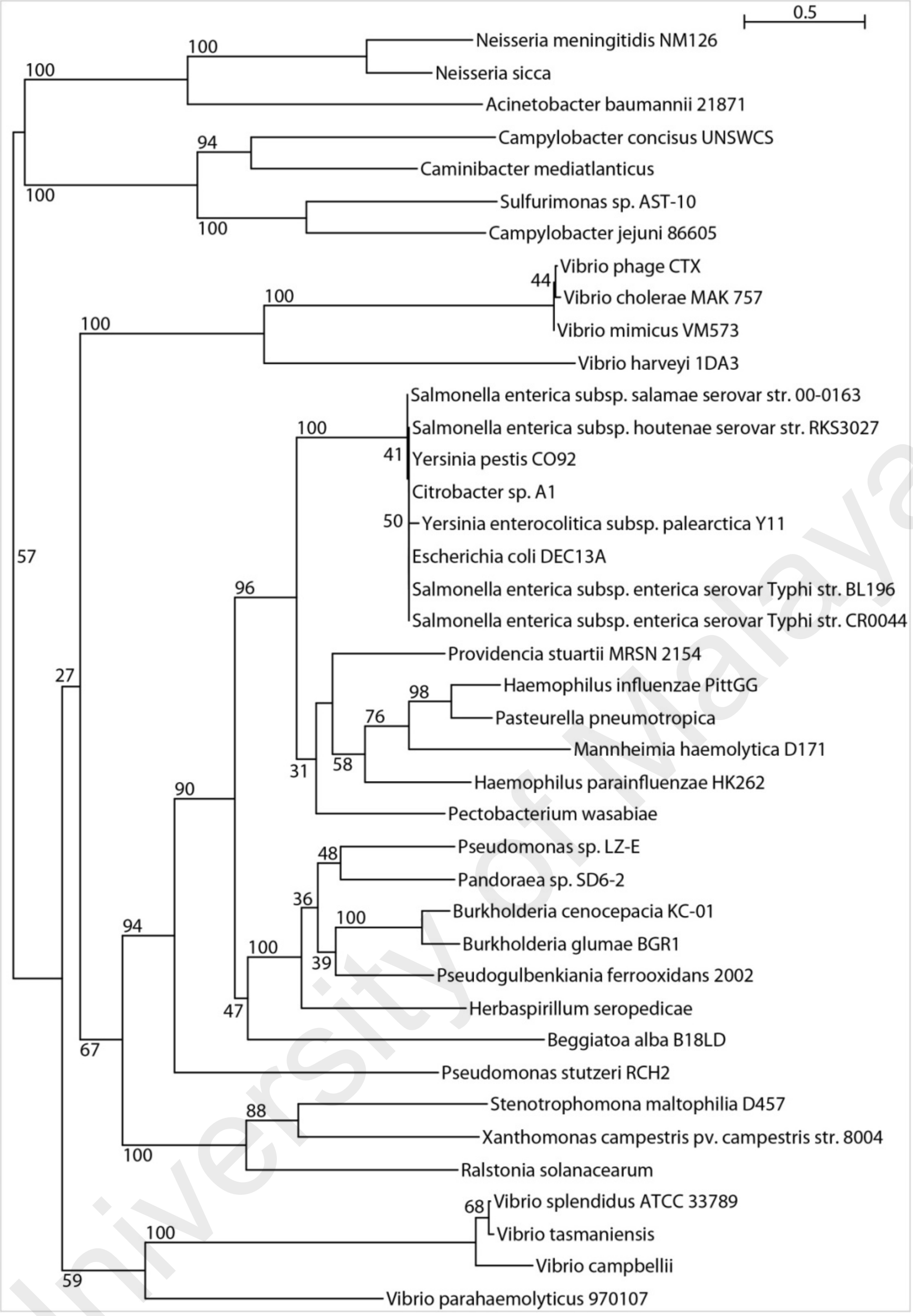
0.5
100
100
Neisseria meningitidis NM126
Neisseria sicca
Acinetobacter baumannii 21871
94
Campylobacter concisus UNSWCS
Caminibacter mediatlanticus
100
Sulfurimonas sp. AST-10
100
Campylobacter jejuni 86605
44
Vibrio phage CTX
Vibrio cholerae MAK 757
100
100
Vibrio mimicus VM573
Vibrio harveyi 1DA3
Salmonella enterica subsp. salamae serovar str. 00-0163
100
Salmonella enterica subsp. houtenae serovar str. RKS3027
41
Yersinia pestis CO92
Citrobacter sp. A1
50
Yersinia enterocolitica subsp. palearctica Y11
57
Escherichia coli DEC13A
96
Salmonella enterica subsp. enterica serovar Typhi str. BL196
Salmonella enterica subsp. enterica serovar Typhi str. CR0044
Providencia stuartii MRSN 2154
27
98
Haemophilus influenzae PittGG
76
Pasteurella pneumotropica
Mannheimia haemolytica D171
31
58
Haemophilus parainfluenzae HK262
90
Pectobacterium wasabiae
48
Pseudomonas sp. LZ-E
Pandoraea sp. SD6-2
36
100
Burkholderia cenocepacia KC-01
94
Burkholderia glumae BGR1
100
39
Pseudogulbenkiania ferrooxidans 2002
47
Herbaspirillum seropedicae
Beggiatoa alba B18LD
67
Pseudomonas stutzeri RCH2
88
Stenotrophomona maltophilia D457
Xanthomonas campestris pv. campestris str. 8004
100
Ralstonia solanacearum
68
Vibrio splendidus ATCC 33789
Vibrio tasmaniensis
100
Vibrio campbellii
59
Vibrio parahaemolyticus 970107

Figure 3.4: Phylogenetic tree of *zot*. Phylogenetic analysis of *S*. Typhi *zot* with *zot* genes of 40 closely and distantly related bacteria strains using the Approximate-Maximum-Likelihood method. The analysis was performed with 1000 replicates with the bootstrapping method. The bootstrap values are indicated on the node of the tree. The branch length is as shown.

3.2.6 Microvariations distinguishing closely related strains

Single nucleotide polymorphisms (SNPs) were determined for the comparisons of the two highly related strains, BL196, and CR0044. The microvariations of the high quality non-synonymous single nucleotide polymorphism (nsSNPs) identified earlier were further investigated. In this study, only the potential functions altering the nsSNP mutations were considered. Despite being highly related and similar, the genomes could be discretely distinguished from each other by 29 nsSNPs (Appendix F). Interestingly, the 29 nsSNPs that were identified, two were found on two highly related virulence determinant genes, RNA polymerase sigma factor *rpoS* and the Vi polysaccharide biosynthesis protein *tviE*. These two mutations are strain-specific and were not detected in any other *S*. Typhi genomes that were analysed. BL196 carried a mutant *rpoS* (P193L), which is 100% similar to *Salmonella* Pullorum S6702, which is a causative agent of fowl typhoid (Lu *et al*., 2012). Interestingly, unusual *rpoS* gene was also observed in P-stx-12. The *rpoS* of P-stx-12 contains an additional 57 bp of a response regulator *gacA* fragment that was fused at the 5’ end of the gene, resulting in a longer RpoS protein. The *gacA* fragment was highly similar to the transposons Tn10d tetA and tetR, which are associated with regulatory and transcription signals, suggesting the occurrence of transposition events and possibly gene regulation. The *rpoS* mutant and its effects on *S*. Typhi were first reported in the Ty2 genome and attenuated strain Ty21a. These mutants, which were partially derived from natural mutations in Ty2, were found to affect the stress response and other related functions significantly (Deng *et al*., 2003; Robbe-saule, Coynault, & Norel, 1995). Mutations in *rpoS* are apparently advantageous to the strain for survival in the host during prolonged stress, allowing for the selection of more efficient transcription factors for survival and fitness during unfavourable conditions (Fang *et al*., 1992). *rpoS* is commonly associated with the virulence regulation of pathogens because it regulates over 30 genes that are related to the stress response, such as the *spv* protein, which is involved in host cell survival, and Vi-polysaccharide biosynthesis proteins in different osmolarities (Santander, Wanda, Nickerson, & Curtiss, 2007).

In contrast, CR0044 carried a mutant *tviE* (H53Y). The *tviE*, which is encoded by SPI-7, is required for virulence capsular formation and acts as a protective antigen. The antibody that responds to the Vi-positive strain has been shown to be more virulent than that targeting the Vi-negative strain (Arricau *et al.*, 1998; Stephen Baker *et al.*, 2005). A role of the RpoS protein in fine-tuning the synthesis of the Vi polysaccharide in *S*. Typhi has also been reported (Santander *et al.*, 2007), suggesting the possible close regulation of these two genes in modulating adaptation and virulence capabilities in different host environments. The nsSNPs that were detected in these two closely related genes may indicate that their adaptive selection is important for host survival. To address false-positive results, the nsSNPs were validated using a high-resolution melt (HRM) analysis and direct sequencing (Appendix G). The unique HRM profiles of the wild-type strains and those containing the SNP transition mutations in the normalised graph are shown in Appendix H and I. Both SNPs showed unique melting profiles for the strains that were tested. For the *rpoS* SNP, the transition mutation from C to T occurred in strain BL196, and the separation of the melting profile began at ~81 ^{0}C and ended at 84 ^{0}C. For the *tviE* SNP, the transition mutation from C to T occurred in strain CR0044, and the separation of the melting profile began at ~74 ^{0}C and ended at 78 ^{0}C. The results of this analysis were further confirmed by the direct sequencing of targeted loci (Appendix H and I). The primers and HRM profiles that were developed may be useful for distinguishing between these two strains and as important markers for future surveillance. It is interesting to note that the independent mutations in both *rpoS* and *tviE* in the closely related strains BL196 and CR0044 supports the aforementioned third postulation parsimoniously that the strains diverged from a common ancestor, which carried features commonly possessed by its descendants. Another 27 high-quality nsSNPs were also identified that mainly encoded for non-virulence factors (Appendix F). Twenty-three genes had well-defined functions, including four that were involved in metabolism and 12 that played roles in cellular processes, signalling, and transport. The remainder were poorly characterised or had unknown functions. Additionally, out of 29 nsSNPs, 24 (83%) mutants (having unique nsSNPs in comparison to all reference genomes) were detected in the carrier strain CR0044 (nsSNPs with reference genomes), suggesting the possible functional adaptation of this carrier strain in the host cell relative to its closely related strain. However, most of the mutant SNPs that were carried by CR0044 were not detected in P-stx-12 with the exception of the gene encoding the trehalose permease IIC component, indicating the adaptation and persistence of the pathogen are strain-specific. In fact, these two carrier strains can be differentiated largely by the presence of 253 SNPs, including 159 nsSNP that potentially resulted in phenotypic effects. Notably, nsSNPs were found in the genes encoding the virulence proteins MsgA, SipD, SsaP, SsrAB, and FimH, whereas the remainder coded for metabolism-related proteins, membrane proteins, regulatory proteins and large numbers of uncharacterised proteins, suggesting that independent genomic factors may have contributed to the carrier state. It is also recognised that this analysis could only have been possible by comparing closely related strains, in which differences that were detected could have contributed to a variety of potential phenotypes. Nevertheless, it is challenging to determine the universality of the genetic signatures of the carrier strains considering the low numbers of representative samples that were used in this study. However, the sets of genes that were detected may be relevant and provide important insights into the adaptive strategies of strains potentially resulting from their varying virulence and persistence capabilities. The genes carrying nsSNPs that were identified and may be under selective pressure deserve further investigation, highlighting the importance of genome-wide association studies for determining the virulence profiles of strains, and particularly carrier-associated strains that could be related to diverse phenotypic effects. The nsSNP analysis sheds light on the plasticity of the genome at the nucleotide level, supporting its dynamic and variable nature, which remarkably contributes to the varying phenotypes and clinical outcomes.

3.2.7 Analysing the molecular effects of nsSNPs, protein structure modelling and molecular dynamics (MD) simulation

Point mutations that cause alterations in amino acids can have profound effects on the structural stability of proteins; hence, a study of their effects is necessary to understand their functionalities. Both of the native and mutant structures of the proteins were modelled. Out of the five native protein structure models that were generated by I-Tasser (Zhang, 2008), the best structure with the highest confidence score (C-scores: RpoS, 0.64; TviE, -0.40) was collected and used for further investigations. The modelled native RpoS and mutant RpoS structures showed good stereochemical properties, with 92.0% and 85.4% of the residues being within the most favourable region of the Ramachandran plot, respectively, whereas the native and mutant TviE showed 86.5% and 80.0% of the residues in the most favourable region of the plot, respectively. All of the structures passed ProSa model quality validation with Z-scores (RpoS native, -5.61; RpoS mutant, -5.57; TviE native, -6.32; TviE mutant, -6.56) falling well within the range of those that are typically reported for native proteins of similar sizes from different sources (X-ray, NMR).

Molecular dynamics (MD) simulation with a realistic aqueous solvent environment was performed to reveal the explicit solvent behaviours of the native and mutant structures, which could elicit the differences in their dynamics and stabilities. The energy minimisation studies were performed for both the native and mutant structures, and the total energies of the native and mutant RpoS achieved were -4115.46 J/mol and -4804.66 J/mol, respectively, whereas those of the native and mutant TviE achieved were -3203.54 J/mol and -2906.31 J/mol, respectively. Energy minimisation assessments provide clues regarding protein stability. The deviation between two structures can be evaluated by the root mean square deviation (RMSD). The higher the RMSD value is, the greater the deviation will be between the native and mutant structures. The RMSD value between the native and mutant RpoS was found to be 0.39 Å, and that between the native and mutant TviE was 0.43 Å. The native, mutant, and superimposed protein structures at their corresponding positions for RpoS and TviE are shown (Figure 4.5). The RMSD values of the native and mutant structures were significantly similar for both RpoS and TviE, suggesting similar levels of protein folding alterations.

The molecular effects and functional modifications were analysed based on multiple predictive tools (refer to Materials and Methods) targeting various aspects of protein dynamics with confidence scores (Appendix J). Out of nine predictive tools used, all found that the nsSNP in RpoS was deleterious (affecting protein structure and function), whereas, in TviE, three predictive tools found that the nsSNP to be deleterious, five to be neutral and one undetermined. The nsSNP in RpoS showed PSIC score of 1.0 (deleterious) with PolyPhen2 (Maathuis *et al.*, 2000) in addition to a probability score of 0 (<0.05, deleterious) with SIFT (Ng & Henikoff, 2003) and a Provean (Choi *et al.*, 2012) score of -9.261 (<-2.5, deleterious), suggesting that the nsSNP could affect the protein drastically and result in functional modifications. It is important to note that the deleterious effects could lead to positive or negative functional modifications. To further validate the results, a method that was based on the hidden Markov model (HMM) from PANTHER (Thomas *et al*., 2003) was used. The nsSNP in RpoS was found to be deleterious, but undetermined for TviE. The predictions of the nsSNP in RpoS by these tools are in agreement and show high correlations among various methodologies. The analysis with I-Mutant 3.0 (Capriotti *et al*., 2008), which is based on the support vector machine (SVM) and DGG stability changes, revealed that the nsSNP in RpoS may have led to its greatly increased stability (DGG value of 0.89 Kcal/mol, >0.5 kcal/mol indicates large increase in protein stability), suggesting that the deleterious effects of the nsSNP are favourable to the protein folding and structure and possibly lead to enhanced functions. The DDG value is calculated from the unfolding Gibbs free energy value of the mutated protein, which is based on a trained cross-validation procedure using a comprehensive experimental database of protein mutations (Capriotti *et al*., 2008).

These predicted results were consistent with the differences in the energy minimisation of MD simulation, suggesting the favourable folding of the mutant structure. Native and mutant structures differ due to the specific properties of the residues that could disrupt the structure and function of the protein. The mutant residue (leucine) of RpoS is molecularly larger in size than the wild-type residue (proline), which is a highly rigid, small molecule that is required to induce a unique backbone conformation. However, the alteration of the small-sized secondary amine structure to a larger-sized primary aliphatic amine at this site can disturb its conformation and lead to bumps in the structure, in which the mutant residue may not be in the energetically favourable position to make the typical hydrogen bond that is formed by the native residue (Figure 4.5). The hydrophobicities of the wild-type and mutant differ because the mutation introduces a very high hydrophobic residue in place of a less hydrophobic residue. This can result in a loss of hydrogen bonding and may disturb proper protein folding. The wild-type proline residue in RpoS is well conserved and is located at the discrete compact three-helical domain within region three of the protein; however, no known mutant residues with similar properties were observed at this position in the other homologous sequences. This region is the specific binding site of bacterial promoters containing an extended -10-promoter element and is primarily involved in the binding of the core RNA polymerase in the holoenzyme. The mutation in this important site could affect protein functioning pertaining to the transcription efficiency.

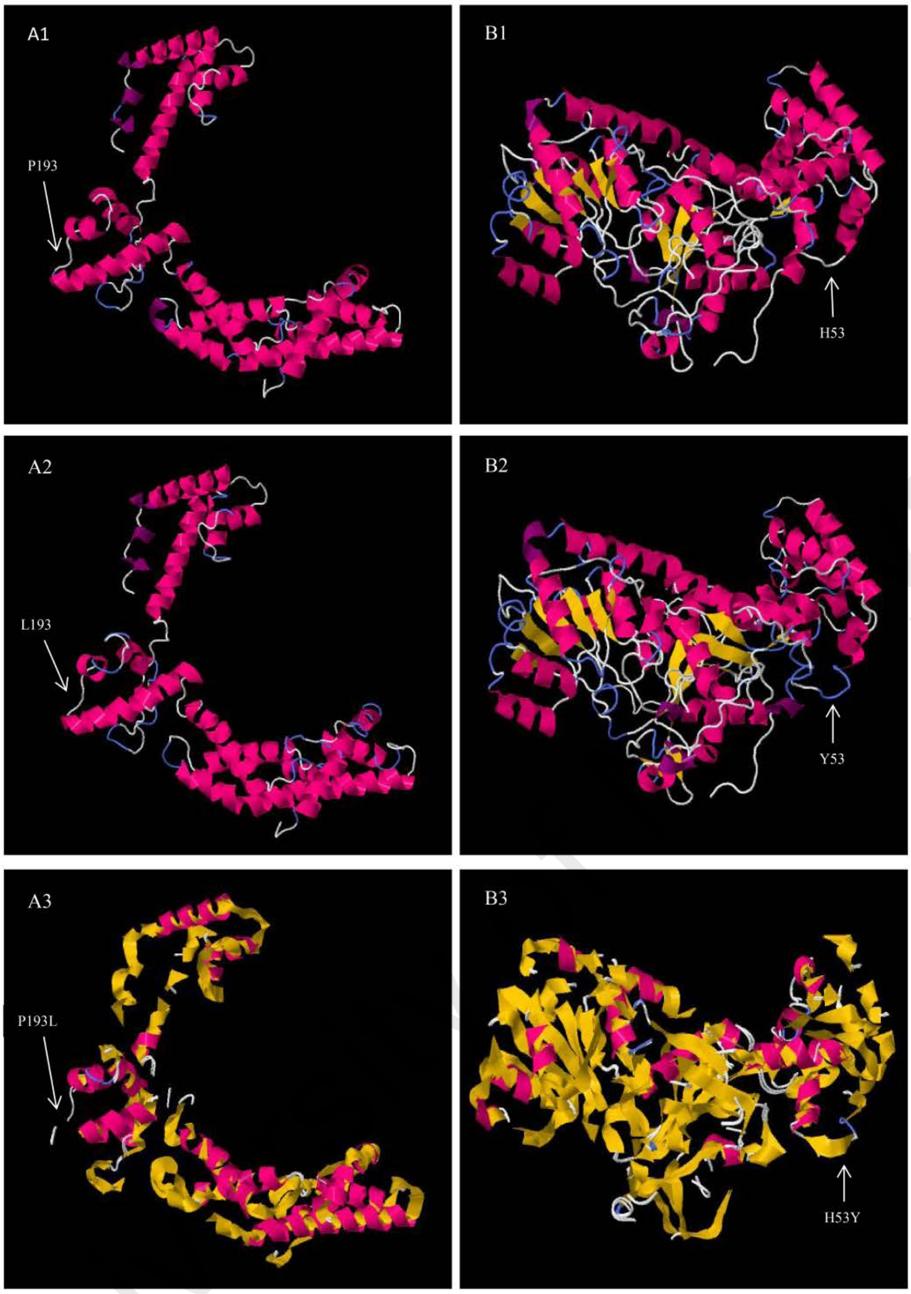

Figure 3.5: Modelled protein structures of RpoS and TviE. A1 showed native RpoS with proline at position 193. A2 showed mutant RpoS with amino acid leucine at position 193. A3 showed the superimposed structure of RpoS native structure (yellow) with mutant structure (pink). B1 showed native TviE with histidine at position 53. B2 showed mutant TviE with amino acid tyrosine at position 53. B3 showed the superimposed structure of TviE native structure (yellow) with mutant structure (pink).

However, the predictions regarding the nsSNP in TviE were contrasting, which suggest more benign effects, indicating that nsSNPs may have more modest effects on the functioning of this protein. The nsSNP in tviE may have decreased its stability (DGG value of -0.54 Kcal/mol, with <-0.5 kcal/mol indicating a large decrease in protein stability) as predicted by I-Mutant 3.0 (Capriotti *et al.*, 2008). Unlike RpoS, the mutant residue (tyrosine) in TviE is not conserved at this position of the helical structure, and other non-similar residues were observed at this position with no known protein binding sites in other homologous sequences. However, the size of the mutant residue (tyrosine) is larger compared with that of the native residue (histidine) despite the fact that both are polar and located at the surface of the protein. This mutation also introduces a more hydrophobic residue at this position, suggesting that it could possibly alter the correct folding of the protein, subsequently affecting hydrogen bonding but with only modest effects. These novel mutations could have important impacts on the pathogenesis and persistence of strains in the host. Although it is challenging to determine the true extent of the effects of the nsSNPs on the protein functions, these data suggest that they alter the protein structures (and possibly their functions) considerably, potentially leading to the enhanced regulation of RpoS and stress response in BL196 and reduced efficiency of TviE in the virulence capsular formation of CR0044. The close regulation of these two genes may be relevant to the virulence and persistence capabilities of the closely related strains that lead to the different clinical outcomes. These data provide essential insights into the underlying molecular mechanisms upon mutations and serve as caveats for future functional gene knock-out studies.

3.3 Conclusions

The genomes of *S*. Typhi in association with three important epidemiological settings were thoroughly dissected. Comparative genomics and phylogeny analyses have revealed that the strain that was associated with the large outbreak was highly related and shared common ancestry with the carrier strain. These findings are supported by their common genomic features and uncommon gene repertoires, including dispensable genes, phages and an additional putative pathogenomic island harbouring virulence-related genes, and *zot* in particular. Apart from these, variations were also identified in T6SS and SPI-related genes, insertion sequences, CRISPRs and nsSNPs among the studied genomes, which may be novel factors that contribute to the varied host adaptations and pathogenicities. Despite being highly similar, BL196 and CR0044 may be distinguished by microvariations in their nsSNPs. Interestingly, the protein modelling and MD simulation of the wild-type and mutant RpoS and TviE suggest that the potential protein structure and functional modification was more stable in *rpoS*, which plausibly leads to enhanced regulation and stress response. On the other hand, the mutation in TviE was less stable than that of the wild type, which could potentially lead to lower capsular formation efficiency. The close association of these virulence-related genes are relevant for long-term host persistence and adaptation, which serve as important caveats for further functional studies. The analysis also revealed that SPI10, which was previously thought to be relatively stable, is possibly prone to excision. Moreover, multiple regions of genomic plasticity were detected. In particular, the discovery of new GIs in the outbreak strain and the highly related carrier strain are of great concern epidemiologically. These results suggest the plasticity and open pan-genome of *S*. Typhi, indicating that the pathogen is more diverse than previously thought and that genes may have been acquired or transferred from one another through HGT, posing a higher risk for effective disease control. The genomic information that was obtained in this study provides novel insights into the pathogenesis and control of *S*. Typhi, essentially, gene targets for vaccine development.

CHAPTER 4

GLOBAL MLST AND COMPARATIVE GENOMICS OF RARE SEQUENCE TYPES OF *Salmonella enterica* SEROVAR TYPHI

4.0 Introduction

Typhoid fever poses a significant health threat to many endemic countries. *Salmonella enterica* serovar Typhi (*S*. Typhi), the etiologic agent, can be transmitted through contaminated food and water via the oral-fecal route. Annually, over 21 million cases and nearly 200, 000 deaths are reported worldwide (Crump & Mintz, 2010). Despite primary treatment and prevention efforts, the global typhoid cases remain very high (Murray, Vos, & Lozano, 2014). The disease is human-restricted, and the infected individuals could persist as long-term and asymptomatic human carriers, which in turn serve as the reservoirs for new infections and outbreaks (Gonzalez-escobedo *et al*., 2011).

The epidemiological investigation of *S*. Typhi is critical for disease control such as during a disease outbreak to trace the potential sources. Over the last few decades, many molecular subtyping methods have been applied to genotype bacterial pathogens. Among these, MLST is the most commonly used genotyping method to determine the ancestral lineages of various bacteria, including *S*. Typhi (Achtman, 2012; Leekitcharoenphon, Lukjancenko, Friis, Aarestrup, & Ussery, 2012). This method allows the discrete characterization of isolates by using the internal fragments of housekeeping genes sequences (Urwin & Maiden, 2003). The seven loci MLST scheme of *S*. Typhi was first utilized by Kidgell *et al.,* (2002) to determine the *S*. Typhi lineages based on the known sequence types (STs). However, MLST is of limited use in the monomorphic pathogen, notably for the *S*. Typhi, as their populations accrue little variations, thus rendering the efforts less useful in a population study. In the recent years, high-throughput whole-genomes sequencing (WGS) has become the ultimate approach to study the bacterial population structure and their phylogeny (Chen *et al*., 2013; Didelot *et al*., 2012; Sherry *et al*., 2013).

Based on the MLST, presumably, the *S*. Typhi of globally most widespread is genetically characterized as ST1 and ST2 in the earlier studies (Dahiya *et al*., 2013; Kidgell *et al*., 2002; Martínez-Gamboa & Silva, 2015; Zhang *et al*., 2011). However, this conclusion was drawn from the analyses of a limited number of samples. There is also a lack of data on the emergence of uncommon STs which may have been circulating, but remain undetected, nor is there a clear answer for the predominance of ST1 and ST2 globally. It is possible that these predominant and clonally related ST1- and ST2-typed strains have high transmissibility and/or enhanced virulence to circumvent killing by the innate immune system and probably acquired the ability to evade the host immune responses, allowing them to establish a long-term carriage stage. To date, there were only a few uncommon STs reported for *S*. Typhi. These include an ST8-typed isolate (422mar92 from Zaire, Africa, 1992) and ST3 (SARB64, Senegal, 1988), which were first reported by Kidgell *et al.*, in 2002. Since then, there was no further record for ST8, at least inferred from all currently available 7,089 *Salmonella enterica* entries in the MLST database (last accessed 10 October 2015). Therefore, it is extremely challenging to gain full evolutionary insights, particularly in regards to virulence and pathogenesis of the pathogen, without first comparing the populations that disseminate widely with the seemingly “near-to-extinct” populations in the evolutionary timescales of *S*. Typhi; and even more challenging if these missing-link populations hardly exist or being identified. To test this hypothesis, MLST was performed on both local and global *S*. Typhi strains, either experimentally derived through conventional MLST or WGS-derived, and scanned for any uncommon STs, and to establish the population structure of this pathogen from the global perspective of *S*. Typhi population. Upon identification, the genomes of the uncommon STs were compared against the predominant ST1- and ST2-typed genomes (three reference genomes and six of previously sequenced genomes). Other publicly available *S*. Typhi genomes were used as background comparison to provide an accurate framework to discern any unique variations carried by these rare populations. These findings provide valuable information that is essential for the understanding of the poor adaptation of *S*. Typhi with uncommon STs-typed, which may be otherwise capable of disseminating globally.

4.1 Material and methods

4.1.1 Local and regional bacterial strains selection

To determine the sequence types (STs) of *S*. Typhi in this region, 19 representative strains have been randomly isolated from various clinical outcomes (outbreaks, sporadic cases, and human carriers). The *S*. Typhi strains were obtained over three decades (34 years; 1983-2008) from 13 distinct geographical locations in three endemic countries (Malaysia, Chile and Papua New Guinea), which represent three large continents (Southeast Asia, Oceania, and South America). These strains include two each from Papua New Guinea and Chile while the remaining strains are from Malaysia. These strains have been previously described (Thong *et al*., 1994; Thong *et al*., 2002, Thong *et al*., 1997; Thong *et al*., 1996a; Thong *et al*., 1995; Thong, *et al*., 1996b) (Appendix L). Although the numbers of strains are relatively limited, they were selected from the best available representatives from the collection. The bacterial cultures were further cryopreserved as glycerol stocks at -80ºC in 50% glycerol for long term storage.

4.1.2 Genomic DNA extraction and MLST

Genomic DNA extracted from the overnight cell cultures in Lysogenic Broth (Oxoid, Hampshire, UK) according to the manufacturer’s protocol (Promega Corporation, Madison, WI, USA) was used as PCR template. Housekeeping genes, *thrA* (aspartokinase+homoserine dehydrogenase*)*, *purE* (phosphoribosylaminoimidazole carboxylase), *sucA* (alpha-ketoglutarate dehydrogenase), *hisD* (histidinol dehydrogenase), *aroC* (chorismate synthase), *hemD* (uroporphyrinogen III cosynthase) and *dnaN* (DNApolymerase III beta subunit) were targeted for the MLST scheme (Kidgell *et al.*, 2002). PCR assays were carried out with ~50 ng of DNA template, 150 µM (each) deoxynucleoside triphosphates, 1 × PCR colourless buffer, 1.2 mM MgCl2, 0.2 µM of primer, and 0.5 U of Go Taq Flexi DNA Polymerase (Promega, Madison, WI, USA) in 25 µl reaction mixtures. PCR conditions used are as follows; initial denaturation at 94°C for 30 s, 30 cycles of denaturation at 95°C for 30 s, primer annealing at 55°C for 30 s, and extension at 76°C for 30 s; and a final extension at 75°C for 2 min. Reactions were performed using a PCR Master Cycler (Eppendorf AG, Hamburg, Germany). Products were separated by 1.5% agarose gel electrophoresis and visualized with GelRed (Biotium Inc., Hayward, CA) staining and UV illumination with a gel documentation system (Gel Doc 2000; Bio-Rad, Hercules, Calif). Primers used for the PCR assays are listed (Appendix N). PCR products obtained were purified using PCR purification Kit (MEGAquick-spin^{TM}, iNtRON Biotechnology, Seongnam, Korea) according to manufacturer's instruction. The purified PCR products were submitted to a commercial sequencing facility (First Base Sdn Bhd) for sequencing.

Primers used for the sequencing reactions are listed in Appendix N. The seven sequenced loci of 19 local and regional *S*. Typhi strains were first trimmed, edited and aligned using MEGA 6 (Tamura, Stecher, Peterson, Filipski, & Kumar, 2013). Sequences were subsequently submitted to the MLST database (http://mlst.warwick.ac.uk) and assigned existing or novel allele type numbers. The composite sequence STs were defined by the database based on the set of allelic profiles derived from each of the seven loci. Allele sequences for each strain were then concatenated to a final sequence length of 3,336 bp.

4.1.3 MLST using whole genome sequences of global strains

To perform MLST for global strains, all accessible 1,814 whole genome sequences of *S*. Typhi (as of 21 September 2015) and 2 *S*. Paratyphi A strains and their background data were retrieved from NCBI Genome databases via anonymous file transfer protocol (FTP) at ftp://ftp.ncbi.nih.gov/genomes/Bacteria/ or manually downloaded from NCBI. The accession numbers of the reference, the previously sequenced, and published genomes are listed (Appendix M; Appendix O). The newly released genomes sequences, raw sequence data, and accession number can be accessed as a batch via European Nucleotide Archive accession number ERP001718 (Wong *et al*., 2015). The results of the ST assignment and unconfirmed allelic profiles were manually evaluated. In brief, all of the whole genome sequences were first re-annotated and validated as previously described (Yap *et al*., 2012a; Yap *et al*., 2012b) for standardization. The nucleotide sequences were then aligned against MLST database (http://mlst.warwick.ac.uk) and assigned existing or novel allele type numbers. For validation, the whole genome sequences were also submitted to Centre for Genomic Epidemiology MLST 1.8 (https://cge.cbs.dtu.dk/services/MLST/) (Larsen *et al*., 2012) to assign ST based on the closest matches against the MLST databases. The composite sequence types (STs) were defined by the database based on the set of allelic profiles derived from each of the seven loci.

4.1.4 MLST Phylogenetic and Data Analyses

All 1,783 concatenated sequences of MLST (local/regional strains and genome-derived) (Appendix O) with 100% allele recovery, including the allelic sequences of two outgroups sequences extracted from *S*. Paratyphi A complete genomes [*S*. Paratyphi A AKU_12601 (FM200053.1) and *S*. Paratyphi A ATCC 9150 (CP000026.1)] were aligned with MAFFT (Katoh & Daron, 2013) with parameter E-INS-i and a phylogenetic tree was constructed by approximate maximum-likelihood using FastTree2 with the generalized time-reversible model and Bayesian mixture model (GTR+CAT) (Price *et al*., 2010). The statistical significance of phylogeny was estimated by bootstrap analysis with 1000 pseudoreplicates. Allele numbers and sequence types (ST) of this study were deposited in the publicly accessible *Salmonella enterica* MLST database (http://mlst.warwick.ac.uk.). To determine the genetic polymorphism of *hemD* among *Salmonella enterica*, all 333 *hemD* (864 bp) trimmed gene sequences were retrieved from MLST *Salmonella enterica* database (http://mlst.warwick.ac.uk). A multiple sequence alignment was then built with CLUSTALW in MEGA 6 (Tamura *et al*., 2013).

4.1.5 Comparative Genomics Analysis of *S*. Typhi with uncommon and predominant STs

After the MLST analyses, the eight genomes subtyped with uncommon ST8 (76-1292, E01-5741, 05-8683, 206926, MDUST177, 627334), ST2233 (np45), ST2359 (ST821/98) were retrieved and studied in detail. The assembled genomes were annotated and mapped to the reference genomes (CT18) as previously described (Ho *et al*., 2012; Yap *et al*., 2012). The tRNA and tmRNA were predicted using Aragorn (Laslett & Canback, 2004) whereas the rRNA was predicted with rRNAMMer (Lagesen *et al*., 2007), and manually validated as described earlier (Suhaimi *et al*., 2014; Osama *et al*., 2012). To compare, three reference genomes and six of the published sequenced genomes (described earlier), representing ST1 (Ty2, P-stx-12, Ty21a, BL196, CR0044) and ST2 populations (CT18, CR0063, UJ308A, UJ816A, ST0208) have been included. The pan-genome data of these genomes was obtained using PGAP (Zhao *et al*., 2012) and compared as previously described (Yap *et al*., 2014). The phage regions of the pan-genomes identified were predicted and confirmed with web server PHAST (Phage Search Tool) (Zhou *et al*., 2011) and intact phage regions were manually examined. For plasmid analyses, the plasmid replicons were detected using PlasmidFinder 1.3 database (Carattoli *et al*., 2014) and plasmid MLST was determined using pMLST1.4 (Carattoli *et al*., 2014) and validated with manual inspection. The acquired antimicrobial genes were identified and determined using ResFinder 2.1 (Zankari *et al*., 2012). To understand other possible genetic events that shaped the rare STs populations, CRISPRs (Clustered regularly interspaced short palindromic repeats) sequences and Restriction-Modification (RM) system prediction analyses using CRISPRFinder (Grissa *et al*., 2007) and Restriction-ModificationFinder1.1 (https://cge.cbs.dtu.dk/services/Restriction-ModificationFinder/), respectively, were performed. For virulence genes and

Salmonella Pathogenicity Islands (SPIs) analyses, the annotated sequences were mapped against Virulence Factor Database (VFDB) and Pathogenicity Island Database (PAIDB) and KEGG pathway (Chelvam *et al*., 2015), respectively using BLASTn (>98% identity and 60% coverage, E-value < $1x^{-10}$) and Artemis (Carver *et al*., 2008).

4.1.6 Robust phylogenomic analyses using 1,808 global *S*. Typhi genomes

Phylogenomic analyses were conducted to understand better the phylogeny of the uncommon STs in the perspective of all *S*. Typhi genomes. To compute large genome samples is computationally demanding, time-consuming and technically challenging. Therefore, a newer approach of alignment-free algorithm was utilized by using the andi v1.4 (Haubold, Klötzl, & Pfaffelhuber, 2015) to compute rapidly the large-scale evolutionary distances among 1,808 *S*. Typhi (~8.6 Gbases) (six genomes were excluded from phylogenetic analyses due to low quality) and bias from mapping against reference genome can be avoided. The approach based on a new distance measure, d_a, which approximates local alignments by anchoring them with long, unique matches of a minimal length. The matches are equidistant in the query, and the subject is equivalent to restricting the analysis to ungapped alignments. These exact matches were then searched efficiently using enhanced suffix arrays. The anchor distances and arrays were computed using the multithreaded UNIX command-line in the andi v1.4 package (Haubold *et al*., 2015). The resulting data was aligned, analyzed, transformed and the phylogenetic tree (NJ method) was constructed with SplitsTree4.12.6 (Huson & Bryant, 2005). A rooted tree was inferred using the *S*. Paratyphi A AKU 12601 and *S*. Paratyphi A ATCC 9150 as outgroup. The approach was repeated for 1,806 genomes without the outgroup to generate an unrooted tree. To enhance the visualization of the gigantic tree, the tree was re-rooted with an *S*. Typhi strain from the earliest *S*. Typhi cluster to the outgroup. The node to *S*. Paratyphi A was also collapsed (a very long branch length) to improve visualization of the massive tree.

4.2 Results

4.2.1 High allele recovery rate in currently sequenced *S*. Typhi genomes and STs identification from both global and local/regional data

Composite allelic profiles were successfully recovered from 1,762 (97%) *S*. Typhi genomes with 100% sequence coverage and 43 (2.3%) genomes using the top BLAST hits criteria (>99 % sequence coverage, E-value cut-off $<1\times10^{-10}$). The remaining profiles were predicted using the top BLAST hits criteria (coverage less than 99%) against MLST database, which accounted for only less than 0.5% of the genomes studied. Out of the 1,827 global and local/regional strains, 1282 (70.2%) and 536 (29.3%) were subtyped as ST1 and ST2, respectively (one unknown ST strain) (Appendix O). The findings showed that ST1 and ST2 have been homogenously circulating for over the period of 109 years and represent the current predominant populations globally. Both STs were detected as early as in the year 1905; the oldest strain included in this study. Likewise, the local/regional strains exhibited the same pattern of ST1/2 dominance throughout a period of 30 years. It was also attempted to detect the geographical structuring of both STs in three levels (continents, region within continents and country), but a very limited evidence was found except for the archipelago of Samoa (116 ST1 and 1 ST2) and Papua New Guinea (1 ST1, 48 ST2) of the Oceania regions, in which a clear skew towards one ST was observed. No association was found between STs and source of isolations in both global and local/regional collection. Interestingly, only six strains (76-1292, E01-5741, 05-8683, 206926, MDUST177, 627334) were subtyped as ST8, a rare ST but common in the African regions (Central, North and South Africa).

Remarkably, ST8 was undetected outside of African continent for more than a century. Although this population was geographically-restricted and seemingly unsuccessful in its dissemination to other parts of the world, the latest ST8-typed strain identified in this study was dated 2012. Since the oldest strain was recovered as earlier as the year 1976, hence, these strains have been circulating in the community for at least 37 years, indicating an establishment of the long-term local reservoir in the African region. It is challenging to confirm whether or not this population has long existed before the aforementioned periods, or represents the ancestral population to ST1 and ST2. Nevertheless, the containment of ST8-typed population only within the African continent but nowhere else is consistent with the notion of the early migration of humans out of Africa, and thus the spread of the pathogens, as proposed by Kidgell *et al.,* (2002). As the bacterium accrues very limited variation, it is highly possible that these two populations (ST1/ST2 and ST8) shared a very recent common ancestor which later diverged into two distinct populations. Apart from ST8, two uncommon STs, ST2233 (np45) and ST2359 (ST821/98) from South Asia (Nepal) and South America (Argentina), respectively, were identified.

4.2.2 High sequence conservation of *S*. Typhi from various endemic regions of typhoid

All seven MLST loci were successfully recovered for the 1,781 *S*. Typhi strains that were obtained from diverse geographical regions (7 continents, 65 countries) and sources of isolation (blood, stool, urine, and environment) over a period of 109 years. The sizes of the trimmed sequenced alleles of all seven loci ranged from 399 bp to 501 bp. The concatenated sequence amounted to 3336 bp, an approximate 0.07% of the size of *S*. Typhi complete genome (with reference to the CT18 complete genome). The pairwise allelic alignments showed that the average sequence divergence of these loci was very low, which were contributed mainly by *hemD*, which delineates two large populations (ST1 and ST2). The *hemD* gene exhibited only two allelic profiles, 1 and 2 (single synonymous nucleotide polymorphism). The SNP of the *hemD* occurs at position 129 of the trimmed partial sequence where cytosine in hemD2 replaces the thymine base of hemD1. Further investigation on the diversity of *hemD* genes by including a total of 333 *hemD* genes from *Salmonella enterica* (retrieved from MLST database (http://mlst.warwick.ac.uk) revealed that the polymorphic site (C129T) of the *hemD* is conserved across all members of the group, except for the polymorphism carried by *hemD*1 of *S*. Typhi (Supplementary File 4).

The global phylogenetic analysis of *hemD* gene from Enterobacteriaceae members indicated that the *hemD*1 allele may have diverged simultaneously or later from the MRCA, given the fact that the *hemD*2 allele is conserved throughout the *Salmonella* spp. and its closely related species, *E. coli* and *Shigella* spp. Although polymorphisms were also detected in *hisD*, *thrA*, and *dnaN* of the uncommon STs (ST8, ST2233, ST2359), but these allelic frequencies were extremely small, in which they differed by only one SNP each (ST8-*hisD*1 to *hisD*3; ST2233-*dnaN*1 to *dnaN*478; ST2359-*thrA*5 to *thrA*545). To understand the phylogeny of these STs, an MLST phylogenetic tree of all the *S*. Typhi strains was inferred using the maximum-likelihood approach from the concatenated sequences. Two main clonal clades were observed to harbor large numbers of ST1 and ST2 strains, which appeared to be a very recent divergence from its common ancestor. The ST8-typed population is phylogenetically close to ST2, whereas ST2359 forms a polytomy with ST1. However, the node containing ST2233 that is closer to the root is polytomically unresolved from ST1/ST2 clades. Although the bifurcation of ST1/ST2 may be the result of the addition of ST2233-typed strain, the internal nodes of ST8 and ST2359 remain unresolved, most likely the result of short divergence and/or low discriminatory power of MLST (Figure 4.1). Such clonal relationships would be expected for a recent bottleneck that allowed only a few clones to survive, possibly through purifying selection, such as those commonly observed in other monomorphic bacteria.

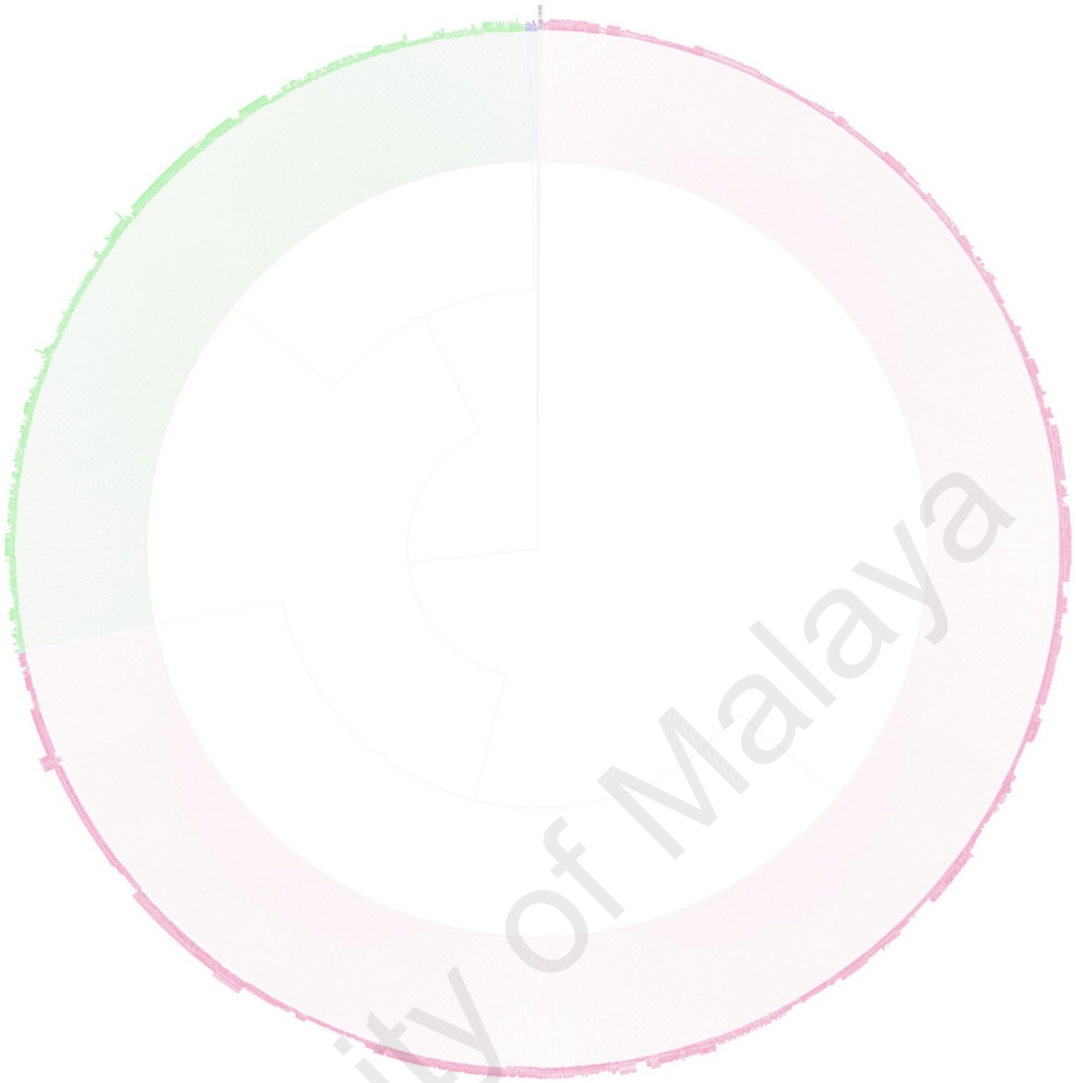

Figure 4.1: *Maximum Likelihood tree shows the genetic relationships of 3,336 bp concatenated MLST genes sequences derived from the seven housekeeping loci of 1,783 *S*. Typhi strains isolated globally. The tree is rooted with S. Paratyphi A. The colors of the strains label represent the STs; red for ST1, green for ST2, blue for ST8, pink for ST2233, light blue for 2359 and black for outgroup. All branches have more than 90% support.

* As the diagram consists of 1,783 nodes and fine lines, the clearer and detailed version of the image (individual nodes with detailed strains labelling) is not included in the text due to extremely large data size (9.7 MB). The original image (with different file formats) is attached with the thesis in softcopy version for better visualization.

4.2.3 Phylogenomic Analyses revealed strong local phylogeographical signals

To view the phylogenetic relationship of these strains with much higher resolution, a whole-genome based phylogenetic tree was built from 1,808 *S*. Typhi genomes (Figure 4.2). This robust phylogenomic tree shows that all *S*. Typhi is monophyletic and distinct from its closest counterpart, *S*. Paratyphi A with long branch length, reiterating the classical phylogeny of these two serovars, which diverged from each other earlier through convergent evolution (Holt *et al*., 2009). In this phylogeny, there is a clear separation from early diverging basal groups and the core groups that diverge later. Remarkably, these earliest basal groups closest to the root consisting of *S*. Paratyphi A are almost entirely isolated from Africa, which supporting the "theory of human-pathogen co-evolution" since the early human migration out of Africa. Moving further from the roots, the pathogen seems to have widely propagated across various regions, evidenced by several sub-clusters isolated from Africa, South America, Asia and Australia/Oceania. The frequencies of these basal groups are relatively low, suggesting that these groups may have been under-sampled or reflect true disappearance, plausibly as a result of purifying selection. In contrast, the more recent core clade experienced a radiation that resulted in at least three large clusters with many sub-clusters observed. Interestingly, strong signals of phylogeographical structuring were detected in which some sub-clusters were recovered from the same geographical regions, despite being separated for years, highlighting the importance of local reservoir in maintaining endemicity in those regions. This observation corroborates with the earlier study (Yap *et al*., 2014) in which a human carrier strain (CR0044-isolated from a food handler) was indeed highly related to the outbreak strain (BL196-isolated from a large outbreak in Malaysia), emphasizing the propensity of human carriers to trigger future outbreaks in the endemic regions. Notably, the ST8-typed population appeared to be ancestrally related to the African and Asia sub-cluster, in which it diverged from the common ancestor to form a distinct terminal cluster. Interestingly, it was observed that ST8-typed population shared the MRCA with a sub-cluster exclusive to Australia/Oceania strains, suggesting that clonally related local strains are circulating in the regions. In contrast, the ST2359-typed ST821/98 was found highly related to the reference genome Ty2 (isolated from Russia) and its derivative vaccine strain Ty21a. Other strains in the sub-cluster of ST821/98 were in fact geographically related, such as those few isolated from Russia and Europe. To note, the close relative of ST821/98; the attenuated Ty21a and its parental strain, Ty2 were both reported to share the same genetic mutant that partly contribute to attenuation of the live vaccine. The ST2230-typed np45 strain was in the major cluster together with the CT18 and ST0208 (isolated from Southeast Asia). The sub-cluster in which np45 located was geographically restraint to mainly South Asian regions (India and Bangladesh). This sub-cluster was ancestrally related to some other smaller clusters that contain uniformly Asian and African strains, to some extent, indicative of population displacement by the relatively recent clones. In comparing the major groups of the phylogeny, the strains were separated by such short internal branch length that they appeared to have diverged nearly simultaneously. An extremely shallow branch separates most of the sub-clusters, indicating shared common ancestors, and apparently was quite successful in its dissemination, illustrated by its appearance over several continents. Notably, few strains from the Australia/Oceania regions were detected in the terminal of the relatively longer branching of the deep node, suggesting divergence of variants from the existing nodes, perhaps arose in response to intense selection pressures.

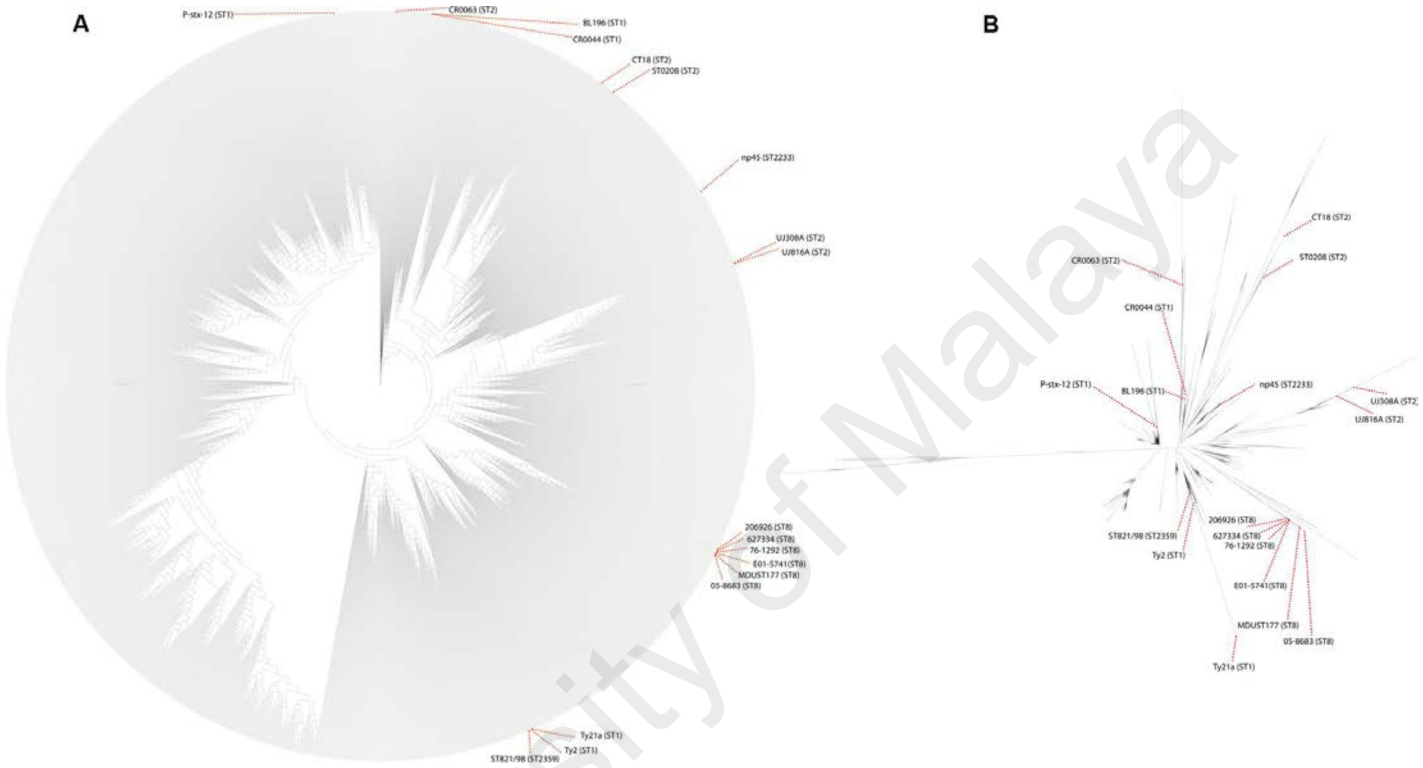

Figure 4.2: *Alignment free-based phylogenomic tree generated from 1,806 *S*. Typhi and 2 *S*. Paratyphi A genomes. All reference, the previously sequenced and the uncommon ST-typed genomes with their STs were labelled. Other strain labels were removed to enhance visualization. A.The phylogenomic tree was rooted with *S*. Paratyphi A. The node to *S*. Paratyphi was collapsed (very long branch length) to enhance visualization of the large tree. B. Unrooted tree generated from 1,806 *S*. Typhi genomes.

* As the diagram consists of 1,806 nodes and fine lines, the clearer and detailed version of the image is not included in the text due to extremely large data size (15 MB). The original image (with different file formats) is attached together with the thesis in softcopy version for better visualization.

2.4 Comparative genomics of rare STs strains revealed unique gene signatures

Because of the limited dissemination of the rare STs, the ST8-typed strains, in particular, which are not found in other countries outside the African region, the studies attempted to elucidate the genomic signatures, which might present potential key factors contributing to their poor dissemination. The differences between the seemingly unsuccessful and successful population may explain the evolutionary events that are required for successful infection, transmission and/or maintenance of carriage state. As the reservoir of *S*. Typhi is mainly human carrier, the low prevalence of this rare STs may reflect its poor adaptation in the host (carriage state), which is required for wide dissemination. This postulation is supported by the fact that the ST 8-typed population was not identified outside of African region, consistent with the route of early human migration out of Africa. In contrast, although ST2233 and ST2359 were rarely identified but their presence outside of African region may reflect new emerging variants. As the current knowledge of *S*. Typhi is solely based on the predominant STs, therefore, comparative genomics analyses were performed to dissect any genetic signatures of this important population. In terms of basic genomic features, the genome sizes of ST8-, ST2233- and ST2359-typed isolates ranged from 4.69 Mb to 4.85 Mb with an average GC content of 52.1%. The CDS contents of these genomes ranged from 4839 to 5046 with an average tRNA and rRNA of 71 and 5, respectively. Every genome also carries one tmRNA each. The basic genetic features of these genomes are similar to many previously described genomes (Deng *et al*., 2003; Holt *et al*., 2008; Parkhill *et al*., 2001). High genomic synteny and conservation, with limited evidence of recombination, were also observed in the uncommon STs-typed strains, which shared similar findings with other previous studies (Holt *et al*., 2008; Yap *et al*., 2014)

SPIs are unique pathogenicity regions with high numbers of virulence factors (VFs) carried by *Salmonella enterica*, which facilitate the bacterium in host colonization, invasion and maintenance of virulence in the host (Baker & Dougan, 2007; Coburn, Grassl, & Finlay, 2007; Dougan & Baker, 2014). SPI-7, the unique virulence region of *S*. Typhi was found intact in all uncommon ST-typed strains. Further, all the strains commonly shared the 227 out of 242 VFs compared. The remaining VFs were shared by at least two genomes except for the strain 76-1292, which carries a unique gene, *astA*, which encode for heat stable enterotoxin. All the compared genomes carried the hypothetical virulence gene, *t0576*, which was absent in CT18. Interestingly, the *tsaC* was only present in the ST8-typed strains and CT18 but lacks in other genomes. Five out of six genomes of ST8-typed strain lack *pilV* but carry the *pilV2*, an alternate and duplicated gene of *pilV*. Notably, four out of six ST 8-typed genomes carried a *seD* gene which was absent in all other genomes. On the contrary, *STY2517* was the only unique gene present in CT18 but absent in others. Notably, Np45 genome which was subtyped as ST2233 lacks VFs needed for secretion machinery such as *sciT*, *ssaV*, and *sthB* although other virulence genes in the respective operons were fully intact, which likely indicates a true absence rather than sequencing artifacts. The VFs with the highest variation were noted in an outer fimbriae gene, *sefC*, although other VFs are highly conserved.

Interestingly, variations in several essential genes were detected, which are involved in virulence and pathogenesis. Synonymous point mutations were identified in three virulence genes; *flgE* (unique SNP), *pipB* (non-unique SNP), *spaO* (non-unique SNP). A numbers of shared non-synonymous point mutations also identified in molecular chaperone *clpB* (C338R), fimbriae-like periplasmic protein *Sfm/fimF* (V64A), E3 ubiquitin--protein ligase/*sseJ* (K713E), *hilD* (K209E) and *tviE* (K266N), although not unique to six of the ST8-typed genomes but was identified only in less than 2% of the 1,808 genomes studied. Notably, all ST8-typed genomes shared the identical unique non-synonymous point mutations in three important virulence genes, the *flhB* (E378G), *sipC* (G176S) and *tviD* (R209Q), which were not identified in all other ST1- and ST2-typed populations. In addition, deletions were also detected in two virulence genes, *fimI* (8 bp deletion) and pseudogene *misL* (15 bp deletion), which may lead to a frameshift and premature stop codon. However, no unique SNPs of VFs were found among all eight genomes, except for the *tviD* which was shared by ST821/98 (ST2359), suggesting that they may have undergone different evolutionary paths.

4.2.5 Antimicrobial resistance genes analyses

Of all six ST8-typed strains that were obtained from the African region, only two (76-1292 and 05-8683) acquired antimicrobial resistance genes. In fact, both the 76-1292 and 05-8683 are predicted multidrug-resistance. The 76-1292 strain harbors *sul1* (sulphonamide resistance), *aadA1* (aminoglycoside resistance), *blaOXA-1*, *catA1* (phenicol resistance) and *tetA* (tetracycline resistance) genes, which were found on IncI1 (pMLST 186) and IncP plasmids. In contrast, 05-8683 carried a rare *ereA*, a gene that is associated with azithromycin resistance (macrolide), apart from other common antimicrobial genes of *catA1* (phenicol resistance), *sul1* (sulphonamide resistance), *tetB* (tetracycline resistance), *dfrA5* (trimethoprim), *blaTEM-1B* (beta-lactam), which were found located in a multireplicon IncF plasmids with pMLST of [F1:A-:B49]. However, both 76-1292 and 05-8683 lacks the IncHI1 plasmid, a replicon type which is commonly associated with the widely disseminated H58-haplotyped *S*. Typhi (Wong *et al*., 2015). On the contrary, no acquired antimicrobial genes were identified for np45 (ST2233) and ST821/98 (ST2359) strains.

4.2.6 Limited evidence of lesion in CRISPRs, Phages, and Restriction-Modification (RM) System

Phage elements were thought to be one the main driving forces of evolution in *S*. Typhi and *Salmonella enterica* (Holt *et al*., 2008); however, limited evidence of novel intact phages was detected in all eight ST8-, ST2230- and ST2359-typed strains. In fact, the shared eight regions are highly similar with only slight sequence variations (attributed mainly to the gain/loss of hypothetical and phage structural genes), as identified in the reference genomes. The CRISPRs, a known prokaryotic adaptive immune system that resists invading nucleic acids, may cause DNA mutations and damage (Li *et al*., 2014). In this analysis, CRISPR spacers and length, direct repeat length, and consensus sequence were commonly shared among genomes, with very little to no variation. An identical six spacers with direct repeats length of 29 bp and with three variants (385 bp, 394 bp or 420 bp) of CRISPR length were detected. Although CRISPR has been found to be correlated with STs in some study for subtyping, at least for *Salmonella enterica* (Fabre *et al*., 2012), very limited evidence of variation was detected in *S*. Typhi to be of valuable use in subtyping using CRISPRs. Besides, the RM systems represent known barriers to the entry of foreign DNA. Yet, the RM system is poorly understood in *S*. Typhi. In this study, it revealed the extensive repertoires of RM systems, which contain four predicted RM systems (Type I, Type II, Type III and Type IV) in all studied genomes. The Type I system of *S*. Typhi, which includes genes encoding methyltransferase (*M.SptAIII*) and specificity subunits (*S.SptAIII*) was highly similar to that of *S*. Paratyphi A ATCC9150, whereas *EcoKI*, which codes for restriction enzyme, is homologous to *E.coli* K-12. The genes (*M.SenAboDcm*, *M.SenSPBIII*, *M.Sen158III*) in Type II code for methyltransferases, which are commonly identified in various *Salmonella enterica*. Type III has both the restriction and methylation domains; the genes *SenAZII* (Restriction enzyme) and *M.SenSPBII* (methyltransferase) were of homologues of *S. enterica* subsp. arizonae serovar 62:z4,z23:-- and *S*. Paratyphi B SPB7, respectively. For Type IV, only *StyLT2Mrr* was identified, which involves in methyl-directed restriction as carried by the closely related *S*. Typhimurium LT2. The comparative genomics data demonstrated that the RM systems were intact with high similarities, suggesting no functional lesions. Notably, two and only plasmid-bearing ST8-typed *S*. Typhi strains (05-8683 and 76-1292) carried an additional Type 1 *EcoR124I* encoding restriction enzyme (*E.coli*) and Type II *M.EcoGIX* encoding methyltransferase (*E. coli* O104: H4 str. C227-11), highlighting the potential role of RM systems in plasmid acquisition.

4.3 Discussion

In the present study, the MLST data collection for *S*. Typhi, to date, is the largest in numbers and were from the most diverse representatives' available (geographical regions, sources) over a period of 108 years. The practicality of scanning thousands of bacterial genomes was successfully demonstrated. With this scale of data, the findings showed unequivocally that ST1 and ST2 are the two most predominant STs globally and are highly successful in dissemination since their emergence. These results concurred with the previous analyses, which relied on limited numbers of *S*. Typhi strains (Dahiya *et al*., 2013; Kidgell *et al*., 2002; Wang, Pan, Zhang, & Zheng, 2009; Zhang *et al*., 2011). The populations of ST1 and ST2 have been expanding drastically since their emergence from the MRCA and are still co-circulating in the population for at least a century.

S. Typhi, a human-restricted pathogen has long been associated with human migration as well as co-evolution with humans (Kidgell *et al*., 2002). In this study, no evidence of bias in the spatial and temporal distribution of STs was identified from the

MLST data, suggesting that the rates of human migrations across continents have been extremely high to disseminate equal genotypes across the human populations. The *S*. Typhi population may have suffered a recent bottleneck during its course of evolution, evidenced by the limited STs and loci recombination. However, the evolutionary signals from the MLST alone are also rather weak, probably owing to its short evolutionary distance which supports the previous age estimates (~50,000 years old) (Kidgell *et al*., 2002) of this relatively "young pathogen". Similar observations have been observed in other monomorphic pathogens such as *Mycobacterium tuberculosis*, the etiologic agent of tuberculosis (Achtman, 2008; Gagneux, 2012).

It is interesting to note that the "primitive" ST3 was completely lost in the population (at least from this data) since the first report of its emergence in Africa in the year 2002 (Kidgell *et al*., 2002). The enigmatic disappearance of this ST, perhaps due to the transient Darwinian selection that rendered the bacteria less fit as a result of harmful mutations accumulated over evolutionary time. However, the possibility of genetic drift or/and the effects of environmental changes such as those observed in its close relative, *S*. Paratyphi (Zhou *et al*., 2014) cannot be ruled out. Indeed, the understanding of such evolutionary dynamics is lacking without constructing a genealogy of *S*. Typhi with the inclusion of all populations over an extended period of time.

Whilst the identification of uncommon STs were reported intermittently, such as ST890 and ST892 in China (Zhang *et al*., 2011) and ST1856 in Indonesia recently (Martínez-Gamboa & Silva, 2015), these events are extremely rare, and possibly reflects the variants from the existing clones. A similar phenomenon was also observed for ST2233 and ST2259 in this study. On the contrary, the population of ST8 is likely to have poor dissemination and adaptation in humans, in view of its low prevalence globally, as well as being geographically-restricted. It is reasonable to speculate that, the population may still maintain the ability to cause human carriage, perhaps poorly, given the fact that the population has been around for at least 37 years. This was further supported by the WGS phylogenomic analyses, which revealed that the ST8 population is likely to diverge from the shared common ancestors of ST1- and ST2-typed population, before substantially diluted by the expansion of ST1 and ST2-typed *S*. Typhi out of Africa.

The findings from the comparative analyses of the ST8-, ST2230-, and ST2359-typed genomes relative to ST1- and ST2-typed genomes could be the key factors underpinning the successful traits of predominant populations. As the study was attempted to interrogate the genomic signatures of these genomes, very limited evidence of marked variations were found in regards to gene gains/loss. The SPI-7, which is implicated as the primary virulence repertoires of *S*. Typhi, was completely intact in all *S*. Typhi with uncommon STs, in addition to other SPIs. This finding is consistent with an earlier study that SPI-7-bearing *S*. Typhi was less invasive than the one with the entire SPI-7 missing (Bueno *et al*., 2004), suggesting that the presence of SPI-7 may be relevant to the fitness of the bacterium. Besides, the genomes harbor very limited numbers of strain-specific genes, which are mainly hypothetical or phages-related. Similar observation was also noted in phage regions, whereby the sequences are highly conserved with little to no variations, except for the few variations punctuated in between the intact phage regions, suggesting the possible hotspots for genome variations. Similarly, the conservation of CRISPRs regions in *S*. Typhi of different STs detected may reflect common defense mechanism acquired by the isolates to fight off incoming foreign DNA. The abovementioned findings are of little surprise as this clonality is well documented in previous studies (Holt *et al*., 2008; Roumagnac *et al*., 2006). To probe further, RM system analyses was performed, to detect for any genomic peculiarity of this population. Although the data indicated no lesion identified in the RM systems; it is worth mentioning that both the plasmid bearing genomes carried an additional RM system in comparison with other studied genomes. The presence of these RM systems may facilitate the intake of plasmids/acquired resistance. Such phenomenon has been observed in various bacteria (Oliveira, Touchon, & Rocha, 2014), but its function in *S*. Typhi is so far, remains poorly understood, particularly its role in moderating virulence of the pathogen in the human host. The presence of plasmids in the 76-1292 and 05-8683, confer these strains with MDR phenotype, which may enhance the virulence of these strains. Nonetheless, these plasmids may grant little fitness advantage since the ST8-typed population with the MDR traits were missing in the subsequent isolations. It is likely that the presence of plasmids for maintaining the MDR traits may incur additional fitness burden on the already-less-fit bacterium, thus impacting the rate of successful infection and transmission. Such notion is also implicated in some *S*. Typhi population in which the plasmids were chromosomally integrated (Wong *et al*., 2015), possibly in the similar efforts to reduce fitness burden.

Intriguingly, the analyses indicated that the salient differences between the ST8- and ST1/ST2-typed populations are the presence of several non-synonymous mutations in few key virulence genes, such as the *flhB*, *sipC*, and *tviD*, which are uniquely present within this ST8-typed population. The affected gene, *sipC*, which encodes for a type III effector protein of the SPI-1 involved in translocation and active modulation, which has been implicated in host invasion. Previous studies have demonstrated that *Salmonella* mutant strains lacking the *sipC* were less invasive (Myeni & Zhou, 2010). An earlier study of *flhB* in *S*. Typhimurium showed that the mutation in the gene (structural gene for hook-associated protein 1; HAP1) resulted in polyhooks and altered flagellar hook length, which may have a severe impact on motility and adhesion of the bacterium (Hirano, Yamaguchi, Oosawa, & Aizawa, 1994). The virulence of *S*. Typhi often correlates well with the presence of the Vi antigen (Arricau *et al*., 1998; Wetter *et al*., 2012). The mutation found in *tviD*, which is required for biogenesis of the Vi polysaccharide, may potentially affect the virulence of the pathogen.

The genomes of ST8 also harbor numbers of non-synonymous mutations in several virulence genes (*clpB*, *Sfm/fimF*, *sseJ*, *hilD*, *tviE*) which are involved in invasion, colonization, virulence and regulation in various stages of pathogenesis. Although these genes are not uniquely acquired, but may be related to the transitionary mutations occur along the evolutionary timeline in gaining fitness advantages. Notwithstanding, it is important to note that although ST1 and ST2 are seemingly successful, the extent of its fitness remains unclear. The mutations may be accumulative rather than occurring at a single event since some of these mutations were strain-specific, but others are commonly shared unique mutations. Further, the loss of function by premature stop codon or truncation from deletion/insertion in the virulence genes of *S*. Typhi is quite rare, but the deletion in *fimI* and *misL* may suggest a selective advantage gained for such aberration. The deletion in *fimI*, a gene in type I fimbrial operon, may affect the biofilm formation ability of *S*. Typhi, as reported in some studies of the closest model, the *S*. Typhimurium (Teplitski, Al-Agely, & Ahmer, 2006). The *misL* (located at SPI-3) gene, which is an important colonization factor in the intestinal persistence during the infection of *S*. Typhimurium was disrupted by deletions. Although some essential genes such as *misL*, may be annotated as a pseudogene, the functionality of these pseudogenes in *S*. Typhi remains unclear. For example, *shdA* which was annotated as a pseudogene, was recently found to be functioning in *S*. Typhi (Urrutia *et al*., 2014). Thus the effect of pseudogenization remains speculative. Interestingly, a very limited number of mutations from ST8-typed *S*. Typhi were shared with np45 and ST821/98, indicated that these populations were not closely related, and apparently, they evolved under distinct evolutionary events. These findings are in line with the phylogenomic analyses and MLST data, in which these strains were distantly related to ST8-population, suggestive of variants from the existing clones. However, very limited evidence is available to imply for its poor dissemination and whether or not this population is newly emerged, re-emerging or on the verge of “extinction”, explained largely by intermittent and short encounters, limited numbers of samples available and phylogeographical incongruence.

It remains elusive as to why strains of ST1 and ST2 are evolutionarily selected, albeit *S*. Typhi is thought to be shaped by weak diversifying selection forces along its evolutionary timeline (Holt *et al*., 2008; Roumagnac *et al*., 2006). It is reasonable to speculate that these successful STs may have predisposed to advantageous genetic variations that drive evolutions, allowing the pathogens to be more adapted to the host as observed hitherto in several genomic studies of *S*. Typhi (Baddam *et al*., 2012a; Baddam, *et al*., 2012b; Deng *et al*., 2003; Hendriksen *et al*., 2015; Ong *et al*., 2012; Parkhill *et al*., 2001; Yap *et al*., 2014). Thus, extensive genomic studies of the *S*. Typhi strains from the endemic regions, particularly with the inclusion of rare subtypes of *S*. Typhi are required to construct full-scale evolutionary history to understand the mutational events that have occurred along the evolutionary timeline.

4.4 Conclusion

In summary, although the *S*. Typhi strains were obtained from diverse geographical locations spanning over a century, extremely low divergence was observed with two predominant STs (ST1 and ST2) being identified, a finding that was unequivocally supported by both analyses (experimentally and genome-derived STs). From the large scale scanning, three highly uncommon STs have been identified, namely the ST8, ST2233, and ST2359. Notably, the ST3 together with two recently reported novel STs (ST890 and ST892) identified in the previous studies are completely absent. The present findings provided strong evidence that *S*. Typhi strains possess a high level of temporal stability and phylogeographical structuring, supported largely by the phylogeographical signals observed in the phylogenomic tree. While many of the genome features have been studied earlier, this work highlighted the genomic signatures of the as-yet-uncharacterized ST8-typed population, which exhibit ‘unsuccessful trait” that may play vital roles in their poor dissemination. The understanding of these traits may have an important impact on disease control and vaccine development as this population may reflect models of attenuation.

CHAPTER 5

GENERAL DISCUSSION

Advances in whole-genome sequencing technology enabled the phylogeny and genetic properties to be determined at higher resolution and finer scale, but as in all comparative genomic analyses, the background information is crucial for meaningful biological interpretation as of how clinical outcomes of the infection are related to the genetic constituents of the pathogen, but this piece of information is often omitted. Furthermore, for a pathogen like *S*. Typhi that demonstrated varied clinical manifestation, the lack of strain provenance might cause bias and provide an incomplete picture of the true variations of the given pathogen. This is further complicated by the lack of genomic data derived from diverse sources, such as those from various geographical regions, sources of isolation and clinical outcomes. Therefore, this study serves to bridge the gap of knowledge by including a set of comparative data of *S*. Typhi associated with various clinical outcomes, both local and global data, which demonstrated the extent of genetic variations pertaining to virulence and pathogenicity, as well as other pivotal genetic features (CRISPRs, Insertion sequences, phages, etc). Understanding of how and why these genetic variations and mutations persist could provide novel information on the adaptation and the emergence of clonal lineages, thus valuable for devising effective control or complete eradication strategies.

Years ago, extracting housekeeping genes from a large number of genomes are technically challenging, given the poor quality of draft genomes and a limited number of genomes available, which either renders the sequence unusable or not significant enough to represent the global picture. With the advent of next-generation sequencing and improved sequencing technology, high-quality draft genomes could be generated at higher scale in much shorter time. In addition, prior to this, MLST of *S*. Typhi relies mainly on a limited set of regional data that lacks in sample diversity and can barely be generalized to a wider population, making the findings less conclusive and convincing.

In the present study, by utilizing an unprecedented MLST data derived from both local and global *S*. Typhi genomes, it demonstrated that the population structure of *S*. Typhi is highly stable and homogeneous, of which only two sequence types, ST1 and ST2 predominate globally for at least 108 years, and still expanding clonally since then. Although there is no information available prior to this period, the relative stability of *S*. Typhi with limited variations over a century strongly suggests that the pathogen is under a weak selective pressures from the host, akin to other monomorphic pathogens such as *Bacillus antracis*, *Yersinia pestis*, and *Mycobacterium tuberculosis*, which also harbor little genetic variations ((Holt *et al*., 2008; Parkhill & Wren, 2011; Wain, House, Parkhill, Parry, & Dougan, 2002). From the MLST data, limited evidence of bias in the spatial and temporal distribution of STs was detected and the population may have suffered a recent bottleneck during its course of evolution, evidenced by the limited STs and loci recombination. The homogeneity may have been also facilitated by high migration rates of human populations since early out-of-Africa events. Despite the sporadic identification of uncommon STs and/or those restricted to a certain geographical area, these occurrence remains rare. In view of that, the comparative genomics analyses of these seemingly less successful with the predominating genomes provided important insights into the global transmission and evolution of *S*. Typhi. This highlighted the earlier hypothesis that the clinical manifestation and severity of the disease may have genetic basis (Ali *et al*., 1992; Khosla, Srivastava, & Gupta, 1977; Parry, Hien, Dougan, White, & Farrar, 2002; Thong *et al*., 1996). This hypothesis is supported by the differential progression of infections of *S*. Typhi such as those in carriage state, which some of these strains successfully hijacked the immune system through several immune-modulation mechanisms. Recent studies demonstrated that *S*. Typhi possesses a modified metabolic and adhesive potential which may infect human tissue via a specialized route, possibly through dendritic cells and down-regulation of several virulence genes to modulate human immune cells, and diminish inflammatory responses (Dougan & Baker, 2014). However, the precise mode of pathogenesis in *S*. Typhi remains elusive at this stage although most of these virulence potentials are known to be encoded by a repertoire of virulence genes and its variants.

To capture the extent of these fine variations, a whole-genome derived phylogenetic tree of *S*. Typhi strains was constructed. Interestingly, *S*. Typhi was shown to exhibit strong phylogeographical signals from both the global and local data. For example, the CR0044 and BL196 were found to be highly related despite isolated over a gap of two years, in two independent cases with an unknown epidemiological link. Similar strong phylogeographical signals were supported with another massive phylogenomic analysis utilizing global data (Chapter 4), demonstrating that the establishment of the local reservoir is an important contributing factor for future outbreaks or vice versa. Although it is challenging to determine the precedence of these events but the findings provided robust empirical evidence that both carrier and outbreak strains are highly related. This is corroborated by the shared genetic features between the two closely related CR0044 and BL196, which contain unique putative pathogenicity islands (~10kb) that harbored a gene encoding a *zot*-like toxin and several other virulence-related genes. These islands appeared to be horizontally transferred, and the *zot* in particular, showed similarity to other *zot*-like genes in diverse bacteria species, ranging from common human pathogens to extremophiles, implying that the ancestral form of this gene might have been widely disseminated. However, little is known for *zot* in *Salmonella* spp., especially in *S*. Typhi. Recent studies have shown that *zot* was also present in *Campylobacter concisus* and potentially involved in causing inflammatory bowel disease (IBD) (Mahendran *et al*., 2013). In this study, sequence alignment of *S*. Typhi *zot* with other *zot* genes provided some evidence that the gene might be functional, based on the fact that several toxins related domains and motif for protein translation remain intact. Although, it is arduous to determine the mechanism of action of this *zot* gene without first performing in-depth proteomic analyses, its mechanism of action is likely to be different from a typical *zot* in *V. cholera*, owing to its incomplete active domain.

One of the most astounding features of typhoid is the carrier state. Human carriers are known to be the reservoir of new typhoid infections, as they continue to shed high levels of *S*. Typhi in the faeces while remains asymptomatic and undetected. The molecular basis involved in the carrier state transformation is poorly understood, although previous studies have linked the colonization of the gall bladder to the maintenance of carrier state. In the present study, the comparative genomics of the carrier strain, CR0044, and the outbreak, Bl196 yielded 29 nsSNPs (confirmed with HRM), of which two virulence genes, *rpoS*, and *tviE* are interrelated and plausibly relevant to carrier transformation and virulence modulation. Interestingly, the carrier strain CR0044 harbored much higher polymorphisms; an approximate 83% (24/29) unique nsSNPs (undetected in reference genomes) than BL196, indicating that the carrier strain may be under stronger selection pressures than strains in other environments. Of all, CR0044 carried an important mutant gene, the *tviE*, which is required for virulence capsular formation and acts as a protective antigen. It has been demonstrated that antibody responding to the Vi-positive strain is more virulent than a Vi-negative strain. In contrast, BL196 carried a mutant *rpoS* which is identical to *Salmonella* Pullorum, the causative agent of fowl typhoid. It is unclear whether or not the *rpoS* variant in *Salmonella* Pullorum enables the pathogen to differentially regulate other genes, as *rpoS* is known to regulate more than 30 genes in *S*. Typhi (Deng *et al*., 2003; Lu *et al*., 2012). Recent studies have shown that *rpoS* is involved in fine-tuning the synthesis of the Vi polysaccharide in *S*. Typhi (Santander *et al*., 2007), highlighting the interplay of *rpos* and *tviE* genes in regulating virulence capabilities of this pathogen in different environments, and apparently, the adaptive selection of these genes are crucial for pathogen survival. In spite of that, it is difficult to ascertain the functionality of the proteins without first accessing the effects of mutations on them. Point mutations could have enigmatic effects on the structural stability of proteins, which require extensive and detailed studies. The modeling and molecular dynamics simulation of native and mutant structures of TviE and RpoS have demonstrated that RpoS in BL196 is stabilizing whereas TviE in CR0044 is destabilising, evidenced by findings from multiple predictive tools and energy minimization studies. Taken together, these data consistently suggest that the mutations may alter the protein structures, and possibly their functions as well, potentially leading to the enhanced regulation of RpoS and stress response in BL196 and reduced efficiency of TviE in the virulence capsular formation of CR0044. Although it is unknown as to what extent these mutations affect the regulation of proteins, but the close interplay of these two genes suggests that they are likely related to the virulence and persistence of two closely related strains that exhibit different clinical outcomes. These findings serve as caveats for future functional studies.

Besides, the evidence for the universality of carrier genotype is scarce as most of the mutant SNPs carried by CR0044 were strain specific, except for the trehalose permease IIC component, the only gene shared with another carrier strain, P-stx-12. In fact, the two carrier strains are highly dissimilar with >200 SNPs, including those of virulence-related, suggesting that carrier transformation may be mediated by diverse strategies and depends on modulation by a unique host environment. Likewise, the major differences between the ST8- and ST1/ST2-typed populations are the unique presence of several nsSNPs in a few key virulence genes, such as the *flhB*, *sipC*, and *tviD*, which were previously implicated as parts of the machinery for virulence phenotype. Besides, the non-unique nsSNPs detected in several virulence genes (*clpB*, *Sfm/fimF*, *sseJ*, *hilD*, *tviE*) may be transitionary mutations but crucial in regulating pathogenesis of *S*. Typhi albeit very limited evidence of marked variations was found in CRISPRs and *IS* with respect to these two populations, consistent with the findings of comparative genomics of *S*. Typhi associated with different outcomes. On the contrary, among the rare populations, ST8-typed shared very limited mutation with np45 (ST2233) and ST821/98 (ST2359), implying that these populations were not distantly related and may have diverged from MRCA under distinct evolutionary events.

Most of the genomic variations in *S*. Typhi are driven by the acquisition of foreign genes through horizontal gene transfers. Although such events are relatively rarer compared to other *Salmonella* spp, but were empirically observed in CR0044 and BL196, of which both acquired an additional unique pathogenicity island that carried a number of putative virulence genes; some of which are related to uropathogenic *E. coli* and *Yersinia pestis*. It is hypothesized that genes acquisitions serve to enrich the gene pool of the pathogen and allows better adaptation in the host. Such observation was reported in few recent studies that showed the chromosomal integration of resistance genes of plasmid origin (Wong *et al*., 2015; Kingsley *et al*., 2009). However, the reason for the incorporation of foreign genes remains unclear, possibly to reduce fitness burden on the already-less-fit bacterium or/and enhance regulation of interrelated genes. Interestingly, the RM systems of *S*. Typhi are seemingly conserved with no noticeable lesions detected, except for the additional RM systems carried by two plasmids bearing ST-8 typed genomes. The presence of these RM systems may aid in the acquisition of plasmids/acquired resistance as such event has been observed previously in various bacteria (Oliveira *et al*., 2014), though undetected in *S*. Typhi. However, the population of ST8 is likely to have very poor dissemination and adaptation in humans, given the facts that it occurs rarely and geographically-restricted. On the other hand, the complete loss of “primitive” ST3 in the population since its emergence prompted more questions than answers. Although these events are extremely rare and hardly detected, it represents an important piece of information that provides the full picture of evolutionary events that shaped the *S*. Typhi of today, and that possibly pave ways for ground-breaking discoveries in disease controls and treatments in the near future.

CHAPTER 6

CONCLUSION AND RECOMMENDATION

In conclusion, this study has laid an important foundation for an in-depth understanding of genome variations and evolution of *S*. Typhi. The *S*. Typhi genomes of Malaysian strains were thoroughly dissected and the detailed genomic features were successfully elucidated. The submission of genome sequences of *S*. Typhi associated with various clinical outcomes to the public domain has great values and profound implications to the scientific communities. The extensive comparative genomics analyses has revealed several key findings underpinning the understanding of the evolution of *S*. Typhi such as the genetic variations that differentiate the outbreak strain from the human carrier strain and the widespread STs-strains from the seemingly unsuccessful STs-strains. The discoveries of novel genetic features such as putative pathogenomic islands and other uncharacterized genes unfold a new avenue for new discoveries pertaining to pathogenicity studies, evolutionary relationships and potential new therapeutic targets and disease control strategies.

The use of large scale MLST scanning over thousands of genomes offered a rapid and effective strategy to narrow the scope of in-depth comparative genomic analysis and consequently provided new insights into the fine scale of pathogen evolution and population and phylogeographical structure such as the successful identification of global sequences types, uncommon sequence types and their respective genomic variations. The utilization of genome-based phylogenetic analysis on *S*. Typhi revealed, to date, the most accurate evolutionary relationship among strains obtained from various geographical regions of the world spanning many years and from diverse sources and clinical representations. The fine phylogenetic relationship of Malaysian outbreak and human carrier strains was successfully determined with respect to global perspectives of *S*. Typhi and further confirmed with the massive phylogenomic tree encompassing 1,808 *S*. Typhi strains, currently the world's largest entero-pathogen WGS data collection available. To the fine level of micro-evolution, the study with protein modelling, MD stimulation, and multiple prediction tools unveiled the potential instability and stability of nsSNP (*tviE* and *rpoS*) associated with strains from different clinical setting which have deepened our understanding on the effects of mutations on these important virulence factors.

Although the work provided an important foundation to the understanding of *S*. Typhi and its evolution the findings needs more validation with the use of different and larger collection of strains and datasets. This will probably become a reality when genome sequencing becomes a routine and indispensable technology in many laboratories in the near future. It is also imperative to perform function validation on the catalogues of proteins and their interaction. Ultimately, the knowledge and understanding of the biological framework of *S*. Typhi generated from these studies will lead to effective disease control strategies that prevent or perhaps, eradicate the spread of typhoid and others diseases, thus the human sufferings, once and for all.

REFERENCES

Achtman, M. (2008). Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annual Review Microbiology, 62*, 53–70.

Achtman, M. (2012). Insights from genomic comparisons of genetically monomorphic bacterial pathogens. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1590), 860–7.

Ali, G., Kamili, M. A., Shah, M. Y., Koul, R. L., Aziz, A., Hussain, A., & Allaqaband, G. Q. (1992). Neuropsychiatric manifestations of typhoid fever. *The Journal of the Association of Physicians of India*, *40*(5), 333–335.

Alikhan, N.-F., Petty, N. K., Ben Zakour, N. L., & Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics*, *12*(1), 402.

Andrio, P., Fenollosa, C., Cicin-Sain, D., Orozco, M., & Gelpí, J. L. (2012). MDWeb and MDMoby: An integrated web-based platform for molecular dynamics simulations. *Bioinformatics*, *28*(9), 1278–1279.

Anjum, M., Marooney, C., Fookes, M., & Baker, S. (2005). Identification of core and variable components of the Salmonella enterica subspecies I genome by microarray. *Infection and Immunity*, *73*(12), 7894–7905.

Arricau, N., Hermant, D., Waxin, H., Ecobichon, C., Duffey, P. S., & Popoff, M. Y. (1998). The RcsB-RcsC regulatory system of Salmonella typhi differentially modulates the expression of invasion proteins, flagellin and Vi antigen in response to osmolarity. *Molecular Microbiology*, *29*(3), 835–850.

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., … Kubal, M. (2008). The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics*, *9*(1), 75.

Baddam, R., Kumar, N., Thong, K. L., Ngoi, S. T., Teh, C. S. J., Yap, K. P., … Ahmed, N. (2012). Genetic fine structure of a Salmonella enterica serovar Typhi strain associated with the 2005 outbreak of typhoid fever in Kelantan, Malaysia. *Journal of Bacteriology*, *194*(13), 3565–3566.

Baddam, R., Thong, K.-L., Avasthi, T. S., Shaik, S., Yap, K.-P., Teh, C. S. J., … Ahmed, N. (2012). Whole-genome sequences and comparative genomics of Salmonella enterica serovar Typhi isolates from patients with fatal and nonfatal typhoid fever in Papua New Guinea. *Journal of Bacteriology*, *194*(18), 5122–5123.

Baker, S., & Dougan, G. (2007). The genome of Salmonella enterica serovar Typhi. *Clinical Infectious Diseases, 45 Suppl 1*, S29–S33.

Baker, S., Holt, K., Van De Vosse, E., Roumagnac, P., Whitehead, S., King, E., … Dougan, G. (2008). High-throughput genotyping of Salmonella enterica serovar Typhi allowing geographical assignment of haplotypes and pathotypes within an urban district of Jakarta, Indonesia. *Journal of Clinical Microbiology*, *46*(5), 1741–1746.

Baker, S., Sarwar, Y., Aziz, H., Haque, A., Ali, A., Dougan, G., … Haque, A. (2005). Detection of Vi-Negative Salmonella enterica Serovar Typhi in the Peripheral Blood of Patients with Typhoid Fever in the Faisalabad Region of Pakistan Detection of Vi-Negative Salmonella enterica Serovar Typhi in the Peripheral Blood of Patients with Typh. *Journal of Clinical Microbiology*, *43*(9), 4418–4425.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Patrick Boyaval, Moineau, S., … Horvath, P. (2007). CRISPRProvides Acquired Resistance Against Viruses in Prokaryotes. *Science*, *315*(5819), 1709–1712.

Bendl, J., Stourac, J., Salanda, O., & Pavelka, A. (2014). PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Computational Biology*, *10*(1), e1003440.

Bertels, F., Silander, O. K., Pachkov, M., Rainey, P. B., & Van Nimwegen, E. (2014). Automated reconstruction of whole-genome phylogenies from short-sequence reads. *Molecular Biology and Evolution*, *31*(5), 1077–1088.

Bingle, L. E., Bailey, C. M., & Pallen, M. J. (2008). Type VI secretion: a beginner's guide. *Current Opinion in Microbiology*, *11*(1), 3–8.

Boyd, E. F., Porwollik, S., Blackmer, F., Mcclelland, M., & Icrobiol, J. C. L. I. N. M. (2003). Differences in Gene Content among Salmonella enterica Serovar Typhi Isolates. *Journal of Clinical Microbiology*, *41*(8), 3823–3828.

Brenner, F. W., Villar, R. G., Angulo, F. J., Tauxe, R., & Swaminathan, B. (2000). Salmonella nomenclature. *Journal of Clinical Microbiology*, *38*(7), 2465-2467.

Bromberg, Y., & Rost, B. (2007). SNAP: Predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Research*, *35*(11), 3823–3835.

Brunet, T. D. P., & Doolittle, W. F. (2015). Multilevel Selection Theory and the Evolutionary Functions of Transposable Elements: Fig. 1.—. *Genome Biology and Evolution*, *7*(8), 2445–2457.

Buckle, G. C., Walker, C. L. F., & Black, R. E. (2012). Typhoid fever and paratyphoid fever: Systematic review to estimate global morbidity and mortality for 2010. *Journal of Global Health*, *2*(1), 10401.

Bueno, S. M., Santiviago, C. a, A, A., Murillo, A. a, Fuentes, J. a, Trombert, a N., … Mora, G. C. (2004). Precise Excision of the Large Pathogenicity Island , SPI7 , in Salmonella enterica Serovar Typhi Precise Excision of the Large Pathogenicity Island , SPI7 , in Salmonella enterica Serovar Typhi. *Journal of Bacteriology*, *186*(10), 3202–3213.

Cady, K. C., & O'Toole, G. A. (2011). Non-identity-mediated CRISPR-bacteriophage interaction mediated via the Csy and Cas3 proteins. *Journal of Bacteriology*, *193*(14), 3433–3445.

Capriotti, E., Fariselli, P., Rossi, I., & Casadio, R. (2008). A three-state prediction of single point mutations on protein stability changes. *BMC Bioinformatics*, *9*(2), 1.

Carattoli, A., Zankari, E., Garciá-Fernández, A., Larsen, M. V., Lund, O., Villa, L., … Hasman, H. (2014). PlasmidFinder and pMLST: in silico detection and typing of plasmid. *Antimicrobial Agents and Chemotherapy*, AAC–02412.

Carver, T., Berriman, M., Tivey, A., Patel, C., B??hme, U., Barrell, B. G., … Rajandream, M. A. (2008). Artemis and ACT: Viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics*, *24*(23), 2672–2676.

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

Casjens, S. R. (2003). Prophages and bacterial genomic: what have we learned so far? *Molecular Microbiology*, *49*(2), 277–300.

Chelvam, K. K., Yap, K. P., Chai, L. C., & Thong, K. L. (2015). Variable responses to carbon utilization between planktonic and biofilm cells of a human carrier strain of Salmonella enterica serovar Typhi. *PLoS ONE*, *10*(5), 1–11.

Chen, C., Zhang, W., Zheng, H., Lan, R., Du, P., Bai, X., … Wang, H. (2013). Minimum Core Genome Sequence Typing for Clinical and Public Health Microbiology Minimum Core Genome Sequence Typing of Bacterial Pathogens : a. *Journal of Clinical Microbiology*, *51*(8), 2582–2591.

Chiou, C. S., Wei, H. L., Mu, J. J., Liao, Y. S., Liang, S. Y., Liao, C. H., … Wang, S. C. (2013). Salmonella enterica serovar typhi variants in long-term carriers. *Journal of Clinical Microbiology*, *51*(2), 669–672.

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., & Chan, A. P. (2012). Predicting the functional effect of amino acid substitutions and indels. *PloS One*, *7*(10), e46688.

Chowdhury, M., Shumy, F., & Anam, A. (2014). Current status of typhoid fever: a review. *Bangladesh Medical, 43*(2), 106–111.

Coburn, B., Grassl, G. A., & Finlay, B. B. (2007). Salmonella, the host and disease: a brief review. *Immunology and Cell Biology*, *85*(2), 112–118.

Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M., & Robles, M. (2005). Gene Ontology Database Blast2GO:A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, *21*(18), 3674–3676.

Crawford, R. W., Rosales-Reyes, R., Ramírez-Aguilar, M. D. L. L., Chapa-Azuela, O., Alpuche-Aranda, C., & Gunn, J. S. (2010). Gallstones play a significant role in Salmonella spp. gallbladder colonization and carriage. *Proceedings of the National Academy of Sciences of the United States of America, 107*(9), 4353–8.

Crosa, J. H., Brenner, D. J., Ewing, W. H., & Falkow, S. (1973). Molecular relationship among the Salmonelleae. *Journal of Bacteriology, 115*(1), 307–315.

Crump, J. a, Crump, J. a, Luby, S. P., Luby, S. P., Mintz, E. D., & Mintz, E. D. (2004). The global burden of typhoid fever. *Bulletin of the World Health Organization, 82*(5), 346–53.

Crump, J. A., & Mintz, E. D. (2010). Global Trends in Typhoid and Paratyphoid Fever. *Clinical Infectious Diseases*, *50*(2), 241–246.

Dahiya, S., Kapil, A., Kumar, R., Das, B. K., Sood, S., Chaudhry, R., … Lodha, R. K. (2013). Multiple locus sequence typing of Salmonella Typhi, isolated in north India - a preliminary study. *The Indian Journal of Medical Research*, *137*(5), 957–962.

Darling, A. E., Mau, B., & Perna, N. T. (2010). Progressivemauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE*, *5*(6), e11147.

Darriba, D., Taboada, G. L., & Posada, D. (2011). ProtTest 3 : fast selection of best-fit models of protein evolution. *Bioinformatics*, *27*(8), 1164–1165.

Delcher, A. L., Phillippy, A., Carlton, J., & Salzberg, S. L. (2002). Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Research*, *30*(11), 2478–2483.

den Bakker, H. C., Moreno Switt, A. I., Govoni, G., Cummings, C. A., Ranieri, M. L., Degoricija, L., … Wiedmann, M. (2011). Genome sequencing reveals diversification of virulence factor content and possible host adaptation in distinct subpopulations of Salmonella enterica. *BMC Genomics*, *12*(1), 425.

Deng, W., Liou, S., Iii, G. P., George, F., Rose, D. J., Burland, V., … Mayhew, G. F. (2003). Comparative Genomics of Salmonella enterica Serovar Typhi Strains Ty2 and Comparative Genomics of Salmonella enterica Serovar Typhi Strains Ty2 and CT18. *Journal of Bacteriology*, *185*(7), 2330–2337.

Di Pierro, M., Lu, R., Uzzau, S., Wang, W., Margaretten, K., Pazzani, C., … Fasano, A. (2001). Zonula occludens toxin structure-function analysis: Identification of the fragment biologically active on tight junctions and of the zonulin receptor binding domain. *Journal of Biological Chemistry*, *276*(22), 19160–19165.

Didelot, X., Bowden, R., Wilson, D. J., Peto, T. E. a., & Crook, D. W. (2012). Transforming clinical microbiology with bacterial genome sequencing. *Nature Reviews Genetics*, *13*(9), 601–612.

Disz, T., Akhter, S., Cuevas, D., Olson, R., Overbeek, R., Vonstein, V., … Edwards, R. a. (2010). Accessing the SEED genome databases via Web services API: tools for programmers. *BMC Bioinformatics*, *11*(1), 1.

Dobrindt, U., & Hacker, J. (2001). Whole genome plasticity in pathogenic bacteria. *Current Opinion in Microbiology*, *4*(5), 550–557.

Doolittle, R., Feng, D., Tsang, S., Cho, G., & Little, E. (1996). Determining Divergence Times of the Major Kingdoms of Living Organisms with a Protein Clock. *Science*, *271*(5248), 470–477.

Doolittle, W. F., & Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature*, *284*(5757), 601–603.

Dougan, G., & Baker, S. (2014). Salmonella enterica serovar typhi and the pathogenesis of typhoid fever. *Annual Review of Microbiology*, *68*, 317–336.

Dutta, U., Garg, P. K., Kumar, R., & Tandon, R. K. (2000). Typhoid carriers among patients with gallstones are at increased risk for carcinoma of the gallbladder. *American Journal of Gastroenterology*, *95*(3), 784–787.

Edwards, R. a, Olsen, G. J., & Maloy, S. R. (2002). Comparative genomics of closely related slmonellae. *Trends Microbiology*, *10*(01), 94–99.

Fabre, L., Zhang, J., Guigon, G., Hello, S. Le, & Guibert, V. (2012). CRISPR typing and subtyping for improved laboratory surveillance of Salmonella infections. *PloS One*, *7*(5), e36995.

Fang, F. C., Libby, S. J., Buchmeier, N. A., Loewen, P. C., Switala, J., Harwood, J., & Guiney, D. G. (1992). The alternative sigma factor katF (rpoS) regulates Salmonella virulence. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(24), 11978–11982.

Faucher, S. P., Viau, C., Gros, P. P., Daigle, F., & Le Moual, H. (2008). The prpZ gene cluster encoding eukaryotic-type Ser/Thr protein kinases and phosphatases is repressed by oxidative stress and involved in Salmonella enterica serovar Typhi survival in human macrophages. *FEMS Microbiology Letters*, *281*(2), 160–166.

Fraser-liggett, C. M. (2005). Insights on biology and evolution from microbial genome sequencing Insights on biology and evolution from microbial genome sequencing. *Genome Research*, *15*(12), 1603–1610.

Fricke, W. F., Mammel, M. K., McDermott, P. F., Tartera, C., White, D. G., LeClerc, J. E., … Cebula, T. A. (2011). Comparative genomics of 28 Salmonella enterica isolates: evidence for CRISPR-mediated adaptive sublineage evolution. *Journal of Bacteriology*, JB–00297.

Gagneux, S. (2012). Host-pathogen coevolution in human tuberculosis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1590), 850–859.

Gibert, I., Barbé, J., & Casadesús, J. (1990). Distribution of insertion sequence IS200 in Salmonella and Shigella. *Journal of General Microbiology*, *136*, 2555–2560.

Gilman, R. H., Terminel, M., Levine, M. M., Hernandez-Mendoza, P., & Hornick, R. (1975). Relative efficacy of blood, urine, rectal swab, bone-marrow, and rose-spot cultures for recovery of Salmonella typhi in typhoid fever. *Lancet*, *305*(7918), 1211 – 1213.

Glynn, J. R., Hornick, R. B., Levine, M. M., & Bradley, D. J. (1995). Infecting dose and severity of typhoid: analysis of volunteer data and examination of the influence of the definition of illness used. *Epidemiology and Infection*, *115*(1), 23–30.

Goldblum, S. E., Rai, U., Tripathi, A., Thakar, M., De Leo, L., Di Toro, N., … Fasano, A. (2011). The active Zot domain (aa 288-293) increases ZO-1 and myosin 1C serine/threonine phosphorylation, alters interaction between ZO-1 and its binding partners, and induces tight junction disassembly through proteinase activated receptor 2 activation. *The FASEB Journal*, *25*(1), 144–158.

Gonzalez-escobedo, G., Marshall, J. M., & Gunn, J. S. (2011). Chronic and acute infection of the gall bladder by Salmonella Typhi: understanding the carrier state. *Nature Reviews Microbiology*, *9*(1), 9–14.

Gopinath, S., Carden, S., & Monack, D. (2004). Shedding Light on Salmonella carriers. *Trends in Microbiology*, *20*(7), 320–327.

Gouy, M., Guindon, S., & Gascuel, O. (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution*, *27*(2), 221–224.

Grissa, I., Vergnaud, G., Pourcel, C., Bland, C., Ramsey, T. L., Sabree, F., … Hwang, J. K. (2007). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Research*, *35*(2), W52–W57.

Groschwitz, K. R., & Hogan, S. P. (2009). Intestinal barrier function: molecular regulation and disease pathogenesis. *Journal of Allergy and Clinical Immunology*, *124*(1), 3–20.

Gunn, J. S., Marshall, J. M., Baker, S., Dongol, S., Charles, R. C., & Ryan, E. T. (2014). Salmonella chronic carriage: Epidemiology, diagnosis, and gallbladder persistence. *Trends in Microbiology*, *22*(11), 648–655.

Hacker, J., Blum-Oehler, G., Mühldorfer, I., & Tschäpe, H. (1997). Pathogenicity islands of virulent bacteria: Structure, function and impact on microbial evolution. *Molecular Microbiology*, *23*(6), 1089–1097.

Hacker, J., & Kaper, J. B. (2000). Pathogenicity Islands and the Evolution of Microbes. *Annual Review Microbiology*, *54*(1), 641–679.

Haubold, B., Klötzl, F., & Pfaffelhuber, P. (2015). Andi: Fast and accurate estimation of evolutionary distances between closely related genomes. *Bioinformatics*, *31*(8), 1169–1175.

Hazen, T. H., Sahl, J. W., Redman, J. C., Carolyn, R., Daugherty, S. C., Chibucos, M. C., … Rasko, D. A. (2012). Draft Genome Sequences of the Diarrheagenic Escherichia coli. *Journal Of Bacteriology*, *194*(11), 3026–3027.

Hendriksen, R. S., Leekitcharoenphon, P., Mikoleit, M., Jensen, J. D., Kaas, R. S., Roer, L., … Hasman, H. (2015). Genomic dissection of travel-associated extended-spectrum-beta-lactamase-producing Salmonella enterica serovar typhi isolates originating from the Philippines: a one-off occurrence or a threat to effective treatment of typhoid fever? *Journal of Clinical Microbiology*, *53*(2), 677–680.

Hirano, T., Yamaguchi, S., Oosawa, K., & Aizawa, S. I. (1994). Roles of FliK and FlhB in determination of flagellar hook length in Salmonella typhimurium. *Journal of Bacteriology*, *176*(17), 5439–5449.

Ho, W. S., Gan, H. M., Yap, K. P., Balan, G., Yeo, C. C., & Thonga, K. L. (2012). Genome sequence of multidrug-resistant Escherichia coli EC302/04, isolated from a human tracheal aspirate. *Journal of Bacteriology*, *194*(23), 6691–6692.

Holt, K. E., Parkhill, J., Mazzoni, C. J., Roumagnac, P., Weill, F.-X., Goodhead, I., … Dougan, G. (2008). High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi. *Nature Genetics*, *40*(8), 987–993.

Holt, K. E., Thomson, N. R., Wain, J., Langridge, G. C., Hasan, R., Bhutta, Z. a, … Parkhill, J. (2009). Pseudogene accumulation in the evolutionary histories of Salmonella enterica serovars Paratyphi A and Typhi. *BMC Genomics*, *10*(1), 1.

Hooft, R. W. W., Vriend, G., Sander, C., & Abola, E. E. (1996). Errors in protein structures. *Nature*, *381*(6580), 272–272.

Hopkins, K. L., Maguire, C., Best, E., Liebana, E., & Threlfall, E. J. (2007). Stability of multiple-locus variable-number tandem repeats in Salmonella enterica serovar Typhimurium. *Journal of Clinical Microbiology*, *45*(9), 3058–3061.

Hopkins, K. L., Peters, T. M., de Pinna, E., & Wain, J. (2011). Standardisation of multilocus variable-number tandemrepeat analysis (MLVA) for subtyping of Salmonella enterica serovar Enteritidis. *Eurosurveillance*, *16*(32), 19942.

Hornick, R. B., Greisman, S. E., Woodward, T. E., DuPont, H. L., Dawkins, A. T., & Snyder, M. J. (1970). Typhoid Fever: Pathogenesis and Immunologic Control. *The New England Journal of Medicine*, *283*(13), 686–691.

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

Huson, D. H., & Bryant, D. (2005). Estimating phylogenetic trees and networks using SplitsTree 4. *Manuscript in Preparation, Software Available from Www. Splitstree. Org*.

Hyatt, D., Chen, G.-L., Locascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*(1), 1.

Jacobsen, A., Hendriksen, R. S., Aaresturp, F. M., Ussery, D. W., & Friis, C. (2011). The Salmonella enterica Pan-genome. *Microbial Ecology*, *62*(3), 487–504.

Jansen, R., Embden, J. D., Gaastra, W., & Schouls, L. M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular Microbiology*, *43*(6), 1565–1575.

Jones, B. D., & Falkow, S. (1996). Salmonellosis: Host immune responses and bacterial virulence determinants. *Annual Review of Immunology*, *14*(1), 533–561.

Kaplan, W., & Littlejohn, T. G. (2001). Swiss-PDB Viewer (Deep View). *Briefings in Bioinformatics* , *2* (2), 195–197.

Katoh, K., & Daron M, S. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780.

Kauffmann, F. (1966). [On the history of salmonella research]. *Zentralblatt Für Bakteriologie, Parasitenkunde, Infektionskrankheiten Und Hygiene. 1. Abt. Medizinisch-Hygienische Bakteriologie, Virusforschung Und Parasitologie. Originale*, *201*(1), 44–48.

Khosla, S., Srivastava, S., & Gupta, S. (1977). Neuro-psychiatric manifestations of typhoid. *The Journal of Tropical Medicine and Hygiene*, *80*(5), 95–98.

Kidgell, C., Reichard, U., Wain, J., Linz, B., Torpdahl, M., Dougan, G., & Achtman, M. (2002). Salmonella typhi, the causative agent of typhoid fever, is approximately 50,000 years old. *Infection, Genetics and Evolution*, *2*(1), 39–45.

Kingsley, R. A., Msefula, C. L., Thomson, N. R., Kariuki, S., Holt, K. E., Gordon, M. A., … Dougan, G. (2009). Epidemic multiple drug resistant Salmonella Typhimurium causing invasive disease in sub-Saharan Africa have a distinct genotype Epidemic multiple drug resistant Salmonella Typhimurium causing invasive disease in sub-Saharan Africa have a distinct genotype. *Genome Research*, *19*(12), 2279–2287.

Kumar, S., Nei, M., Dudley, J., & Tamura, K. (2008). MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. *Briefings in Bioinformatics*, *9*(4), 299–306.

Kurazono, H., Yamamoto, S., Nakano, M., Nair, G. B., Terai, A., Chaicumpa, W., & Hayashi, H. (2000). Characterization of a putative virulence island in the chromosome of uropathogenic Escherichia coli possessing a gene encoding a uropathogenic-specific protein. *Microbial Pathogenesis*, *28*(3), 183–189.

Kusumoto, M., Ooka, T., Nishiya, Y., Ogura, Y., Saito, T., Sekine, Y., … Hayashi, T. (2011). Insertion sequence-excision enhancer removes transposable elements from bacterial genomes and induces various genomic deletions. *Nature Communications*, *2*, 152.

Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H. H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, *35*(9), 3100–3108.

Lam, S., & Roth, J. R. (1983). IS200: A salmonella-specific insertion sequence. *Cell*, *34*(3), 951–960.

Langille, M. G. I., & Brinkman, F. S. L. (2009). IslandViewer: An integrated interface for computational identification and visualization of genomic islands. *Bioinformatics*, *25*(5), 664–665.

Larsen, M. V., Cosentino, S., Rasmussen, S., Friis, C., Hasman, H., Marvig, R. L., … Lund, O. (2012). Multilocus Sequence Typing of Total Genome Sequenced Bacteria. *Journal of Clinical Microbiology*, *50*(4), 1355–1361.

Laskowski, R. a, Rullmannn, J. a, MacArthur, M. W., Kaptein, R., & Thornton, J. M. (1996). AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *Journal of Biomolecular NMR*, *8*(4), 477–486.

Laslett, D., & Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research*, *32*(1), 11–16.

Lawrence, J. G., Ochman, H., & Ragan, M. A. (2002). Reconciling the many faces of lateral gene transfer. *Trends in Microbiology*, *10*(1), 1–4.

Le Minor, L., Veron, M., & Popoff, M. (1982). [A proposal for Salmonella nomenclature]. In *Annales de microbiologie* (Vol. 133, pp. 245–254).

Lee, A., White, N., & Van Der Walle, C. F. (2003). The intestinal zonula occludens toxin (ZOT) receptor recognises non-native ZOT conformers and localises to the intercellular contacts. *FEBS Letters*, *555*(3), 638–642.

Leekitcharoenphon, P., Lukjancenko, O., Friis, C., Aarestrup, F. M., & Ussery, D. W. (2012). Genomic variation in Salmonella enterica core genes for epidemiological typing. *BMC Genomics*, *13*, 88.

Li, H., Li, P., Xie, J., Yi, S., Yang, C., Wang, J., … Song, H. (2014). New clustered regularly interspaced short palindromic repeat locus spacer pair typing method based on the newly incorporated spacer for Salmonella enterica. *Journal of Clinical Microbiology*, *52*(8), 2955–2962.

Lin, a W., Usera, M. a, Barrett, T. J., & Goldsby, R. a. (1996). Application of random amplified polymorphic DNA analysis to differentiate strains of Salmonella enteritidis. *Journal of Clinical Microbiology*, *34*(4), 870–876.

Lindahl, E., Azuara, C., Koehl, P., & Delarue, M. (2006). NOMAD-Ref: Visualization, deformation and refinement of macromolecular structures based on all-atom normal mode analysis. *Nucleic Acids Research*, *34*(2), W52–W56.

Lipman, S F Altschul, W Gish, W Miller, E W Myers, D. J. (2014). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410.

Liu, F., Barrangou, R., Gerner-Smidt, P., Ribot, E. M., Knabel, S. J., & Dudley, E. G. (2011). Novel virulence gene and CRISPR multilocus sequence typing scheme for subtyping the major serovars of Salmonella enterica subspecies enterica. *Applied and Environmental Microbiology*, *77*, 1946–1956.

Liu, S. L., Sanderson, K. E., & Anderson, K. E. E. S. (1996). Highly plastic chromosomal organization in Salmonella typhi. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(19), 10303–10308.

Liu, Y., Lee, M. A., Ooi, E. E., Mavis, Y., Tan, A. L., & Quek, H. H. (2003). Molecular typing of Salmonella enterica serovar typhi isolates from various countries in Asia by a multiplex PCR assay on variable-number tandem repeats. *Journal of Clinical Microbiology*, *41*(9), 4388–4394.

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: A program for inproved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, *25*(5), 955–964.

Lu, Y., Chen, S., Dong, H., Sun, H., Peng, D., & Liu, X. (2012). Identification of genes responsible for biofilm formation or virulence in Salmonella enterica serovar pullorum. *Avian Diseases*, *56*(1), 134–143.

Maathuis, M., Colombo, D., Kalisch, M., & Bühlmann, P. (2000). A method and server for predicting damaging missense mutations. *Ann. Stat. Cell Statist. Soc. Ser. B J. Roy. Statist. Soc. Ser. B Biol*, *37*(16), 3133–3164.

Mahendran, V., Tan, Y., Riordan, S., & Grimm, M. (2013). The prevalence and polymorphisms of zonula occluden toxin gene in multiple Campylobacter concisus strains isolated from saliva of patients with. *PloS One*, *8*(9), e75525.

Marks, F., Adu-Sarkodie, Y., Hunger, F., Sarpong, N., Ekuban, S., Agyekum, A., … May, J. (2010). Typhoid fever among children, ghana. *Emerging Infectious Diseases*, *16*(11), 176–178.

Martínez-Gamboa, A., & Silva, C. (2015). IS200 and multilocus sequence typing for the identification of Salmonella enterica serovar Typhi strains from Indonesia. *International Microbiology*, *18*(2), 99–104.

McClelland, M., Sanderson, K. E., Spieth, J., Clifton, S. W., Latreille, P., Courtney, L., … Wilson, R. K. (2001). Complete genome sequence of Salmonella enterica serovar Typhimurium LT2. *Nature*, *413*(6858), 852–856.

Medini, D., Donati, C., Tettelin, H., Masignani, V., & Rappuoli, R. (2005). The microbial pan-genome. *Current Opinion in Genetics & Development*, *15*(6), 589–594.

Mohd Suhaimi, N. S., Yap, K.-P., Ajam, N., & Thong, K.-L. (2014). Genome sequence of Kosakonia radicincitans UMEnt01/12, a bacterium associated with bacterial wilt diseased banana plant. *FEMS Microbiology Letters*, *358*(1), 11–13.

Moreno Switt, A. I., Orsi, R. H., den Bakker, H. C., Vongkamjan, K., Altier, C., & Wiedmann, M. (2013). Genomic characterization provides new insight into Salmonella phage diversity. *BMC Genomics*, *14*(1), 481.

Murray, C. J. L., Vos, T., & Lozano, R. (2014). Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010 (vol 380, pg 2197, 2012). *Lancet*, *384*(9943), 582.

Myeni, S. K., & Zhou, D. (2010). The C terminus of SipC binds and bundles F-actin to promote Salmonella invasion. *Journal of Biological Chemistry*, *285*(18), 13357–13363.

Nagaraja, V., & Eslick, G. D. (2014a). Letter: Chronic Salmonella typhi carrier status and gall-bladder cancer - Authors’ reply. *Alimentary Pharmacology and Therapeutics*, *39*(12), 1440.

Nagaraja, V., & Eslick, G. D. (2014b). Systematic review with meta-analysis: The relationship between chronic Salmonella typhi carrier status and gall-bladder cancer. *Alimentary Pharmacology and Therapeutics*, *39*(8), 745–750.

Nair, S., Alokam, S., Kothapalli, S., & Porwollik, S. (2004). Salmonella enterica serovar Typhi strains from which SPI7, a 134-kilobase island with genes for Vi exopolysaccharide and other functions, has been deleted. *Journal of Bacteriology*, *186*(10), 3214–3223.

Nath, G., Maurya, P., & Gulati, A. K. (2010). ERIC PCR and RAPD based fingerprinting of Salmonella Typhi strains isolated over a period of two decades. *Infection, Genetics and Evolution*, *10*(4), 530–536.

Navarro, F., Llovet, T., Echeita, M. A., Coll, P., Aladuena, A., Usera, M. A., & Prats, G. (1996). Molecular typing of Salmonella enterica serovar typhi. *Journal of Clinical Microbiology*, *34*(11), 2831–2834.

Ng, P. C., & Henikoff, S. (2003). SIFT : predicting amino acid changes that affect protein function. *Nucleic Acids Research*, *31*(13), 3812–3814.

Niyaz Ahmed, Ulrich Dobrindt, J. H. and S. E. H., Mayne, S. T., Wright, M. E., & Cartmel, B. (2008). Genomic fluidity and pathogenic bacteria applications in diagnostics Epidemiology and Intervention Trials. *Nature Reviews Microbiology*, *6*(5), 387–394.

Ochiai, R. L., Acosta, C. J., Danovaro-Holliday, M. C., Baiqing, D., Bhattacharya, S. K., Agtini, M. D., … Jodar, L. (2008). A study of typhoid fever in five Asian countries: Disease burden and implications for controls. *Bulletin of the World Health Organization*, *86*(4), 260–268.

Ochman, H., & Groisman, E. A. (1996). Distribution of pathogenicity islands in Salmonella spp. *Infection and Immunity*, *64*(12), 5410–5412.

Oliveira, P. H., Touchon, M., & Rocha, E. P. C. (2014). The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Research*, *42*(16), 10618–10631.

Olsen, J. E., Skov, M. N., Angen, Ø., Threlfall, E. J., & Bisgaard, M. (1997). Genomic relationships between selected phage types of Salmonella enterica subsp. enterica serotype typhimurium defined by ribotyping, IS200 typing and PFGE. *Microbiology*, *143*(4), 1471–1479.

Ong, S. Y., Pratap, C. B., Wan, X., Hou, S., Rahman, A. Y. A., Saito, J. A., … Alama, M. (2012). Complete genome sequence of Salmonella enterica subsp. Enterica serovar typhi P-stx-12. *Journal of Bacteriology*, *194*(8), 2115–2116.

Ooka, T., Ogura, Y., Asadulghani, M., Ohnishi, M., Nakayama, K., Terajima, J., … Hayashi, T. (2009). Inference of the impact of insertion sequence (IS) elements on bacterial genome diversification through analysis of small-size structural polymorphisms in Escherichia coli O157 genomes. *Genome Research*, *19*(10), 1809–1816.

Orgel, L. E., & Crick, F. H. (1980). Selfish DNA: the ultimate parasite. *Nature*, *284*(5757), 604–607.

Osama, A., Gan, H. M., Teh, C. S. J., Yap, K. P., & Thong, K. L. (2012). Genome sequence and comparative genomics analysis of a vibrio cholerae O1 strain isolated from a cholera Patient in Malaysia. *Journal of Bacteriology*, *194*(24), 6933–6933.

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., … Stevens, R. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research*, *42*(D1), D206–D214.

Pallen, M. J., Loman, N. J., & Penn, C. W. (2010). High-throughput sequencing and clinical microbiology: progress, opportunities and challenges. *Current Opinion in Microbiology*, *13*(5), 625–31.

Pang, S., Octavia, S., Feng, L., Liu, B., Reeves, P. R., Lan, R., & Wang, L. (2013). Genomic diversity and adaptation of Salmonella enterica serovar Typhimurium from analysis of six genomes of different phage types. *BMC Genomics*, *14*(1), 1.

Parkhill, J., Dougan, G., James, K. D., Thomson, N. R., Pickard, D., Wain, J., … Barrell, B. G. (2001). Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18. *Cellular and Molecular Immunology*, *413*(6858), 848–852.

Parkhill, J., & Wren, B. W. (2011). Bacterial epidemiology and biology--lessons from genome sequencing. *Genome Biology*, *12*(10), 230.

Parry, C. M., Hien, T. T., Dougan, G., White, N. J., & Farrar, J. J. (2002). Typhoid Fever. *New England Journal of Medicine*, *347*(22), 1770–1782.

Parsons, D. A., & Heffron, F. (2005). sciS, an icmF homolog in Salmonella enterica serovar Typhimurium, limits intracellular replication and decreases virulence. *Infection and Immunity*, *73*(7), 4338–4345.

Pérez-Losada, M., Cabezas, P., Castro-Nallar, E., & Crandall, K. A. (2013). Pathogen typing in the genomics era: MLST and the future of molecular epidemiology. *Infection, Genetics and Evolution*, *16*, 38–53.

Pickard, D., Wain, J., Baker, S., Line, A., Chohan, S., Fookes, M., … Dougan, G. (2003). Composition, acquisition, and distribution of the Vi exopolysaccharide-encoding Salmonella enterica pathogenicity island SPI-7. *Journal of Bacteriology*, *185*(17), 5055–5065.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2-Approximately Maximum-Likelihood Trees for Large Alignments. *Plos One*, *5*(3), e9490.

Pukatzki, S., Mcauley, S. B., & Miyata, S. T. (2009). The type VI secretion system : translocation of effectors and. *Current Opinion in Microbiology*, *12*(1), 11–17.

Radnedge, L., Agron, P. G., Worsham, P. L., & Andersen, G. L. (2002). Genome plasticity in Yersinia pestis. *Microbiology*, *148*(6), 1687–1698.

Raffatellu, M., Chessa, D., Wilson, R. P., Akçelik, M., Bäumler, A. J., & Ba, A. J. (2006). Capsule-mediated immune evasion: a new hypothesis explaining aspects of typhoid fever pathogenesis. *Infection and Immunity*, *74*(1), 19–27.

Rahman, T., Hosen, I., & Chakraborty, S. (2013). A Rapid Glimpse on Typhoid Fever : An Updated Mini Review. *Journal of Life Medicine*, *1*(3), 83–92.

Rescigno, M., Urbano, M., Valzasina, B., Francolini, M., Rotta, G., Bonasio, R., … Ricciardi-Castagnoli, P. (2001). Dendritic cells express tight junction proteins and penetrate gut epithelial monolayers to sample bacteria. *Nature Immunology*, *2*(4), 361–367.

Robbe-saule, V., Coynault, C., & Norel, F. (1995). The live oral typhoid vaccine Ty21a is a rpoS mutant and is susceptible to various environmental stresses. *FEMS Microbiology Letters*, *126*(2), 171–176.

Rosso, M.-L., Chauvaux, S., Dessein, R., Laurans, C., Frangeul, L., Lacroix, C., … Marceau, M. (2008). Growth of Yersinia pseudotuberculosis in human plasma: impacts on virulence and metabolic gene expression. *BMC Microbiology*, *8*(1), 1.

Roumagnac, P., Weill, F.-X., Dolecek, C., Baker, S., Brisse, S., Chinh, N. T., … Achtman, M. (2006). Evolutionary history of Salmonella Typhi. *Science*, *314*(5803), 1301–4.

Rowe, B., & Gibert, I. (1994). Insertion sequence IS200 fingerprinting of Salmonella typhi: an assessment of epidemiological applicability. *Epidemiology and Infection*, *112*(2), 253–261.

Sabbagh, S. C., Forest, C. G., Lepage, C., Leclerc, J. M., & Daigle, F. (2010). So similar, yet so different: Uncovering distinctive features in the genomes of Salmonella enterica serovars Typhimurium and Typhi. *FEMS Microbiology Letters*, *305*(1), 1–13.

Santander, J., Wanda, S. Y., Nickerson, C. A., & Curtiss, R. (2007). Role of RpoS in fine-tuning the synthesis of Vi capsular polysaccharide in Salmonella enterica serotype Typhi. *Infection and Immunity*, *75*(3), 1382–1392.

Schell, M. A., Ulrich, R. L., Ribot, W. J., Brueggemann, E. E., Hines, H. B., Chen, D., … DeShazer, D. (2007). Type VI secretion is a major virulence determinant in Burkholderia mallei. *Molecular Microbiology*, *64*(6), 1466–1485.

Schofield, M. J., & Hsieh, P. (2003). DNA mismatch repair: molecular mechanisms and biological function. *Annual Reviews in Microbiology*, *57*(1), 579–608.

Schreiber, F., Kay, S., Frankel, G., Clare, S., Goulding, D., van de Vosse, E., … Baker, S. (2015). The Hd, Hj, and Hz66 flagella variants of Salmonella enterica serovar Typhi modify host responses and cellular interactions. *Scientific Reports*, *5*, 7947.

Sherry, N. L., Porter, J. L., Seemann, T., Watkins, A., Stinear, T. P., & Howden, B. P. (2013). Outbreak Investigation Using High-Throughput Genome Sequencing within a Diagnostic Microbiology Laboratory. *Journal of Clinical Microbiology*, *51*(5), 1396–1401.

Shukla, V. K., Singh, H., Pandey, M., Upadhyay, S. K., & Nath, G. (2000). Carcinoma of the gallbladder--is it a sequel of typhoid? *Digestive Diseases and Sciences*, *45*(5), 900–903.

Sippl, M. W. and M. J. (2007). ProSA-web: interactive web service for the recognition of erros in three-dimensional structures of proteins. *Nucleic Acids Research*, *35*(2), W407–W410.

Siriken, B. (2013). Salmonella Pathogenicity Islands. *Mikrobiyoloji Bulteni*, *47*(1), 181–188.

Sorek, R., Kunin, V., & Hugenholtz, P. (2008). CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nature Reviews in Microbiology*, *6*(3), 181–186.

Swart, a L., & Hensel, M. (2012). Interactions of Salmonella enterica with dendritic cells. *Virulence*, *3*(7), 660–667.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular Biology and Evolution*, (30), 2725–2729.

Tatusov, R. L., Koonin, E. V, & Lipman, D. J. (2012). A Genomic Perspective on Protein Families. *Science*, *278*(1997), 631–637.

Teplitski, M., Al-Agely, A., & Ahmer, B. M. M. (2006). Contribution of the SirA regulon to biofilm formation in *Salmonella enterica* serovar Typhimurium. *Microbiology*, *152*(11), 3411–3424.

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., … Narechania, A. (2003). PANTHER: A library of protein families and subfamilies indexed by function. *Genome Research*, *13*(9), 2129–2141.

Thomson, N., Baker, S., Pickard, D., Fookes, M., Anjum, M., Hamlin, N., … Dougan, G. (2004). The role of prophage-like elements in the diversity of Salmonella enterica serovars. *Journal of Molecular Biology*, *339*(2), 279–300.

Thong, K. L., Cheong, Y. M., Puthucheary, S., Koh, C. L., & Pang, T. (1994). Epidemiologic analysis of sporadic Salmonella typhi isolates and those from outbreaks by pulsed-field gel electrophoresis. *Journal of Clinical Microbiology*, *32*(5), 1135–1141.

Thong, K. L., Cordano, A. M., Yassin, R. M., & Pang, T. (1996). Molecular analysis of environmental and human isolates of Salmonella typhi. *Applied & Environmental Microbiology*, *62*(1), 271–274.

Thong, K. L., Goh, Y. L., Yasin, R. M., Lau, M. G., Passey, M., Winston, G., … Reeder, J. C. (2002). Increasing genetic diversity of Salmonella enterica serovar Typhi isolates from Papua New Guinea over the period from 1992 to 1999. *Journal of Clinical Microbiology*, *40*(11), 4156–4160.

Thong, K. L., Passey, M., Clegg, A., Combs, B. G., Yassin, R. M., & Pang, T. (1996). Molecular analysis of isolates of Salmonella typhi obtained from patients with fatal and nonfatal typhoid fever. *Journal of Clinical Microbiology*, *34*(4), 1029–1033.

Thong, K. L., Puthucheary, S. D., & Pang, T. (1997). Genome size variation among recent human isolates of Salmonella typhi. *Research in Microbiology*, *148*(3), 229–235.

Thong, K. L., Puthucheary, S., Yassin, R. M., Sudarmono, P., Padmidewi, M., Soewandojo, E., … Pang, T. (1995). Analysis of Salmonella typhi isolates from Southeast Asia by pulsed-field gel electrophoresis. *Journal of Clinical Microbiology*, *33*(7), 1938–1941.

Tien, Y. Y., Ushijima, H., Mizuguchi, M., Liang, S. Y., & Chiou, C. S. (2012). Use of multilocus variable-number tandem repeat analysis in molecular subtyping of Salmonella enterica serovar Typhi isolates. *Journal of Medical Microbiology*, *61*(2), 223–232.

Tiong, V., Thong, K. L., Yusof, M. Y. M., Hanifah, Y. A., Sam, J. I. ching, & Hassan, H. (2010). Macrorestriction analysis and antimicrobial susceptibility profiling of Salmonella enterica at a university teaching hospital, Kuala Lumpur. *Japanese Journal of Infectious Diseases*, *63*(5), 317–322.

Touchon, M., & Rocha, E. P. C. (2010). The small, slow and specialized CRISPR and anti-CRISPR of Escherichia and Salmonella. *PLoS ONE*, *5*(6), e11126.

Underwood, A. P., Dallman, T., Thomson, N. R., Williams, M., Harker, K., Perry, N., … Wain, J. (2013). Public health value of next-generation DNA sequencing of enterohemorrhagic Escherichia coli isolates from an outbreak. *Journal of Clinical Microbiology*, *51*(1), 232–237.

Urrutia, I. M., Fuentes, J. A., Valenzuela, L. M., Ortega, A. P., Hidalgo, A. A., & Mora, G. C. (2014). Salmonella Typhi shdA: Pseudogene or allelic variant? *Infection, Genetics and Evolution*, *26*, 146–152.

Urwin, R., & Maiden, M. C. J. (2003). Multi-locus sequence typing: a tool for global epidemiology. *Trends in Microbiology*, *11*(10), 479–487.

Vaishnavi, C., Kochhar, R., Singh, G., Kumar, S., Singh, S., & Singh, K. (2005). Epidemiology of typhoid carriers among blood donors and patients with biliary, gastrointestinal and other related diseases. *Microbiol and Immunology*, *49*(2), 107–112.

Van Belkum, A., Tassios, P. T., Dijkshoorn, L., Haeggman, S., Cookson, B., Fry, N. K., … Gerner-Smidt, P. (2007). Guidelines for the validation and application of typing methods for use in bacterial epidemiology. *Clinical Microbiology and Infection*, *13*(s3), 1–46.

Waddington, C. S., Darton, T. C., Jones, C., Haworth, K., Peters, A., John, T., … Pollard, A. J. (2014). An outpatient, ambulant-design, controlled human infection model using escalating doses of salmonella typhi challenge delivered in sodium bicarbonate solution. *Clinical Infectious Diseases*, *58*(9), 1230–1240.

Wain, J., House, D., Parkhill, J., Parry, C., & Dougan, G. (2002). Unlocking the genome of the human typhoid bacillus. *Lancet Infectious Diseases*, *2*(3), 163–170.

Wang, H., Pan, J., Zhang, W., & Zheng, W. (2009). Multilocus sequence typing of Salmonella typhi isolates in Hangzhou. *Chinese Journal of Health Laboratory Technology*, *8*, 018.

Welch, R. A., Burland, V., Plunkett, G., Redford, P., Roesch, P., Rasko, D., … Blattner, F. R. (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. *Proceedings of the National Academy of Sciences*, *99*(26), 17020 –17024.

Werner, G., Klare, I., & Witte, W. (2007). The current MLVA typing scheme for Enterococcus faecium is less discriminatory than MLST and PFGE for epidemic-virulent, hospital-adapted clonal types. *BMC Microbiology*, *7*(1), 1.

Wetter, M., Goulding, D., Pickard, D., Kowarik, M., Waechter, C. J., Dougan, G., & Wacker, M. (2012). Molecular Characterization of the viaB Locus Encoding the Biosynthetic Machinery for Vi Capsule Formation in Salmonella Typhi. *PLoS ONE*, *7*(9), e45609.

WHO. (2012). Book Review: Working to Overcome the Global Impact of Neglected Tropical Diseases. *Perspectives in Public Health*, *132*(4), 192–192.

Wong, V. K., Baker, S., Pickard, D. J., Parkhill, J., Page, A. J., Feasey, N. A., … Dougan, G. (2015). Phylogeographical analysis of the dominant multidrug-resistant H58 clade of Salmonella Typhi identifies inter- and intracontinental transmission events. *Nature Genetics*, *47*(6), 632–639.

Wren, B. W. (2000). Microbial genome analysis: insights into virulence, host adaptation and evolution. *Nature Reviews Genetics*, *1*(1), 30–39.

Wyant, T. L., Tanner, M. K., & Sztein, M. B. (1999). Salmonella typhi Flagella Are Potent Inducers of Proinflammatory Cytokine Secretion by Human Monocytes. *Infection and Immunity*, *67*(7), 3619–3624.

Xu, D., & Zhang, Y. (2011). Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophysical Journal*, *101*(10), 2525–2534.

Yap, K. P., Gan, H. M., Teh, C. S. J., Baddam, R., Chai, L. C., Kumar, N., … Thong, K. L. (2012). Genome sequence and comparative pathogenomics analysis of a salmonella enterica serovar typhi strain associated with a typhoid carrier in Malaysia. *Journal of Bacteriology*, *194*(21), 5970–5971.

Yap, K. P., Teh, C. S. J., Baddam, R., Chai, L. C., Kumar, N., Avasthi, T. S., … Thonga, K. L. (2012). Insights from the genome sequence of a Salmonella enterica serovar typhi strain associated with a sporadic case of typhoid fever in Malaysia. *Journal of Bacteriology*, *194*(18), 5124–5125.

Yap, K.P., Gan, H. M., Teh, C. S. J., Chai, L. C., & Thong, K. L. (2014). Comparative genomics of closely related Salmonella enterica serovar Typhi strains reveals genome dynamics and the acquisition of novel pathogenic elements. *BMC Genomics*, *15*(1), 1007.

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., … Larsen, M. V. (2012). Identification of acquired antimicrobial resistance genes. *Journal of Antimicrobial Chemotherapy*, *67*(11), 2640–2644.

Zerbino, Daniel, & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using the Brujin graphs. *Genome Research*, *18*(5), 821–829.

Zhang, H., Zhang, X., Yan, M., Pang, B., Kan, B., Xu, H., & Huang, X. (2011). Genotyping of Salmonella enterica serovar Typhi strains isolated from 1959 to 2006 in China and analysis of genetic diversity by genomic microarray. *Croatian Medical Journal*, *52*(6), 688–693.

Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, *9*(1), 40.

Zhao, Y., Wu, J., Yang, J., Sun, S., Xiao, J., & Yu, J. (2012). PGAP: Pan-genomes analysis pipeline. *Bioinformatics*, *28*(3), 416–418.

Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J., & Wishart, D. S. (2011). PHAST: a fast phage search tool. *Nucleic Acids Research*, *gkr485*.

Zhou, Z., McCann, a., Weill, F.-X., Blin, C., Nair, S., Wain, J., … Achtman, M. (2014). Transient Darwinian selection in Salmonella enterica serovar Paratyphi A during 450 years of global spread of enteric fever. *Proceedings of the National Academy of Sciences*, *111*(33), 12199–12204.

LIST OF RELATED PUBLICATIONS AND CONFERENCE PROCEEDINGS

Publications (ISI)

Yap, K. P., Ho, W. S., Gan, H. M., Chai, L. C., & Thong, K. L. (2016). Global MLST of *Salmonella* Typhi Revisited in Post-genomic Era: Genetic Conservation, Population Structure, and Comparative Genomics of Rare Sequence Types. *Frontiers in Microbiology,* 7.

Yap, K. P., Gan, H. M., Teh, C. S., Chai, L. C., & Thong, K. L. (2014). Comparative genomics of closely related *Salmonella enterica* serovar Typhi strains reveals genome dynamics and the acquisition of novel pathogenic elements. *BMC Genomics*, *15*(1), 1007.

Yap, K.P., Gan, H.M., Teh, C.S.J., Baddam, R., Chai, L.C., Kumar, N., … Thong, K.L. (2012). Genome Sequence and Comparative Pathogenomics Analysis of a *Salmonella enterica* Serovar Typhi Strain Associated with a Typhoid Carrier in Malaysia. *Journal of Bacteriology*, *194*(21), 5970.

Yap, K.P., Teh, C.S.J., Chai, L.C., Baddam, R., Kumar, N., Avasthi, T.S., ... Thong, K.L. (2012). Insights from the genome sequence of a *Salmonella enterica* serovar Typhi strain associated with a sporadic case of typhoid fever in Malaysia. *Journal of. Bacteriology*, *194*(18), 5124.

Baddam, R., Thong, K.L., Avasthi, T.S., Shaik, S., **Yap, K.P**., Teh, C.S., Chai, … Ahmed, N. (2012). Whole-genome sequences and comparative genomics of *Salmonella enterica* serovar Typhi isolates from patients with fatal and nonfatal typhoid fever in Papua New Guinea. *Journal of Bacteriology*, *194*(18), 5122-5123.

Baddam, R., Kumar, N., Thong, K.L., Ngoi, S.T., Teh, C.S.J., **Yap, K.P**., ... Ahmed, N. (2012). Genetic fine structure of a *Salmonella enterica* serovar Typhi strain associated with the 2005 outbreak of typhoid fever in Kelantan, Malaysia. *Journal of Bacteriology*, *194*(13), 3565-3565.

Chelvam, K. K., **Yap, K. P**., Chai, L. C., & Thong, K. L. (2015). Variable Responses to Carbon Utilization between Planktonic and Biofilm Cells of a Human Carrier Strain of *Salmonella enterica* Serovar Typhi. *Plos One,* e0126207.

Yap, K. P., Thong, K. L. The Plethora of *Salmonella Typhi* Genomes in Typhoid Research: Revolutionizing Genomics, Envisaging Future. (*Under submission*)

Conference Abstract & Proceedings

Yap, K.P. (2014). Comparative genomics study of *Salmonella enterica* serovar Typhi strain isolated from a large typhoid outbreak in Malaysia using whole next generation sequencing approach. *High Impact Research Student Seminar*, University of Malaya, Malaysia. (Oral Presentation, National)

Yap, K.P., Chai, L.C., Thong, K.L. (2015). Epidemiological Investigation of a large typhoid outbreak: Whole Genome Sequencing Reveals Close Ancestry with a Human Carrier Strain. *Postgraduate Colloquium on Medical Sciences*, Faculty of Medicine, UiTM, Malaysia.(Oral Presentation, National) – Recipient of Young Scientist Award, Best Speaker Award, Champion)

Yap, K.P., Ho, W.C., Chai, L.C., Thong, K.L. (2015). MLST of *Salmonella* Typhi: Genetic Conservation, Population Structure, and WGS Phylogeny Comparison. *22nd Malaysian Society for Molecular Biology and Biotechnology Scientific Meeting*. (Poster Presentation, National)

Yap, K.P., Ho, W.C., Chai, L.C., Thong, K.L. 2015. The population structure of *Salmonella* Typhi is highly homogeneous: Evidence from MLST and Phylogenomic Analyses. *National Postgraduate Seminar*. (Poster Presentation, National)

Yap, K. P., Gan, H. M., Chai, L. C., Thong, K. L. (2014). Sporadic strain revisited: The genome dissection of *Salmonella enterica* serovar Typhi strain with whole-genome sequencing. *National Postgraduate Seminar*. (Poster Presention, National) – Best Poster Award

Yap, K. P., Gan, H. M., Chai, L. C., Thong, K. L. (2014). Genome-wide study of an outbreak associated *Salmonella enterica* serovar Typhi strain in Malaysia unveiled potential function-altering nsSNPs using next-generation sequencing approach. *21st Malaysian Society for Molecular Biology and Biotechnology Scientific Meeting*. (Poster Presentation, National)

Yap, K. P., Gan, H. M., Chai, L. C., Thong, K. L. (2013). Unveiling the genome dynamics of sporadic-associated *Salmonella enterica* serovar Typhi strain in Malaysia using whole genome next generation sequencing approach. *20th Malaysian Society for Molecular Biology and Biotechnology Scientific Meeting*. (Poster Presentation, National) – Best Poster Award

Yap, K.P. (2012) Genome architecture and comparative genomics of outbreak-associated *Salmonella enterica* serovar Typhi strain in Malaysia. *13th Biological Science Graduate Congress (BSGC)*, University of Chulalongkorn, Bangkok, Thailand. (Oral Presentation, International) – Best Speaker Award (Silver Medalist)

Yap, K. P., Gan, H. M., Chai, L. C., Thong, K. L. (2012). Comparative genomics of *Salmonella enterica* serovar Typhi strain P-stx-12 with CT18 and Ty2. *19th Malaysian Society for Molecular Biology and Biotechnology Scientific Meeting*. (Poster Presentation, National)

Yap, K. P., Teh, C.S.J., Chai, L. C., Thong, K. L. (2012). Genome sequence and comparative genomics of a *Salmonella enterica* serovar Typhi strain in Malaysia. *National Postgraduate Seminar*. (Poster Presentation, National)

Thong, K. L., **Yap, K. P**., Chai, L. C. (2015). MLST scheme of *Salmonella* Typhi revisited: Subtyping Showed Concordance Clustering with Whole Genome Phylogeny. 9th *International Conference on Typhoid and Invasive NTS Disease,* Bali. Indonesia. (Poster Presentation, National)

Thong, K. L., **Yap, K. P**., Teh, C.S.J., Chai, L. C. (2013). Whole genome sequence analysis of an outbreak strain of *Salmonella enterica* serovar Typhi reveal non-synonymous SNPs. *American Society of Microbiology Conference on Salmonella*, Boston, Massachusetts, USA. (International, Oral Presentation)

Thong, K. L., **Yap, K. P**., Gan, H. M., Teh, C. S., Chai, L. C., (2012). Unveiling the genome dynamics of large outbreak-associated *Salmonella enterica* serovar Typhi strain in Malaysia. *31st Symposium of the Malaysian Society for Microbiology*, Sabah, Malaysia. (International, Plenary/Keynote Talk)

Thong, K. L., **Yap, K. P**., Gan, H. M., Teh, C. S., Chai, L. C., (2012). Comparative genome sequence of a *Salmonella enterica* serovar Typhi strain associated with a sporadic case of typhoid fever in Malaysia. *International Conference on Clinical Microbiology and Microbial Genomics,* Hilton San Antonio Airport, USA. (International, Invited Speaker)

LIST OF NON-RELATED PUBLICATIONS AND CONFERENCE PROCEEDINGS

Publications (ISI)

Ho, W. S., Gan, H. M., **Yap, K. P**., Balan, G., Yeo, C. C., Thong, K. L. (2012). Genome sequence of multidrug-resistant *Escherichia coli* EC302/04, isolated from a human tracheal aspirate. *Journal of Bacteriology*, *194*(23), 6691-6692.

Suhaimi, N. S. M., **Yap, K. P**., Ajam, N., Thong, K. L. (2014). Genome sequence of *Kosakonia radicincitans* UMEnt01/12, a bacterium associated with bacterial wilt diseased banana plant. *FEMS Microbiology Letters*. *358*(1), 11-13.

Osama, A., Gan, H. M., Teh, C. S. J., **Yap, K. P**., Thong, K. L. (2012). Genome sequence and comparative genomics analysis of a *Vibrio cholerae* O1 strain isolated from a cholera patient in Malaysia. *Journal of Bacteriology, 194*(24), 6933-6933.

Ho, W. S., **Yap, K. P**., Yeo, C. C., Rajasekaram, G., Thong, K. L. (2015). The Complete sequence and comparative analysis of a multidrug-resistance and virulence multireplicon IncFII plasmid pEC302/04 from an extraintestinal pathogenic *Escherichia coli* EC302/04 indicate extensive diversity of IncFII plasmids. *Frontiers in microbiology*, 6.

Lim, S. Y., **Yap, K. P**., Teh, C. S. J., Jabar, K. A., Thong, K. L. (2016). Comparative genome analysis of multiple vancomycin-resistant Enterococcus faecium isolated from two fatal cases. *Infection, Genetics and Evolution*. *49*,55-65.

Lim, S. Y., **Yap, K. P**., Thong, K. L. (2016). Comparative genomics analyses revealed two virulent Listeria monocytogenes strains isolated from ready-to-eat food. *Gut Pathogens*, *8*(1), 65.

Chuah, L. O., **Yap, K. P**., Thong, K. L., Liong, M. T., Ahmad, R., Shamila-Syuhada, A. K., & Rusul, G. (2016). Genome sequence of “Anthococcus,” a novel genus of the family Streptococcaceae isolated from flowers. *Genome Announcements*, *4*(6), e01410-16.

Conference Abstracts and Proceedings

Thong, K.L., **Yap, K.P**., Ho, W.S, Ngoi, S. T., (2015). Microbial Genomics of Bacterial Pathogens: insights into pathogen evolution, host adaptation, resistance mechanism and phylogeography, *40th Annual Conference of the Malaysian Society for Biochemistry & Molecular Biology,* Malaysia (National)

Lim, S.Y., **Yap, K.P**., Teh, C.S.J., Thong, K.L. (2015). Comparative Genomic Analysis of Clinical Vancomycin-resistance *Enterococcus faecium*. *40th Annual Conference of the Malaysian Society for Biochemistry & Molecular Biology*, Malaysia (National)

Lim, S.Y., **Yap, K.P**., Teh, C.S.J., Thong, K.L. (2015). Genome Plasticity Reveals Genetic Variation of Clinical Vancomycin-resistant *Enterococcus faecium*. International Congress of the Malaysian Society for Microbiology, Malaysia. (International)

Ung, E.H., Thong, K.L., Yew, S.M., Choo, S. W., Wee, W.Y., **Yap, K.P**. (2014). ZOT-proteins occur in conjunction with E-family virulence factors in AHPND-causing Vibrio parahaemolyticus associated with either of three prophage elements in their genome. 9th Symposium on Diseases in Asian Aquaculture. Ho Chi Minh, Vietnam. (International)

Ung, E.H., Thong, K.L., Yew, S.M., Choo, S. W., Wee, W.Y., **Yap, K.P**. (2014). An AP1, 2 & 3 PCR positive non-Vibrio parahaemolyticus bacteria with AHPND histopathology. 9th Symposium on Diseases in Asian Aquaculture. Ho Chi Minh, Vietnam. (International)