Molecular typing and evolutionary relationships of Salmonella enterica serovar Typhi

Author:
Octavia, Sophie

Publication Date:
2008

DOI:
https://doi.org/10.26190/unsworks/17952

Molecular typing and evolutionary relationships of *Salmonella enterica* serovar Typhi

Sophie Octavia

School of Biotechnology and Biomolecular Sciences

The University of New South Wales

Australia

A thesis submitted for the degree of Doctor of Philosophy

March 2008

PLEASE TYPE

1.3.1 THE UNIVERSITY OF NEW SOUTH WALES

Thesis/Dissertation Sheet

Surname or Family name: **Octavia**

First name: **Sophie** Other name/s:

Abbreviation for degree as given in the University calendar: **PhD in Microbiology**

School: **Biotechnology and Biomolecular Sciences** Faculty: **Science**

Title: **Molecular typing and evolutionary relationships of *Salmonella enterica* serovar Typhi**

Abstract 350 words maximum: (PLEASE TYPE)

The evolutionary relationship between *Salmonella enterica* serovar Typhi, other typhoid-like enteric fever causing serovars and 10 non-Typhoid serovars from *S. enterica* subspecies I, could not be determined by comparative nucleotide sequences of six genes. Phylogenetic analyses of the dataset showed that the genes of interest underwent frequent recombination, suggesting a low level of clonality within subspecies I of *S. enterica*.

To establish the evolutionary relationships within serovar Typhi, genome-wide Single Nucleotide Polymorphism (SNP) was explored as a marker for both typing purposes and phylogenetic analysis. Thirty eight SNPs were typed in 73 global Typhi isolates, including 18 isolates expressing the special flagellar antigen z66, using restriction enzyme digestion method. The isolates were differentiated into 23 SNP profiles and grouped into four distinct clusters. The z66 isolates were divided into four SNP profiles and were all grouped into one cluster, suggesting a single origin.

An alternative SNP typing method using the hairpin real time PCR assay was investigated to type four additional SNPs, termed as biallelic polymorphisms (BiP). These BiPs were found to classify 481 global Typhi isolates into five major clusters (Roumagnac *et al.*, 2006). Typing four BiPs resulted in the identification of four additional SNP profiles. We proposed nine SNPs were required to type Typhi isolates into 13 subclusters for global epidemiology.

An enzymatic-based method using *Cel*I nuclease was evaluated to discover more SNPs from other Typhi genomes. The efficiency of the *Cel*I was shown to be unsatisfactory and we were unable to demonstrate the effectiveness of the proposed method.

Nine Variable Number of Tandem Repeats (VNTRs) were typed in the 73 Typhi isolates using fluorescent-labelled universal primers, and analysed on an automated DNA sequencer. Five isolates were unable to give PCR products in one or more VNTR loci. Nine VNTRs could differentiate 68 Typhi isolates into 65 MLVA profiles, suggesting a higher discriminating power than SNP typing. SNPs were shown to be a more appropriate marker for phylogenetic tracing for Typhi while VNTRs were highly discriminating but could not be used to establish the evolutionary relationships of diverse Typhi isolates..

Originality statement

‘I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.’

Signed …….....

Date …….....

Acknowledgement

I would like to thank my supervisor, Dr. Ruiting Lan, for his invaluable advice, infinite patience and continuous guidance since my PhD year. I also would like to thank my PhD committees, Prof. Hazel Mitchell and Dr. Mark Tanaka, for their constructive feedback and comments during my PhD progress reviews.

I would like to thank members of Evolutionary Microbiology Lab and all the people in Lab 301 (Helicobacter and Campylobacter Research Lab and Epsilon Proteobacteria Lab) for making the lab enjoyable. Special thanks to Alfred for being a considerate, resourceful and entertaining neighbour both in the lab and in the office. Thank you to the 2007 Honours Students (Kyi, Nat and Stan) for an extremely fun year.

I am very grateful to the proofreaders (Alvin, Arash, Connie, Garry, Krisa, Luis, Nadeem, Phebe, Ram, Reg, Stan and Winnie) of my thesis (voluntarily and involuntarily); my friends, for "all the best" wishes, prayers and believing in me; and last but not least to my family for their continuous moral supports, a big thank you to my loving and understanding parents who regularly provide me nutritious meals. This thesis is dedicated to you!

"I can do all things through Christ who strengthens me." **Phil 4:13**

Abstract

Salmonella enterica serovar Typhi (simply referred to as Typhi) causes typhoid fever, a serious systemic infection spread by the faecal-oral route. This thesis investigated the relationships of Typhi to other serovars belonging to *S. enterica* subspecies I and between isolates within the Typhi clone.

Initial attempt was made to determine the evolutionary relationship between Typhi and the other typhoid-like enteric fever causing serovars including serovars Paratyphi A, Paratyphi B clone b1, Paratyphi C and Sendai. The nucleotide sequences of six genes were analysed in these serovars (excluding Sendai) as well as in 10 non-typhoid serovars from *S. enterica* subspecies I. The relationships between the serovars were unable to be established as the phylogenetic analyses of the dataset showed that the genes studied underwent frequent recombination. This suggested a low level of clonality within subspecies I of *S. enterica* and this new finding has challenged the current notion that *S. enterica* is highly clonal.

Current DNA-based typing methods such as pulse field gel electrophoresis and ribotyping are not useful to establish the evolutionary relationships of Typhi. Population structure studies using Multilocus Enzyme Electrophoresis and Multilocus Sequence Typing (MLST) have shown that Typhi is highly homogeneous. For epidemiological investigation and understanding the evolution of the pathogen, a method that can fulfil both purposes will be ideal. Therefore, genome-wide Single Nucleotide Polymorphism (SNP) was explored as a marker for both typing purposes and phylogenetic analysis. The comparison of two complete genomes for serovar Typhi strains Ty2 and CT18 allowed 38 SNPs to be selected and these were typed using restriction enzyme digestion.

Seventy-three isolates from around the world, isolated between 1981-1995, were SNP-typed. They were differentiated into 23 SNP profiles, 12 profiles were represented by a single isolate and 11 profiles were shared by more than one isolate. The overall relationships of these SNP profiles were visualised using a minimum spanning network. The relationships of most SNP profiles could be resolved and these profiles divided into four major clusters, cluster I to cluster IV. We have shown

that SNPs were useful markers for establishing the relationships of Typhi isolates and were more discriminating than MLST.

The 18 isolates expressing the z66 flagellar antigen included in the study were divided into four SNP profiles. These SNP profiles were grouped into one cluster, suggesting a single origin. Our evolutionary analysis suggested that Typhi was originally monophasic having only an H1 antigen and then gained a new phase-2-flagellin like operon only recently.

An alternative SNP typing method using the hairpin real time PCR assay was investigated to type four additional SNPs. These SNPs, termed as biallelic polymorphisms (BiP), were found by Roumagnac *et al.* (2006, Science **314:**1301-1304) to classify global Typhi isolates into five major clusters. These BiPs, BiP 36, BiP 48, BiP 56 and BiP 33, were selected to type the 73 Typhi isolates used in this study. Two BiPs, BiP 36 and 33, were found to be unique in one isolate each. Typing four BiPs resulted in the identification of four additional SNP profiles and identified one new cluster. Based on the data of four BiPs, the SNP profiles were distinguished into six clusters where each cluster was supported by a combination of four BiPs. BiP 48 appeared to have undergone parallel or reverse changes which led to conflicting phylogenetic signals (homoplasy). This contradicted previous findings by Roumagnac *et al.* where no homoplasy was observed in all of the BiPs analysed. The results suggested that there was recombination within the Typhi clone.

This study has identified subsets of SNPs for global epidemiology and typing of Typhi at three different levels: 1) the major clusters; 2) 13 subclusters; and 3) individual SNP profiles. Three SNPs and four BiPs are required to identify the major clusters for our clustering scheme and Roumagnac *et al.* clustering scheme respectively. Nine SNPs (five SNPs from our study and four BiPs from Roumagnac *et al.*) can divide the isolates into 13 subclusters and a minimum of 19 SNPs (15 SNPs and four BiPs) can differentiate Typhi isolates into 27 SNP profiles.

The SNPs selected for our SNP typing scheme were recognised by comparing two completed genomes of Typhi strains, CT18 and Ty2 respectively. However, the identified SNPs only revealed polymorphisms and their locations along the evolutionary path between strains CT18 and Ty2. By comparing only two genomes, any SNPs developed before the divergence of these two Typhi

strains or other isolates diverged from them could not be detected. This phenomenon is known as phylogenetic discovery bias. An enzymatic-based method using a mismatch specific *Cel*I nuclease was evaluated to discover more SNPs from other Typhi isolates. Most of the steps in the method have been shown to be successful. However, the cleavage activity of *Cel*I nuclease was found to be insufficient for our purpose of discovering new SNPs.

Although SNPs were valuable markers for studying the evolutionary genetics in Typhi, SNP typing clearly still has a limited discriminatory power. Several SNP profiles contained multiple isolates. Another molecular marker, Variable Number of Tandem Repeat (VNTR) was employed and has been shown to be highly discriminating in two MLVA studies of Typhi isolates. We explored the genome of Typhi CT18 to find more polymorphic VNTRs. Forty-six potential VNTRs were identified, however only two were found to be polymorphic. Nine VNTRs, with seven most polymorphic VNTRs from the previous two MLVA studies, were used for typing the 73 global Typhi isolates used for the SNP typing. Five isolates failed to give PCR products in one or more VNTR loci, therefore a complete data for nine VNTRs was only available for 68 Typhi isolates.

VNTR typing was shown to be more discriminating than SNP typing. The 68 Typhi isolates could be differentiated into 65 MLVA profiles. Typing four of the nine loci could give the same level of discrimination. No clustering was observed when VNTR data alone were used to construct the phylogenetic relationships of the Typhi isolates analysed. Minimum spanning trees were generated to analyse the MLVA data at a cluster level according to the four major clusters defined by SNP typing. The relationships were congruent for the majority of the profiles from cluster I, II and IV. However, in the most divergent cluster, cluster III, the relationships revaled by SNP data were conflicting to the VNTR data. Therefore, VNTRs could be used to determine the relationships of closely related isolates. However, in more diverse Typhi isolates, VNTRs were only useful to reveal the extent of genetic diversities. SNPs are required to determine the phylogenetic relationships of these diverse isolates. In conclusion, SNP typing is useful to determine the evolutionary relationships of the Typhi isolates but has a limited discriminatory power. On the other hand, VNTR typing has a high discriminatory power but it is unable to establish the relationships of diverse isolates. A combination of these two methods offers the best approach for local and global epidemiology and the evolutionary analysis of Typhi.

Publications and Conference proceedings arising from this thesis

Publications

Octavia, S., and R. Lan. 2007. Single Nucleotide Polymorphism typing and genetic relationships of *Salmonella enterica* serovar Typhi isolates. Journal of Clinical Microbiology **45:**3795 - 3801.

Octavia, S., and R. Lan. 2006. Frequent recombination and low level of clonality within *Salmonella enterica* subspecies I. Microbiology **152:**1099-1108.

Octavia, S. 2007. Origins and molecular evolution of *Salmonella typhi*. *in* Focus: Australian Society for Microbiology Syntrophy **8**:1, 11.

Oral Presentations

Octavia, S., and R. Lan. 2007. Molecular evolution of *Salmonella enterica* serovar Typhi *in* Australian Evolution Society 5th Annual Conference. Sydney, Australia.

Octavia, S., and R. Lan. 2006. Molecular molecular evolution and Single Nucleotide Polymorphism typing of *Salmonella enterica* serovar Typhi *in* Australian Society of Microbiology Annual Conference. Gold Coast, Australia.

Poster Presentations

Octavia, S., and R. Lan. 2007. Evolutionary origins of *Salmonella enterica* serovar Typhi *in* Australian Society of Microbiology Annual Conference. Adelaide, Australia.

Octavia, S., and R. Lan. 2006. Frequent recombination and low level of clonality within *Salmonella enterica* subspecies I *in* Australian Society of Microbiology Annual Conference. Gold Coast, Australia.

Octavia, S., and R. Lan. 2005. Single Nucleotide Polymorphism typing of *Salmonella enterica* serovar Typhi *in* Australian Society of Microbiology Annual Conference. Canberra, Australia.

Octavia, S., and R. Lan. 2004. How clonal is *Salmonella enterica*? *in* Australian Society of Microbiology Annual Conference. Sydney, Australia.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Amp	ampicillin
AFLP	amplified fragment length polymorphism
BiP	biallelic polymorphism
bp	base pair
Ct	cycle treshold
dNTP	deoxynucleotide triphosphate
EDTA	ethylenediaminetetra acetic acid
ET	electrophoretic type
HP	hairpin
hr	hour
IEC	intestinal epithelial cell
IPTG	isopropyl thiogalactoside
IS	insertion sequence
kb	kilobase
LB	luria bertani
LPS	lipopolysaccharide
M	molar
MDR	multidrug resistant
min	minute
ML	maximum likelihood
MLEE	multilocus enzyme electrophoresis
MLST	multilocus sequence typing
MLVA	multilocus VNTR analysis
MST	minimum spanning tree
NA	nutrient agar
NJ	neighbour-joining
PAUP	phylogenetic analysis using parsimony
PFGE	pulse field gel electrophoresis
RE	restriction enzyme

RNase	ribonuclease
R-T PCR	Real time PCR
SARA	*Salmonella* reference collection A
SARB	*Salmonella* reference collection B
SARC	*Salmonella* reference collection C
sec	second
SNP	single nucleotide polymorphism
SPI	*Salmonella* pathogenicity island
TBE	tris-borate EDTA
TE	tris EDTA
U	unit
UPGMA	unweighted pair group method with arithmetic means
VNTR	variable number of tandem repeat

Chapter 1: Literature review

1.1 Clones and the concept of clonality

As bacteria reproduce asexually by binary fission, their populations are considered to be groups of clones deriving from a common ancestor. This implies that genetic transmission occurs vertically from parent to daughter cell and any genetic differences between them are a result of mutation. Recombination in the form of horizontal gene transfer and lateral genetic transfer could also occur and contribute to the diversification of bacterial clones. The proportion of mutational forces versus recombination affects the population structure of bacteria. It could range from strongly clonal, as in the case of *Bacillus anthracis* where little or no recombination has occurred during the evolution of the species (260), to non-clonal or panmictic, as in *Helicobacter pylori* (301) where the clones are relatively unstable due to frequent recombination and the alleles being randomly associated. Many species of bacteria, such as *Escherichia coli*, are believed to have an intermediate population structure where recombination contributes more towards diversification than that of mutation but it is not sufficient to prevent the emergence of clonal complexes (98). Population structure studies of bacterial pathogens integrate epidemiological, phylogenetic and evolutionary relationships of pathogens to provide a better understanding regarding the behaviour of the bacterial pathogens.

1.2 Salmonella enterica

The genus *Salmonella* comprises of a group of Gram-negative bacteria belonging to the family Enterobacteriaceae and has been assigned into more than 2,500 different serovars (240). The classification and identification of these serovars were originally based on the Kauffman-White serotyping scheme. This scheme accounts for the difference in antigenic properties of *Salmonella*, including the somatic lipopolysaccharide (O antigens) and the flagellin proteins (flagella H1 and H2 antigens).

The O antigen is a polysaccharide side chain that is located on the cell surface. It is a linear polymer, which consists of oligosaccharide repeats of three to six sugar residues and is linked through an oligosaccharide core to lipid A. Antigenic variation is determined by the type and the arrangement of the sugar residues and about 60 types have been recognised (241). The polymorphism in the O antigen is mainly resulted from the genes of the O antigen gene cluster.

Flagella H1 and H2 are encoded by the *fliC* and *fljB*, respectively. Most serovars of *S. enterica* can undergo phase variation as they have alternative expression of the two flagellin genes (diphasic). This phase variation is mediated by the reversible site-specific inversion of a DNA fragment, *hin* (294). However, many serovars, for example serovar Typhi could only have one active H1 flagellin gene (monophasic) while serovars Gallinarum and Pullorum are aflagellate.

Each distinctive combination of O and H antigens is formally recognised as a serovar. Originally, these serovars were designated by Latin binomial species names. However, because of their close relatedness, the species names were then retained as the serovar names of the single *Salmonella* species known as *Salmonella enterica* (159). For example, the name *Salmonella typhi* refers to *S. enterica* serovar Typhi or simply Typhi (the latter convention is used in this thesis).

1.2.1 Population structure and genetic relationship between subspecies

DNA hybridisation and biotyping have classified the *S. enterica* serovars into seven subspecies: I, II, IIIa, IIIb, IV, V and VI (51, 157). Multilocus enzyme electrophoresis (MLEE), which differentiates isolates on the basis of the relative electrophoretic mobilities of the selected metabolic enzymes, has defined an eighth group, designated as subspecies VII, consisting of only five isolates from two serovars that were initially allocated to subspecies IV on the basis of biochemical characteristics (261). Furthermore, based on MLEE data, subspecies I, IIIa, IIIb and VI were separately clustered from subspecies II, IV and VII. Subspecies V was clustered separately from the other subspecies which further confirms that it is the most divergent *S. enterica* subspecies (261). It has also been considered as a separate species of *Salmonella* and is named *S. bongori* although this designation is not universally accepted.

MLEE has been extensively used to estimate the genetic relatedness and diversity within *S. enterica* natural populations. Based on large-scale MLEE studies, three reference collections have been established by Selander's group: *Salmonella* reference collection A (SARA), which consists of 72 strains of serovar Typhimurium and its closely related serovars Heidelberg, Muenchen, Paratyphi B and Saintpaul, isolated from a variety of hosts and environmental sources (22); *Salmonella* reference collection B (SARB), consisting of 72 strains of 37 subspecies I serovars (33); and *Salmonella* reference collection C (SARC), which consists of 16 strains representing the eight subspecies (34).

MLEE has shown that many serovars vary genetically and have multiple electrophoretic types (ETs) (21, 22, 261, 285). Some serovars have been shown to be genotypically heterogeneous. For example, serovars Derby and Newport (21) include divergent isolates where the ETs are clustered distantly in MLEE trees, while other serovars are confined within a single cluster of closely related ETs and have a predominant widely distributed ET (21, 22, 261, 285).

Nevertheless, the population structure of *S. enterica* is considered to be clonal, with strong linkage disequilibrium displayed by non-random associations between the alleles of the 24 metabolic enzyme loci studied (21, 22, 261, 285). The low recombination rate has also been demonstrated by the sequence data of six housekeeping genes: proline permease (*putP*) (214), glyceraldehyde-3-phosphate (*gapA*) (217), 6-phosphogluconate dehydrogenase (*gnd*) (215), malate dehydrogenase (*mdh*) (31) and isocitrate dehydrogenase/phosphatase (*icd* (333) and *aceK* (216)) from the 16 SARC strains where gene trees for the six housekeeping genes have been shown to be largely congruent. A phylogenetic tree derived from these gene trees could be used to demonstrate the relationship between subspecies, and to illustrate the virulence factors that were acquired with the emergence of the subspecies (Figure 1.2-1).

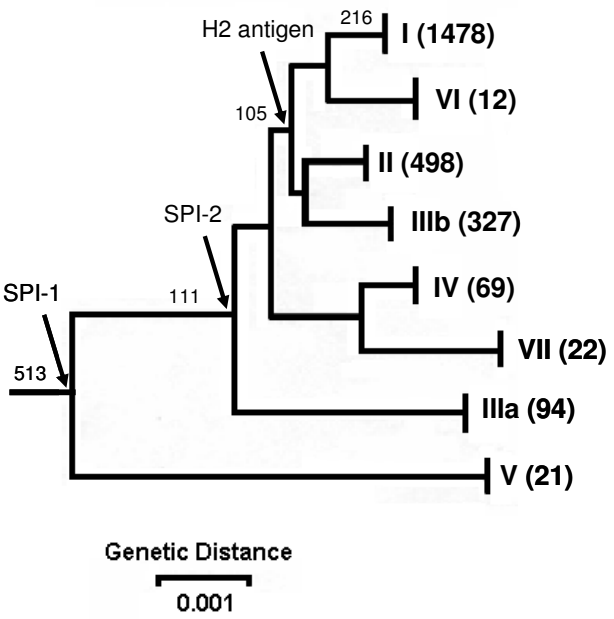


Figure 1.2-1. Phylogenetic tree of *S. enterica* showing the evolutionary relationships of the subspecies determined by sequences of five housekeeping genes as adopted from Lan *et al.* (156).The number of serovars reported for each subspecies is specified in parentheses, after the subspecies number. The arrows indicate the gain of SPI-1, SPI-2 and H2 antigen genes. The numbers at each node correspond to the number of genes gained as determined by microarray analysis of the genome of *S. enterica* serovar Typhimurium strain LT2.

Most subspecies of *S. enterica* are not usually associated with disease and some are commensals in cold-blooded animals (18). However, serovars belonging to subspecies I, which comprises approximately 60% of the total identified serovars, can cause intestinal infections in warm-blooded animals (salmonellosis), and are responsible for 99% of *Salmonella*-related infections in humans (286). The diseases caused by this species range from mild gastroenteritis to a more serious systemic disease called enteric fever, including typhoid fever and paratyphoid fever, a milder form of typhoid fever. Some serovars could be host generalists. For example, the widely prevalent serovar Typhimurium causes gastroenteritis in humans but mainly asymptomatic chronic infection in chickens, while a number of serovars have a restricted host range; for example, Typhi, which causes typhoid fever, exclusively infects humans.

1.2.2 *Salmonella* Pathogenicity Island

S. enterica contains a large number of genetic segments which vary in size and are thought to have been acquired by horizontal gene transfer (95). These regions are associated with several virulence determinants enabling *S. enterica* to colonise, invade and persist in a variety of different hosts, and thus are designated *Salmonella* pathogenicity island (SPI). At present, there are at least 10 different SPIs - five (SPI-1 to SPI-5) of which were discovered by genetic analysis (95) and five others (SPI-6 to SPI-10) by genome sequencing (231) - that have been described (Table 1.2-1) and they have lower G+C composition than the rest of the chromosome. Typically, the percentage of G+C content of *S. enterica* except *S. bongori* is 52% (http://www.sanger.ac.uk) while the percentage of these SPIs ranges from 38.1% to 56.7%. All of SPIs except SPI-1 and SPI-9 are located adjacent to tRNA loci, suggesting the role of these loci as hotspots for acquisition of DNA elements by horizontal gene transfer. Some of these SPIs are conserved throughout *S. enterica* species while others are specific only to certain serovars. The varying distribution of SPIs suggests that horizontal gene transfer was involved in the adaptation of a prototrophic free living *Salmonella* species to a pathogen. *S. enterica* could exist as an intestinal commensal or invade the intestinal mucosa of different hosts.

Table 1.2-1. Characteristics of Salmonella pathogenicity islands

Name	%G+C	Location of insert	Distribution	Function	References
SPI-1	47	*flhA-mutS*	*Salmonella* spp.	Type III secretion system (T3SS) and iron uptake	Reviewed by Lostroh *et al.* (177)
SPI-2	44.6	tRNA *valV*	*S. enterica*	T3SS	Reviewed by Kuhle at al. (152)
SPI-3	47.3	tRNA *selC*	*Salmonella* spp.	Magnesium uptake	(7, 26, 27)
SPI-4	44.8	tRNA like structure	*Salmonella* spp.	Type I secretion system (T1SS) and survival in macrophage	(143, 346)
SPI-5	43.6	tRNA *serT*	*Salmonella* spp.	T3SS effectors	(347)
SPI-6	51.5	tRNA *aspV*	subsp. I and some in IIIB, IV, VII	Fimbriae	(84, 231)
SPI-7	49.7	tRNA *pheU*	subsp. I	Vi antigen, pilus assembly and SopE	(238)
SPI-8	38.1	tRNA *pheV*	sv. Typhi	Bacteriocin related resistance	(231)
SPI-9	56.7	Prophage	subsp. I	T1SS and putative toxin	(231)
SPI-10	46.6	tRNA *leuX*	subsp. I	Sef fimbriae	(231)

spp. – species, subsp. – subspecies and sv. – serovar

1.3 Evolution of the human-adapted clones causing enteric fever

The serovars that can cause enteric fever in humans include serovar Typhi, the causative agent of typhoid fever, and serovars Paratyphi A, Paratyphi B (ET Pb1), Paratyphi C and Sendai, which cause paratyphoid fever. MLEE has been used to analyse the genetic diversity and evolutionary relationships among 761 *S. enterica* isolates. These include human-adapted enteric fever-causing serovars as well as 22 other non-enteric fever-causing serovars belonging to subspecies I (285). Serovar Typhi has been shown to be very distinct in comparison to other paratyphoid causing serovars. MLEE has suggested that there is no close relationship among paratyphoid fever-causing serovars except between Paratyphi A and Sendai even though the two serovars differ in both serotype and biotype. Both of these serovars have also been shown to be closely related to the host-generalist non-invasive serovar Panama. Paratyphi B clone Pb1 was closely related to other clones that cause gastroenteritis, suggesting a recent evolution and this serovar, and was shown to be closely related to serovars Typhimurium, Heidelberg and Saintpaul (285). Lastly, serovar Paratyphi C was shown to be closely related to serovar Choleraesuis although they differ in disease symptoms.

A comparative genomic analysis has been done using *S. enterica* serovar Typhimurium strain SL1344 spotted DNA microarray to determine certain genetic contents which are associated with host adaptation, how they evolved and the phylogenic relationships between various serovars and strains of both *S. enterica* (subspecies I and IIIa) and *S. bongori* (43). The genomic differences between 24 isolates of 12 serovars of *S. enterica* and two strains of *S. bongori* suggest that strains of the same serovar could differ in the presence of absence of genes. The differences in genetic content in the isolates examined were mainly due to prophages, IS elements and genetic compositions in SPIs. A study using an LT2-based array, which was supplemented with annotated chromosomal open reading frames from CT18 that were 10% divergent from LT2, comparing 79 isolates representing all main disease-causing serovars of *S. enterica* also showed similar differences (242).

Serovars Paratyphi C and Choleraesuis were clustered together suggesting a common origin. Typhi was shown to share common genetic features with serovar Paratyphi A and Sendai, which clustered

them together (43). It was shown that there was a cluster of genes absent, including the Lpf fimbrial operon, the *sodC-1* gene encoding for a periplasmic superoxide dismutase and the *avrA* gene, which is a homolog of *Yersinia pseudotuberculosis* secreted effector protein YopJ, all of which have been associated with the virulence of serovar Typhimurium (19, 46, 68). The absence of this gene cluster, unique to the enteric fever-causing serovars, suggests either that they originated from a common ancestor or that there may have been a convergent evolution of these serovars to their host specialisation. It is likely that enteric fever-causing serovars did not evolve by a vertical descent since the genetic distances between the enteric fever-causing serovars are not closer than the genetic distances to the other serovars.

1.4 Pathogenesis and epidemiology of typhoid fever

Typhoid fever is a serious systemic disease involving the reticuloendothelial system and the gall bladder, usually characterised by extended fever, abdominal discomfort, malaise, headache, constipation, rose-coloured spots on the chest, hepatomegaly and splenomegaly. It is caused by *Salmonella enterica* serovar Typhi. Although this typhoid bacillus has been described in the mid-1800s by William Jenner, typhoid fever could only be treated in 1948 after the discovery of the antibiotic chloramphenicol (348).

Infection develops following ingestion of food or water that is contaminated with 10^3-10^6 cfu/ml of Typhi (118). Upon ingestion, Typhi initially traverses through the gastric acid-rich stomach to reach and colonise the intestine. The bacteria then adhere to and invade the epithelial cells of the small intestine, possibly in the distal ileum, and microfold (M) cells. M cells, the specialised epithelial cells overlaying the Peyer's patches, and dendritic cells internalise the bacteria. Subsequently, the bacteria will be distributed to the lymphatic and reticuloendothelial systems in the small intestine, liver and spleen and remain there for a few days before being transported back to the bloodstream (65, 120, 132, 234, 337). Typhoid fever may progress to bacteremia, particularly in patients with underlying diseases such as cancer (318) and immunosuppression (187). Bacteremia could subsequently cause localised tissue infections such as endocarditis, meningitis, renal failure and pneumonia (186, 218, Caers, 2006 #3862, 335). There is approximately 10% chance of mortality if the patients are untreated with antibiotics however, this is decreased to only

1% if treated (http://www.who.int). After antibiotic treatment, the bacteria are usually cleared from the patients, however in up to 3% of patients, asymptomatic chronic infection of the gallbladder develops and subsequently results in persistent colonisation of the intestine (234). These carriers may excrete 10^6 Typhi cells per gram of faeces (283) and could continually spread the disease to other individuals and cause outbreaks (226, 245). Typhoid fever carriers have also been shown to be approximately 8.5 times more likely to develop gallbladder cancer (153, 293).

While typhoid fever has been regarded as 'the disease of history' in most industrialised countries, it is still a devastating disease and is endemic in countries where hygiene and sanitation are poor. There are more than 17 million cases of typhoid fever worldwide, and approximately 600,000 deaths are reported annually (http://www.who.int). Typhi can cause large outbreaks, for example, the outbreak that occurred in Kinshasa, Democratic Republic of Congo in late 2004 resulted in 42,564 cases and 214 deaths (340).

1.5 Differential host responses and disease manifestations between host restricted and host generalist serovars

Although many *S. enterica* serovars could infect a broad range of animal hosts, Typhi is among the few serovars that only infects humans. Currently there is no optimal animal model available to demonstrate the pathogenesis of Typhi. The factors believed to contribute to the host-pathogen interactions during Typhi infection and immunopathology of typhoid fever are mainly identified from studies of Typhimurium infection in mice. Serovar Typhimurium causes gastroenteritis in humans. However, in mice it could cause systemic infection that resembles typhoid fever in humans. A study by Mills *et al.* (200) showed that the mechanisms of invasion and intracellular trafficking in human intestinal epithelial cells (IEC) are similar but there are clearly differences between the two serovars, including the interaction to IEC and survival in the macrophages.

The insertion of a chromosomal region encoding for invasion determinants of Typhi enabled non-invasive *E. coli* strain to enter IEC while insertion of the homologous region from Typhimurium did not (63). During interaction with the IEC, Typhi establishes a transient infection without

significant inflammation. Typhi is transported earlier through the polarised human epithelial cell monolayers and in larger numbers than serovar Typhimurium (147). Nonetheless, Typhi does not trigger the migration of neutrophils across the monolayers unlike serovar Typhimurium (195). This could be due to the fact that the numbers of Typhi cells internalised by the IEC, via the M cells overlying the Peyer's patches, were significantly lower than that observed for Typhimurium. Moreover, Typhi were cleared more rapidly with minor damage to the M cells and the IEC, as shown from the analysis of ileal loop infection in mice (235). The ability of *S. enterica* serovars to cause systemic diseases depends on its ability to survive within macrophages after phagocytosis. Interestingly, serovar Typhi could not persist in the murine macrophages unlike serovar Typhimurium, which could survive in both murine and human macrophages, suggesting a possible role of macrophage interaction in host restriction (328).

1.6 Factors contributing to the pathogenesis of Typhi

Human colonic epithelial cell lines were used to model early interactions of serovar Typhi with IEC and macrophage-like cell lines. These cell lines could be used to model interactions that occur in the lamina propria and have been used to identify pathogen-associated bacterial components and host genetic factors that contribute to the outcome of Typhi infection.

1.6.1 Bacterial factors

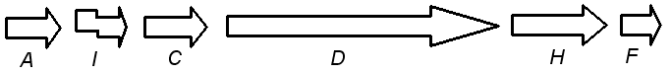
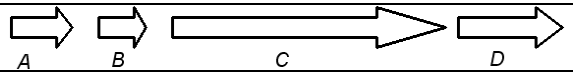
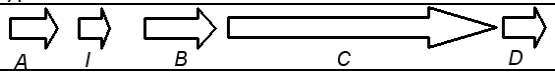
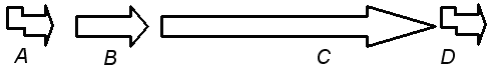
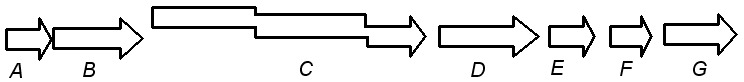
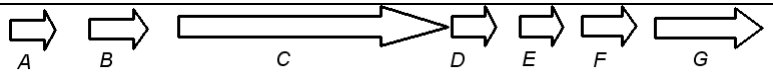
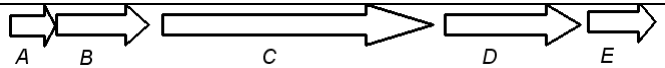
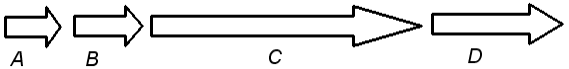
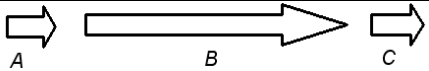
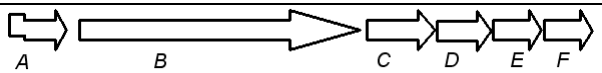
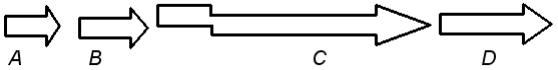
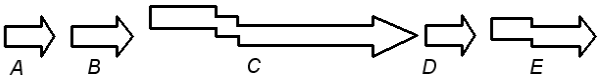
1.6.1.1 *stg* and Type IV pili

Initial colonisation depends on the adhesion of Typhi to the IEC and is mediated by fimbriae or pili. From the genomic analysis of Typhi strain CT18, there are 13 different types of putative fimbrial operons and a type IV pilus encoded by a *pil* gene located on SPI-7 (231). The fimbrial operons are chaperone/usher dependent including (Figure 1.6-1): *bcf* (bovine colonisation factor), *agf* (aggregative fimbriae), *fim* (type 1 fimbriae), *saf* (*Salmonella* fimbriae), *sef* (serovar Enteritidis fimbriae), *sta*, *stb*, *stc*, *std*, *ste*, *stg*, *sth* and *tcf* (Typhi colonisation factor) genes. It is still unknown what selective pressures are involved in generating and maintaining these fimbrial operons.

However, some of the genes within the fimbrial operons of Typhi strain CT18 contain a stop codon or frameshift mutation (Figure 1.6-1). Many of the fimbrial gene sequences identified in serovar Typhi are also found on the genomes of other serovars of subspecies I (315). The presence of a large number of fimbrial sequences appears to be common in the genomes of *S. enteric*, although different serovars may contain a different combination of fimbrial sequences. It is likely that multiple fimbrial operons may compensate the loss of an individual operon as illustrated in serovar Typhimurium. The simultaneous inactivation of the *lpf*, *pef* and *agf* operons results in 26 fold reduction in the virulence of serovar Typhimurium in mice in comparison to inactivation of each individual operon (321).

None of the fimbrial operons were found to be restricted to enteric fever causing serovars (315). Early Southern hybridization has suggested that *tcf* was restricted to serovar Typhi and was thought to be involved in the adaptation to human host (83). Further hybridisation analyses on a larger number of serovars indicate that typhoidal serovars Paratyphi A, Sendai (but not Paratyphi B and Paratyphi C) and nontyphoidal serovars, including Heidelberg, Choleraesuis, Montevideo, Typhisuis and Muenchen also contain this operon (315). This suggests that *tcf* operon is not exclusive to Typhi and it is not a unique characteristic of enteric fever causing serovars.

Figure 1.6-1. The feature of 12 fimbrial operons found in Typhi CT18

Operon Name	ORF Name	Genetic Organisation[1]	Location[2]	Other information
fim	*STY0589, STY0590* and *STY0592-STY0595*		5' of *argF*	stop codon at *fimI*
tcf	*STY0345-STY0348*		3' of *sinR*	
saf	*STY0332-STY0337*		5' of *sinR*	
sef	*STY4836A-STY4839*		3' of *leuX*	stop codons at *sefA* and *sefD*
bcf	*STY0024-STY0026* and *STY0029-STY0032*		20 kb 3' *thrB*	stop codons at *bcfC*
sta	*STY0201-STY0207*		5' of *panB*	
stb	*STY0369-STY0373*		3' of *thrW*	
stc	*STY2378-STY2381*		5' of *thiM*	
std	*STY3175-STY3177*		3' of *glyU*	
ste	*STY3984* and *STY3086-STY3090*		3' of *relA*	stop codon at *steA*
stg	*STY3918-STY3920* and *STY3922*		3' of *glmS*	stop codon at *stgC*
sth	*STY4938, STY4940-STY4941* and *STY4943-STY4944*		3' of *creD*	stop codon and frameshift mutation at *sthC* and frameshift at *sthE*

[1] The designation of the genes within the operon are indicated

[2] The operon is located either upstream (5' end) or downstream (3' end) of a known gene

Currently, there are only two Typhi determinants that have been suggested to confer adherence to the IEC: the *stg* and type IV pili. *stg* is located between orthologues of the *glmS-pstS* intergenic region in Typhi (315). It consists of five ORFs referred to as *stgABCC'D* (85). The gene encoding for the putative usher of the *stg* operon, *stgC*, is predicted to be a pseudogene due to the presence of a TAA stop codon at codon 171 in the coding sequence (315). This stop codon could also be observed in the sequenced genomes in serovar Typhi strains CT18 and Ty2 (57, 231).

The expression of *stg* is not affected by changes in NaCl concentration, iron availability and pH (85). The *stg* gene cluster of serovar Typhi is a serovar-specific adhesin, which may be involved in the initial stages of typhoid fever pathogenesis by mediating the adherence of Typhi to IEC and inhibiting phagocytosis. A Typhi isolate mutated in the *stg* operon has been shown to adhere 80% less than the wild type, and has a higher level of phagocytosis by macrophages. However, the survival rate within the macrophages was similar and there was no significant elevation in invasion of IEC. This suggests that *stg* may promote the inhibition of phagocytosis to evade the inflammatory cells of the IEC and enable the invasion of deeper tissues (85).

In contrast, type IV pili have been shown to increase the uptake of Typhi by macrophages. The type IV pili are encoded by the *pil* operon. This operon contains a shufflon that is simpler than other shufflons, which contain seven (146) and six (142) 19-bp repeats respectively, encoded by the plasmids R64 and R721. The shufflon in the *pil* operon could form the corresponding PilV, the minor pilus proteins, by DNA inversion mediated by the recombinase gene product (Rci). PilV terminates the function of the *pil* operon, thus bacterial self-association mediated by type IV pili could only be achieved when PilV proteins are not expressed (203). The *pilS* mutants only retain 5 to 25% of their adherence and invasion abilities in IEC as compared to the wild type (358), while no significant reduction is observed in *pilV* mutants (203).

1.6.1.2 Lipopolysaccharide

As a pathogen it is necessary for Typhi to be able to respond to the changing environment, either during the invasion of IEC or within macrophages. One of the adaptation strategies during host infection is the remodeling of the bacterial surface involving the modification of outer membrane

proteins and lipopolysaccharides (LPS). The structure of LPS consists of lipid A, a core oligosaccharide and the O antigen (278). LPS is a major virulence factor in *S. enterica* and it is the most abundant component on bacterial surface (64, 131). Moreover, the structural variability of O antigen, which is the most surface-exposed component of LPS, allows the emergence of large varieties of *S. enterica* serovars (262).

In the pathogenesis of Typhi, LPS acts as one of the necessary ligands for binding to Cystic Fibrosis Transmembrane Receptor (CFTR), which will be discussed later, and the structures of LPS affect the efficiency of bacterial ingestion by the phagocytic cells. Phenotypic alteration of LPS exists in which the amounts of O-antigen chains are diminished after interaction with IEC (183). The genes for the synthesis of the core oligosaccharides of LPS are located in the *waa* gene cluster. Comparison of genes within this cluster between serovar Typhi strain Ty2 and Typhimurium LT2 shows 99% sequence similarity and identical gene organisation, supporting the notion that LPS core structures in these serovars are identical (112). The transcription of genes involved in the synthesis and modification of the LPS core and the O-side chain is positively regulated by the RfaH elongation factor (269). Normal expression of LPS requires the presence of *rfaH* genes and is transcriptionally regulated during the late exponential and stationary phase of the bacteria by the alternative sigma factors RpoN and RpoS (269).

Previously, it has been thought that a complete LPS is necessary for adhesion to and penetration into HeLa cell monolayers (206). Lyczak *et al.* (183) have suggested that the efficiency of internalisation of Typhi strain Ty2 by IEC was not affected by the loss of expression of the LPS O antigen. A further study of Typhi mutants with defined deletions in genes involved in the synthesis and polymerisation of the O antigen and the assembly of the outer core has suggested that the invasiveness of Typhi strain Ty2 was not affected by the presence or the length of the O antigen. However, the terminal glucose residue of the outer core structure composed of Glc I-Gal I-Glc II (Figure 1.6-2) is required for the interaction and efficient internalisation of Typhi into IEC (112).

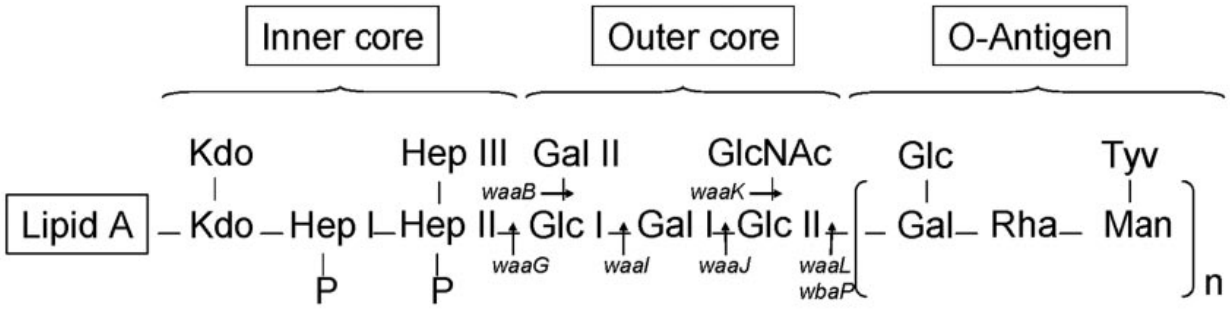


Figure 1.6-2. Structural representation of the inner core, outer core and O antigen of serovar Typhi. Adopted from Hoare *et al.* (112)

1.6.1.3 Vi antigen

Another important virulence factor is the Vi antigen, a capsular polysaccharide antigen made of α-1->4-galacturonic acid with an N-acetyl located at C-2 and variable O-acetylation at C-3 (Figure 1.6-3) (306). This antigen has been used for the differentiation of Typhi isolates by phage typing (24); serological detection of Typhi in typhoid carriers (221) and clinical samples (16); molecular-based detection by PCR amplification of the *viaB* locus (289); and for vaccine development against typhoid fever (353).

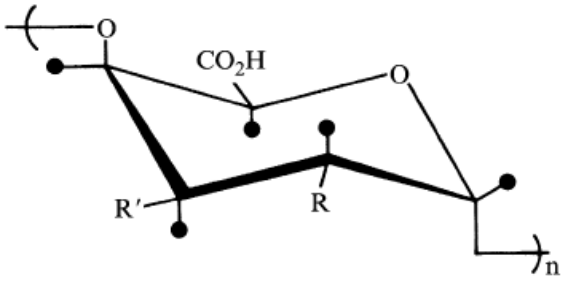


Figure 1.6-3. Diagrammatic representation of the Vi antigen unit with the N-acetylated C-2 (R) and O-acetylated C-3 (R'). Adopted from Szu *et al.* (305).

The Vi antigen is encoded by the *viaB* operon consisting of 10 genes, including genes for biosynthesis (*tviBCDE*) and export (*vexABCDE*) of the Vi antigen. *tviA* (327) (Figure 1.6-4) is the regulator of the *viaB* gene cluster and is itself regulated independently by *ompR-envZ* (237), *rcsB-*

rcsC two component regulatory systems (10) and *rpoS* (275), all of which are controlled by osmolarity. During low osmolaritity, which is encountered in tissue or blood, these regulatory systems activate the Vi antigen expression, suggesting the possible role of the Vi antigen for interaction with human tissue or blood.

The *viaB* locus is located in an unstable region of the Typhi genome, a 134-kb DNA region termed the SPI-7 (Figure 1.6-4) and spontaneous loss of the whole island has been previously reported (208). Moreover, the amount of Vi antigen produced by Typhi isolates could decrease following multiple subculturing (76). Isolates lacking the *viaB* operon have been found to be responsible for the Indian multidrug-resistant typhoid fever epidemic in the year 2000, although these isolates are less infectious than Vi antigen positive isolates (198). These Vi antigen negative mutants have also been shown to be more invasive (201, 359), although studies on human volunteers have suggested that they are less virulent than the Vi-positive isolates (118). The disease rates were significantly higher in volunteers infected with Typhi Vi positive strains than derivatives lacking the Vi antigen. Furthermore, the infective doses were 100-fold higher for Vi negative strains than for the isolates expressing the Vi antigen in order to cause the same rate of disease. Although the *in vivo* study demonstrated that the loss of the Vi antigen results in considerable reduction in the virulence of serovar Typhi, these Vi negative strains are still able to cause typhoid fever (118), and therefore vaccine development targeting the Vi antigen may not be effective against Vi negative strains (11, 137, 198).

Even though some isolates may appear Vi-negative serologically, PCR typing has shown that these isolates may still retain the *viaB* locus, and some isolates may restore Vi antigen expression. A study showed that 12 of 2,222 Typhi isolates from Pakistan were Vi negative by slide agglutination, however 11 of Vi agglutination-negative isolates were Vi positive by PCR typing (331). It remains unclear why the expression of Vi antigen is suppressed in these isolates upon recovery from clinical samples, even though the operon is still present.

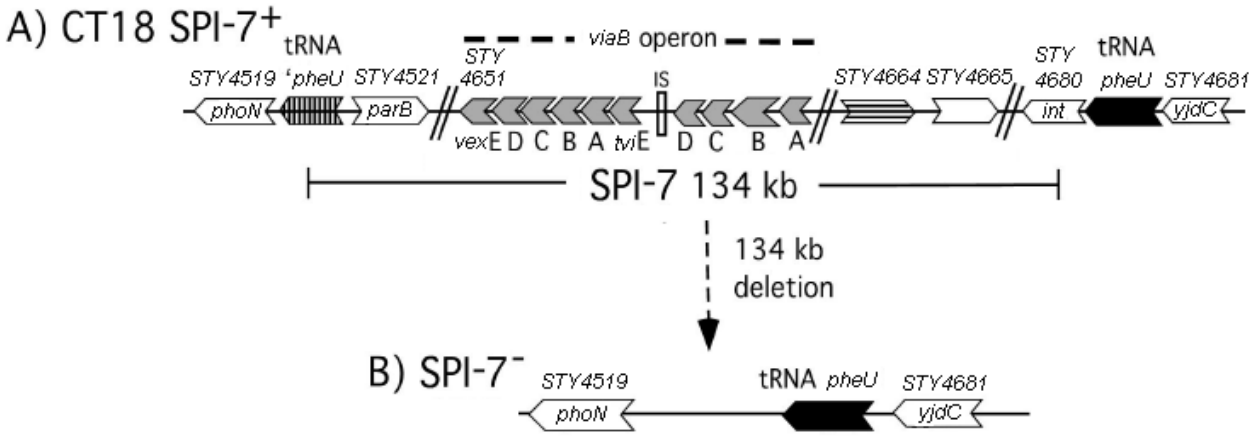


Figure 1.6-4. Genes located in the SPI-7 region of serovar Typhi strain CT18. Adopted from Nair *et al.* (208). (A) Vi-positive SPI-7+ Typhi strain CT18. (B) Predicted structure if SPI-7 is deleted due to recombination between *'pheU* and *pheU*

Vi antigen may also function to evade the innate immune recognition in the intestinal mucosa. Sharma *et al.* (288) have suggested that the interaction of Vi is through two major proteins of the prohibitin family, which inhibits the early inflammatory response. It is hypothesised that Typhi does not induce neutrophil influx in human IEC due to the presence of the *viaB* locus, which is absent in Typhimurium genome. The presence of *viaB* locus has been shown to reduce the production of toll-like receptor 4 (TLR-4)-dependent IL-8 in human colonic tissue (255). Recently, the TviA regulatory protein located in the *viaB* locus has also been shown to mediate the reduction in flagellin secretion thereby reducing the TLR-5-mediated IL-8 expression in epithelial cells (343). The decrease in IL-8 expression results in the reduced circulation of neutrophils in patients with typhoid fever and is in concordance to an earlier study by Raffatellu *et al.* (255). Serovar Typhimurium triggers the migration of neutrophils across human colonic epithelial cells while serovar Typhi is unable to induce this response (195).

Another cytokine, IL-17, is also downregulated during the infection by wild type Typhi expressing the Vi capsular antigen (256). IL-17 is a cytokine which recruits neutrophils in response to bacterial and viral lung infections (317, 341, 354) and other severe inflammatory diseases such as rheumatoid arthritis (41), asthma (345) and multiple sclerosis (191). Introduction of the *viaB* locus into serovar Typhimurium resulted in the reduced production of IL-17, less severe histopathological changes and reduced fluid accumulation in the bovine ligated ileal loop model.

Typhi strain Ty2 with a deletion in the *viaB* locus resulted in increased inflammation and fluid secretion in the calf intestine, similar to serovar Typhimurium (256). Furthermore, isolates expressing Vi antigen enhanced the survival of Typhi in mouse and human macrophage cell lines (111) and avoided lysis by the serum complement (176).

Other than serovar Typhi, the *viaB* locus is also present in serovar Paratyphi C and some isolates of serovar Dublin. It appears that the locus is absent in other serovars that are only associated with gastroenteritis. However, Vi antigen is also absent in the genomes of serovar Paratyphi A, Paratyphi B and Sendai although these serovars could cause typhoid fever-like disease in humans. This indicates that serovars causing enteric fever may not share unique genetic determinants that distinguish them from other serovars causing gastroenteritis.

1.6.1.4 Other possible genetic determinants

Two methods have been used to identify Typhi proteins that are uniquely expressed during infection in humans, including selective capture of transcribed sequences (SCOTS) and in-vivo-induced antigen technology (IVIAT). SCOTS is a method which could identify the *in vivo* global transcription profiles during the course of infection based on microarray analysis. Two studies have been done to determine which genes are expressed in macrophages (70, 71). The first study using SCOTS has identified 36 genes, which are only expressed by Typhi and not Typhimurium, when infecting human macrophage-like cells THP-1. These genes are mostly located in the SPI and bacteriophages elements (70). A more recent study by the same group (71) suggests that at different time points, the level of expression of the genes located on the SPI is different, in particular SPI-1 and SPI-2 (Figure 1.6-5). For example, some of the genes encoded in SPI-2, which are required for the survival inside macrophage were immediately upregulated after the bacterial uptake by macrophages while most were only upregulated after 2 h of infection. These could be due to the requirements at different stages of internalisation or survival within macrophage. Inside the THP-1, Typhi stress response is only upregulated for antimicrobial peptides while the SOS response or the oxidative stress response did not show elevated expression. Furthermore, genes encoding for flagellar machinery, chemotaxis and iron transport systems were downregulated *in vivo* (71).

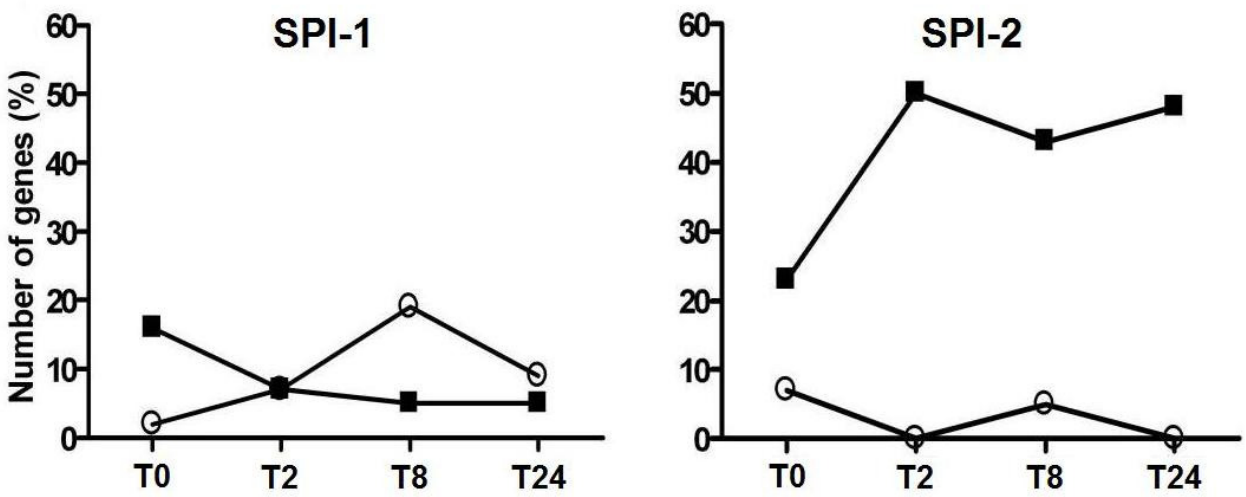


Figure 1.6-5. The percentage of genes that were significantly upregulated (■) or downregulated (○). Adapted from Faucher *et al.* (71)

IVIAT is an immunoscreening technique, which could be used to identify proteins produced during infection by Typhi. A study by Harris *et al.* (103) showed that there were 35 immunogenic Typhi proteins expressed *in vivo*. These proteins were shown to be reactive with convalescent-phase sera purified from humans with serovar Typhi bacteremia. The majority of proteins were associated with fimbrial structure, antimicrobial resistance, heavy metal transport, bacterial adhesion, extracytoplasmic substrate trafficking and hydrolases (103). However, there were only two genes, *tcf* (Typhi colonisation factor) and *STY3683* (putative membrane protein) that were neither present in serovar Typhimurium LT2 nor produced seroreactive activity in humans, which had not been exposed to Typhi. These genes may contribute to the pathogenesis of Typhi, though not to host specificity as *tcf* has been previously shown to be present in other serovars (83).

Recently, a liquid chromatography-mass spectrometry (LC-MS)-based proteomics strategy was implemented to study the proteome of Typhi strain Ty2, which was grown in low pH/low magnesium condition to mimic the intracellular environment within infected macrophages (9). The comparison to Typhimurium strain LT2 revealed a subset of proteins which were highly expressed only in Typhi. These proteins include conserved hypothetical proteins, a toxin-like protein (CdtB), hemolysinE (HlyE) and putative pertussis-like subunit (9). Whilst the functions of hypothetical proteins and pertussis toxin in Typhi are unclear, CdtB and HlyE may be required for virulence and

survival in human macrophages. CdtB has a DNase I-like activity and is a subunit of cytolethal distending toxin (CDT), which could cause DNA damage leading to arrest during lifecycle of the host. The expression of this protein could only occur when Typhi is phagocytosed (99). HlyE or also known as ClyA is a pore forming toxin which is conserved in serovar Typhi and Paratyphi A but absent in many other serovars, including Typhimurium (228). Taken together, these proteins may contribute to the host specificity of Typhi and its pathogenesis within human macrophages.

1.6.2 Environmental factors

1.6.2.1 Anaerobiosis

As Typhi colonises the intestine, one of the environmental challenges is low oxygen concentration or anaerobiosis. The adaptive response of Typhi includes changes in the gene expressions, which may be necessary for virulence. (47, 135). Typhi strain Ty2 has been shown to enhance its ability in entering and proliferating within mammalian epithelial cell lines in anaerobic conditions (47). Mutants containing fusions of three operons in oxygen-regulated genes but defective in nitrate respiratory system were characterised. These mutants were unable to use nitrate and fumarate - vital components for terminal electron acceptors during anaerobiosis - and had a reduced ability to enter and proliferate within Hep-2 epithelial cells lines. Therefore, these anaerobically-induced genes are required for Typhi invasiveness indirectly by providing the most energetically-favourable metabolic process for bacterial survival and proliferation (47). Another study by Kapoor *et al.* (135) demonstrated that anaerobiosis affects the expression of outer membrane proteins (OMP) and increasing Typhi cell surface hydrophobicity by altering cell membrane protein lipid composition. Increased levels of superoxide dismutase and catalase were also observed, suggesting anaerobiosis induces antioxidant defence mechanisms (135).

1.6.2.2 Normal flora

Typhi requires the presence of commensal microbes in the range of 10^8 cfu/ml in the small intestine and 10^{11}-10^{12} cfu/ml in the large intestine for infection (181). Lyczak *et al.* (181) have suggested

that the normal flora aids the translocation of cystic fibrosis transmembrane regulator (CFTR), the receptor for Typhi, by redistributing the preformed protein from the intracellular storage to the epithelial plasma membrane. The presence of other microbes, such as protozoa, has also been shown to affect the invasiveness of serovar Typhimurium phage type DT104 (258). Rumen protozoans appear to transiently enhance enteroinvasive capabilities but not the intracellular survival of DT104 isolates bearing the plasmid SGI-1 in HEp-2 cell lines (258).

1.6.3 Host factors

Variations in hosts' genes have also been linked to the susceptibility to typhoid fever, including polymorphisms in cystic fibrosis transmembrane conductance regulator (CFTR) protein and PARK2 and PACRG, a gene cluster associated with ubiquitination and proteasome-mediated protein degradation.

1.6.3.1 Cystic fibrosis transmembrane conductance regulator (CFTR)

CFTR mutants have been extensively studied for their effects in causing the autosomal recessive disorder cystic fibrosis. The normal function of the CFTR protein is to regulate the secretion of ions across epithelia. Impairment of CFTR leads to defects in pancreatic function, nutrient absorption, the development of the male reproductive system and the structure of skin and airway secretions (182). CFTR is essential for the uptake of Typhi by intestinal epithelial cells (IEC) (182, 239). Human IECs expressing wild-type CFTR was able to ingest significantly more Typhi than CFTR mutants and this ingestion was inhibited by the addition of anti-CFTR monoclonal antibodies and also synthetic peptides mimicking a CFTR domain (239). Furthermore, humans carrying heterozygous genes for CFTR have significantly reduced amounts of Typhi internalisation. In contrast, Typhimurium did not use CFTR to gain entry to IECs suggesting that there may be other epithelial cell receptors, which are used by different serovars for translocation into the intestinal submucosa (239).

Lyczak *et al.* (2002) (182) further demonstrated that Typhi increased the cell surface localisation of CFTR on IEC. Increased IEC expression of CFTR also resulted from the treatment with water-soluble extract of whole Typhi cells, sensitive to heat and protease, prior to addition with live Typhi cells. Regulation of CFTR localisation or expression was not observed in Typhimurium (182). CFTR heterozygotes conferred resistance to infection and development of typhoid fever (182, 239). Furthermore, polymorphisms in the CFTR gene have also been linked to variable susceptibility for Typhi infection (320) and indicates that CFTR plays an important role in the initial adherence of Typhi to the IEC leading to successful Typhi infection.

1.6.3.2 PARK2/PACRG polymorphisms

The PARK2/PACRG gene cluster involved in ubiquitination and proteasome-mediated protein degradation has been associated with Typhi infection (5). *in vitro* studies have suggested that these gene clusters are regulated during Typhimurium pathogenesis as they are associated with the intracellular evasion mechanism (89, 150). Serovar Typhimurium co-localise with ubiquitinated proteins in macrophages and IEC. Invasion of the IEC by Typhimurium requires the reversible activation of Rho-family GTPases Cdc 42 and Rac1 (89) by the bacterial encoded proteins, SopE and SptP, which are functional at different times during the bacterial uptake (150).

Unfortunately, the direct effects of SopE and SptP in Typhi pathogenesis are yet to be determined. Not much is known about SptP and it is absent in serovars Paratyphi A and Typhi. In Paratyphi A, genes encoding SopE are located on a prophage region, SPA-2 (193); in Typhi, the genes are located within the SPI-7 (238). However, in strains CT18 and Ty2, the gene encoding for SopE2, a homologue of SopE, is a pseudogene (56, 231). Interestingly, the gene clusters for SopE are absent in serovar Paratyphi C, which only has a partial SPI-7 (39). Nevertheless, a polymorphism in PARK2/PACRG was shown to have a significant role in the susceptibility to typhoid and paratyphoid fever (5). Additional larger studies are required to determine the direct effect of these polymorphisms in the epidemiology of enteric fever.

1.6.3.3 Interleukin-6 (IL-6)

Cytokine production may be one of the earliest responses to downregulate the local inflammatory response in the IEC that occurs after infection with Typhi. Infection by enteric bacteria has been associated with the production of cytokines, in particular IL-6. The cytokine IL-6 is involved in regulating acute-phase response to injury and infection. It also plays a crucial role in growth and differentiation of various cell types such as haematopoiesis, liver and neuronal regeneration, embryonal development and fertility (105). This cytokine is differently regulated by different *S. enterica* serovars. Small numbers of Typhi cells are sufficient to induce significant production of IL-6 early in the infection without bacterial invasion (338). Typhi is also able to induce significantly higher levels of IL-6 in IEC than serovars Typhimurium and Dublin in murine and human small intestine cell lines (337). This may explain why during initial stages of infection, Typhi rarely cause diarrhoea and could colonise the deeper tissues of the body while avoiding triggering early inflammatory response in the gastrointestinal tract (GIT) (337).

1.7 Evolutionary relationship

1.7.1 Genetic diversity

Two population structure studies have been conducted for Typhi isolates using Multilocus Enzyme Electrophoresis (MLEE) (285) and Multilocus Sequence Typing (MLST) (141). These methods have been widely accepted as powerful tools to examine the population genetics of bacterial species. Unfortunately there is not sufficient variation in this species for MLEE or MLST to be useful for determination of relationships between isolates, nor can it be used for epidemiological studies. This is because Typhi is highly homogeneous.

MLEE has been used to analyse the electrophoretic mobilities of 24 metabolic enzyme on 334 isolates from a diverse global distribution, from which only two electrophoretic types (ET) were identified, ET1 and ET2 (285). The two ETs differed in two of the 24 enzyme loci analysed with ET1 representing 82% of the examined isolates and globally distributed, while ET2 is restricted to

Senegal and Togo in Africa (285). A more sensitive method is MLST, which analyses the isolates based on the sequence diversity of seven housekeeping genes. Unlike MLEE, MLST could identify all the sequence changes, including synonymous changes that do not result in amino acid replacements. This has been performed on 26 global Typhi isolates: three each from Chile, Vietnam, Hong Kong, the Indian subcontinent and Nigeria; two each from Malaysia, Thailand, Senegal, and Zaire; one each from Indonesia, Peru and Russia; and two isolates representing the two ETs, strains SARB63 and SARB64. Four sequence types (ST) designated as ST1, ST2, ST3 and ST8 have been defined. Six isolates from Eurasia, South America and Africa isolated between 1918 and 1999 belonged to ST1. ST2 consisted of 18 isolates that were isolated between 1981 and 2000 from Eurasia, South America, Africa and India. This suggests that ST1 and ST2 are globally distributed. SARB63, the representative isolate of ET1, corresponded to ST2, while SARB64, an ET2 isolate was typed to be ST3. ST8 was only represented by one isolate from Zaire and together with ST3, which was also represented by one isolate it is highly likely that these two isolates originated from ST2 as they differ only in one polymorphism (141). None of the isolates, except SARB63 and SARB64, were from Senegal and Togo and it is likely that additional isolates from West Africa could have been typed as ST3. Similarly, only one African isolate was found to belong to ST8 (141). Even though only seven African isolates were analysed using MLST, they could be distinguished into four STs, thus a larger sample from Africa is required to determine if more STs could be identified.

1.7.2 Genomes of *S. enterica* serovars

Currently there are 17 *Salmonella* genomes that have been fully or partially sequenced, including three strains of enteric fever causing serovars Typhi (CT18 and Ty2) and Paratyphi A (SARB42) and three strains of serovar Typhimurium (LT2, DT104 and SL1344). A brief comparison of the four published genomes (serovar Typhi strains CT18 and Ty2, serovar Paratyphi A strain SARB42 and serovar Typhimurium strain LT2) is summarised in Table 1.7-1. G+C% contents are 1% higher in serovars Typhimurium and Paratyphi A than in serovar Typhi. All of these genomes have seven copies of rRNA, consistent with the copy number of rRNA in *S. enterica* lineages (172).

Table 1.7-1. The general overview of four completed genomes of *S. enterica*

	Serovar Typhimurium		Serovar Paratyphi A	Serovar Typhi			
	LT2				CT18		
	Chromosome	Plasmid pSLT	SARB42	Ty2	Chromosome	Plasmids pHCM1	Plasmids pHCM2
Size (bp)	4,857,432	93,939	4,585,229	4,791,961	4,809,037	218,150	105,516
G+C content	53%	53%	53%	52.05%	52.09%	47.58%	50.60%
rRNA clusters	7	0	7	7	7		
tRNAs	85	0	82		78		
tRNA pseudogene	1	0					
Structural RNAs	11	1	36				
CDS (including pseudogenes)	4,489	108	4,263	4,545	4,599	249	131
CDS pseudogenes	39	6	173	206	204	8	0

1.7.2.1 Typhi CT18 and Ty2

There are two fully sequenced Typhi genomes including strain Ty2, a strain frequently used in experimental studies and in vaccine development (57) and strain CT18, a multi-drug resistant isolate, first isolated in 1993 from a child with typhoid fever in Vietnam (231). These strains belong to different ST as defined by MLST (141). There are 4,599 and 4,545 coding sequences on CT18 and Ty2 respectively of which, 4,195 genes, accounting for 98% of the genome sequences, are identical between them (57). The average length of genes in Ty2 is 910 bp, which covers 88% of the genome, while the length in CT18 is 958 bp and covers only 87.6% of the genome (56).

The differences between these two strains include 282 single point mutations and unique open reading frames (ORFs). Twenty-nine and 84 ORFs are unique to Ty2 and CT18, respectively (57). The unique ORFs in Ty2 are mostly associated with putative prophages. It is unclear whether these strain specific genes have any functional significance or simply a result of chance infection by phages. Strain variation has been associated with the presence of prophage genome sequences and the difference in lineage specific islands, although it is also unclear if these resulted from genetic insertions or deletions (307). The number of insertion sequence elements is also different between the two strains. Ty2 contains 26 copies of IS*200F* and one copy of each IS*1230B*, IS*285* and IS*1351* while CT18 contains three copies of IS*1* in addition to those four IS elements, which are present in Ty2.

The two Typhi strains differ in the arrangement of the rRNA operons. In Ty2, the homologous recombination of *rrnG-rrnH* caused the rRNA operons to be located on the opposite replichores. This major inter-replichore inversion spans the terminus of replication and almost half of the Ty2 genome resulting in uneven sized replichores (57).

Further differences between these two strains could be observed in the nitrate reductase pathway, NR-Z and *rpoS* gene. Ty2 has fused a *narW* gene fused with an *narV* in the NR-Z operon resulting from an in-frame deletion. Typhi isolates mutated in anaerobic respiration are less efficient in replicating intracellularly. However, this deletion in Ty2 may not have any effect since the function of NR-Z could be replaced by its homolog, the NR-A. On the other hand, the abnormal *rpoS* in Ty2 caused by a frameshift has been associated with poor survival under starvation and other stresses and this property is incidentally beneficial for vaccination purposes (56).

In addition, Ty2 does not possess plasmids, whereas CT18 carries two plasmids, pHCM1 and pHCM2. pHCM1 is the larger plasmid, with a length of 218 kb and shares a 168 kb region with more than 99 percent similarity with plasmid R27, an incH1 plasmid that was first isolated from *S. enterica* (290). pHCM2 is approximately 107 kb and is considered to be phenotypically cryptic. It shares 97% DNA sequence identity to the virulence-associated plasmid, pFra, of *Yersinia pestis* (140). The plasmid encodes putative genes directly related to DNA metabolism and replication, suggesting that it may improve the function of replication under stressful conditions (246).

However, the impact to Typhi of carrying the plasmid in different environments is yet to be determined.

1.7.2.2 Genomic comparison between serovars

The other completed genome sequences include the genomes of serovar Typhimurium strain LT2 (194) and serovar Paratyphi A strain ATCC9150 (193). *S. enterica* serovar Typhimurium strain LT2 was first isolated in the 1940s and has been used in a study on phage-mediated transduction. Paratyphi A strain ATCC9150 is well known as SARB42 and is also the representative of electrophoretic Pa1 based on MLEE (33). Among the three serovars, Paratyphi A has the smallest genome, approximately 200 kb smaller than those of Typhi and Typhimurium. This could be due to the fewer prophage and mobile elements (193) associated with this strain.

When Typhi CT18 was compared to Typhimurium LT2, 601 genes (13.1%) were found to be unique to CT18 while 479 genes (10.9%) were found to be unique to LT2 relative to CT18. In LT2, genes essential for survival in macrophages are encoded in phage regions, whereas in both CT18 and Ty2 they are encoded in nonphage regions. LT2 possesses a plasmid named pSLT, a 94 kb virulence plasmid containing 108 CDS encoding genes for adhesion and virulence genes of serovar Typhimurium (194).

Paratyphi A strain SARB42 and Typhi strain CT18 share more genes than either of them do with serovar Typhimurium strain LT2. A spotted DNA microarray using Typhimurium strain LT2 was used to analyse the relationships between enteric fever-causing serovars and it was shown that the three serovars Typhi, Paratyphi A and Sendai are most closely related (43). There are 154 genes which are present in both Typhi and Paratyphi A but absent in LT2, including 14 pseudogenes in both Typhi and Paratyphi. Nevertheless, each of them has many genomic features not shared by the other.

Typhi CT18 has more genes covering 24 regions of two or more genes including SPI-7 and three large putative phage or phage remnants (193). On the other hand, most of the genes present in Paratyphi A but absent in Typhi are located in the two prophage regions SPA-1 and SPA-3 and an

additional prophage, SPA-2 is located at a different site in Typhi. In both of these serovars, only a few clusters encoding for fimbrial synthesis are fully functional. Serovars Typhi and Paratyphi A are both monophasic, expressing only a flagellar antigen encoded by *fliC* at the H1 locus. However, the mechanism which obstructs their ability to express the antigen at H2 locus is different. In Typhi, the genes responsible for phase variation are deleted whereas in Paratyphi A, a frameshift mutation in *hin* permanently causes this serovar to express the phase 1 flagellar antigen, which is confirmed by whole genome sequencing (193).

Seventy five percent of the gene contents and nucleotide divergence between serovars Typhi and Paratyphi A genomes suggest that they are not closely related serovars (59). Any similarities that are observed between these two serovars could be a result of convergent evolution. Frequent recombination may have occurred over a short period during the evolution of serovars Typhi and Paratyphi A (59). Comparison of whole genomes from three other enteric fever causing serovars, Paratyphi B, Paratyphi C and Sendai, will enable common genomic features to be identified and to establish the evolutionary processes involved in their emergence of these pathogens.

1.7.2.3 Pseudogenes

1.7.2.3.1 Comparison of pseudogenes in serovars Typhi, Paratyphi A and Typhimurium

The most striking difference between Typhimurium to Typhi and Paratyphi A is the presence of large number of pseudogenes despite the similar genome size and coding sequences. Typhimurium LT2 only has 39 pseudogenes whereas Typhi strains CT18 and Ty2 have 204 and 206 pseudogenes, respectively, and Paratyphi A strain ATCC9150 has 173 pseudogenes. Pseudogenes are genes that have lost functions due to one or a few irregularities such as small insertions, deletions and substitutions forming stop codons. Others have extensive degradation and are recognised as pseudogenes only by comparison to closely related species (231).

In CT18, there are 124 pseudogenes inactivated by the introduction of a single frameshift or stop codon suggesting recent origins while 45 resulted from a change in the homopolymeric tracts. 27

pseudogenes include remnants of IS, transposase, integrases and prophages while 75 of them are involved in housekeeping functions in the other serovars. Additionally, 46 genes are potentially involved in virulence or host interactions including fimbrial operons, flagellar methylation, host range specificity and type three secretion system effector proteins (231). This suggests that Typhi has lost the ability to express several virulence-associated genes. The lost of these genes may have contributed to the host restriction of this serovar.

Of the pseudogenes identified in Typhi, 195 are common to both strains while 9 and 11 are intact in Ty2 and CT18, respectively. The large numbers of pseudogenes with the same mutations suggest that Typhi has evolved once and only relatively recently. Genes inactivated in both Typhi genomes include genes in seven of the 12 fimbrial operons as well as *shdA* and *ratB*, which are putative fimbrial-like genes. In serovar Typhimurium, both *shdA* and *ratB* have been associated with the ability to persist in the intestine. Some of the genes located on SPI-1 to SPI-5 have also been inactivated. The complete list of pseudogenes unique to Typhi strain CT18 and Ty2 respectively is summarised in Table 1.7-2.

The differences in pseudogene content between Ty2 and CT18 fall into no discernible pattern or functional relationship. They may also have arisen due to variations in stresses applied by human host defence systems that contributed to the pathogenesis of both strains. For example, the three genes that are intact in Ty2 include the following: *ttrS*, encoding for sensor for tetrathionate, an alternative electron acceptor for anaerobic growth in the presence of vitamin B12; *sopE*, an effector secreted by type III secretion system and is involved in actin rearrangements of the epithelical cells; and *wcaA*, a glycosyltransferase involved in the formation of exopolysaccharide capsule that could provide protection from dehydration and acid stress (57). They may also reflect the need to adjust the balance of metabolic capabilities in optimising virulence achievable by more than one possible combination of genes.

The Paratyphi A genome contains 173 pseudogenes (193). Comparison to the whole genome sequence of Typhi CT18 has shown that only 166 of these pseudogenes can align to orthologs in Typhi. Almost all of the pseudogenes in Paratyphi A have different inactivating mechanisms to those in Typhi (193). Only 28 of the pseudogenes between these two serovars have identical

mutations (Table 1.7-3). The genome sequence of a Paratyphi A strain has shed further insight into the evolution of enteric fever causing clones. As there are fewer pseudogenes in Paratyphi A, it may be inferred that this serovar appeared more recently than Typhi (193). These serovars are niche-adapted clones and it is likely that the loss of genes common to both serovars is advantageous and can enhance the pathogenic lifestyle of enteric-fever causing serovars. They invade the systemic tissue and are transmitted through excretion via the gall bladder, instead of colonising the IEC.

Information on the distribution of pseudogenes in all human-restricted *S. enterica* serovars are required to determine if gene inactivation plays a role in the host specificity of these serovars. If certain gene inactivations occur in some but not all human-restricted serovars, the genes are more likely to be unnecessary and hence inactivated. This is a result of trade offs pertaining to resource allocation or because selection does not disfavour the degeneration of functional genes into their pseudogene equivalents ("use it or lose it"). The hypothesis of antagonistic pleiotropy explains this as happening through ecological specialisation, in which the non-redundant genes for catabolic functions will be inactivated (48). On the other hand, if a gene is inactivated in all human adapted clones of *S. enterica*, the genes may be incompatible with living in human host and are under selective pressure. This further emphasises the need for other enteric fever serovars, including Paratyphi B, Paratyphi C and Sendai to be completely sequenced.

Table 1.7-2. Pseudogenes in serovar Typhi which are different between strain CT18 and Ty2 (57)

Genes that were pseudogenes in CT18 but intact in Ty2				Genes that were pseudogenes in Ty2 but intact in CT18			
Gene	Product	ORF		Gene	Product	ORF	
		Ty2	CT18			Ty2	CT18
ybaD	Conserved hypothetical protein	t2448	STY0453	*stbC*	Outer membrane fimbrial usher protein	t2523	STY0371
fimI	Fimbrin-like protein FimI	t2319	STY0590	*aroM*	AroM protein	t2474	STY0423
ttrS	Sensor kinase TtrS protein	t1254	STY1735	-	Hypothetical bacteriophage protein	t1913	STY1027
sopE2	Putative invasion-associated secreted protein	t1023	STY1987	-	Putative exported protein	t1549	STY1423
wcaA	Putative glycosyltransferase	t0757	STY2328	*narV*	Respiratory nitrate reductase 2 gamma chain	t1490	STY1485
-	Putative transcriptional regulator	t0589	STY2504	*narW*	Respiratory nitrate reductase 2 delta chain	-	STY1486
-	Phosphoenolpyruvate-protein phosphotransferase	t0425	STY2668	-	Putative d-mannonate oxidoreductase	t1429	STY1553
	Bacteriocin	t3035	STY3280	*astA*	Arginine *N*-succinyltransferase	t1183	STY1810
torC	Cytochrome *c*-type protein	t3695	STY3955	-	Putative regulatory protein	t1166	STY1829
				stcC or *yehB*	Putative outer membrane usher protein	t0706	STY2379
				gabP	GabA permease (4-aminobutyrate transport carrier)	t2688	STY2913
					Bacteriophage P4 DNA primase	t4529	STY4832

- No gene name has been assigned

Table 1.7-3. Pseudogenes in serovar Paratyphi A which have the same inactivating mutations in serovar Typhi (193)

ORF	Gene	Product
SPA0197	*fhuA*	outer membrane protein receptor / transporter for ferrichrome, colicin M, and phages T1, T5, and phi80
SPA0337	-	putative anaerobic dimethylsulfoxide reductase
SPA0350	*sinH*	putative invasin
SPA0353	*ratB*	putative outer membrane protein
SPA0354	*shdA*	similar to the C-terminal region of AIDA; IcsA; subspecies I specific; Peyer's patch colonization and shedding factor
SPA0466	-	-
SPA0662	*mglA*	ABC superfamily (ATP bind), galactose (methyl-galactoside) transport protein
SPA0765	*wcaK*	putative galactokinase in colanic acid gene cluster
SPA0805	*sopA*	Secreted effector protein of *Salmonella Dublin*
SPA0809	*dacD*	DD-carboxypeptidase, penicillin-binding protein 6b
SPA0912	*fliB*	N-methylation of lysine residues in flagellin
SPA1012	-	-
SPA1087	*hyaA*	uptake hydrogenase small subunit
SPA1189	-	-
SPA1314	-	putative transport protein
SPA1471	*orf408*	putative regulatory protein, *deoR* family
SPA1647	*fhuE*	outer membrane receptor for Fe(III)-coprogen, Fe(III)-ferrioxamine B and Fe(III)-rhodotrulic acid uptake
SPA1769	-	-
SPA1826	*sopD2*	possible secreted protein
SPA1952	*slrP*	leucine-rich repeat protein
SPA2171	-	-
SPA2201	-	-
SPA2628	*hin*	H inversion: regulation of flagellar gene expression by site-specific inversion of DNA

SPA2762	-	-
SPA3108	*yhaO*	putative HAAAP family transport protein
SPA3605	-	-
SPA3627	-	-
SPA4123	*dmsA*	putative anaerobic dimethyl sulfoxide reductase, subunit A

- No gene name has been assigned

Of the 204 pseudogenes discovered in CT18, 145 are intact in Typhimurium strain LT2. Twenty three pseudogenes in Typhi CT18 are also pseudogenes in Typhimurium LT2, however only 15 of them have the same inactivating mutations (Table 1.7-4) (194). The remaining 16 pseudogenes in LT2 include genes that are involved in maltose, trehalose and histidine metabolism. Some of the genes could be removed due to redundancy. For example, the sodium ion pump and copper homeostasis protein. However, the consequence of functional losses of other pseudogenes in LT2 is still unclear (194).

Table 1.7-4. The list of pseudogenes shared between serovar Typhi and Typhimurium (231)

ORF	Gene	Product	Note
STY3029	-[1]	transposase	
STY4037	-	conserved hypothetical protein	
STY0602b	-	probable IS element transposase fragment	
STY0609a	*cusS*	putative copper-ion sensor protein fragment	Different[2]
STY0610	*silA*	putative inner membrane proton/cation antiporter SilA	
STY1995	-	transposase	
STY3025	-	insertion sequence transposase	
STY2139	*araH*	L-arabinose transport system permease	
STY2468	-	bacteriophage tail fiber assembly protein	
STY2469	-	putative bacteriophage tail protein	
STY3800	*cdh*	CDP-diacylglycerol pyrophosphatase	
STY3801	-	hypothetical protein	
STY4034	-	putative IS1351 transposase	
STY4035	-	putative transposase	
STY4146	-	putative transposase	
STY4174	-	putative trans-sulfuration enzyme	
STY0329	-	IS element transposase	
STY0339	-	probable transposase remnant	
STY0602	-	phage integrase fragment	
STY0602a	-	probable IS element transposase fragment	
STY0603	-	probable IS element transposase fragment	
STY1657	*malX*	PTS system, maltose and glucose-specific IIABC component	
STY1657a	*malY*	putative aminotransferase (fragment)	

[1] No gene name has been assigned

[2] The inactivating mutation is different between the two serovars

1.7.2.3.2 Pseudogenes in other specialised clones

Losing genomic functions in pathogenic bacterial species has been progressively more evident due to increasing adaptation to the host. This trend has been demonstrated in many obligate pathogens, such as *Rickettsia prowazekii* (8) and *Mycobacterium leprae* (202), both of which have undergone massive genome downsizing made evident by the presence of a large number of pseudogenes. Pseudogenes are also abundant in other specialised clones, including *Yersinia pestis* and *Shigella* species (Figure 1.7-1).

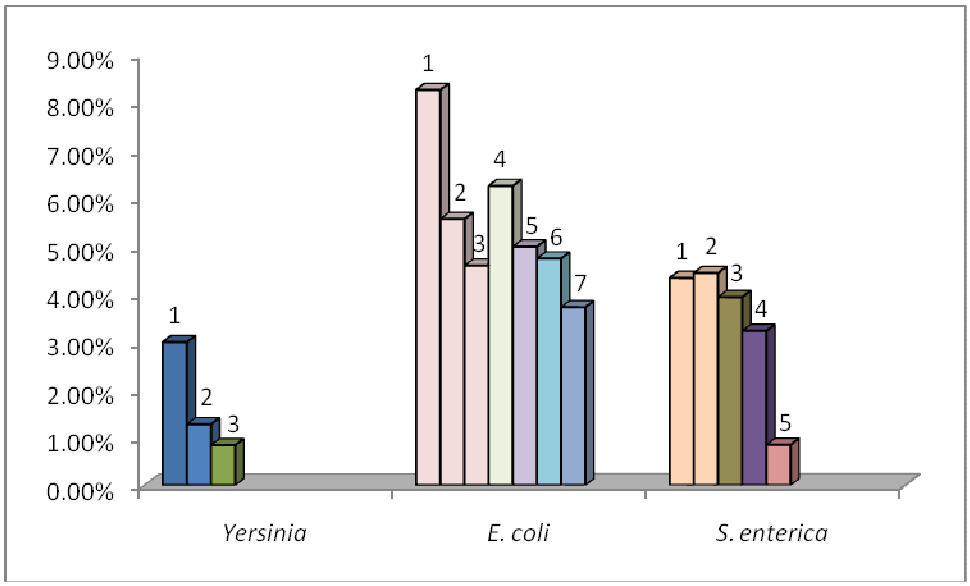


Figure 1.7-1. Comparison of the percentage of pseudogenes present between different specialised clones of *Yersinia* (1. *Y. pestis* strain CO92; 2. *Y. pestis* strain KIM; and 3. *Y. pseudotuberculosis* strain IP31758), *E. coli* (1-3. *S. flexneri* strains 2457T; 301 and 8401 respectively; 4. *S. dysenteriae* strain 197; 5. *S. boydii* strain 227; 6. *S. sonnei* strain 046; 7. *E. coli* strain K12) and *S. enterica* (1-2. Typhi strains CT18 and Ty2 respectively; 3. Paratyphi A strain ATCC 9150; 4. Choleraesuis strain SC-B6; and 5. Typhimurium strain LT2).

1.7.2.3.2.1 Yersinia pestis

Y. pestis is a clone that evolved from *Y. pseudotuberculosis* 1,500-20,000 years ago (3). The selective pressures that led to the recent evolution of *Y. pestis* are as yet unknown (2). The pathogen causes plague and was responsible for the Black Death, in which some 50% of the population of

Europe was eliminated during the second pandemic in 1346 (61)**.** *Y. pseudotuberculosis* causes a much milder enteric disease. *Y. pestis* has undergone genomic downsizing by insertional and point mutations that lead to non-functional genes in comparison to *Y. pseudotuberculosis* (56, 233). Two *Y. pestis* strains CO92 and KIM have approximately twice as many pseudogenes in comparison to *Y. pseudotuberculosis* strain IP32935. It is likely that the inactivation or deletion of genes in *Y. pseudotuberculosis* may have provided an advantageous strategy during the divergence of this species. The reduction in genome contents may possibly enhance virulence when the respective lineage was evolving to *Y. pestis* to infect mammals by subcutaneous injection through infected fleas rather than by gastrointestinal tract route. (61). For example, the *yadA* gene, eoncoding an adhesion protein involved in attachment to epithelial cells, is a pseudogene in *Y. pestis* due to a 1-bp deletion (42). Moreover, *Y. pestis* has more IS elements disrupting the genes but uninterrupted and functional in *Y. pseudotuberculosis* for example the *inv* and *ail* genes. There are only 20 IS in IP32935 strain but there are 117 and 138 IS elements in KIM and CO92 respectively.

1.7.2.3.2.2 Shigella

Shigella causes bacillary dysentery and was estimated to have originated 35,000-270,000 years ago from *E. coli* (248). It is known to be biochemically less active than *E. coli*, due to loss of gene functions. To date, there are six *Shigella* genomes that have been published: *S. boydii* serotype 4 strain 227 (Sb227) (352), *S. dysenteriae* serotype 1 strain 197 (Sd197) (352), *S. flexneri* serotype 2a strain 2457T (Sf2457T) (336), *S. flexneri* serotype 2a strain 301 (Sf301) (130), *S. flexneri* serotype 4 strain 8401 (Sf8401) (220) and *S. sonnei* strain 046 (Ss046) (352).

Shigella strains lack the catabolic pathways, which are thought to be attributable to their niche adaptation that allows them to invade the intestinal epithelial cells (248). Comparison of the four *Shigella* strains, Sf2457T, Sd197, Ss046 and Sb227, to *E. coli* indicates that *Shigella* have a large number of pseudogenes. Some of the pseudogenes could be detrimental to the pathogenic lifestyle in *Shigella* if active, for example *ompT* and *cadA*, while others are inactivated because they are not needed, for example flagellum production, its associated machinery and lactose utilisation.

Both *ompT* and *cadA* genes, which encode for surface protease and lysine decarboxylase respectively, are inactivated in the genomes of *Shigella* (352). A mutation in *ompT* enables the surface expression of actin-polymerisation factor allowing *Shigella* to spread into adjacent epithelial cells (211). *cadA* is required for lysine decarboxylation and the by-product of lysine metabolism is cadaverine, which is a potent inhibitor of *Shigella* enterotoxin activity (54). Thus, inactivation of *cadA* allows a more efficient pathogenesis of *Shigella* during infection. Interestingly, *cadA* was disrupted differently in at least two *Shigella* strains; in Sd197, it was inactivated by a frameshift mutation while in Ss046 it was by an insertion sequence (352). *Shigella* also lacks motility and lactose fermentation which are both positive in *E. coli* (155), allowing differentiation between *Shigella* and *E. coli*. *Shigella* uses actin-based motility for intracellular movement, thus flagella became redundant. This is reminiscent of lactose utilisation that is not required in its intracellular life cycle.

1.7.2.4 Implications from genome sequence: pathogenicity and host specificity

Following the availability of the genome sequences of serovar Typhimurium and Typhi, the genetic basis of invasiveness and human host-specificity of Typhi has been examined. A study was performed using a microarray constructed with the sequences from the annotated ORFs of Typhimurium LT2 added with annotated ORFs from Typhi CT18, which are divergent from Typhimurium (32). The array consisted of 417 Typhi-specific genes and an additional 4,442 Typhimurium genes including 104 genes from pSLT plasmid (32). The genome coverage in the microarray was 96.6% and 94.5% for Typhimurium and Typhi respectively. This was used to explore the gene content differences among nine diverse Typhi isolates including the two Typhi strains representing two MLEE types and the strain CT18. Only 13 genomic regions, nine of which were only found in Typhi CT18 but not in Typhimurium LT2, were varying (Figure 1.7-2) while more than 4,000 of the CT18 genes were shared by all of the Typhi strains examined.

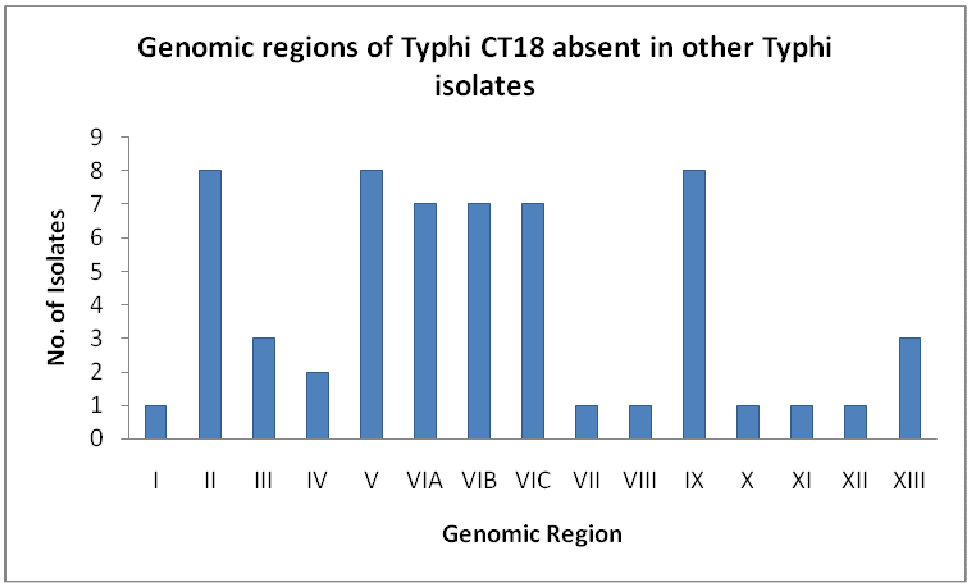


Figure 1.7-2. Typhi CT18 genomic regions that were absent in other Typhi isolates analysed

The genome regions absent in the majority of Typhi strains examined were also absent in Typhimurium LT2, including regions II (STY1048-STY1071), V (STY2038-STY2077), VIA-VIC (STY2419-STY2420, STY4124-STY4125 and STY4657-STY4658), IX (STY3188-STY3193) and XIII (STY4821-STY4834). All but region IX were mobile genetic elements including prophage, IS elements and P4-like phage, suggesting that these regions are unstable and are involved in strains variation (32). Altogether, the differences in genetic content indicate that even though Typhi is highly homogeneous, its genomic reservoir is unstable.

1.8 Typing

Epidemiological studies of pathogens, such as Typhi, are essential for public health management, including for surveillance, to identify origins of outbreak, to identify factors contributing to the persistence and spread of typhoid fever in endemic areas, and ultimately, to devise methods for rapid diagnosis and to improve treatment, control and prevention of typhoid fever infections. A variety of techniques are used for epidemiological typing of strains. However, a number of factors must be considered for a typing method to be efficacious, including: typability, the percentage of distinct isolates able to be typed for any assigned marker; reproducibility, the percentage of

identical results from repeated typing; and discrimination, the ability to distinguish unrelated strains, which is represented by a Simpson's index of diversity (D value) (124).

Traditionally, Typhi isolates were only characterised by phage typing, based on the sensitivity or resistance to Vi phages. The phages used to type Typhi isolates are derived from one parent phage, Vi phage II, and are uniquely adapted to differentiate Vi types (50). Unfortunately, knowledge of the relationships between phage types is very limited and the genetic bases for this strain variation remain largely unknown. Phage typing is still used for differentiating Typhi isolates to describe their global epidemiology, but is not useful in localised outbreaks because of the lack of discrimination power between closely-related isolates (109).

The developments of molecular typing methods have significantly assisted the epidemiological analyses and differentiation of clinical isolates. Methods such as pulse-field gel electrophoresis (PFGE) (145, 151, 209, 309, 310), IS*200* typing (311), ribotyping (79, 166, 219), random amplified polymorphic DNA (RAPD) (252, 287) and amplified fragment length polymorphism (AFLP) (210) have been employed for these purposes.

1.8.1 Principles of molecular typing methods

Currently available molecular techniques for typing of Typhi rely on electrophoretic separation of DNA fragments, represented by a pattern of bands on a gel. Ribotyping and IS*200* typing only detect variations in specific regions of the genome while PFGE, RAPD and AFLP detect genetic variations in the whole genome, although the sequence variations could not be determined.

Ribotyping utilises the Southern blotting technique. The probes used for hybridisation are derived from the 16S and 23S rRNA sequences found on the seven rRNA operons. Bacterial isolates are differentiated based on the restriction polymorphisms, termed as ribotypes, surrounding the rRNA operons (125, 300). This method could effectively detect changes in the genome either by rearrangements of the existing DNA or addition and deletion of DNA and recombinations but not inversions (219). Different *S. enterica* serovars produce unique banding patterns and are simple to interpret since only a small number of bands are observed, however it has a limited ability to

distinguish closely related strains of the same serovar (210). IS*200* typing also employs the Southern blotting method to differentiate isolates based on the hybridisation of DNA fragments to the IS*200* probe (154).

PFGE uses rare-cutting restriction endonuclease digestion of chromosomal DNA followed by comparative visualisation of large-sized fragments electrophoresed by pulse electric fields (197, 280). RAPD, also known as arbitrarily primed PCR utilises short oligonucleotides 9-10 bases in length to hybridise with sufficient affinity to chromosomal DNA sequences. The oligonucleotides are allowed to anneal at a defined temperature to amplify the regions of bacterial genome (342). AFLP is a genome fingerprinting method that selectively amplifies a subset of DNA fragments generated by restriction enzymes, usually with *Eco*RI and *Mse*I (329).

These DNA-based typing approaches are highly sensitive, however they are unable to identify the basis of similarities and differences in the genetic information of closely related bacterial pathogens. Furthermore, the discriminatory ability of these methods varies and their usefulness in surveillance and outbreak investigations is limited. It is therefore necessary to compare the currently available techniques, which have been utilised for epidemiological study of Typhi isolates from outbreaks and sporadic cases.

1.8.2 Comparison of different molecular typing methods for epidemiologic investigations of typhoid fever

The efficiency of any molecular typing method chosen to analyse the strains isolated from outbreaks or sporadic cases that have occurred in different countries, is always compared to PFGE. The latter has been regarded as the most appropriate method for outbreak analysis and is the "gold standard" of molecular typing methods. PFGE could differentiate Typhi isolates associated with different disease severities and could also distinguish multi-drug resistant (MDR) isolates from different endemic regions (308).

Two or more methods are usually used concurrently for epidemiological studies to ensure that the strains responsible for the disease could be accurately identified. A study by Navarro *et al.* (213) compared phage typing, ribotyping, IS*200* typing and PFGE to analyse a total of 83 Typhi isolates, 48 from sporadic cases and 35 from a large outbreak in Spain. Of these, ribotyping and PFGE were shown to the most discriminating with D values of 0.974 and 0.988 respectively.

Both PFGE and ribotyping are the most commonly used methods and they have been demonstrated to be useful to distinguish Typhi isolates in several studies. Nevertheless, the discriminatory power of these methods, especially in ribotyping, evidently relies on the restriction enzyme selected as observed in the study by Hermans *et al.* (106). A total of 78 Typhi strains isolated from patients living in different areas in Bangladesh from 1994 to 1995 were analysed using phage typing, ribotyping, IS*200* typing and PCR fingerprinting (106). Ribotyping was performed using two restriction enzymes, *Pst*I and *EcoR*I, where eight and one fingerprint/s were observed respectively, suggesting that *EcoR*I ribotyping could not be used to discriminate isolates from different geographic origins in Bangladesh (106).

Unfortunately, PFGE and ribotyping do not always achieve the same level of strain discrimination. There were cases where PFGE was better than ribotyping. For example, in the study by Hosoglu *et al.* (119), both PFGE and ribotyping were used to analyse the isolates responsible for sporadic cases of invasive Typhi infection and to identify the presence of identical strains in Turkey. There were 78 isolates, 65 of which were clinical samples while the remaining 13 were environmental isolates (119). In this study, PFGE had better discrimination among isolates than ribotyping. PFGE using enzymes *Xba*I and *Bln*I gave 30 digestion patterns for both enzymes while ribotyping only yielded 18 different patterns and was found to be less discriminating than PFGE (119). PFGE was shown to be useful to provide evidence of the clonality of strains of Typhi within a specific geographic region, which will provide a better understanding of transmission patterns.

On the other hand, ribotyping was shown to be able to discriminate isolates from sporadic cases as well as homogeneous clones related to outbreaks as observed in the study by Le *et al.* (161). A collection of isolates responsible for sporadic cases and minor outbreaks in Vietnam between 1995 to 2002, which were resistant to ampicillin, chloroamphenicol, tetracyclines, streptomycin, and

cotrimoxazole, were analysed using plasmid fingerprinting, phage typing, PFGE using *Xba*I and ribotyping using *Pst*I (161). It was regarded as the best method to use because it displayed the highest discriminatory power, even though it could only detect changes at restriction sites on different copies at the rRNA gene (161).

AFLP is an alternative to PFGE and ribotyping. The study by Nair *et al.* (210) compared the genetic diversity of six Typhi isolates from diverse geographic areas using AFLP, ribotyping and PFGE. It was found that more variants could be differentiated using AFLP with a D value of 0.88 while ribotyping has a D value of 0.63 and PFGE has a D value of 0.74. The closer the value is to 1, the better is the method for epidemiological purpose. The isolates previously identified as identical by both ribotyping and PFGE could be distinguished and thus showed its discriminatory ability to very closely related strains. On the other hand, it has downsides such as cost and technical complexity (210).

1.8.3 Chromosomal rearrangement affects typing using PFGE and ribotyping

S. enterica, like many other enteric bacterial species, has a highly conserved chromosomal organisation (149). The endonuclease I-*Ceu*I, encoded in the mobile intron in the choloroplast 23s ribosomal RNA gene of *Chlamydomonas eugametos*, could specifically recognise and cut the 26 bp sequences present in the rRNA operons of bacterial genomes. The pattern of the fragments is known as the genome type, which is used to differentiate the genomes of *S. enterica* (90). Cleavage of the *S. enterica* genome generates seven fragments corresponding to the seven *rrn* genes present in all *S. enterica* lineages (172). The number and distribution of I-*Ceu*I recognition sites are identical within *S. enterica*. However, differences in the fragment lengths exist among *S. enterica* lineages, due to independent insertions or deletions.

Typhi genomes have undergone rapid changes and variations in *rrn* restriction patterns (170), in contrast to serovar Typhimurium where almost 90% of the strains analysed display identical I-*Ceu*I patterns and lengths (171). The order of *rrn* gene fragments in the rRNA operon is also rearranged in other host specialised *S. enterica* serovars, including: Paratyphi A (193), Paratyphi C (173),

Gallinarum (350) and Pullorum (168). Chromosomal rearrangements have been shown to affect ribotyping and PFGE patterns, as they could alter restriction endonuclease recognition regions. Previously, variations in ribotypes were thought to be due only to point mutations (212). However, it has been found that homologous recombination, including duplications, deletions, transpositions and inversions, between the seven copies of *rrn* genes also contributes to new ribotypes (169). Therefore, to determine the nature of new ribotypes, it has been suggested that ribotyping should be performed concurrently with genome typing by partial I-*Ceu*I digestion of the rRNA operons (219).

By employing ribotyping and PFGE, Echeita *et al.* (62) have shown that chromosomal rearrangements existed in strains from three of eight different outbreaks of typhoid fever in Spain between 1989 to 1994. This suggests that genetic rearrangement could also occur in isolates from the same outbreak.

There are many reported cases where survival rates are significantly lower in infections for bacterial pathogens with chromosomal rearragements (40, 110, 282). Nevertheless, despite the high susceptibility of genetic reorganisation, the stability and survival of Typhi is still maintained, suggesting that the constraints of the gene order have been partially relaxed during Typhi evolution (169).

Chromosomal rearrangements have also been observed in relatively younger pathogens, such as *Bordetella pertussis* (232) and *Yersinia pestis* (42). In *B. pertussis*, chromosomal rearrangement resulted from homologous recombination of the IS*481* which is present in a high number of copies in the *B. pertussis* chromosome (36). Similarly, the significant numbers of genome rearrangement in *Y. pestis* also resulted from inversions of genome segments at insertion sequences (254). This suggests that strain variation in these pathogens is achieved by chromosomal rearrangement.

1.8.4 Variable number of tandem repeats

The number of fully sequenced bacterial genomes is increasing rapidly. Sequence comparison reveals the presence of repeats that vary in length, position and nature and are often unique for a single strain. Repeats may consist of simple homopolymeric tracts of a single nucleotide or of

several multimeric classes of repeats that can be homogeneous, heterogenous or degenerate repeat sequence motifs (Figure 1.8-1). Repeat regions are variable in length, where the number of repeat copy number could change significantly and these repeats are referred as variable number of tandem repeats (VNTR).

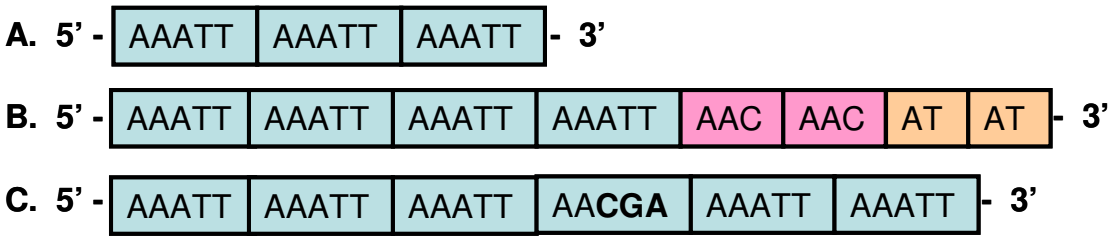


Figure 1.8-1. Schematic representation of different repeat types. (A) Homogeneous sequence motif consisting of 5-nucleotides in length. (B) Heterogenous repeat consisting of two 5-nucleotide units, two 3-nucleotide units and three 2-nucleotide units. (C) Degenerate repeat of AAATT with the discrepancy in bold. Adapted from van Belkum *et al.* (319).

Slippage strand misalignment (SSM) during DNA synthesis (162) together with inadequate DNA mismatch repair has been suggested to cause the variability observed in VNTRs. Briefly, SSM is initiated with denaturation and displacement of the DNA strands followed by mispairing of complementary bases at the site of an existing repeat. Mutational changes including base substitutions, insertions or deletions of one or more repeat units may generate new motifs that could be multiplied by subsequent SSM giving rise to VNTR (Figure 1.8-2). The possibility of duplication for longer VNTR increases as the regions of the repeats become longer.

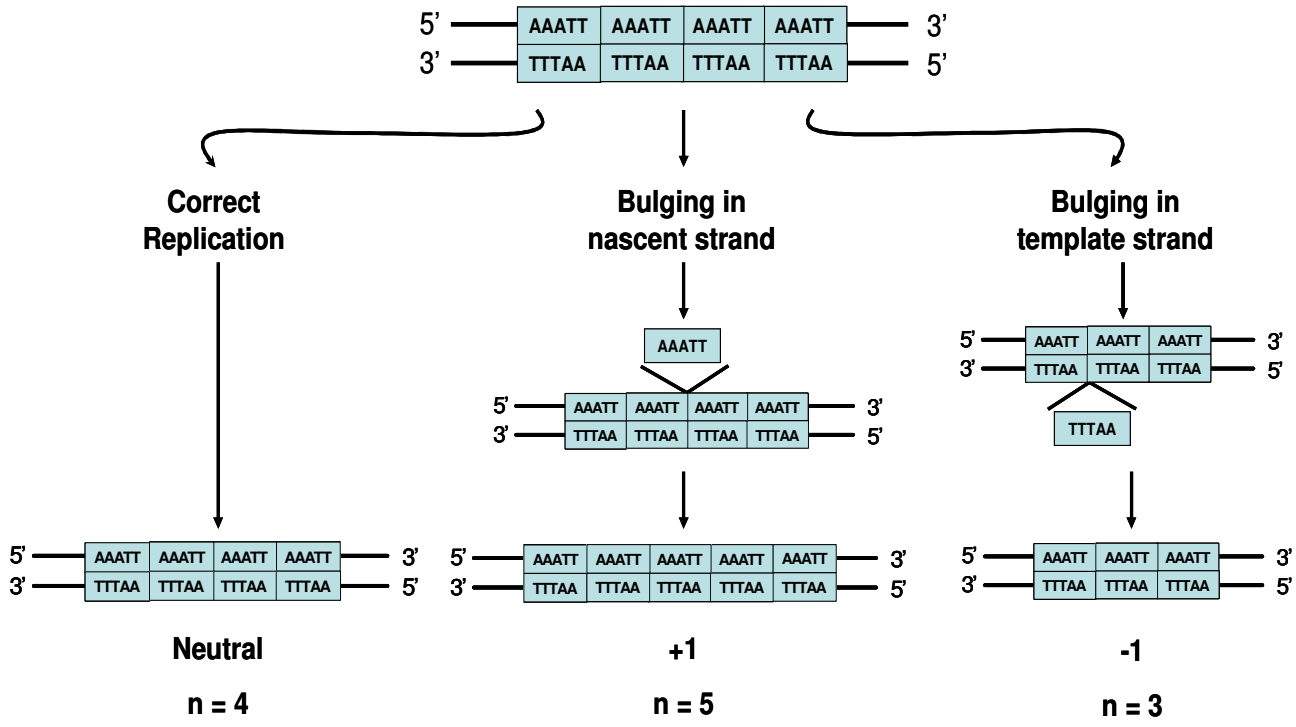


Figure 1.8-2. Schematic diagram representing the mechanism of SSM during replication that may result in either shortening or lengthening of VNTRs. Adapted from van Belkum *et al.* (319).

1.8.4.1 Functional roles of VNTRs in bacterial pathogenesis

Repeat sequences could be found in the non-coding intergenic regions (known as interspersed repeats) or within coding regions as part of open reading frames or on promoter regions (known as contiguous repeats) (319). Interspersed repeats are widely distributed throughout the genome and examples of these include the repetitive extragenic palindrome (REP) elements (92), the enterobacterial repetitive intergenic consensus (ERIC) (123) and the BOX repeats (188). Contiguous repeats usually vary between strains and the diversity in bacterial populations lead to phenotypic differences due to differential gene transcription and translation.

The availability of complete bacterial genome sequences allows for the VNTRs to be identified and compared. There are more abundant shorter-length repeats than the longer ones. However, VNTRs are not generally present at high frequency except the repeats that appear to be functional for regulating genes (80). Analyses of repeats in more than 300 prokaryotic genomes suggest that there

are significant differences in the distribution and types of repeats (205). Short repeats of 1-4 nucleotides are more often found in the host-adapted pathogens with reduced genomes, and are presumably involved in enhancing antigenic variance. Repeats with 5-11 nucleotides are usually found in non-pathogenic bacteria and opportunistic pathogens with large genomes and could be due to the tendency for expansion of genomes (205).

These repeats in bacterial genomes may affect the regulation and structure of DNA, while polymorphic repeats in pathogenic bacteria have been associated with improved virulence. The variation in the repeats also allows new transcriptions for genes coding for various surface exposed proteins leading to antigenic shifts and thus better evasion of the host immune system. The use of VNTRs to facilitate the survival of pathogens has been demonstrated in several pathogens, for example *Haemophilus influenzae*, *Neisseriae*, *Mycobacteria*, *Staphylococcus aureus* and *Helicobacter pylori*.

1.8.4.1.1 Haemophilus influenzae

H. influenzae colonises the human respiratory tract and could also cause meningitis. As an opportunistic pathogen this bacterium has developed strategies to evade the host immune systems, including polymorphisms in VNTRs. The genome of *H. infuenzae* has been explored for the presence of VNTRs and it was found that tetranucleotide repeats were abundant. However, it also has many different classes of repeats including mono-, di-, penta- and heptanucleotide motifs. These repeats have been associated with virulence genes involving LPS biosynthesis (116, 128, 339), haemoglobin receptors (265), putative adhesion molecules (53) and a methyltransferase of the type III restriction modification system (20, 55).

1.8.4.1.2 Neisseriae

DNA motifs consisting of pentamers are present in a large number of copies in two important pathogenic *Neisseriae* species, *N. gonorrhoea* and *N. meningitis* (207). The number of repeats in this adhesin varies continuously at low frequency in vivo. However, it has been demonstrated that

environmental stresses will enhance the growth of a repeat with a particular number of copies (299). H.8, the surface-exposed lipid and protein-based macromolecules associated with the virulence of *Neisseriae*, is also controlled by the presence of pentameric repeats (12). These suggest that pentameric VNTRs are one of the important factors in the pathogenicity of *Neisseriae*.

1.8.4.1.3 Mycobacteria

Analyses of genomes of *M. tuberculosis* and *M. bovis* genomes reveal the presence of polymorphic VNTRs that are located on the genic regions and have undergone frame-shift mutations leading to variation in the copy numbers. These genes encode for membrane surface lipoproteins, transporters, cell-wall synthesis proteins and hypothetical proteins (297). In these pathogens, VNTRs have been shown to be involved in the variations of virulence, surface antigen variations and adaptations to the host.

1.8.4.1.4 Staphylococcus aureus

Genes encoding membrane bound proteins called microbial surface components, which recognise adhesive matrix molecules (MSCRAMMs) in *S. aureus* have repeat motifs ranging from 1 to 81 bp units. The repeats in MSCRAMMs are on the *cna* gene involved in collagen-biding (267), the *clfA* gene encoding the clumping-factor (196), the *coa* gene encoding coagulase protein (93) and the *fnb* gene encoding for fibronectin binding protein (133).

1.8.4.1.5 Helicobacter pylori

As the genome of *H. pylori* strain HP26695 was completely sequenced (312), analysis has shown the presence of 27 VNTRs-dependent and thus, putative phase-variable genes. These repeats are on the genes encoding products involved in lipopolysaccharide (LPS) biosynthesis, cell-surface-associated proteins and DNA restriction/modification systems (277). In this pathogen, the size of repeats also determines whether particular genes are switched on or off. This is the case, for example, with the *cagA* gene, which has been associated with virulence (351) and the *fliP* gene,

which encodes for a basal body and is involved in the motility of *H. pylori* (134), suggesting the involvement of the repeats in antigenic variation or adaptive evolution.

In general, these examples highlight the involvement of VNTR elements in bacterial pathogenesis. VNTRs allow them to respond to varieties of environmental stress, and many of them are involved in the antigenic variation and surface-exposed membrane proteins. As whole genome sequences become increasingly available, more comprehensive studies of the involvement of repeat variation in bacterial pathogenicity and adaptation to the host could be performed.

1.8.4.2 VNTRs as markers for epidemiological studies

Due to their polymorphism, VNTRs are useful markers to subtype bacterial pathogens. The technique that involves the amplification and analysis of fragment size at DNA regions containing VNTR is known as multiple-locus VNTR analysis (MLVA). MLVAs for more than dozens of bacteria including pathogenic bacteria, which are homogeneous including *S. aureus*, *M. tuberculosis*, *Yersinia pestis* and *Bacillus anthracis*, have been proven to be a rapid method for determining their genetic diversities.

1.8.4.2.1 Staphylococcus aureus

As previously mentioned, the genome of *S. aureus* carries a variety of DNA repeats. The first use of MLVA for typing of 34 clinical isolates of *S. aureus* was reported in 2003 based on five VNTRs, with repeat motifs ranging from 9 to 24 bp, located in the *sdr*, *clfA*, *clfB*, *ssp* and *spa* genes (273). The assay utilised multiplex PCR methods and the analysis was done using standard agarose gel electrophoresis separation. However, the coagulase gene that has been known to contain repeats (93) was excluded from the assay as it was unable to differentiate the isolates. The study showed that the MLVA was able to distinguish and determine the relationship between the tested isolates and its discriminatory power was comparable to PFGE (273). Unfortunately, the difference between isolates was only observed based on the banding pattern, therefore the individual numbers of repeats per locus were not easily determined.

Other repeat-based typing systems have been developed for *S. aureus*, including the use of Staphylococcal interspersed repeat units (SIRUs). SIRUs consist of seven VNTRs with repeat units ranging from 48 to 159 bp that are present in all seven genomes of *S. aureus* (101). The VNTRs were typed in 16 isolates from the United Kingdom and the assay was compared with PFGE and MLST. The highest discrimination was achieved by PFGE followed by MLVA and MLST, suggesting that SIRUs could not replace the MLVA scheme described earlier (101).

1.8.4.2.2 Mycobacterium tuberculosis

Tandem repeats were first used in 1991 to identify *M. tuberculosis* using the flanking regions of an insertion element composed of 36 bp direct repeat copies interspersed by spacer elements, which are non-repetitive sequences of equal length (107). MLVAs have been used in various comprehensive studies whereby four typing sets of novel VNTR loci were identified. The first set contains 11 novel VNTRs of 15 bp repeats that were amplified in 48 *M. tuberculosis* strains. However, the resolution was lower than the IS*6110* typing assay (88). A second set of markers consisted of 12 polymorphic Mycobacterial interspersed repetitive units (MIRUs) that were 46-100 bp sequences dispersed within intergenic regions, and two other tandem repeats (302, 303). This panel was applied on 31 *M. tuberculosis* complex strains and the discriminatory power was found to be similar to IS*6110* typing (302, 303).

The third set contained eight new VNTRs as well as 13 previously described loci that were assessed on 90 *M. tuberculosis* complex strains (*M. tuberculosis* (64 strains), *M. bovis* (9 strains), *M. africanum* (17 strains) (157). These repeats were units of a multiple of three base-pairs; fifteen of which were within putative genes. Lastly, in the fourth set, the MLVA involved the initial identification of 87 VNTR loci from the genome strain H37Rv (296). Nine loci were variable and using six of the nine loci, the VNTRs were able to discriminate seven VNTR types of 34 *M. tuberculosis* strains, which were isolated from members of the Beijing family that represented 14 different IS*6110* RFLP types (296). In addition, five loci could distinguish isolates that have a low number of IS*6110* copies, suggesting that MLVA was superior to IS*6110* typing and was independent from the IS*6110* copy number (296).

1.8.4.2.3 Bacillus anthracis

B. anthracis could infect both animals and humans and is the causal agent of anthrax. The first MLVA approach was done using a 12 bp tandem repeat in the *vrrA* gene (126). This VNTR divided 198 *B. anthracis* isolates into five groups. A well-defined MLVA was established using an additional six loci, named $vrrB_1$, $vrrB_2$, *vrrC1*, *vrrC2* and CG3. These were found and characterised from the AFLP studies, together with two loci on the virulence plasmid pXO1 and pXO2 (138). A new set of VNTR regions in 14 loci were described and have distinguished the 31 isolates typed into 27 genotypes (158). There are many published studies using MLVAs using *vrrA* and seven of the above mentioned loci: to distinguish 426 global isolates into 89 distinct genotypes and two major clonal lineages (A and B) (138), to demonstrate that 98 isolates from the Kruger national park in South Africa had the greatest genetic diversity (295) and to confirm that 135 *B. anthracis* isolates from an anthrax outbreak in USA in 2001 were of the same genotype as the strain Ames that was used in laboratories (113). The MLVA method has become the current standard for molecular typing of *B. anthracis* isolates and has since been used in studies to describe the genetic diversity of isolates in France (86), Poland (91), Italy (69), Korea (272), Georgia (199), and Chad (184).

1.8.4.2.4 Yersinia pestis

As an etiologic agent of plague that has killed many millions of people in three separate pandemics, *Y. pestis* has been regarded as a possible agent of bioterrorism. It is a highly monomorphic pathogen characterised by a lack of sequence diversity (3). The establishment of a rapid and highly discriminating typing method is important and thus there is a need for an excellent MLVA. The first VNTR identified was a 4-bp unit repeat located in the intergenic region and nine alleles were identified when this was analysed in 35 diverse *Y. pestis* strains (4). Using the Tandem Repeats Finder software, 49 of the 76 identified VNTR loci were found to be polymorphic among five isolates (158). Twenty five of these loci were used to differentiate 180 isolates into 61 genotypes and a subset of seven markers was proposed for rapid comparison (244). Klevytska *et al.* (144) also

examined the allelic diversity of 42 chromosomal VNTR loci, identified using Genequest software, in 24 selected isolates including 12 global isolates and 12 Californian isolates. Furthermore, using VNTRs, the human case isolates could be associated to isolates from environmental sources allowing rapid epidemiological differentiation between bioterrorism and naturally occurring plague (178).

1.8.4.3 The use of VNTRs for typing of *Salmonella enterica*

Currently, the most sophisticated MLVA uses a set of eight VNTRs with repeat units ranging between six to 198 bp (257). This assay was able to differentiate 99 isolates of *S. enterica* subspecies *enterica* serovars Typhi, Paratyphi A and Typhimurium from human sources in France into 52 genotypes and classify them into four distinct groups (257). The polymorphism of the VNTRs were able to categorise the isolates into their respective serovars and subtype 25 and eight isolates of serovars Typhi and MDR Typhimurium, respectively (257). The discriminatory power of MLVA was much higher than any of PGFE, AFLP or integron profiling. The markers described in the study hold the potential for standardised MLVA to investigate outbreaks of salmonellosis. However, independent useful VNTR markers have also been described for typing of isolates from serovars Typhimurium, Newport, Enteritidis and Typhi, respectively.

1.8.4.3.1 Serovar Typhimurium

In 2003, VNTR loci were first used to type *S. enterica* serovar Typhimurium phage type DT104 (164). Although PFGE has also been regarded as the "gold-standard" to type Typhimurium, phage type DT104 is highly homogeneous where unrelated strains show identical PFGE profiles and strains of this phage type are epidemiologically related (266). Eight VNTR loci with repeat motifs ranging from six to 189 bp were chosen to type 37 serovar Typhimurium phage type DT104 isolates. Five of the loci had a high discriminatory power and could be used to differentiate the isolates into 28 VNTR patterns (164). Fluorescent Capillary Electrophoresis (CE) was used to determine the size of the VNTR loci. However, one-coloured dye is costly and inefficient and the sizes of some PCR amplicons were too large for the capillary instrument. In a subsequent study, the

problems were resolved by introducing two new VNTR loci that were shorter and using multiple fluorescent dyes, which improved the MLVA by allowing it to analyse a total of 106 Typhimurium isolates from different hosts including 16 DT104 isolates (165).

MLVA has been shown to have a discrimination capability at least as good as that of PFGE. Using the genome of Typhimurium LT2, 54 sequences containing VNTR loci were identified and 10 loci, with repeat units ranging from 6 bp to 232 bp, were analysed in 30 and 20 isolates of serovars Typhimurium and Newport respectively, from Arkansas (344). Only six showed polymorphisms. Both VNTR and PFGE showed identical discrimination of the isolates (344). In a study of 1019 Typhimurium isolates, which were collected between the year 2003-2005 in Denmark from routine surveillance, MLVA was shown to be a better typing method than PFGE (314). The isolates were distinguished into only 148 PFGE types while MLVA differentiated them into 373 VNTR types (314). Another study also compared PFGE and MLVA using five VNTR loci to type 195 epidemiologically unrelated Typhimurium isolates from pigs in the period of 1997-2004 (25). An additional 190 Typhimurium isolates from poultry and 186 from human cases of gastroenteritis were also analysed. Better discrimination was achieved using VNTR where only 34 PFGE profiles identified were obtained, compared with 96 different MLVA profiles for the pig isolates. Most common PFGE type was also further differentiated into 56 different VNTR types (25).

1.8.4.3.2 Serovar Enteritidis

MLVA has also been developed for subtyping of the serovar Enteritidis. In a study employing this technique, 153 isolates were obtained during 1998-2003, including 40 isolates of serovar Enteritidis from four food-borne disease outbreaks (30). The MLVA utilised 10 loci located in both coding and intergenic regions of the chromosome and varying in size from 6 to 117 bp repeat units. Its discrimination ability and epidemiological value was compared with PFGE and phage typing. There were 57 MLVA types, 33 PFGE types and 15 phage types, suggesting that MLVA had greater discrimination among non-epidemiologically linked isolates than both PFGE and phage typing (30). Seven of the ten VNTR loci with the highest discrimination were optimised for multiplexing and were used to type 34 isolates of serovar Enteritidis from 1978 to 2004 (45). The isolates were both from human and non-human sources and MLVA was able to associate the

isolates with their sources. Here too, MLVA was shown to have a better discriminatory power than PFGE, phage typing or MLEE, suggesting its usefulness for molecular epidemiologic study of serovar Enteritidis infections (45).

1.8.4.3.3 Serovar Typhi

The Tandem Repeats Finder program was used to explore the genome sequence of serovar Typhi CT18, while five VNTR loci with repeat units ranging from 6-16 bp motifs were selected to type 61 Typhi isolates from various Asian countries from the years 2000-2001 (174). The sizing of the VNTRs was done on normal agarose gels and only three VNTRs showed variation, resulting in 49 MLVA types being identified (174). These VNTRs, designated as TR1, TR2 and TR3, were multiplexed and the study has shown the presence of genetic diversity among Typhi isolates between different geographical areas (174). The locus TR1 was also included in the study by Ramisse *et al.* (257) as previously mentioned and showed variations in the 99 isolates analysed.

1.9 Multidrug-resistant Typhi are clonal and antibiotic resistance is plasmid-borne

PFGE has suggested that most outbreaks in Asia and Africa have been due to a single or closely-related serovar Typhi strain/s. MDR Typhi isolates have been shown to be distinct and could independently co-exist with sensitive strains (308). Six MDR Typhi strains isolated from Korea in 1999 were analysed for their plasmid size and ribotyping pattern (230). The resistance to antibiotics ampicillin, chloramphenicol, trimethoprim-sulfamethoxazole, streptomycin, tetracycline and gentamicin is believed to be the result of a complex class 1 integron containing six resistance gene cassettes (230). The isolates were found to have the same patterns both by ribotyping and plasmid size, and it was therefore concluded that these isolates were clonal.

MDR Typhi isolates obtained from the blood of patients in three different parts of Kenya, reported from recent outbreaks in a two year period, were characterised using plasmid and chromosomal DNA typing (136). Of 102 Typhi isolates analysed, only 14 isolates were fully susceptible to the 11

antibiotics tested. Four strains/isolates showed different susceptibilities to antibiotics, while the remaining 84 isolates were all uniformly resistant to ampicillin, streptomycin, chloroamphenicol, tetracycline and cotrimoxazole, which are the first antibiotics administered in Kenya. The isolates also had higher tolerance to the antibiotics, indicated by a 5- to 10-fold increase in nalidixic acid and ciprofloxacin minimal inhbitory concentrations (MICs) respectively, although it remained within the sensitive range (136). Plasmid analyses showed that all MDR strains contained a single 110-kb plasmid, which was most likely acquired by horizontal transfer, and six isolates contained one to two additional plasmids of 4-10 kb in size (136). Analysis by PFGE, using *Xba*I and *Spe*I restriction enzymes, characterised two different patterns: fragment pattern I which is represented by 75.5% of the isolates studied and fragment II, which consisted of 24.5% Typhi isolates. All MDR isolates from South Africa had pattern I, again showing these isolates are from a single clone. However, neither antibiotic sensitive isolates from South Africa nor MDR isolates from Hong Kong and Pakistan had either PFGE patterns identified in that study (136). It was shown that MDR and sensitive serovar Typhi isolates shared similar PFGE patterns and were not specific to any region in Kenya, suggesting that there are multiple MDR clones in Kenya and they coexist with the sensitive isolates.

1.10 Aims of the study described in this thesis

Only a few of the more than 2,400 serovars of *S. enterica* exclusively infect humans and are medically important. These serovars include Typhi, the agent of human typhoid fever, and serovars Paratyphi A, Paratyphi B clone b1, Paratyphi C and Sendai, all of which cause the milder form of typhoid-like enteric fevers. Although the relationships between *S. enterica* subspecies are well established (34), little is known about the relationship among enteric fever causing serovars. Consequently, the evolution of host adaptation and pathogenicity of these serovars is not well-constructed. It is yet to be determined whether these serovars are related to one another or whether they resulted from convergent evolution in multiple phylogenetic ancestries. Previously, MLEE has shown that there is no close evolutionary relationship between clones of different enteric fever causing serovars except for serovar Paratyphi A and Sendai (285). However, microarray study has shown that Typhi is closely related to serovars Paratyphi A and Sendai (43). Completed genome

sequence of serovar Paratyphi A strain ATCC9150 also reveals a high similarity in gene contents to serovar Typhi strain CT18 (193). This suggests that comparison at nucleotide sequence level is necessary to further confirm the MLEE study. Thus, **the first aim of the project** was to determine the genetic relationships of *S. enterica* serovars closely related to serovar Typhi, as measured by MLEE, and other enteric fever causing serovars using the sequencing of six housekeeping genes (Chapter 3).

Typhoid fever remains as a major global health problem that could be controlled by implementing adequate food handling practices and proper management of safe water supplies. The major needs are reliable, highly discriminatory and inexpensive typing methods for epidemiology study, rapid diagnostic tests, and a cost effective treatment and prevention strategy. The emergence, spread and persistence of Typhi, especially the MDR strains and Vi negative Typhi strains, in endemic countries challenge current therapeutic and public health managements.

Current typing methods are DNA based, which include PFGE and ribotyping, and have been shown to be useful for typing of isolates from sporadic and outbreak cases. Unfortunately, these methods could not be used to establish the evolutionary relationships of Typhi. The ongoing genome sequencing projects as well as the fully completed ones that include two strains of serovar Typhi, provide a better understanding of the gene content and the genome organisation of *S. enterica* serovars. Comparative study of the two Typhi genomes has revealed high conservation of genetic information and that the differences were contributed by strain specific genes and single nucleotide polymorphisms (SNP). SNPs have been shown to be a valuable molecular marker in other species. **The second and third aim of the project** was to use genome-wide SNPs as a marker to differentiate Typhi isolates and to establish the genetic relationships between these isolates (Chapter 4 and Chapter 5).

SNPs identified from comparison of only two Typhi genomes had a limit in that only mutations between the two genomes can be discovered. SNPs in the other isolates could not be detected. For best resolution to establish the relationships of Typhi isolates using SNPs, genomes from many different lineages are required. However, until the platform for whole genome sequencing is fully automated and cost efficient, it will not be plausible to sequence all Typhi isolates. Thus, a new

method to discover SNPs without the need to fully sequence the genomes is important. This comes to the **fourth aim of this project** which was to discover new SNPs in Typhi isolates using an enzymatic based approach (Chapter 6) and large scale cloning.

The genomes have also been shown to contain repeat units that are variable between strains, termed as variable number of tandem repeats (VNTRs). These VNTRs have also been used as a marker for typing, including for Typhi. However, no study has been previously done in Typhi to determine if this marker is also suitable for ascertaining phylogenetic relationships and for global epidemiology of Typhi. **The fifth and the last aim** of this project was to use VNTR for typing and to compare it to SNP-based typing (Chapter 7).

Ultimately, the knowledge acquired could shed light on the evolution of host adaptation and pathogenicity of serovar Typhi. This will provide a better picture of the evolutionary processes of typhoid fever and the factors that contribute to the variation in the severity of disease and diverse clinical outcomes of infection.

Chapter 2: General Materials and Methods

2.1 Strains

2.1.1. List of strains

Fifteen SARB strains (Table 2.1-1) and seventy-three worldwide Typhi isolates (Table 2.1-2), differing in localities and year of isolations, were obtained from *Salmonella* Genetic Stock Centre, University of Calgary, Canada except one, Typhi strain 422Mar92 which was obtained from Imperial College London, UK (Table 2.1-2). The bacteria were maintained on nutrient agar plates [13 g/l nutrient broth (Oxoid) and 15 g/l agar (Ajax)] and stored in glycerol solution [every 1 ml solution contains 0.4 ml of 60% glycerol (Ajax) and 0.6 ml of 1% peptone (Sigma)] aliquoted into a glass vial (ProSciTech) at -80^{o}C.

Table 2.1-1. *S. enterica* strains belonging to subspecies I used in this study (Chapter 3)

Strain Name	SARB No.	Serovar	Source	Locality	MLEE Difference[1]	Serogroup
Cs6	SARB5	Choleraesuis	Unknown	Switzerland	16	C_1
De1	SARB9	Derby	Avian	Oklahoma, 1986	17	B
In1	SARB26	Infantis	Human	North Carolina	11	C_1
Mo1	SARB30	Montevideo	Human	Georgia	10	C_1
Np8	SARB36	Newport	Human	North Carolina	10	C_2-C_3
Pa1	SARB42	Paratyphi A	Human	Lab Strain	12	A
Pb1	SARB43	Paratyphi B	Human	France, 1976	14	B
Pb7	SARB47	Paratyphi B	Human	Africa, 1981	11	B

Pc2	SARB49	Paratyphi C	Human	France, 1988	13	C_1
Pc4	SARB50	Paratyphi C	Human	France, 1977	9	C_1
Pn1	SARB39	Panama	Human	Italy	10	D
Sf1	SARB59	Senftenberg	Chicken	Maryland, 1987	11	-[2]
Sw1	SARB57	Schwarzengrund	Unknown	Scotland, 1988	10	D
Tm1	SARB65	Typhimurium	Human	Mexico	14	B
Tp2	SARB64	Typhi	Human	Senegal, 1988	-	D

[1] Based on MLEE data from Boyd *et al.* (33) and the allelic difference was between this strain and Typhi Tp2

[2] The serogroup of serovar Senftenberg (Serogroup E_4) could not be confirmed as no appropriate primer pair was available.

Table 2.1-2. List of 73 Typhi isolates used in this study (Chapters 4 to 7)

Strain Name	Genotype	Phage Type	Locality	Year	z66 Flagellar antigen[1]
3123	3		Chile	1983	-
3125	3	46	Chile	1983	-
3126	3	46	Chile	1983	-
25T-36	29	E1	Alberta	1993	-
25T-40	4	E1	BC	1993	-
25T-44	2	E1	Ontario	1993	-
26T12	6	O	Manitoba	1994	-
26T17	4	B1	BC	1994	-
26T19	5	A	Alberta	1994	-
26T24	2	E1	Ontario	1994	-
26T30	3	I+IV	Quebec	1994	-
26T32	24	I+IV	Quebec	1994	-
26T37	2	I+IV	BC	1994	+
26T38	14	E1	BC	1994	-
26T40	19	M3	BC	1994	-

26T49	11	B1	BC	1994	-
26T50	8	I+IV	Alberta	1994	-
26T51	28	DVS	BC	1994	-
26T56	23	F1	Quebec	1994	-
26T6	30	UT	BC	1994	-
26T9	16	B1	Manitoba	1994	-
414Ty	3	I + 1V	Australia	1981	+
415Ty	3	UT	Netherlands	1982	+
416Ty	3	UT	Japan	1982	+
417Ty	22	I+IV	New Caledonia	1982	+
418Ty	3	I+IV	Netherlands	1988	+
419Ty	3	I+IV	Netherlands	1988	+
420Ty	3	UT	Japan	1982	+
421Ty	3	UT	France	1984	+
422Mar92			Zaire	1992	-
423Ty	3	I+IV	Australia	1981	+
425Ty	3	I+IV			+
444Ty	3	I+IV			+
445Ty	3				+
446Ty	3	I+IV			+
701Ty	27				+
702Ty	3				+
CC6	7	A	Thailand	1995	-
CC7	7	A	Thailand	1995	-
CDC1196-74	6	A	Mexico		-
CDC1707-81	3	UT	Liberia		-
CDC3137-73	6	K1	India		-
CDC3434-73	22	G1	Peru		-
CDC382-82	13	M1	Marshall Island		-
CDC9032-85	23	UT	Taiwan		-
CT18			Vietnam	1994	-

In15	3	D2	Indonesia	1994	-
In20	9	A	Indonesia	1992	+
In24	3	C3	Indonesia	1992	-
IP.E88 353		UT	Darkar		-
IP.E88 374		UT	Dakar		-
PL27566	26	M1		1994	-
PL73203	2	A			-
PNG32	2	D2	Papua New Guinea	1994	-
R1167	19	A			-
R1637	14	E2			-
R1962	1	UT	Alberta	1993	-
ST1	18	I+IV	Indonesia		-
ST1002	9	E1	Malaysia	1987	-
ST1106	4	D1	Malaysia	1987	-
ST145	3	I+IV	Malaysia	1994	+
ST24A	3	DVS	Malaysia	1986	-
ST24B	3	DVS	Malaysia	1986	-
ST309	3	E1	Malaysia	1987	-
ST60	2	C4	Malaysia	1986	-
T189	11	N	Thailand	1990	-
T202	3	UT	Thailand	1990	-
Tp1	25	A	Dakar	1988	-
Tp2	17	UT	Dakar	1988	-
Ty2	9	E1			-
TYT1668	21	M1	Chile		-
TYT1669	6	UT	Chile		-
TYT1677	5	F8	Chile		-

[1] Strains ST145, 26T37 and In20 were found to carry z66 by PCR using primer pairs adapted from Huang *et al.* study (122)

2.2. PCR serogrouping for identification of S. enterica strains

Serogrouping by PCR was performed on all *S. enterica* strains used in this study. The PCR primers were designed to identify the isolates to serogroup level by amplifying the *rfb* gene cluster (117, 180). This approach could accurately determine major *S. enterica* serogroups including serogroups B, C_1, C_2-C_3 and D. Serogroup A and D only differ at one nucleotide and therefore the same primer pair was used to confirm the identity of isolates belonging to these serogroups. The primers (Table 2.2-1) were designed to yield PCR products with different sizes for better discrimination on the agarose gel (Figure 2.2-1) (117, 180).

Table 2.2-1. The primers used for serogrouping of *S. enterica* isolates adopted from Hoorfar *et al.*(117) and Luk *et al.* (180) studies.

Serogroup	Primer Pair	Sequence 5' -> 3'
B	9044	AGAATATGTAATTGTCAG
	9045	TAACCGTTTCAGTAGTTC
C_1	9046	GGTTCCATAAGTATATCT
	9047	CTGGATACGAACCCGTAT
C_2-C_3	9048	ATGCTTGATGTGAATAAG
	9049	CTAATCGAGTCAAGAAAG
A/D	9050	TCACGACTTACATCCTAC
	9051	CTGCTATATCAGCACAAC

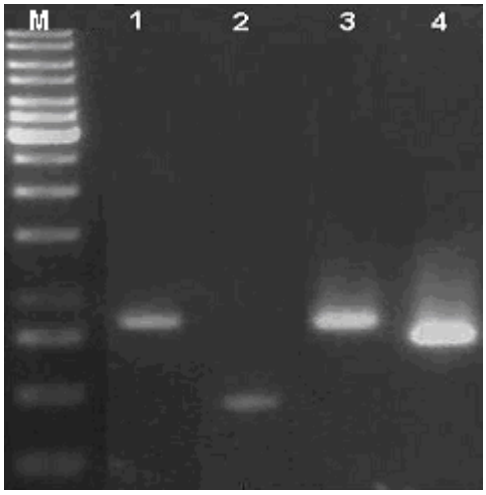


Figure 2.2-1. PCR-products corresponding to different serogroups on 1% agarose gel electrophoresis. Lane M – 1 kb marker [from the largest size to the smallest size visible: 10 kb, 8 kb, 6 kb, 5 kb, 4 kb, 3.5 kb, 3 kb, 2.5 kb, 2 kb, 1.5 kb, 1 kb, 750 bp, 500 bp, 250 bp]; Lane 1 – serogroup B; Lane 2 – serogroup C_1; Lane 3 – serogroup C_2-C_3 and Lane 4 – serogroup A or serogroup D.

2.3. Confirmation of the z66 flagellar antigen

All the Typhi isolates used in this study were typed for the presence of the z66 flagellar antigen gene by PCR using primers of Huang *et al.* (122) (Table 2.3-1 and Figure 2.3-1). Among the 73 Typhi isolates, 15 were known to express z66 by serotyping and were all confirmed by PCR. An additional three isolates were found to carry the z66 flagellar antigen gene by PCR.

Table 2.3-1. The primer pairs used to type for z66 flagellar antigen adopted from Huang *et al.* (122) study

Target	Primer Pair	Sequence 5' -> 3'
z66[1]	9292	CAACCGCTAGTGATTTAGTTT
	9293	CTGTCCCTGTAGTAGCCGTAC
fliC[2]	9294	ATGCCTACACCCCGAAAGAA
	9295	ACCCTCTTTTGTTACTTCAG

[1] Primer pair 9292/9293 was used to amplify a 375 bp fragment from the central region of the new z66 flagellin gene.

[2] Primers pair 9294/9295 was designed from the specific central region of *fliC* of Ty2 to amplify a 340 bp fragment of *fliC*. All Typhi isolates show successful PCR amplifications when this pair of primers is used.

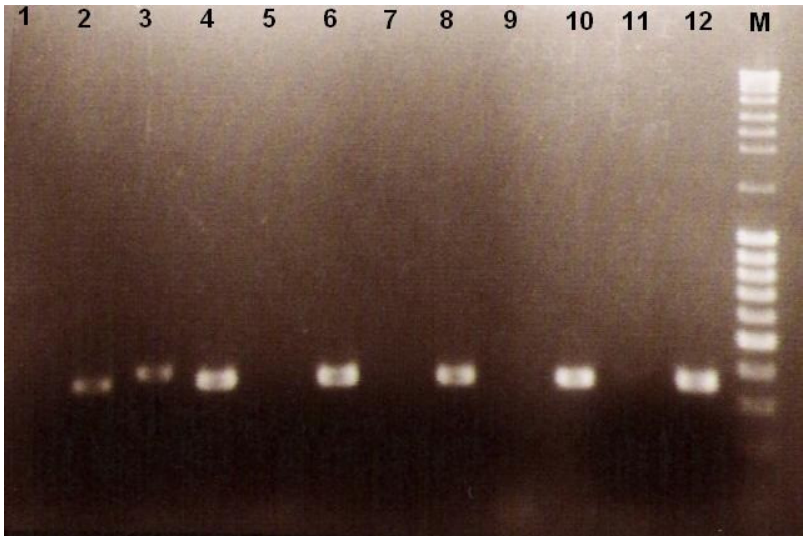


Figure 2.3-1. PCR amplification using z66 and *fliC* primers. M – marker; Lane 1 and 2 - 414Ty; Lane 3 and 4 – CT18; Lane 5 and 6 – Ty2; Lane 7 and 8 – 3123; Lane 9 and 10 – 3125; and Lane 11 and 12 – 3126.

2.4. Phenol/Chloroform DNA Extraction

DNA extraction by this method was modified from Bastin *et al.* study (17). A single subcultured bacterial colony was inoculated into a 10 ml nutrient broth and incubated on shaker at 37˚C overnight. From the 10 ml overnight culture, 1 ml was transferred into a microcentrifuge tube and centrifuged at 14,000 rpm for 30 sec. The supernatant was removed and the step was repeated. The bacterial cell pellet was resuspended in 250 µl of 50 mM Tris-HCl (Promega) pH 8.0 and spun down at 14,000 rpm for 30 sec. Another 250 µl Tris-HCl was added along with 10 µl 0.5 M EDTA (Sigma) followed by incubation at 37°C for 20 min. After that, 10 µl (20 mg/ml) lysozyme (Sigma) was added and incubated again at 37˚C for 20 min. Subsequently, 1.5 µl (20 mg/ml) proteinase K (Sigma) and 15 µl (10%) SDS (Promega) were added, mixed gently and incubated at 50˚C for 2 h. Thereafter, 0.5 µl (20 mg/ml) RNase (Sigma) was added and further incubated for at 65°C for 15min. The supernatant was transferred into a phase-lock gel dividing tube (Eppendorf). The solution was mixed with 250 µl of phenol (Ajax):chloroform (Ajax):isomyl alcohol (Sigma) in the ratio of 25:24:1 for 2 min and centrifuged at 14,000 rpm for 4 min and the step was repeated twice. A 250 µl of chloroform:isoamyl alcohol (24:1) was then added into the tube and followed by

centrifugation at 14,000 rpm for 4 min. The top aqueous phase was transferred into a new microcentrifuge tube. Approximately 600 µl of cold 100% ethanol (EtOH) (Ajax) was added and the tube was slowly inverted to mix and centrifuged at 14,000 for 5 min to precipitate the DNA. The precipitated DNA was spooled and rinsed in 70% ethanol and finally dissolved in a new microcentrifuge tube containing 100 µl TE buffer [10 mM Tris-HCl pH 8.0 and 1 mM EDTA pH 8.0]. The tube was heated with lid opened for 10 min at 65˚C to evaporate EtOH.

2.5. PCR

The PCR reaction contained a mixture of 0.2 µl DNA template (~20 ng), 0.2 µl (30 pmol/µl) each forward and reverse primers (Sigma-Aldrich), 0.2 µl 10 mM dNTPs, 2 µl 10x Taq polymerase PCR buffer (New England Biolabs), 0.125 µl (1.25 U) Taq polymerase (New England Biolabs) and MilliQ water to adjust to the final volume to 20 µl. PCR cycles were performed in a Hybaid PCR Sprint Thermocycler (Thermo Analysis Biocompany, Hybaid Limited, UK) with the following conditions: initial DNA denaturation for 2 min at 94 °C; followed by DNA denaturation for 15 sec at 94 °C, primer annealing for 30 sec at 50 °C and polymerisation for 90 sec at 72 °C for 35 cycles, with a final extension of 5 min at 72 °C. PCR products were verified on ethidium bromide-stained 2% TBE agarose gel in 1xTBE buffer [every litre contains: 10.8 g Tris Base (Promega), 5.5 g Boric acid (Analar Analytical Reagent) and 4 ml 0.5 mM EDTA], before purification using sodium acetate/ethanol precipitation.

2.6. Sodium acetate/ethanol precipitation of PCR product

The precipitation of PCR product is carried out in a microcentrifuge tube and 2:1 of 80% EtOH and 1:10 of 3 M sodium acetate (Ajax) were added into the samples that need to be purified. The mixture was slightly vortexed and left at room temperature for at least 30 min. The tube was then centrifuged at 14,000 rpm for 15 min and the supernatant was removed. The tube was dried at 65°C on a heating block for 10 min to evaporate residual EtOH. Finally, the precipitated DNA was dissolved in 10 µl of MilliQ water.

2.7. Cloning: pGEM-T Easy Vector Ligation

The pGEM-T easy Vector system (Promega) was used for the cloning of PCR products. The pGEM-T easy vector is a high copy vector which contained T7 and SP6 RNA polymerase promoters flanking a multiple cloning region within the α-peptide coding region of the enzyme β-galactosidase. Insertion of the fragment would result in the inactivation of the α-peptide and this will allow recombinant clones to be identified by blue/white colour screening on selective media. The reaction includes 5 µl of 2x rapid ligation buffer, 1 µl pGEM-T easy vector (50 ng), x µl of purified PCR product, 1 µl of T4 DNA ligase (3 Weiss units/µl) and top up to a final volume of 10 µl with milliQ water. The reaction was incubated at room temperature overnight. The volume of purified PCR product was calculated based the concentration required for the insert on the following formula:

$$\frac{\text{50 ng of vector} \times \text{0.25 kb (average size of insert)}}{\text{3 kb (size of vector)}} \times \text{3 (Insert) : 1 (vector molar ratio)} = \text{12.5 ng of purified PCR product}$$

Adapted from pGEM -T and pGEM -T Easy Vector Systems Technical Manual

2.7.1. Preparation of competent cells

A 25 ml of Luria-Bertani Broth [Each 500 ml contains 5 g tryptone (Oxoid), 2.5 g yeast extract (Amersco) and 2.5 g NaCl (Ajax Chemicals)] was cultured with 1 ml of *E. coli* strain DH5α grown overnight and this was incubated at 30°C on a shaking waterbath until OD_{600} has reached 0.3. This was then incubated on ice for 15 min followed by centrifugation at 5000 rpm for 6 min at 4°C. The supernatant was discarded and the cells were resuspended in 8 ml transformation buffer 1 [10 mM MES (2-n-morpholino-ethane sulfonic acid) (Sigma), 100 mM RbCl (Sigma), 10 mM $CaCl_2$ (Sigma) and 50 mM $MnCl_2$ (Sigma)] and incubated on ice for 15 min. The cells suspension was centrifuged at 5000 rpm for 6 min at 4°C and the supernatant was discarded. The cells were again resuspended in another 8 ml of transformation buffer 1 and the steps were repeated. Finally, after the supernatant was discarded, the cells were resuspended in 800 µl of transformation buffer 2 [10 mM MOPS (Sigma), 10 mM RbCl, 75 mM $CaCl_2$ and 15% (v/v) glycerol (Ajax Chemicals)] and

incubated on ice for 15 mins before 80 µl were aliquoted into microcentrifuge tubes and stored at -80°C.

2.7.2. Heat Shock Transformation

The competent cells *E. coli* strain DH5α, prepared by $CaCl_2$ and RbCl, were incubated on ice for 5 min prior to transformation. Two µl of vector containing purified DNA after PCR amplification using *BsaH*I adaptor-specific and *Cel*I adaptor-specific primers was mixed with 35 µl of competent cells in a microcentrifuge tube. The cells were incubated on ice for 30 min prior to a heat shock at 42°C for 90 sec and immediately incubated on ice for 3 min. The cells were then added with 300 µl of LB broth which was preheated to 37°C followed by incubation at 37°C for 1 hr with agitation. After incubation, the cells were centrifuged at 14,000 rpm for 15 sec. The supernatant was removed and resuspended in a 100 µl LB. A total of 10 µl, 20 µl and 30 µl of the cells were each plated onto LB/Amp/IPTG/X-gal agar plates [LB added with 100 µg/ml ampicillin (Sigma), 0.5 mM IPTG (Progene) and 80 µg/ml X-gal (Promega)]. As a control, 30 µl of cells were also plated onto LB plate. The plates were all incubated overnight at 37°C.

2.7.3. DNA Extraction following cloning by Boiling Method

White colonies were picked and patched onto a new NA plate, and incubated at 37°C overnight. The next day, white patched colonies were then dissolved in 50 µl of TE buffer in a microcentrifuge tube. The TE buffer was then heated at 98°C for 10 min and after which, placed on ice for 3 min. The solution was then centrifuged at 14,000 rpm for 5 min. The supernatant was transferred into a fresh microcentrifuge tube and stored at -20°C.

2.8. DNA Sequencing

The 20 µl PCR sequencing reaction contained 1 µl BigDye™ (v. 3.1, Applied Biosystem), 20 ng purified PCR product, 3.5 µl 5x PCR sequencing buffer (Applied Biosystem), 1 µl 3.2 pmol/µl of

forward primer (Sigma-Aldrich) and MilliQ water. The sequencing reaction was run on a thermocycler with the conditions as recommended by the manufacturer, as the following: initial DNA denaturation for 2 min at 96^{o}C; and 25 cycles of denaturation at 96^{o}C for 10 sec, annealing at 50^{o}C for 5 sec and extension at 60^{o}C for 4 min. The ramp rate was set at 1^{o}C/sec as recommended by the BigDye™ manufacturer.

Unincorporated dye was removed by ethanol precipitation. A total volume of 20 µl of PCR sequencing sample was transferred into a microcentrifuge tube containing 16 µl MilliQ water and 64 µl 95% EtOH. The mixture was vortexed slightly and incubated at room temperature for 15 min. The mixture was subsequently centrifuged at 14,000 rpm for 20 min followed by removal of supernatant. The precipitated DNA was resuspended in 200 µl 70% ethanol and centrifuged at 14,000 rpm for 15 min. The supernatant was removed and the tube was dried on a heating block at 65°C for 10 min to evaporate residual EtOH. The sequencing reactions were resolved on an Automated DNA Sequence Analyser ABI3730 (Applied Biosystem) at the sequencing facility of the School of Biotechnology and Biomolecular Sciences, the University of New South Wales.

Chapter 3: Frequent recombination and low level of clonality within Salmonella enterica subspecies I

3.1 Introduction

Salmonella has been assigned to more than 2,500 different serovars (240). The classification into different serovars is based on the serotyping scheme which accounts for the difference in antigenic properties of the lipopolysaccharide (O antigen) and the flagellin (H antigen). These serovars were assigned Latin Binomial species names. However, because of their close relatedness, the species names were then retained as the serovar names of the single *Salmonella* species known as *Salmonella enterica* (35, 159). For example, the name for *S. typhi* refers to *S. enterica* serovar Typhi or simply Typhi (the latter convention is used in this paper). Based on DNA hybridisation and biotyping studies, the *Salmonella* serovars were classified into seven subspecies (I, II, IIIa, IIIb, IV, V and VI) (52, 160). Multilocus Enzyme Electrophoresis (MLEE) analysis has defined an eighth group, designated as subspecies VII, which consists of only five isolates of two serovars initially allocated to subspecies IV on the basis of biochemical characteristics (34).

Most subspecies of *S. enterica* are not commonly associated with disease and may behave like commensals in cold-blooded animal (18). However, subspecies I strains cause intestinal infections in warm-blooded animals and are responsible for 99% of *Salmonella*-related infection in humans (240, 286). The widely prevalent serovar Typhimurium causes gastroenteritis in humans but mainly asymptomatic chronic infection in chickens. A few serovars have a restricted host range, for example, Typhi exclusively infects humans, causing typhoid fever.

MLEE has been extensively utilised to study the extent of genetic diversity within *S. enterica* natural populations. It has shown that many serovars vary genetically and are represented by multiple electrophoretic types (ETs) (21, 22, 261, 284, 285). Some serovars are genotypically heterogeneous, for example Derby and Newport (21) include divergent isolates with ETs clustered distantly in MLEE trees while others could be confined within a single cluster of closely related ETs where each has a predominant widely distributed ET (21, 22, 261, 284,

285). From large scale MLEE studies, three reference collections were established by Selander's group: *Salmonella* Reference Collection A (SARA), which consists of 72 strains of serovar Typhimurium and its closely related serovars (22); *Salmonella* Reference Collection B (SARB), of 72 strains of 37 subspecies I serovars (33); and *Salmonella* Reference Collection C (SARC), of 16 strains representing the eight subspecies (34).

Based on MLEE data, the population structure of *S. enterica* is considered to be clonal with strong linkage disequilibrium, noted by non-random associations between the alleles of the 24 metabolic enzyme loci studied (21, 22, 261, 284, 285). A low recombination rate has also been demonstrated by the sequence data of six housekeeping genes from the 16 SARC strains. Gene trees for the six housekeeping genes are largely congruent (31, 34, 214-217, 333). These findings lead to the conclusion that *S. enterica* is one of the species with the highest level of clonality among bacterial species.

In this study we sequenced four genes from a selected number of SARB strains to determine the genetic relationships of strains belonging to subspecies I, in particular looking to see if there exists a serovar closely related to Typhi. Instead, we found that recombination is frequent in subspecies I, revealing a low level of clonality within the subspecies, and we were unable to resolve the relationships of the isolates studied.

3.2. Materials and Methods

3.2.1. Bacterial isolates

Fifteen SARB strains were chosen (Chapter 2, Table 2.1-1). The strains were obtained from the *Salmonella* Genetic Stock Centre (SGSC), University of Calgary, Canada. The strain names designated by Boyd *et al.* (38) have been used instead of the SARB numbers for convenience. SARB contains two Typhi strains, Tp1 and Tp2. Only Tp2 was selected for this study since Tp1 is identical to genome sequence strain CT18 based on multilocus sequence typing (MLST) by Kidgell *et al.* (141). We further confirmed the identity of Tp2 by sequencing three (*hemD*, *hisD* and *thrA*) MLST genes previously shown to vary among the Typhi isolates studied by Kidgell *et al.* (141). The other SARB strains were selected because they have the least allelic differences to the Tp2 according to MLEE data (33) or because

they cause enteric fever in humans. The identity of all other strains used in this study was confirmed by PCR serogrouping (117, 180), targeting the O antigen gene clusters. Strain Pc4 was purified from the original stock which was contaminated with other *S. enterica* strains. Chromosomal DNA was prepared using the phenol/chloroform precipitation method.

3.2.2. Gene fragments and primer sequences

Four genes were selected on the basis that they are unlikely to be under selection pressure. The genes used were, *mglA* (galactoside transport ATP-binding protein MglA), *proV* (glycine betaine/L-proline transport ATP-binding protein), *speC* (ornithine decarboxylase) and *torC* (cytochrome C-type protein) (231). These genes are functional in Typhimurium LT2 but pseudogenes in Typhi CT18. The primer pairs used in this study were designed based on the Typhi CT18 genome sequence (Table 3.2-1) and synthesised commercially (Sigma-Aldrich).

Table 3.2-1. Genes and Primers used in this study

Gene	Chromosomal location (bp)[1]	Direction	Size (bp)	Sequence 5'-3'	Position[2]
torC	3,824,187	Forward	18	GGTCATTGTCGGGATTGT	60-77
		Reverse	18	TCCGTCCAGCCTTCGATT	728-745
speC	3,129,168	Forward	17	AAAATCGGGCATCTCTG	892-909
		Reverse	17	CGCCTCGCTGATACGCA	1876-1893
proV	2,808,847	Forward	18	GGCTCGGGTAAATCCACA	190-207
		Reverse	18	TTCATCGACAACCGGCAC	1090-1107
mglA	2,251,323	Forward	20	GTCTTTTCGGTATTTATCAA	173-190
		Reverse	18	AATAGCCAGCGACCAATG	1229-1246

[1] According to Typhi CT18 genome (Accession No: NC_003198)

[2] Relative to the first base of the initiation codon.

3.2.3. PCR assay and DNA sequencing

Each genes were sequenced for both forward and reverse directions. Some of the PCR assay and sequencing works were done as completion for an Honours degree at the University of New South Wales. The sequences reported in this paper have been deposited in the GenBank database (Accession Nos DQ285482-DQ285541).

3.2.4. Bioinformatic Analysis

CONSED version 8.0 (94) program package accessed through the Australian National Genomic Information Service was utilised for sequence editing. PILEUP from the GCG package (60) and MULTICOMP (263) were used for multiple sequence alignment and comparison. PHYLIP (78) was used to generate phylogenetic trees and bootstrap values. SPLITTREE version 3.2 (14) was used to create network structures using the distance method. Overall compatibility of informative sites was measured by using the RETICULATE program (127), which gives a measure of phylogenetic concordance between two sites with values ranging from 0% (fully incompatible) to 100% (fully compatible). This method was used to obtain a measure of recombination within and between loci and for comparison with other datasets. Maximum Likelihood (ML) analysis of the congruence of gene trees as described by Feil *et al.* (72, 75) was done using PAUP version 4.0 beta (304) with the parameters of the HKY85 model of DNA substitutions, estimation of transitions to transversion ratio (Ti/Tv) and α parameter assuming gamma distribution. ML generates scores for comparison of one gene tree against another based on the 99^{th} percentile of the distribution of scores for 200 trees from random topology. Two gene trees are considered to be significantly congruent if the difference between the likelihood scores of the trees of the two genes (Δ-lnL) is lower than that any of the 200 random trees, as the second gene tree should be of better fit to the data from first gene than the 200 random trees (72, 75). Calculation of the linkage disequilibrium index (I_A) (192) from MLEE data was done using an in-house program MLEECOMP (249).

3.3. Results

3.3.1. Sequence variation in the four genes

The 15 SARB strains were sequenced for the four genes, *mglA*, *proV*, *speC* and *torC*. The total length of sequences obtained was 2985 bp with 743 bp, 818 bp, 820 bp and 604 bp for *mglA*, *proV*, *speC* and *torC* respectively. The average pairwise percentage difference for all genes and strains was 1.06 (Table 3.3-1). A total of 133 sites were polymorphic (sites at which more than one type of nucleotides exists) but only 66 were parsimony informative (at least two types of nucleotides at the site, each represented in at least two of the sequences), with 19, 12, 24 and 11 sites for *mglA*, *proV*, *speC* and *torC* respectively. Sequence data of two genes, *mutS* and *mdh*, available from the Brown *et al.* (38) study for the same SARB strains used in this study, were included for comparison (Table 3.3-1) and subsequent analyses.

Table 3.3-1. **Pairwise nucleotide difference**

Gene	Size (bp)	Pairwise Percentage Difference			No. Polymorphic Sites	No. Informative Sites
		Average	Minimum	Maximum		
mglA	743	1.35	0.27	2.02	37	19
proV	818	0.78	0.12	1.59	26	12
speC	820	1.27	0.00	2.32	49	24
torC	604	0.83	0.17	1.82	21	11
mdh[1]	831	0.81	0.00	1.56	24	13
mutS[1]	1098	0.81	0.09	1.28	35	21

[1] Genes sequenced by Brown *et al.* (38).

Comparison of Typhi Tp2 with the two genome sequence strains, CT18 and Ty2 (57), reveal that Tp2 was identical to Ty2 in all four genes sequenced but differed from CT18 by one base in *torC*. While most SARB strains had functionally intact sequences, four cases of gene inactivation were observed. Two strains, Pc2 and Pc4 had the same deletion as Typhi strains of a CG repeat in *mglA*. The changes in these two strains must be independent as their sequences were very different from each other. Strain Pc2 also had a C to T substitution

forming a stop codon in *torC*. Strain Pa1 had a substitution from base C to T leading to a stop codon in *mglA*.

The sequence alignment for informative sites is shown in Table 3.3-2. It was clear that no strains were consistently similar in all six genes. Only two pairs of strains, Pn1 and Mo1, and Sw1 and Pc4, shared similarity in two or more genes. Mo1 and Pn1 shared similarity over the entire sequence in three genes (*proV*, *torC* and *mdh*) but only parts of the sequences in another two genes (*speC* and *mutS*). Sw1 and Pc4 had almost identical sequences in two genes (*proV* and *mdh*) but only some segments similar in another three genes (*mglA*, *speC*, and *mutS*).

	mglA	***proV***	***speC***	***torC***	***mdh***	***mutS***
			11111111111111111111111			11111111111122222222
	23334444445556677888	223444558889	02222233444555566666677	22333555666	234446666677	456778888889902333444
	90330224723706560008	665258275581	71112528158115663577892 9	03135012689	7170461346759	625171118898953079567
	43168891458033924 88	288774340936	12584563355257235614988 1	71678510461	8928728059265	149632351769885497768
Consensus	**CACAGCGGCATCCCCTTCT**	**TCGCGTGGTCCT**	**TCCGTACGGGTCCTGGGGCCAGGG**	**CCCGCGCCCCA**	**CCTTTATATAATC**	**GTGGATCGGGCCTCGGAACGT**
De1	T.T.A..A...GTTTCA..	.TAA.C....T.	ATTA....A....C..........	T...T.TT...	.T.C.........	A...G.......C.A......
Pa1	T.T....A....T..CA..	..A..C.....C	.TTA.G.....TT..A.A...A..	..T........	...C..CG.GGC.	.A...C.....TC........
Sw1	...G.....G.TA..CC..	....A.A....C	..............A....T...T	.T..T....T.	T.C.CG......T	.A...C...A.......G.AC
Pc4	.G...--..G.TA..CC..	....A..A....	......TT......A....T....	.....A.T.T.	T.C.CG.......	.A...C...A........G..
CT18	...G.--.T..........	.T........T.	.TTA.G.....TT....A...A..	.....A....G	...C.GCG...C.	...A....A....G......C
Sf1	T.T....A..CG.......	.T...C.....C	A...C....A...C..........	T.....TT...	..C.CG.......	.....C..A........G...
In1	T.T. ...T.........C	C....C...AT.	A...C.............T..AA.	..........G	....G..GA....	..A..C.........A.....
Pb1	TG.G....T..........	C....C....T.	A............CA.A.......	....	.TC..........	......TAA.T.C..A.G...
Pb7	...G....T..........	..........TC	....C...A......A....GA..	.....A.T...	....	...A....A.......G....
Pn1	...G........TT.CAAC	....AC.A....	......TT......A....T...T	T.TA...TT.G	T.C.CG......T	..A.....A.........G..
Mo1	....A.AA........C..	....A..A....	..............A....T....	T..A...TT.G	..C.CG......T	AA...C...A........G..
Cs6	....A.AA....A...AAC	..AA....CATC	A................AT..AA.	.T.........	..C....TA....	...A.......TC....G.AC
Tm1	....A.AA...GTTTCC..	.....C..CA.C	.TTA.G...AC....A....GA..	.....A.....	....GGCG.GGC.	......TAA.T.C...G....
Np8	..T....A..C........	C.....A....C	ATTA.G...AC....A....GA..	..........G	.T....CG.....	.A..G.......CGA......
Pc2	.....--A..C....CAA.	C........AT.	A............CA.A.......	T.........G	...C..CG.....	..A.G.......C.......C

Table 3.3-2. Informative sites of the four genes sequenced in this study and two additional genes (*mdh* and *mutS*) from the study by Brown *et al.* (2003). The numbers on the top of the alignment, reading vertically, are base positions.

3.3.2. Phylogenetic relationships

Evolutionary trees were constructed by the neighbour-joining (NJ) method for each of the six genes and as well as the concatenated sequences for all six genes (Figure 3.3-1). The individual gene trees did not resemble one another in their topologies, and inconsistent clustering of strains could be observed. However, there were three cases where strains were grouped closely together in two or more genes: In1 with Np8 in *mglA* and *torC*, Pn1 and Mo1 in *torC* and *mdh*, and Mo1 and Pc4 in *proV*, *mdh* and *mutS*. The groupings of Pn1, Mo1 and Pc4 were also apparent in the combined tree while In1 and Np8 appeared to be on separate clusters. Most of the branching orders were poorly supported statistically as bootstrap values including those for the combined six-gene tree were low. We believe that this was due to conflicting signals resulting from recombination, which will be discussed later and not because of the low phylogenetic signal in our data. The combined six-genes tree was then compared with an MLEE tree using data from Boyd *et al* (33) which was reconstructed to include only the strains used in this study. Except for three strains, Sw1, Pn1, and Pc4, broadly falling within the same cluster, other strains were inconsistently clustered in the two trees.

Split decomposition (14) was then used to visualise the relationship of the strains. The method displays conflicting phylogenetic signals resulted from recombination as network structures. As shown in Figure 3.3-1, using concatenated six-gene sequences, the relationships of five strains, Mo1, Pc4, Pn1, Sw1 and Cs6, were resolved with network structures. However, other strains showed a star phylogeny radiating from the same central point. This suggested that recombination is extensive and the strain relationships were not well represented by a splits graph.

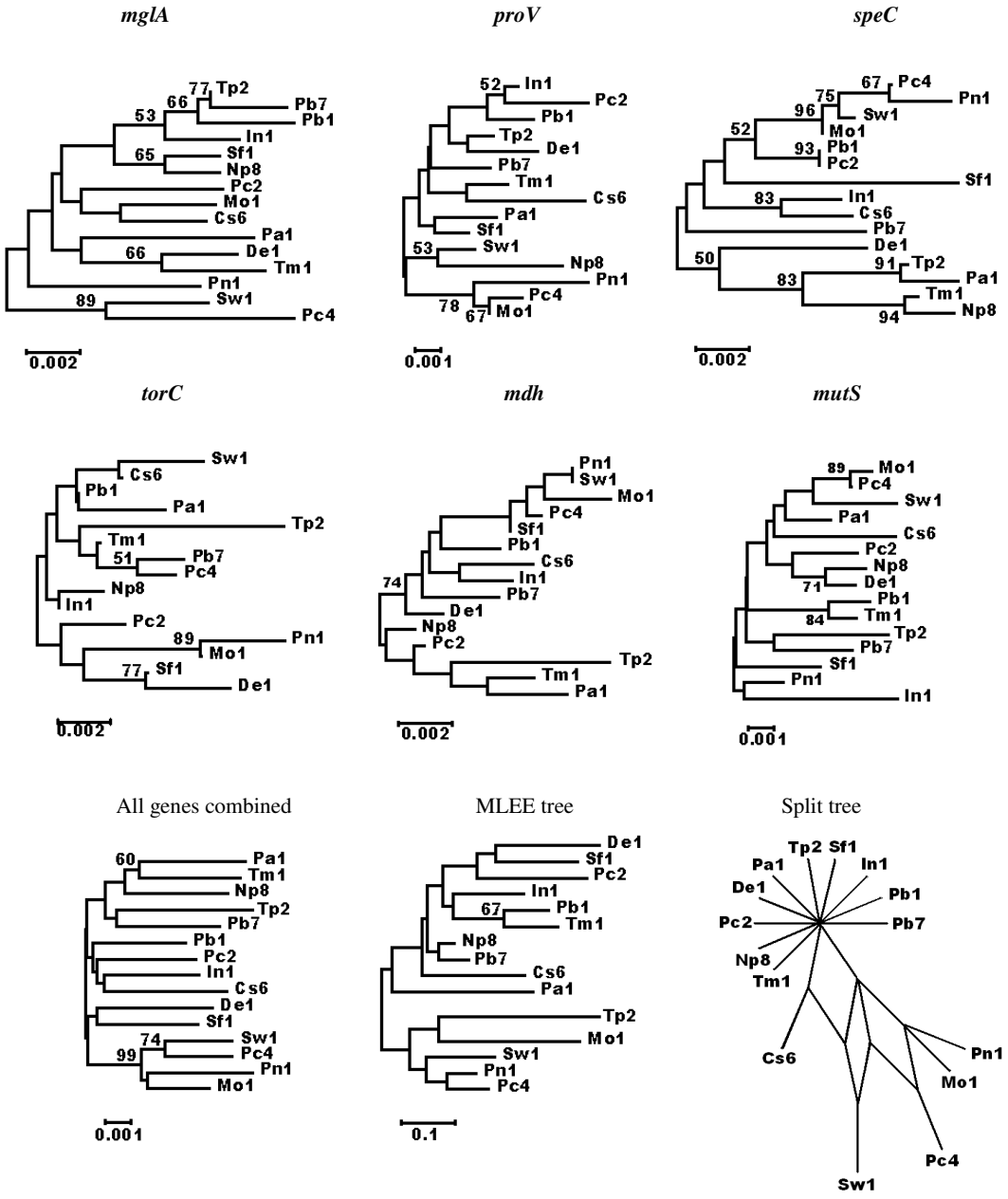


Figure 3.3-1. Phylogenetic trees. Shown are neighbour-joining trees of the individual genes, the concatenated sequences of six genes, and MLEE data; and the Splits tree of the concatenated sequences of six genes. Bootstrap values, if greater than 50%, are presented at nodes of the neighbour joining trees.

3.3.3. Congruence analysis

To establish the degree of incongruence among the six genes trees, ML analysis (72, 75) was carried out and the results are summarised in Table 3.3-3. None of the six gene trees was congruent to all the other gene trees. Gene trees with the largest number of congruencies were those of *mdh* and *mutS* which were congruent to three other gene trees. *proV* and *torC* trees were only congruent to two other gene trees while the *speC* tree was only to one other gene tree. The *mglA* tree was congruent to none of the other gene trees. Overall only 37% of the gene tree comparisons were congruent among the SARB strains. To compare between subspecies data, we also analysed the six house keeping gene trees (*gapA*, *icd*, *mdh*, *putP*, *gnd* and *aceK*) of SARC strains sequenced by Selander's group (31, 34, 214-217, 333). All the gene trees were congruent to each other (data not shown).

Table 3.3-3. Maximum likelihood analysis for congruence between each gene tree of the SARB strains analysed in this study

Gene	α[1]	Ti/Tv[2] ratio	-lnL	-lnL (99th) [3]	Δ-lnL score of gene tree[4]					
					mglA	***proV***	***speC***	***torC***	***mdh***	***mutS***
mglA	0.61	4.42	1416	28		46	44	38	45	42
proV	0.02	5.09	1389	25	40		35	**25**	**25**	30
spec	0.69	4.80	1482	64	110	72		89	**44**	77
torC	0.29	14.03	1016	30	48	**27**	34		**28**	31
mdh	0.49	5.63	1386	39	62	**36**	**23**	47		**38**
mutS	0.02	5.74	1876	46	65	**33**	**36**	48	**34**	

[1] Nucleotide substitution rate variation between sites with gamma distribution as the parameter.

[2] Estimated transition/transversion ratio.

[3] Difference in –lnL score from –lnL column (reference data) and the 99th percentile from random topology.

[4] Difference in –lnL score from reference data to each calculated data from other genes. Trees are deemed as congruent if Δ-lnL is equal to or lower than the 99th percentile of random trees when compared to reference data (72, 75). Congruent gene trees are highlighted bold.

3.3.4. Compatibility analysis

We further assessed the level of recombination in *S. enterica* subspecies I by compatibility analysis of the six genes using the program RETICULATE developed by Jakobsen and Eastal

(127). We calculated compatibility values both within a gene and between genes (Figure 3.3-2). *mglA* has the lowest average within locus compatibility, at only 52%, followed by *proV* (53%), *mutS* (65%), *torC* (67%), *mdh* (73%) and *speC* (78%); while for between loci comparison, *torC* and *speC*, both at 53%, are the most compatible followed by *mdh* (51%), *mutS* (50%), *proV* (47%) and *mglA* (40%).

We compared the within-subspecies I values from this study with those between *S. enterica* subspecies calculated using data of the six housekeeping genes, *gapA*, *icd*, *mdh*, *putP*, *gnd* and *aceK* from the 16 SARC strains sequenced by Selander's group (31, 214, 216, 217, 286, 333). As shown in Figure 3.3-2, the compatibility values were much higher for between subspecies than within subspecies I. We also compared these values with those of the closely related species *E. coli,* using data from Reid *et al.* (264) for seven housekeeping genes from 14 strains representing common clones of pathogenic *E. coli.* The *S. enterica* subspecies I values were lower than those for *E. coli* (Figure 3.3-2).

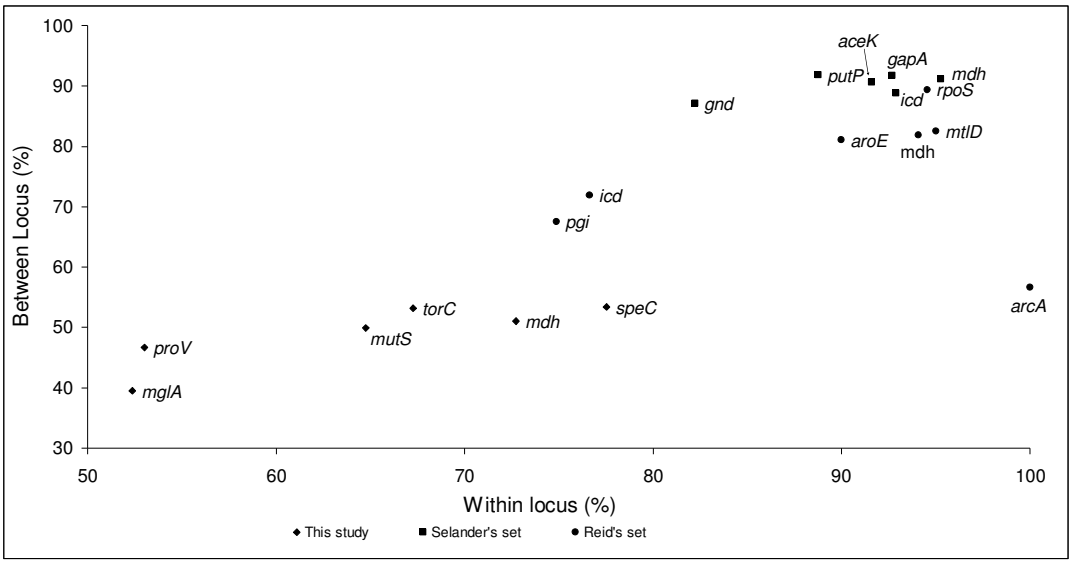


Figure 3.3-2. Comparison of average compatibility values within and between loci of *S. enterica* and *E. coli* strains. Selander's set is the sequence data for six housekeeping genes of 16 *S. enterica* strains representing different subspecies from Selander's group (31, 34, 214-217, 333). Reid's set is the sequence data for seven housekeeping genes of 14 common pathogenic *E. coli* strains from Reid *et al.* (264).

3.4. Discussion

3.4.1. Recombination and clonality within S. enterica subspecies I

This study examined six genes from 15 SARB strains of subspecies I and found that recombination is frequent. ML analysis showed incongruence between the six gene trees studied. The incongruence was less likely to be a result of low sequence variation, as ML analysis was relatively insensitive to sequence variation and has been applied to other species with comparable low level of variation (72). Compatibility analysis was consistent with ML analysis showing that recombination occured frequently within *S. enterica* subspecies I. Altogether, the results suggested that the level of clonality within subspecies I is low.

It was apparent from our analysis that there were different levels of clonality within *S. enterica* species. Comparison of our data with that from the SARC set representing the eight different subspecies by both ML and compatibility analyses showed that recombination occured far more frequently within *S. enterica* subspecies I than between *S. enterica* subspecies. This situation was rather similar to the case of *Rhizobium melitoti* (192). There were two major divisions of the species where recombination was rare between the divisions but common within (192).

The level of recombination in *S. enterica* subspecies I could be compared with that in other species. By compatibility analysis, we have shown above that the frequency of recombination in subspecies I was higher than in *E. coli*. The ML analysis allows comparisons with a number of species to which the method has been applied (72, 75). The percentage of gene tree comparisons which are congruent is 88, 75, 55, and 7 for *E. coli* (75), *Haemophilus influenzae* (75), *Staphylococcus aureus* (72), and *Neisseria meningitidis* (75) respectively. In our study, 37% of the comparisons were congruent among the SARB strains (Table 3.3-3). Therefore, the level of clonality of subspecies I was on the lower end of the spectrum in comparison to the other four species.

It was interesting to note that Brown *et al.* (38) recently reported that *mutS,* a gene involved in mismatch repair and a strong mutator, undergoes frequent recombination in SARB strains in comparison to SARC strains. The study used *mdh* for comparison and attributed

incongruence of the two gene trees to recombination in *mutS*. Using only two genes, in fact, one could not determine which gene has undergone recombination. *mdh* was regarded as "non-recombinant", under the assumption that it did not undergo recombination in this "highly clonal" species. From the comparison of the six gene trees, however, *mdh* has undergone frequent recombination, and so does *mutS*. Interestingly, the *mutS* tree was not the most incongruent to the other gene trees as suggested otherwise because of its potential mutator properties. It seemed that *mutS* may not be particularly more recombinogenic although the implication is not yet clear.

3.4.2. Reexamination of the MLEE data uncovers the myth of high clonality at all levels in S. enterica

The results from this study drawed a sharp contrast to the long held view that *S. enterica* has a highly clonal population structure (192, 283). In the landmark paper on bacterial population structures by Maynard Smith *et al.* (192) which, for the first time, ranked the recombination rate and hence level of clonality of different species, *S. enterica* was found to be clonal at all levels, from individual serovars to the species as a whole. The study used the Index of Association (I_A) to measure the extent of linkage disequilibrium from MLEE data. I_A values for *S. enterica* were significantly greater than zero. The sequence data seemed to be in conflict with the MLEE data.

We looked into the MLEE data to seek an explanation. We first checked whether the discrepancy came from the use of the I_A index as a relative measure of clonality. The MLEE data for *S. enterica* used by Maynard Smith *et al.* (192) was in fact data for 14 serovars of subspecies I, originated from Selander *et al.* (285). The dataset was thus representing subspecies I rather than the whole species. We obtained the MLEE data for 80 ETs that represent all eight subspecies from Boyd *et al.* (34). The I_A for the 80 ETs is 3.219±-0.156, which is almost two and half times more than that for the subspecies I data of 106 ETs (1.393 ± 0.135). Note that the number of enzymes used was the same in the two datasets eliminating its effect on the scale of the I_A. The difference in I_A for data between subspecies I and the whole species seemed to reflect their difference in the level of clonality and was consistent with the sequence data.

We further examined the subspecies I MLEE data, from Selander *et al.* (285). We tested whether removing closely related ETs, which potentially corresponded to clonal complexes, affected the I_A. We used eBURST (73) to identify closely related ETs and one ET was selected to represent each cluster. When ETs differing by 1, 2 and 3 loci (out of 24) were removed successively, the I_A values dropped progressively from 0.783 ± 0.223 to 0.289 ± 0.296, and then to 0.036 ± 0.371. Thus, clonal structure disappeared when closely related ETs were treated as a unit. This change in I_A resembled that which occured in an organism with an epidemic population structure such as *N. meningitidis* (192). This analysis showed that the subspecies I MLEE data gave no support to a strongly clonal population structure for *S. enterica.*

The conclusion reached by Maynard Smith *et al.* (192) that *S. enterica* was clonal at all levels was based on that the I_A values for individual serovars are equal to or higher than that for the whole data set (see Table 1 of their study). Their interpretation assumed that a serovar represents a real genetic group. However, a number of serovars including two, Paratyphi C and Choleraesuis, used in Maynard Smith *et al.* (192) are known to be not clustered together with a single origin (285). We suspected this may have contributed to the high I_A values and examined the data for Paratyphi C and Choleraesuis from Selander *et al.* (285). When we excluded the divergent Pc4 from the Paratyphi C data, the I_A value dropped from 4.157 ± 0.465 to -0.444 ± 0.459. Note that we considered only ETs for the I_A calculations. Similarly when we took out the two divergent isolates (Cs6 and Cs13) from the Cholerasuis data, the I_A dropped from 1.432 ± 0.419 to -0.313 ± 0.455. Taking out other isolates only slightly altered the I_A. Therefore, treating a heterogeneous serovar as a single population artificially inflated the I_A, which led to the misinterpretation of high clonality at the serovar level, at least, in these two cases.

In a series of studies of the SARC set using six housekeeping genes (31, 34, 214-217, 333), it has been shown that in all six gene trees, the two isolates for each of the eight subspecies are consistently grouped, although some recombination is detectable (37). Additionally, in most cases the branching patterns among the subspecies were also consistent, suggesting only low levels of recombination. The sequence data were interpreted as a strong support of the MLEE data that *S. enterica* is highly clonal but, as one could now see, wrongly reinforced the misinterpretation of the I_A analysis from the subspecies I MLEE data. Furthermore, the

housekeeping gene studies have only used two isolates to represent a subspecies, which would not allow identification of recombination events within a subspecies.

3.4.3. Predominance of intra-subspecies recombinational exchange

Based on the level of variation in the six genes, it seemed that recombinational exchange occurred only within subspecies I. Among the SARB strains the level of sequence divergence in the six genes had a maximum of 2.32%. In contrast, sequence divergence between subspecies based on the data of the six housekeeping genes, *aceK*, *gapA*, *icd*, *mdh*, *putP* and *gnd*, of the 16 SARC strains (31, 34, 214-217, 333) averaged 5.69%; with divergence between subspecies I and the other subspecies ranging from 2.71% to 10.07%. We further compared levels of divergence of *mdh* and *mut*S between SARC and SARB strains sequenced by Brown *et al.* (38). For *mdh*, the difference between strains of subspecies I and the other subspecies ranges from 2.31% to 8.66%. In contrast, *mdh* from the 15 SARB strains of this study has an average of 0.81% and the maximum of 1.56%. Similarly for *mut*S, no strain within subspecies I have a level of difference equal to or higher than that between subspecies. Predominance of intra-subspecies recombination may be a result of a number of factors. MutS creates a barrier for recombination of divergent DNA (38, 253, 259, 330, 349) which may block recombination with other subspecies. There could also be a niche barrier (190). *S. enterica* strains of subspecies I usually share the common niche, the warm blooded animals, while other subspecies are commonly isolated from reptiles. It remains to be determined from data for other subspecies whether niche is a significant barrier to recombination in *S. enterica*.

3.4.4. Relationships of subspecies I isolates

Typhi has been shown to be a homogeneous clone and was suggested to have arisen about 50,000 years ago (141). We initially wished to determine which SARB strain is the closest relative of Typhi using sequence data. In the MLEE tree of 72 SARB strains, the two Typhi strains, Tp1 and Tp2, were grouped together and clustered with Derby De1 (33). However, Typhi Tp2 differed from the other serovars by at least nine of the 24 enzyme loci studied by

Boyd *et al.* (33), with the least allelic difference to Pc4, rather than to De1 of 17 differences. The neighbour joining (NJ) trees (Figure 3.3-1) showed that Typhi is placed inconsistently in the six gene trees. Typhi was clustered together with Pb7 in *mglA* and *mutS*, and with Pa1 in *speC*. This relationship was also reflected in their near identical sequences in *mglA* and *speC* (Table 3.3-2). In *proV*, Typhi was clustered with De1 although it had a higher sequence similarity to Pb7. Typhi was not closely clustered with any other strain in *torC* and *mdh.* Thus there was no clear indication of the closest relative of the Typhi clone in the 15 SARB strains analysed.

Spotted DNA microarray study using Typhimurium LT2 suggests that Typhi is most closely related to serovars Paratyphi A and Sendai (43). Full genome comparison of serovar Paratyphi A ATCC 9150 indicates that it shares more gene contents to Typhi CT18 than to Typhimurium LT2 (193). In contrast, MLST has shown that Typhi was the most distinct serovar which did not have identical alleles to the other subspecies I strains in all seven loci analysed. Falush *et al.* (67) developed computer modelling to visualise the patterns of divergence and genetic exchange in Typhi using the MLST data (141). The assumption that Typhi is the most divergent is debatable as a large genetic distance between Typhi and the other serovar was resulted from *purE*. It is most likely that this gene was imported into Typhi from another species although the sequence is most closely related to subspecies VI (67). This provides further evidence that considerable recombination within *S. enterica* species exists. A new question is raised whether genetic exchange plays a role in driving the genetic diversity within *S. enterica* species, but the genetic barrier controlling the exchange has not yet fully developed. Nevertheless, frequent recombination observed in the analysed genes has impeded the attempt to determine which serovar, among 12 serovars of *S. enterica* subspecies I analysed that is closely related to Typhi.

For the 15 SARB strains studied, the only strains that appeared to have a clear relationship were the four more closely related strains, Sw1, Pc4, Pn1 and Mo1. Both the splits tree and the combined six-gene NJ tree showed that the four strains form one group, where Sw1 and Pc4 appeared to be more closely related**.** High level of recombination appeared to have eliminated most of the phylogenetic signals from the gene trees. This was also evident in the bootstrap values which were low in most of the interior branches for the combined six-gene NJ tree.

The MLEE tree of SARB has been widely used to represent the strain phylogeny of these strains (229, 313). Although we could only make a comparison of 15 of the 72 strains, no consistency of clustering of strains were observed between the MLEE tree and the combined sequence tree (Figure 3.3-1), suggesting that the MLEE tree does not necessarily represent the true phylogeny of the strains and its use for mapping and inferring genetic events may not be warranted.

3.4.5. Whole genome sequences should be used to infer phylogenetic relationships

Present study has used multiple housekeeping gene sequences to analyse the genetic relatedness of strains belonging to subspecies I. However, the number of genes and the size of the fragments of these genes which are needed to provide the most accurate phylogenetic relationships have not been thoroughly explored. Examining more loci will presumably increase the ability to resolve the relationships of serovars belonging to subspecies I. Unfortunately, it is still imperative, which of the genes could be used to accurately determine the evolutionary relationships, due to the presence of frequent recombination within this subspecies. Currently there are 12 genomes for serovars belonging to subspecies I which are either fully sequenced or almost completed (www.salmonella.org and www.sanger.ac.uk). Comparisons of these genomes will reveal the degree of nucleotide divergence allowing the extent of recombination to be determined.

Vernikos *et al.* (323) constructed a phylogenetic tree based on the whole-genome sequence based alignment of 11 *S. enterica* strains. These strains include five serovars of subspecies I: two strains of serovar Typhi; two of serovar Paratyphi A; three of serovar Typhimurium; and one strain of each serovar Enteritidis and Gallinarum respectively and one strain of each *S. arizonae* and *S. bongori* representing subspecies IIIa and V respectively. Three *E. coli* strains and one *S. flexneri* strain were included as outgroups. The distribution of putative horizontally acquired (PHA) genes was compared between these serovars and the relative time of insertion of these PHA was determined. A large number of 434 PHA genes separated *E. coli* from *S. enterica* lineage suggesting the role of horizontal gene transfer on the evolution of *S. enterica* (326). All strains belonging to subspecies I were grouped separately from the strains of subspecies IIIa and V.

The typhoidal serovars including serovar Typhi and Paratyphi A respectively were shown to be located on the same node but were separately placed to the non typhoidal serovars including serovar Typhimurium, Enteritidis and Gallinarum respectively. Serovar Typhi and Paratyphi A were differed by 484 PHA, belonging to either prophage structures or of phage origin, where 24% of the genes have unknown function; 26% are related to cell surface components; 11% are pseudogenes; and 24% are associated with pathogenesis (326).

Despite serovars Typhi and Paratyphi A have been shown to be most closely related, there are approximately 1,500 *S.enterica* serovars belonging to subspecies I and only five serovars were analysed in that study (326). Comparison to other serovars from subspecies I is needed to determine if there is any serovar other than serovar Paratyphi A that is most closely related to Typhi. However, approximately 30% and 25% of protein coding sequences in Typhi CT18 Paratyphi A strain ATCC 9150 respectively were horizontally acquired (326), further supporting that different regions within the genomes have different evolutionary histories. These genes show a low G+C content of 43.3% suggesting that they have been acquired from distantly related genomes (326). By using this knowledge, the phylogenetic trees generated need to be established using genes that could reliably portray the evolutionary history of serovar Typhi without being complicated by frequent homologous recombination or horizontal gene transfer.

3.5. Conclusion

This study has contributed to a better understanding of the population structure of *S. enterica*. Previous sequence studies using strains of different subspecies (286) showed largely congruent gene trees, leading to the general conclusion that *S. enterica* is highly clonal. In contrast, using SARB strains of subspecies I, this study suggested that recombination has occurred at a frequency sufficiently high to have eliminated much of the phylogenetic signals. Statistical analyses using compatibility, split decomposition and maximum likelihood provided further evidence that recombination is frequent in *S. enterica* subspecies I. These findings revealed that the clonality of *S. enterica* varies within the species. Further studies will be required to quantify recombination and mutation parameters in subspecies I and to ascertain these parameters in the other subspecies.

Housekeeping genes are not expected to be under positive selection pressure. Recombination in these genes occurs as a result of population dynamics of the serovars belonging to *S. enterica* subspecies I. Thus, the recombination events seen are due to random genetic drift and are likely to be neutral mutations. The observation of high level of recombination within subspecies I stipulates further work on the evolution of *S. enterica* clones. Nearly 1,500 serovars in subspecies I have been reported, comprising 60% of known *S. enterica* serovars (240). Only the top 2% subspecies I serovars have been studied at population genetic level, largely by MLEE, which provided a limited picture of evolutionary origins of the specialised (eg host adapted) clones and the diversity of the subspecies. The findings of frequent recombination from this study have now blurred that picture as the MLEE relationships between more distantly related ETs are no longer considered reliable. It is much needed to determine the relationships, at sequence level, of clones encompassing the whole subspecies in addition to those frequently encountered in causing human or domestic animal infections, which will provide a better framework within which to study the evolution of pathogenicity and host adaptation.

Chapter 4: Single nucleotide typing polymorphism of *S. enterica* serovar Typhi

4.1 Introduction

Typhoid fever, a serious systemic disease, is caused by *Salmonella enterica* serovar Typhi (or simply Typhi in this paper) and is endemic in countries where hygiene and sanitation still remain as unresolved problems. Annually there are more than 17 million cases of typhoid fever, with around 600,000 deaths (340). Genetic diversity within Typhi has been studied extensively using molecular techniques such as pulse-field gel electrophoresis (PFGE) (145, 151, 209, 309, 310), ribotyping (79, 166, 219), IS200 typing (311), amplified fragment length polymorphism (AFLP) (210) and random amplification of polymorphic DNA (RAPD) (252, 287). However, a major drawback of these techniques is that the relationships derived from such data do not necessarily reflect true evolutionary relationships of the isolates.

Two population genetic studies showed that Typhi is a highly homogeneous clone. Multilocus Enzyme Electrophoresis (MLEE) of 24 metabolic enzymes revealed only two major electrophoretic types (ETs) (261) and Multilocus Sequence Typing (MLST) of seven housekeeping genes found only 3 base substitutions in 3336 bp analysed and divided 26 Typhi isolates into four sequence types (STs) (141). Therefore, there is insufficient variation for either MLEE or MLST to be useful for determination of relationships among isolates or for global epidemiological studies.

To facilitate global epidemiology study and to establish the evolutionary relationships within the Typhi clone, there is a need for a molecular method that is inexpensive, discriminative, simple and reproducible for large scale typing of isolates. Single Nucleotide Polymorphisms (SNPs) are potential markers and have been used to type several pathogens including *Escherichia coli* O157:H7 (356), *Bacillus anthracis* (236) *Mycobacterium tuberculosis* (81, 97) and *Yersinia pestis* (2). The discovery of SNPs is facilitated by the sequencing of more than one genome of the same clone. The two completed Typhi genomes of strains CT18 and Ty2 (57, 231) allowed us to explore the differences between them and to identify SNPs suitable for typing. We selected 37 SNPs that could be differentiated by the presence or

absence of a restriction enzyme site to analyse a collection of worldwide Typhi isolates and showed that SNP typing is a good tool for genotyping and determining evolutionary relationships of global Typhi isolates.

4.2. Materials and Methods

4.2.1. Bacterial isolates

Seventy-three worldwide serovar Typhi isolates, differing in localities and years of isolation, were obtained from the *Salmonella* Genetic Stock Centre, University of Calgary, Calgary, Canada, and one isolate was obtained from Imperial College London, London, United Kingdom (Chapter 2, Table 2.1-2).

4.2.2. Genomic Analyses

A pairwise gene comparison of Typhi CT18 (Accession No: AL513382) and Ty2 (Accession No: AE014613) to identify polymorphic genes between the two genomes was performed using BLAST tools available in Australian National Genetic Information Service (ANGIS). The full genome sequences are available from NCBI website (www.ncbi.nlm.nih.gov) and the files which are in genebank (.gbk) format were downloaded. Genebank files contain all information of the ORFs from the genome sequences. The ORFs were then separated into each file where each file contains each individual gene. A database containing individual ORFs of Ty2 was created to compare them to the ORFs of CT18 using the blastn program. Individual pairwise comparisons between the two Typhi strains were collected in each multicomp (.multi) files. ORFs which are specific to either of Ty2 or CT18 were eliminated from analyses.

Fully sequenced genomes from eight strains of other *S. enterica* serovar, seven of which were of subspecies I: Choleraesuis strain SC-B67 (GenBank Accession No: AE017220); Paratyphi A strain SARB42 (Accession No: CP000026); Typhimurium strain LT2 (Accession No: AE006468); Paratyphi B strain SPB7 (ftp://genome.wustl.edu/pub/seqmgr/bacterial/salmonella); Enteritidis strains PT4 and

Gallinarum strain 287/91 (ftp://sanger.ac.uk/pub/pathogens/Salmonella); and Pullorum (http://www.salmonella.org) and *S. enterica* subspecies V strain 12149 (ftp://ftp.sanger.ac.uk/pub/pathogens/Salmonella), were included for analyses to deduce the likely ancestral base of each SNP in Typhi and these strains were also used as outgroups for phylogenetic analysis. The relationships of the isolates were determined using PAUP (304) to construct a maximum parsimony tree from the SNP data and using Arlequin v 3.1 (66) to generate a minimum spanning tree (MST). Parsimony uses characters-based data to create a matrix for calculating the best estimations of phylogenetic relationships between the taxa included (77) while MST connects taxa based on minimum pairwise differences (66).

4.2.3. Selection of SNPs and design of primers

The sequences of the genes found to be polymorphic from comparison of the two genomes was concatenated and screened for restriction enzyme (RE) cutting sites using NIP from staden package (28) accessible from ANGIS. Both 6-base and 4-base REs with a maximum cost of AU$0.25 per enzyme was searched for suitable SNPs. Primers were designed to amplify genes encompassing the SNPs selected based on Typhi CT18 genome sequence (231) (Table 4.2-1) and synthesised through Sigma-Aldrich. The PCR amplicons ranged from 100 to 890 bp. The size difference between undigested and digested fragments was at least 50 bp to allow easy detection on an agarose gel. If a polymorphism is too close to either end of the targeted gene, the primer was designed using the sequence of its neighbouring gene.

4.2.4. PCR, restriction enzyme digestion and DNA sequencing

Genes containing the SNPs were PCR amplified and subsequently, the PCR product (15 µl) was digested with 1 U restriction enzyme RE, at 37^{o}C for 2 hr followed by gel electrophoresis on a 2% agarose in TBE buffer (274).

Table 4.2-1. List of primers used for SNP typing

SNP No.	Gene Name	Encoding	Positions of chromosome (bp)		Primer Pair	Position[1]	Sequences 5' -> 3'	Expected Fragments Size (bp)		Enzyme
			Start	Finish				CT18	Ty2	
1	*bcfD*	fimbrial subunit	28429	29436	9217	35[a]	GGAGTGTCGTATCGCAGT	609	256/344	*hae*III
					9218	644	GGTTGAAATTGCCGCTGT			
2	*leuO*	probable activator protein in leuABCD operon	135983	136927	9175	704	CGTGGCTTATCAGGGCAT	670	37/633	*hha*I
					9176	1315	ACACCTGCTTTACGCCTT			
3	*secA*	preprotein translocase SecA subunit	160107	162812	9150	20	CCAAAGTATTCGGTAGCC	166/49	215	*alu*I
					9151	235	GAAGTGACGCATCCCAAA			
4	*gcd*	glucose dehydrogenase	200408	202798	9152	698	CTCACCAGCGTCTGTTCG	107/45	152	*alu*I
					9153	867	TTGACCGGGAGGATAATG			
5	*malZ*	maltodextrin glucosidase	448758	450134	9117	425	ATCAGGACCGAGTGTATT	164	112/52	*hha*I
					9118	589	TACCGGATTCAGATATAG			
6	*STY0568*	probable metabolite transport protein	574305	575552	9119	23	GGATTGATTACTGGAAAC	294	138/156	*hha*I
					9120	317	TTAGCGACCCTGTATAGA			
7	*fimI*	fimbrin-like protein FimI	596706	597239	9227	204	GGCGATTGGTGATACGAC	564/60/6	624/6	*alu*I
					9228	834	AAATGGAACGCTGACGGG			
8	*sdhB*	succinate dehydrogenase iron-sulfur protein	775282	776001	9137	146	GCCTTTCTTTCCGCCGTT	439	178/361	*hae*III
					9138	585	GCATCGCTCATCCCTTCC			
9	*STY0917*	possible transport protein	908948	910291	9235	663	CGCCGAAAACCTGAAACT	528	376/152	*nla*III
					9236	1191	GCCCATACCGTACCGAAA			
10	*STY1395*	invasin-like protein	1347328	1349310	9193	759	AGGTGCGGCTCTGGTCTG	402	118/284	*rsa*I
					9194	1161	AGCGGAAGTTGAGGAAGG			
11	*STY1397*	putative thiol peroxidase	1350263	1350820	9221	449	CTTCAGCGTCAGGGTACA	552/342	894	*alu*I
					9222	1343	AACTGGCGGCAAGCGTAG			
12	*STY1583*	putative secreted protein	1531158	1531979	9121	210	GGCAACGCTTATCTCCCC	386	273/113	*rsa*I
					9122	596	CGGTCACGACGCAATCCT			

13	*ydhD*	conserved hypothetical protein	1615107	1615454	9133	41	AAAACCCGATTCTCCTGT	66/79	145	*hae*III
					9134	186	CAGTTGGCGTATTTCGGT			
14	*ssrA*	putative two-component sensor kinase	1647071	1649833	9223	1371	CGAGTCTGGTCATTTCAC	851	595/256	*alu*I
					9224	2222	CATCTATTTCTGGCATTC			
15	*STY1889*	conserved hypothetical protein	1786795	1787337	9123	14	AAATCTTTGCAGCTATCC	276/83/152	276/235	*rsaI*
					9124	525	TACAATCTTACCCGTTCC			
16	*pduX*	conserved hypothetical protein	2089486	2090388	9183	521	GCACGCAAGATTACCACC	76/171	247	*hpa*II
					9184	768	TCCAGCATCAGACCCACC			
17	*hisH*	amidotransferase	2110842	2111432	9107	220	CTGGGTATCTGCTTAGGG	137/53/38	190/38	*hha*I
					9108	448	GGCGATAGTCCACGGGTT			
18	*yehU*	putative two-component system sensor kinase	2220042	2221727	9154	2340	ATTGCCTGCGTTTACATT	283/616	799	*alu*I
					9155	3139	ACCCCTCTTTCCCTTCAC			
19					9231	16	CTGGTGTTGCTGCTGCTT	132/167	299	*nla*III
					9232	315	ATACAGCTCAGCGCCGTC			
20	*STY2406*	glutathione-S-transferase-family protein	2238562	2239206	9177	1194	TGGTCATTCTCCTTTTCA	104	50/54	*hha*I
					9178	1298	ATTTATCAGGCTATCCGC			
21	*STY2408*	putative gentisate 1,2-dioxygenase	2239932	2240969	9213	676	CTGCGGCTGGAATACATA	66/782	848	*hae*III
					9214	1524	GGGTGCTGGTGCTGGAAA			
22	*nfo*	endonuclease IV	2269017	2269874	9158	549	CTTTGCCGCTGGATACGA	273/227	500	*nla*III
					9159	1049	GGTCTGTTTGTTCTGCTT			
23	*STY2514*	anaerobic glycerol-3-phosphate dehydrogenase subunit B	2349474	2350733	9125	417	TTGCGTCGTCGGCGTTAG	117/80/202	117/282	*rsa*I
					9126	816	TCATCGCCCGGCATCCAG			
24	*ligA*	DNA ligase	2499609	2501624	9215	372	GTATGAAAACGGCGTGCT	456	164/292	*hae*III
					9216	828	AAACCGAGCGTGGGGCGA			
25	*talA*	transaldolase A	2542737	2543687	9179	261	TATTCCCGGTCGCGTCTC	99/34	133	*hha*I
					9180	394	TTCCCAGGTCGCGGCGAG			
26	*STY2713*	putative exported protein	2546128	2547171	9197	228	GCCGCCCGACGCTAAAGA	66/146	212	*taq*I
					9198	440	CGATAAACAGTAATGCCG			
27	*STY2873*	conserved hypothetical protein	2741767	2742243	9185	256	CGCATAAGCCACTTCAAC	30/295	325	*hpa*II

					9186	581	GAGTCATCGCCAGCACAG			
28	*STY3039*	conserved hypothetical protein	2907509	2908459	9233	119	GGTTATCAGATTCCCCTC	613	126/487	*nla*III
					9234	732	AGCATCTCGCCCACGGTT			
29	*STY3341*	conserved hypothetical protein	3183611	3184066	9225	170	CGTTAGCGGACCTGATGC	704	240/464	*alu*I
					9226	874	AGTGAAGATGTGGTGGTG			
30	*gltD*	glutamate synthase (NADPH) small chain	3353514	3354932	9115	1036	GTTCAACCGCTGGGCATT	203	62/141	*hha*I
					9116	1239	TGCGAGTCCAGTTCCACG			
31	*STY3542*	conserved hypothetical protein	3382461	3382724	9156	29	TAAGCATTCTGTCCGTTC	256/224	480	*alu*I
					9157	509	CATATCCATTGGCGTACT			
32	*rffT*	probable 4-alpha-l-fucosyltransferase	3482496	3483854	9211	69	ACAAACGAATCCAGCCCT	228/72	300	*hae*III
					9212	369	TATTCCTACCTGGTTATG			
33	*STY3991*	hypothetical protein	3856485	3857498	9187	411	CGAGTATGTGATGGGGTT	284	185/99	*hpa*II
					9188	695	CGTTGAGCGTTGCCAGCG			
34	*STY4105*	putative PTS system protein	3963524	3966847	9127	60	CCTCGGCTACACCTTCTC	559	153/406	*rsa*I
					9128	619	GATAAACCACGGCAACAG			
35	*STY4435*	hypothetical protein	4306483	4306593	9160	1303	TTATCTGGCGTCGTCACT	134/76/34	210/34	*nla*III
					9161	1557	CGGTCATATCGTCATCGG			
36	*STY4529*	hypothetical protein	4417838	4418617	9189	155	GTCGTCAGAGTCTTATTT	224/344	568	*hpa*II
					9190	723	TTTGCCGCCAGTAGAAGT			
37	*treB*	PTS system, trehalose-specific IIBC component	4652879	4654297	9219	1026	GCTGATTGCGCTGTCGAA	189	122/67	*hae*III
					9220	1215	AGCAGACCGGCAAGGCCA			

[1] Relative to the first base of the initiation codon of the gene

4.3. Results

4.3.1. Selection of SNPs for typing

From comparison of the full genome sequences of Typhi CT18 (Accession No. AL513382) and Ty2 (Accession No. AE014613) (57, 231) using BLAST tools available in Australian National Genetic Information Service (ANGIS), 253 single copy genes carrying SNPs with non-indel variation were found (Table 4.3-1). The polymorphisms within these genes varied from one base to eight bases accounting for 285 base substitutions, with the majority of the genes (239) having a single base substitution. Of the 285 polymorphisms, 111 were synonymous SNPs (sSNP) and 174 non-synonymous SNPs (nsSNP).

Table 4.3-1. The distribution of non synonymous and synonymous SNPs from genomes comparison of strain CT18 and Ty2

Base Change	Total Genes	Substitution		Total
		Non Synonymous	Synonymous	
1	239	147	92	239
2	8	6	10	16
3	2	6	0	6
4	1	3	1	4
6	2	9	3	12
8	1	3	5	8
	Total	174	111	285

Using the MLST terminologies, a single nucleotide change is considered to result from either a mutation or recombination. Any substitution with more than one base changes is due to recombination as multiple nucleotide changes are unlikely to a result of independent mutations (74). A total of 14 genes (Table 4.3-2) were found to have more than a single base change. The genes were located at different places ranging from 130 kb to 4.6 mb from the origin of the Typhi strain CT18 genome. None, except *yehU* and *yehT*, were from the same operon. Eleven of these genes were hypothetical proteins or only have putative functions while two encode for possible metabolic functions and one is a transposase element.

Table 4.3-2. Information of the genes which have more than a single base substitution

Total Base Change	ORF/gene name	Product	Position (bp)[1]	
			Start	End
2	*STY1630*	hypothetical protein	1,563,846	1564730
	rffT	probable 4-alpha-l-fucosyltransferase	3,482,496	3,483,854
	STY0649	conserved hypothetical protein	131,760	132,995
	yehT	putative two-component system response regulator	2,219,326	2,220,045
	STY3342	conserved hypothetical protein	3,184,202	3,186,373
	STY3405	probable membrane transport protein	3,250,996	3,252,240
	STY3542	conserved hypothetical protein	3,382,461	3,382,724
3	*gph*	phosphoglycolate phosphatase	4,192,593	4,193,351
	ydhB	putative transcriptional regulator	1,618,808	1,619,740
	treB	PTS system, trehalose-specific IIBC component	4,652,879	4,654,297
4	*STY2575*	putative transcriptional regulator	2,412,405	2,413,424
6	*yehU*	putative two-component system sensor kinase	2,220,042	2,221,727
	yfbB	conserved hypothetical protein	2,372,462	2,373,220
8	*tnpA*	transposase for insertion sequence element IS*200*	2,475,976	2,476,434

[1] The position corresponds to the genome of Typhi strain CT18

Thirty-six genes were selected for SNP typing using seven most economical 4-base REs (Table 4.3-3) to discriminate 37 SNPs of which 17 were sSNPs and 20 were nsSNPs. *Alu*I was utilised to differentiate eight SNPs, *Hha*I and *Hae*III seven each, *Hpa*II four, *Nla*III and *Rsa*I five each and *Taq*I one. All except two of the 37 SNPs were singly present in a gene. SNP 18 and SNP 19, sSNP and nsSNP respectively, were located in the same gene, *yehU*, encoding putative two-component system sensor kinase. There were six polymorphic sites scattered along the 1,686 bp *yehU*, four of which resulted in a change of amino acid.

Table 4.3-3. Seven most economical 4-bp cutter restriction enzymes

Enzyme	Recognition Site	No of SNPs detected[1]	Price (AUD/unit)
*Alu*I	AG'CT	8	0.11
*Hha*I	GCG'C	7	0.053
*Hae*III	GG'CC	7	0.035
*Hpa*II	C'CGG	4	0.053
*Nla*III	CATG'	5	0.23
*Rsa*I	GT'AC	5	0.11
*Taq*I	T'CGA	1	0.026

[1] Total number of SNPs that could be detected in a pool of 285 SNPs identified from the comparison of Typhi strains CT18 and Ty2

4.3.2. SNP typing

The 73 Typhi isolates (Chapter 2, Table 2.1-2) were typed for the 37 SNPs and they were scored to be either CT18- or Ty2-liked (Figure 4.3-1). From the 37 SNPs typed, 17 SNPs of them were shared by two or more isolates while 16 and four were found to be unique to CT18 and Ty2, respectively. An additional SNP was discovered upon analysing SNP 18. Sequencing of the PCR product revealed that two isolates (CC6 and CC7) had the same base as Ty2 at the site of SNP 18 but a single base change 360-base downstream created a new RE site. This SNP was designated as SNP 38, making a total of 38 SNPs. In SNP 37 (G<->A substitution), CT18 has an A base according to the genome data (231) but had the same digestion pattern as Ty2 indicating that it has the same base (G) as Ty2, which was confirmed by sequencing. This could be due to an error in the CT18 genome sequence (231) or strain CT18 that was used in this study had a mutation. Nevertheless, the base A allele was present in eight isolates.

The 73 Typhi isolates were grouped into 23 SNP profiles (Table 4.3-5). Twelve profiles were represented only by one isolate including the profiles of CT18 and Ty2. The other 11 profiles were shared by two or more isolates with SNP profile 10 being the largest and shared by 23 isolates. The z66 isolates had SNP profiles 1, 2, 4 or 5. It is interesting to note that isolate 422Mar92, belonging to a unique MLST sequence type (ST8) previously thought to be restricted to African isolates (141), also fell into this largest SNP profile.

Figure 4.3-1. The digestion patterns of Ty2 and CT18 for 37 SNPs typed

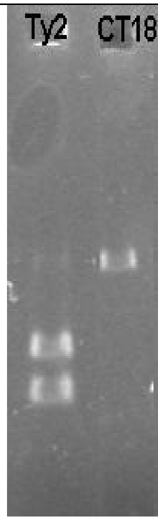
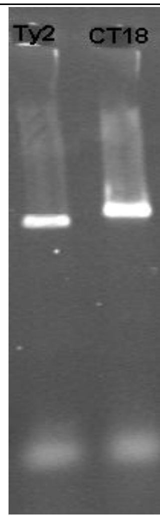
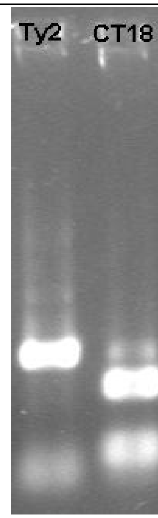
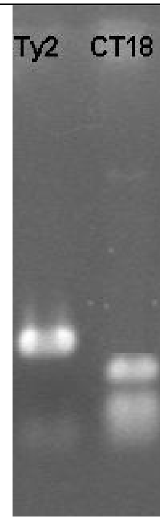
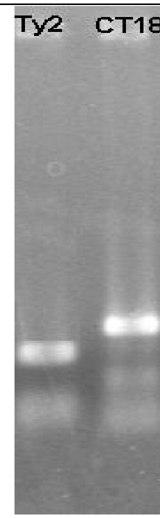
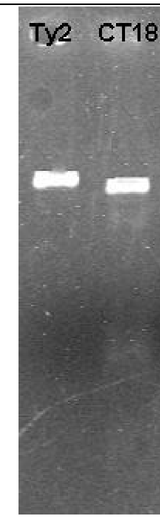
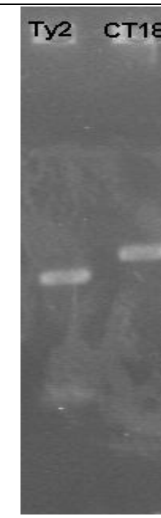
SNP No.	1		2		3		4		5		6		7		8	
Ty2 and CT18 on the gel																
Expected Size (bp)	256/344	609	37/633	670	215	166/49	152	107/45	112/52	164	138/156	294	624/6	564/60/6	178/361	439

Figure 4.3.1 (Cont.)

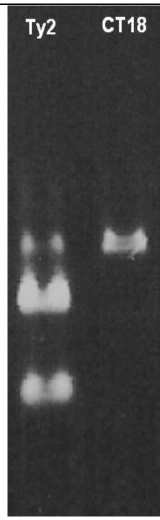
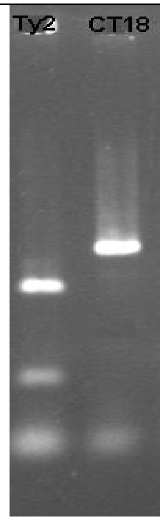
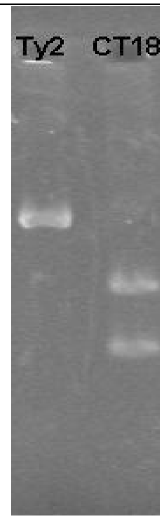
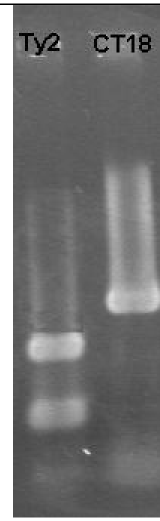
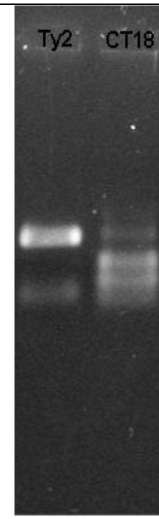
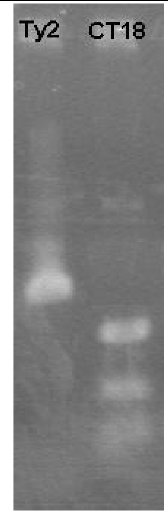
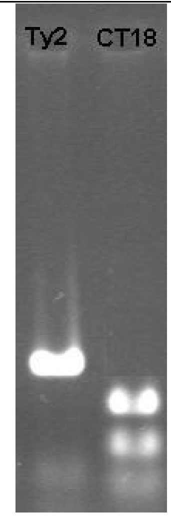
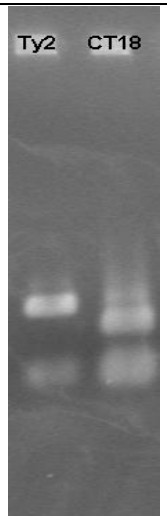
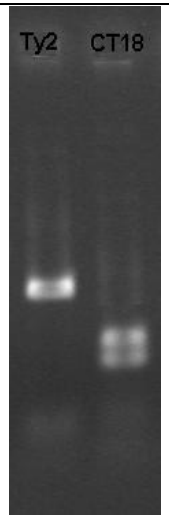
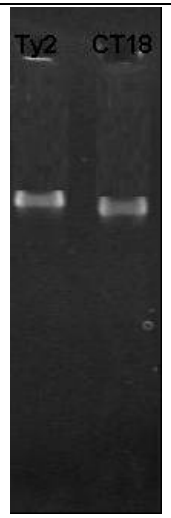
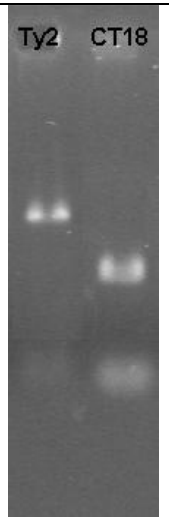
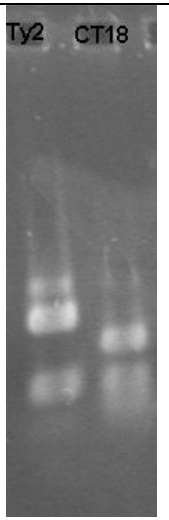
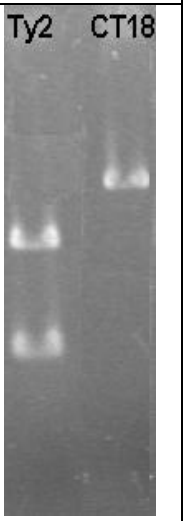
<table>
<tr><td>SNP No.</td><td colspan="2">9</td><td colspan="2">10</td><td colspan="2">11</td><td colspan="2">12</td><td colspan="2">13</td><td colspan="2">14</td><td colspan="2">15</td><td colspan="2">16</td></tr>
<tr><td>Ty2 and CT18 on the gel</td>
<td colspan="2">  </td>
<td colspan="2">  </td>
<td colspan="2">  </td>
<td colspan="2">  </td>
<td colspan="2">  </td>
<td colspan="2">  </td>
<td colspan="2">  </td>
<td colspan="2">  </td></tr>
<tr><td>Expected Size (bp)</td><td>376/152</td><td>528</td><td>118/284</td><td>402</td><td>894</td><td>552/342</td><td>273/113</td><td>386</td><td>145</td><td>66/79</td><td>595/256</td><td>851</td><td>235</td><td>83/152</td><td>247</td><td>76/171</td></tr>
</table>

Figure 4.3.1 (Cont.)

SNP No.	17		18			19		20		21		22		23		24	
Ty2 and CT18 on the gel																	
Expected Size (bp)	190	137/53	799	283/616	New[1]	299	132/167	50/54	104	848	66/782	500	273/227	117/282	117/80/202	164/292	456

[1] Two isolates were shown to have different digestion pattern which was neither CT18- nor Ty2-liked. This was a result of another base substitution downstream of the targeted SNP and therefore was designated SNP 38

Figure 4.3.1 (Cont.)

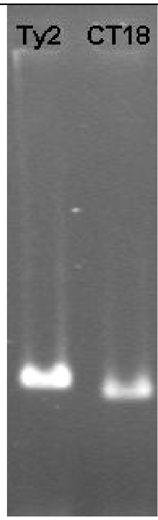
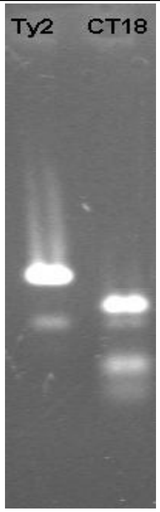
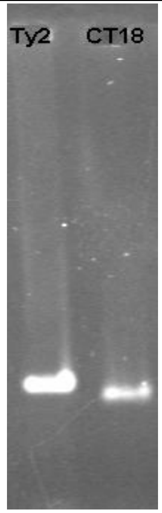
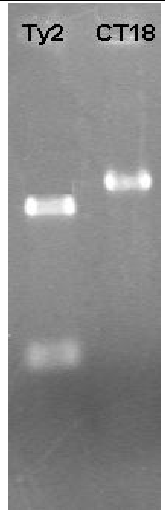
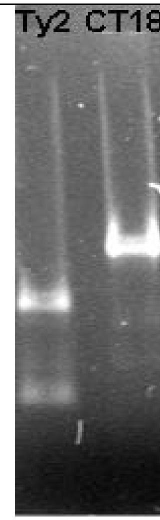
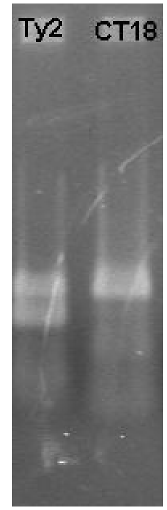
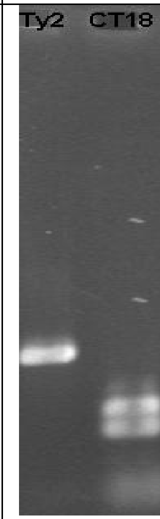
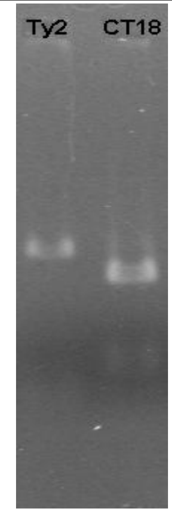
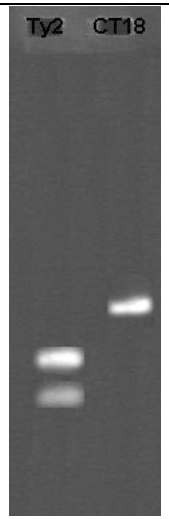
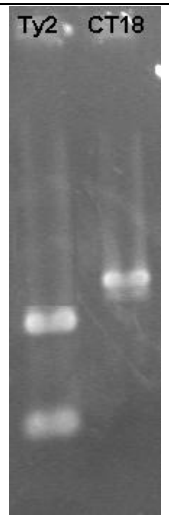
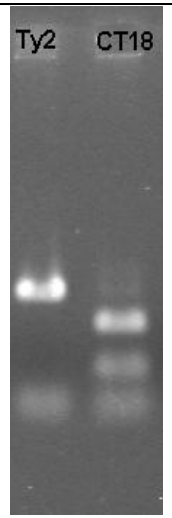
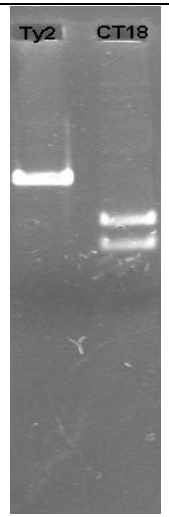
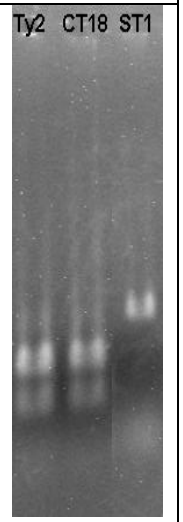
SNP No.	25		26		27		28		29		30		31		32	
Ty2 and CT18 on the gel																
Expected Size (bp)	133	99/34	212	66/146	325	30/295	126/487	613	240/464	704	62/141	203	480	256/224	300	228/72

Figure 4.3.1 (Cont.)

SNP No.	33		34		35		36		37	
Ty2 and CT18 on the gel										
Expected Size (bp)	185/99	284	153/406	559	210/34	134/76/34	568	224/344	122/67	189[2]

[2] Both Ty2 and CT18 had the same digestion patterns. Several isolates remained undigested.

Table 4.3-4. SNP profiles of the Typhi isolates

Cluster	SNP Profile	Total No. Isolates	SNP No.																																									
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
			A	C	C	T	G	C	C	G	G	C	T	G	G	C	G	C	G	C	C	C	T	G	G	G	C	A	C	C	A	C	A	A	C	G	C	G	A	C
I	1	2	.	.	.	.	.	.	.	.	.	.	**G**[b]	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	**T**	.	.	.
	2	12	.	.	.	.	.	.	.	.	.	.	**G**	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	**T**	.	G	.
	3	3	.	.	.	.	.	.	.	A[a]	.	.	**G**	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	**T**	.	G	.
	4	1	.	.	.	.	.	.	.	.	.	.	**G**	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	**T**	A	.	.
	5	4	.	.	.	.	.	.	.	.	.	.	**G**	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	**T**	A	G	.
II	6	1	.	**T**	T	C	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	**A**	.	.	.	.	.	.	.	**T**	.	.	.	.	.	.	G	.
	7	2	.	**T**	T	C	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	**A**	.	.	.	.	.	.	.	**T**	.	.	.	.	.	.	.	.
	8	1	.	**T**	.	.	A	.	.	A	.	.	.	.	.	.	.	.	A	.	.	.	.	**A**	.	.	.	.	.	.	.	**T**	.	.	.	.	.	.	.	.
	9 (CT18)	1	.	**T**	T	C	A	T	T	A	A	T	G	.	.	T	T	.	.	T	T	T	C	**A**	.	A	.	G	.	T	.	**T**	G	.	T	.	T	.	G	.
III	10	23	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.
	11	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	T
	12	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.
	13	4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	14	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	G	.
	16	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	A	.	.
	23	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	G	.
IV	15	1	.	.	.	.	.	.	.	.	.	.	.	A	A	.	.	T	**A**	.	.	.	.	.	.	.	**T**	.	T	.	.	.	.	.	.	A	.	A	.	.
	17	1	.	.	.	.	.	.	.	.	.	.	.	A	A	.	.	.	**A**	.	.	.	.	.	.	.	**T**	.	T	.	.	.	.	.	.	A	.	.	.	.
	18	4	.	.	.	.	.	.	.	.	.	.	.	A	A	.	.	T	**A**	.	.	.	.	.	.	.	**T**	.	T	.	.	.	.	.	.	A	.	A	G	.
	19 (Ty2)	1	G	.	.	.	.	.	.	.	.	.	.	A	A	.	.	T	**A**	.	.	.	.	.	A	.	**T**	.	T	.	C	.	.	G	.	A	.	A	G	.
	20	1	.	.	.	.	.	.	.	.	.	.	.	A	A	.	.	.	**A**	.	.	.	.	.	.	.	**T**	.	T	.	.	.	.	.	.	A	.	A	G	.
	21	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	**A**	.	.	.	.	.	.	.	**T**	.	.	.	.	.	.	.	.	.	.	.	G	.
	22	1	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	T	**A**	.	.	.	.	.	.	.	**T**	.	.	.	.	.	.	.	.	.	.	.	G	.
		Informative[c]		*	*	*	*			*			*	*	*			*	*					*			*		*			*				*	*	*	*		
		Type[d]	N	N	S	S	S	S	N	S	N	N	N	S	S	N	S	N	S	S	N	N	N	S	S	N	N	N	N	N	S	S	S	N	S	S	N	N	S
		Ancestral[e]	A	C	C	T	G	C	C	G	G	C	T	T	G	C	G	C	G	C	C	C	T	G	G	G	C	A	C	C	A	C	A	G	C	G	C	G	G	C
		Minimum Set[f]	*	*	*		*	*		*			*		*			*	*	*							*	*										*	*		

[1] Underlined bases were confirmed by sequencing of the representative isolate for each SNP profile

[2] Bases in bold support the corresponding cluster

[3] The site contains two alleles, each shared by two or more isolates

[4] N refers to non-synonymous SNP and S refers to synonymous SNP

[5] The ancestral base for each SNP was derived from consensus of eight non Typhi strains: Choleraesuis strain SC-B67 (GenBank Accession No: AE017220); Paratyphi A strain SARB42 (Accession No. CP000026); Typhimurium strain LT2 (Accession No. AE006468); Paratyphi B strain SPB7 (ftp://genome.wustl.edu/pub/seqmgr/bacterial/salmonella); Enteritidis strains PT4 and Gallinarum strain 287/91 (ftp://sanger.ac.uk/pub/pathogens/Salmonella); and Pullorum (http://www.salmonella.org/genomics/spu.dbs) and S. enterica subspecies V strain 12149 (ftp://ftp.sanger.ac.uk/pub/pathogens/Salmonella)

[6] The minimal set of SNPs required to identify all SNP profiles assigned in this study (See text for details)

Table 4.3-5. Information of the strains used in this study

Cluster	SNP profile	Strain Name	Genotype	Phage Type	Locality	Year	z66 Flagellar antigen[1]	Haplotype[2]
I	1	ST1106	4	D1	Malaysia	1987	-	
		414Ty	3	I + 1V	Australia	1981	+	59
	2	ST145	3	I+IV	Malaysia	1994	+	
		26T37	2	I+IV	BC	1994	+	
		In20	9	A	Indonesia	1992	+	59
		417Ty	22	I+IV	New Caledonia	1982	+	59
		418Ty	3	I+IV	Netherlands	1988	+	
		420Ty	3	UT	Japan	1982	+	59
		425Ty	3	I+IV			+	
		444Ty	3	I+IV			+	
		445Ty	3				+	
		446Ty	3	I+IV			+	
		701Ty	27				+	
		702Ty	3				+	
	3	CDC3137-73	6	K1	India		-	42
		26T30	3	I+IV	Quebec	1994	-	
		26T32	24	I+IV	Quebec	1994	-	
	4	423Ty	3	I+IV	Australia	1981	+	
	5	415Ty	3	UT	Netherlands	1982	+	
		416Ty	3	UT	Japan	1982	+	59
		419Ty	3	I+IV	Netherlands	1988	+	
		421Ty	3	UT	France	1984	+	
II	6	3125	3	46	Chile	1983	-	50
	7	Tp1	25	A	Dakar	1988	-	39
		ST24A	3	DVS	Malaysia	1986	-	16
	8	Tp2	17	UT	Dakar	1988	-	39
	9	CT18			Vietnam	1994	-	1
III	10	CDC3434-73	22	G1	Peru		-	52
		CDC1707-81	3	UT	Liberia		-	81
		CDC382-82	13	M1	Marshall Island		-	

		R1167	19	A			-	
		ST60	2	C4	Malaysia	1986	-	50
		ST24B	3	DVS	Malaysia	1986	-	
		In15	3	D2	Indonesia	1994	-	8
		PL27566	26	M1		1994	-	
		PL73203	2	A			-	
		26T6	30	UT	BC	1994	-	
		26T9	16	B1	Manitoba	1994	-	
		26T17	4	B1	BC	1994	-	
		26T19	5	A	Alberta	1994	-	
		26T40	19	M3	BC	1994	-	
		26T49	11	B1	BC	1994	-	
		26T50	8	I+IV	Alberta	1994	-	
		26T51	28	DVS	BC	1994	-	
		26T56	23	F1	Quebec	1994	-	
		3126	3	46	Chile	1983	-	
		PNG32	2	D2	Papua New Guinea	1994	-	8
		TYT1668	21	M1	Chile		-	76
		TYT1677	5	F8	Chile		-	
		422Mar92			Zaire	1992	-	6
	11	CC6	7	A	Thailand	1995	-	50
		CC7	7	A	Thailand	1995	-	
	12	26T12	6	O	Manitoba	1994	-	
	13	CDC1196-74	6	A	Mexico		-	11
		CDC9032-85	23	UT	Taiwan		-	50
		T189	11	N	Thailand	1990	-	42
		In24	3	C3	Indonesia	1992	-	14
	14	R1962	1	UT	Alberta	1993	-	42
		IP.E88 353		UT	Darkar		-	
		IP.E88 374		UT	Dakar		-	
	16	TYT1669	6	UT	Chile		-	52
	23	26T38	14	E1	BC	1994	-	
		3123	3		Chile	1983	-	
IV	15	ST1	18	I+IV	Indonesia		-	52
	17	T202	3	UT	Thailand	1990	-	52

18	25T-40	4	E1	BC	1993	-	
	R1637	14	E2			-	
	ST1002	9	E1	Malaysia	1987	-	52
	ST309	3	E1	Malaysia	1987	-	
19	Ty2	9	E1			-	10
20	25T-36	29	E1	Alberta	1993	-	
21	26T24	2	E1	Ontario	1994	-	
22	25T-44	2	E1	Ontario	1993	-	

[1] Strains ST145, 26T37 and In20 were found to carry z66 by PCR

[2] Haplotype according to Roumagnac *et al.* (271) study

4.3.3. Phylogenetic relationships

The relationships of the isolates were determined through a number of analyses. Initially, we used PAUP (304) to construct a maximum parsimony tree from the SNP data. There were 574 most parsimonious trees of equal length which differed mostly in the branching patterns, and the consensus tree from these trees revealed four major clusters (Figure 4.3-2). The division between clusters is supported by alleles unique to or uniformly present in the cluster (Table 4.3-4), suggesting that these were genuine genetic groups. Cluster I was supported by two SNPs, base G and T alleles in SNP 11 and 35 respectively; cluster II by four SNPs (SNP 2, 8, 22 and 30); and cluster IV by two SNPs (SNP 17 and 25). However there were no unique SNPs to support cluster III.

eBURST (73) was then used to identify the most closely related SNP profiles as clonal complexes (CCs) using the MLST terminology, based on the principle that SNP profiles differing by one SNP arose by a single mutation and thus originated from the same clone. eBURST revealed four CCs (Figure 4.3-3) and six ungrouped profiles, 8, 9, 16, 17, 19 and 22. The latter were considered as singletons as they had differences of more than one SNP from the four CCs or from one another. CC1 consisted of SNP profiles 10-14, 21 and 23 with SNP profile 10 identified as the founder of the CC. CC2 consisted of SNP profiles 1-5 and CC3 consisted of SNP profile 15, 18 and 20, with SNP profile 2 and 18 predicted as the founder for each of the CCs respectively. CC4 was the smallest having only two SNP profiles, 6 and 7, and it was not determinable which SNP profile was the founder.

A Minimum Spanning Tree (MST) was constructed using Arlequin v 3.1 (66) to visualise the overall relationships of the profiles (Figure 4.3-4). The MST groupings were consistent with the four clusters observed in the maximum parsimony consensus tree. The MST showed that SNP profile 10 was the ancestral profile connecting to the outgroup by two changes, indicating that cluster III arose first and the other three clusters emerged from cluster III. Most SNP profiles were linked to only one other SNP profile. However, SNP profiles 4, 16, and 17 showed equal distance to two or more SNP profiles, and alternative connections for these SNP profiles were represented on the tree as networks. The z66 isolates were all grouped into cluster I. It appeared that isolates expressing flagellar antigen z66 had a single

origin in cluster I. However, one isolate from SNP profile 1 and all SNP profile 3 isolates do not have the z66 flagellar antigen. Presumably these isolates have lost the gene.

The four CCs identified by eBURST were consistent with the phylogenetic clustering with the exception of SNP profile 21. SNP profile 21 was assigned to CC1 by eBURST but belongs to cluster IV. However there is no real conflict as SNP profile 21 was the founding member of cluster IV. Cluster I was more homogeneous than the other clusters, as it was represented by CC1 only while the other clusters contained more divergent members in addition to CCs.

The phylogenetic tree allowed us to determine whether there was any association of phylogenetic clustering with genome types, defined by the arrangement of I-*Ceu*I fragments (219), and/or phage types, based by sensitivity or resistance to Vi phages (51) (Figure 4.3-4 and Table 4.3-5). Phage type is largely independent from genome type as shown in other studies (148). Nevertheless, the occurrence of a particular combination of genome type and phage type were predominant in two phylogenetic clusters. Most of the isolates belonging to cluster I had genome type 3 and phage type I+IV. Although genome type 3 dominated, there were also other genome types in the cluster. For example, SNP profile 2 contained genome types 22 and 27. However, these two genome types were likely to have been derived from the predominant genome type 3, as each required only a single genomic rearrangement (169). Cluster IV contained all isolates with phage type E1 and its variant E2, which were used in this study but had no dominant genome type. Clusters II and III had no apparent association with particular genome or phage types. Cluster II contained three genome types and four phage types while cluster III had the most variable characteristics with 18 and 16 different genome types and phage types respectively. No association of phylogenetic clusters with year of isolation and/or localities of isolation was found. The 19 Canadian isolates analysed were scattered in three of the four clusters. Isolates from cluster III spanned all five regions: Africa, America, Asia, Europe and Oceania.

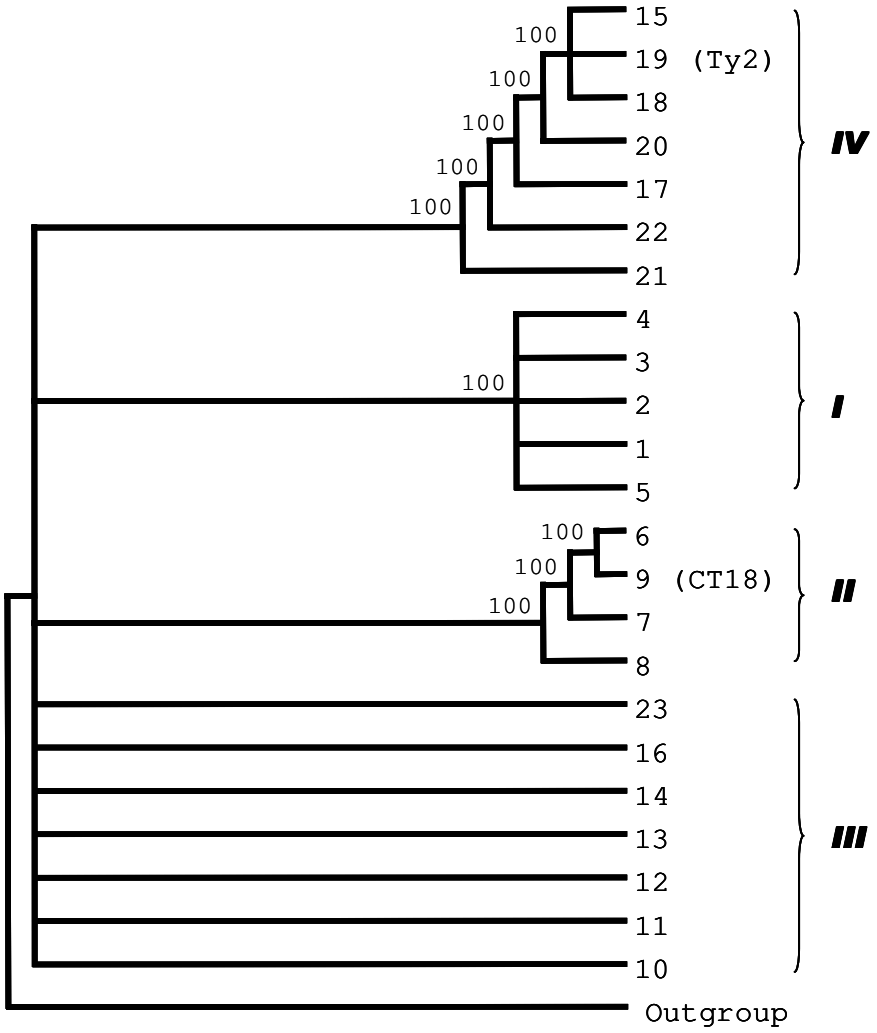


Figure 4.3-2. Consensus tree derived from the 574 maximum parsimony trees found through a heuristic search. The frequencies in percentage, of the branching orders found in all trees, are presented at nodes of the consensus tree.

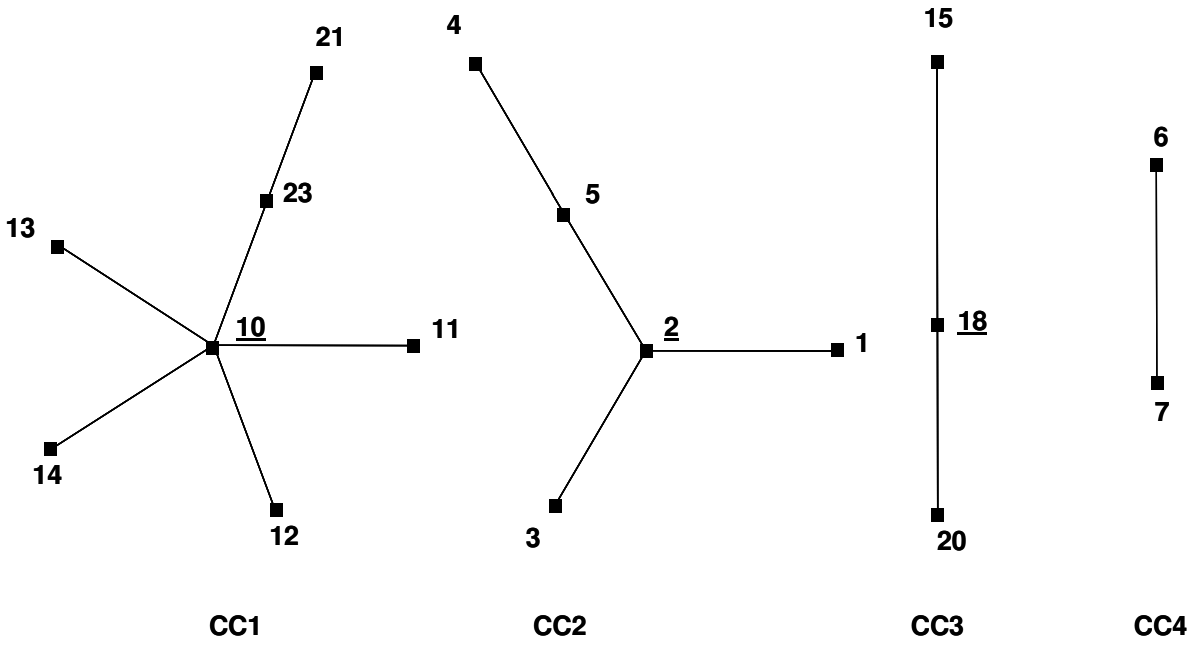


Figure 4.3-3. eBURST clonal complexes (CCs). The numbers on the nodes are SNP profile numbers. The underlined numbers represent the founder of each CC.

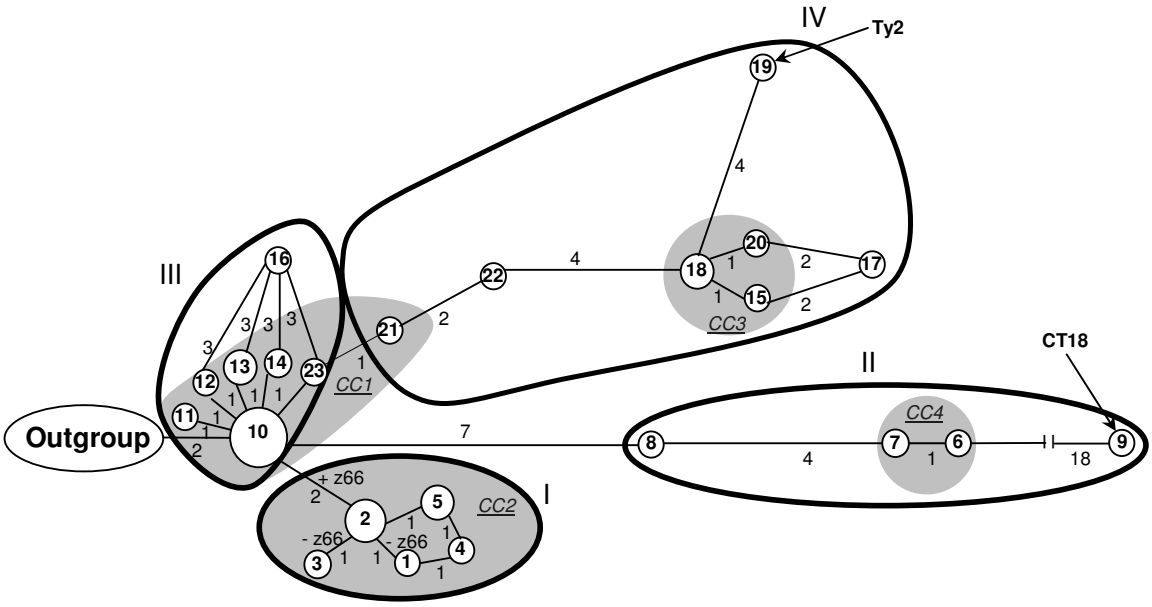


Figure 4.3-4. Phylogenetic relationships of Typhi SNP profiles. Within the circles are SNP profile numbers. The relative size of the circles (not to scale) illustrates size (the number of isolates) of the SNP profile. The numbers on the branches are the total SNP differences between the connecting nodes. The gain or loss of the z66 antigen is marked on the branch. Clusters are labelled with roman numerals. Each shaded area represents a clonal complex (CC). SNP profiles of CT18 and Ty2 are labelled and indicated with an arrow.

4.3.4. The discriminatory power of SNP typing

The ability of SNP typing to discriminate isolates was determined using Simpson's index of diversity (D) calculated using an in-house program MLEECOMP (available upon request) (249). The D value for this study was 0.870. We compared the D value of SNP typing to those of MLST (141) and ribotyping (148), both of which used global isolates, as no comparable dataset is available for PFGE, the "gold standard" for the comparison of the powers of typing methods (270). MLST and ribotyping had D values of 0.503 and 0.873, respectively.

4.3.5. Minimal SNP set required for differentiating the SNP profiles

To reduce the cost of genotyping and/or for large scale typing, it would be useful to define a minimal SNP set that could identify all SNP profiles as assigned in this study. We derived 16 SNPs that could be used to classify the 73 Typhi isolates into the same 23 SNP profiles (Table 4.3-4). These 16 SNPs utilise only four, *Alu*I, *Hae*III, *Hha*I and *Hpa*II, of the seven enzymes to type 3, 4, 5 and 3 SNPs, respectively.

4.3.6. Comparison of approaches to SNP discovery for typing

A similar study by Roumagnac *et al.* (271) surveyed 200 gene fragments from over 100 Typhi isolates for variation and found 88 informative SNPs. These SNPs were used as markers to differentiate 481 global Typhi isolates into 85 haplotypes, which were then grouped into five major clusters. Twenty-nine of the isolates analysed in this study were also studied by Roumagnac *et al.* (271). These isolates were distinguished into 17 SNP profiles in this study and into 14 haplotypes in the Roumagnac *et al.* (271) study (Table 4.3-5). Overall, the 38 SNPs offered a slightly higher differentiation for these 29 isolates. However, SNPs from the two studies gave different resolution to different groups. The SNPs tested in this study were able to divide haplotypes 39, 42, 50 and 59 further, while their SNPs distinguished the six isolates of SNP profile 10, the largest profile, into individual haplotypes.

From the three BiPs which differentiated 29 isolates into four major clusters according to Roumagnac *et al.* (271), BiP 36 further differentiated two isolates of SNP profile 10 into a separate entity while BiP 48 and BiP 56 were contradictory. BiP 56 grouped an isolate from each of the SNP profile 10 from cluster III and SNP profile 19 from cluster IV into one cluster. Unfortunately, only one isolate from cluster IV was represented and therefore no deduction could be made over this observation. BiP 48 was ambiguous whereby it grouped seven isolates from all of four clusters into one. If the clusters defined in this study were rearranged to include BiP 48 as the informative base, more parallel events were expected to take place. At cluster level, it appeared that clusters II and III were subdivisions of their cluster II as the six SNP profiles falling into the cluster II of Roumagnac *et al.* (271) were grouped in two separate clusters in this study (SNP profiles 6, 7 and 8 in cluster II and 10, 11 and 13 in cluster III).

According to the study by Roumagnac *et al.* (271), the z66 isolates have haplotypes 11, 52 and 59. Although most of these isolates were in a branch of one cluster, haplotype 11 and 52 were from very distinct clusters. Nevertheless, it was suggested that the z66 cluster radiated from haplotype 59, similar to current finding where 18 of the z66 isolates studied were divided into four SNP profiles and were all grouped into cluster I, indicating a single origin.

4.4. Discussion

4.4.1. Identification of SNPs through pairwise comparison of two Typhi strains

Typhoid fever is a major health problem especially in developing countries where sanitation and hygiene still remain poor. We have developed a molecular typing method using genome wide SNPs as markers. From the comparison of two complete genomes for serovar Typhi strains Ty2 (57) and CT18 (231), the difference between the two strains resulted from 285 SNPs in 253 genes. However, it is interesting to note that the number of nsSNPs is 30% greater than sSNPs since deleterious nsSNPs are expected to be eliminated from populations rather quickly. From 37 selected SNPs, there were four times more unique SNPs in CT18 than in Ty2 and the evolutionary process involved in this observation remains to be

elucidated. This seems to be a general phenomenon as similar observations were made in other bacterial clones (2, 81, 97, 236). In *M. tuberculosis*, 65% of the SNPs were nsSNPs (81) while 58% were nsSNPs in *B. anthracis* (236). The likely explanation is that the time frame is too short for many of the nsSNPs, in particular mildly deleterious ones, to be removed by purifying selection (268).

In this study we took the advantage of available genome sequences to select SNPs used for typing. As these SNPs were derived from the comparison of only two genomes, they could only reveal the evolutionary path separating Ty2 and CT18, due to phylogenetic discovery bias (236). Nevertheless, the SNPs used in this study allowed determination of the position of the last common ancestor of these Typhi isolates and node position for the SNP profiles. However, the true branch length of the SNP profiles, except for the two representing Ty2 and CT18, could not be determined. A recent study by Roumagnac *et al.* (271) took a different approach to obtain SNPs aiming to circumvent the phylogenetic discovery bias problem. Despite the large number of SNPs used, each of the five clusters was supported by a single SNP only and there is little resolution of relationships within a cluster. No parallel or reverse changes were detected contrary to what we observed in this study. Another recent study by Holt *et al*. (114) used high throughput sequencing to obtain genome wide SNPs for 17 Typhi strains. A total of 1,700 SNPs were used to generate a parsimony tree to infer the genetic relationships of these strains, in addition to the strains CT18 and Ty2 which genome data are available. The parsimony tree fully supported the division of strain CT18 and Ty2 into different clusters.

4.4.2. Homoplastic loci result from parallel or reverse changes

Six SNPs, 8, 11, 17, 35, 36, and 37 seemed to have undergone parallel or reverse changes across two or more independent lineages. For example, the allele base A of SNP 8 supporting cluster II was also present in SNP profile 3 of cluster I, and the two alleles of SNP 37 were present in all four clusters. Note that, although the T allele of SNP 25 was shared by all SNP profiles of cluster IV and two profiles of cluster III, this allele was not considered a result of parallel or reverse change because the two SNP profiles of cluster III are in direct line to the emergence of cluster IV.

As alleles were initially deduced from RE digestion, we confirmed the base changes of the alleles concerned by sequencing the representative isolates. Thus, the polymorphisms observed were a result of parallel or reverse changes due to either mutation or recombination. However, it will be difficult to determine whether the changes were due to mutation or recombination. We have recently shown that recombination is frequent within *S. enterica* subspecies I (Chapter 3). Recombination within a serovar may also be frequent and the parallel or reverse changes observed are likely to be due to recombination. It is interesting to note that four of the six SNPs involved were non-synonymous and selection pressure may play a role in driving some of these parallel or reverse changes. However none of the genes is known to be related to virulence.

4.4.3. Geotemporal distribution of isolates

A collection of seventy three global Typhi strains, isolated between the years 1981 to 1995 were SNP-typed using restriction enzyme digestion. They were distinguished into 23 SNP profiles and four distinct clusters. No clustering of isolates by year of isolation or locality was observed suggesting that major Typhi clones have spread globally. The isolates typed in this study were selected to represent the global diversity. It was not surprising that the major clones spread throughout the globe. Nevertheless, the sample size was too small to illustrate dominant clones circulating in a particular country or region. The largest sample we included from a single country was the Canadian sample of 19 isolates. Extensive epidemiological surveillance especially, regions where Tyhi is endemic will be required to further address the issue of spatial and temporal clustering. Unfortunately, there has been no method suitable for this purpose until now. Large scale typing of isolates from different regions using a genotyping method such as the SNP typing developed in this study will assist to further elucidate any spatial or temporal clustering of Typhi clones.

4.4.4. SNP typing is more discriminating than ribotyping and MLST

The SNP typing method developed in this study had a considerably higher discriminatory power than MLST but a similar power to ribotyping. However, the power of SNP typing could be increased by incorporating more SNPs, while ribotyping is constrained to detect variation in the 7 regions containing *rrn* operons only. Furthermore, variation detected by

ribotyping is mostly resulted from genome rearrangement due to *rrn* recombination rather than mutational variation (219), which could not be used to determine true relationships. The number of SNPs typed could also be reduced to 16 SNPs utilising only four of the seven REs. The total enzyme cost for typing an isolate is very small, far more economical in comparison to MLST. Further enhancement such as automation and PCR multiplexing could give additional advantage to this approach.

4.4.5. Origin of Typhi isolates expressing z66 flagellar antigen

Although most Typhi isolates are monophasic with the expression of the flagellar antigen encoded by the *fliC* gene at the H1 locus, some Indonesian isolates have an additional z66 flagellar antigen which was first described in 1981 (96) and was thought to be encoded by *fljB* at the H2 locus (204). The z66 flagellar antigen is now known to be encoded by a gene in the *fljBA*-like operon not located in the H2 locus (122).

The 18 z66 isolates studied were divided into four SNP profiles and were all grouped into cluster I indicating a single origin. Surprisingly, some of the isolates within this cluster do not carry z66 and it is highly likely that they lost the antigen. The presence of z66 flagellar antigen only in cluster I suggested that Typhi was originally monophasic having only an H1 antigen and then gained a new phase-2-flagellin like operon only recently during the divergence of cluster I. This new flagellin gene is more similar to H27 *fliC* of *E. coli* than to other H antigen genes of *S. enterica* (122) and is located on a linear plasmid as shown recently(13), which further supports the hypothesis of recent acquisition of the z66 antigen through lateral transfer. The findings in this study suggested that the earlier hypothesis is less likely to be correct.

Typhi was thought to be first adapted to humans in Indonesia and initially diphasic with an H2 locus encoding the z66 flagellar antigen (87). Our findings suggest otherwise, in that some monophasic Typhi isolates gained the z66 antigen and became diphasic. The H antigen is a part of the cell surface and one of the targets by the host immune system, and thus is under intense selection pressure for change (334). Since there is a limited spread of the z66 strains outside Indonesia (324, 325), it is possible that the z66 strains arose in Indonesia. It is unclear whether isolates carrying the z66 antigen have any advantages over the others in the

environment or the host. There is a high incidence of typhoid fever in Indonesia and it is possible that the coexistence of both monophasic and diphasic serovar Typhi isolates in Indonesia is a result of balancing selection to maintain the genetic diversity of this serovar (286).

4.5. Conclusion

We have shown that SNPs obtained from genome comparison are valuable markers for typing Typhi isolates and also to determine relationships of isolates in the homogeneous Typhi clone. The SNPs used were able to distinguish and resolve the relationship between Typhi isolates better than previously reported molecular typing methods. We differentiated the 73 global isolates studied into 23 SNP profiles using 38 SNPs and a subset of 16 SNPs could be used to achieve the same level of differentiation. The distinctive advantage of SNP based method is that true genetic relationships could be established. We have identified four clusters within the Typhi clone and revealed the origin of the isolates expressing z66 flagellar antigen. However, it should be noted that the SNPs used were derived from comparison of only two genomes and could only reveal the evolutionary path separating Ty2 and CT18, due to phylogenetic discovery bias (236). Nevertheless these SNPs allowed determination of the position of the last common ancestor of Typhi and node position for the SNP profiles despite the true branch length of the SNP profiles, except for the two representing Ty2 and CT18 could not be determined.

The SNP typing method designed in this study had considerable discriminatory power and will be a valuable tool for global epidemiological studies of Typhi. SNP typing is a flexible tool as many more SNPs are available to further increase the discriminatory power. We used only a small set of 37 SNPs from the 285 genic SNPs identified and chose to use economical restriction enzymes for detection of the SNPs. This approach had the advantage of minimal cost in consumables and the need for only basic laboratory equipment. A constraint was that many of the SNPs identified could not be used. However, other approaches such as real time PCR (104) may be employed to type the other SNPs when higher differentiation is needed.

Chapter 5: Hairpin real time PCR typing of four SNPs dividing major clusters

5.1 Introduction

The study in Chapter 4 has demonstrated the potential of genome wide Single Nucleotide Polymorphisms (SNPs) for molecular typing and determining relationships among Typhi isolates. Thirty eight SNPs were typed to differentiate 73 global *Salmonella enterica* serovar Typhi isolates into 23 SNP profiles (Chapter 4). Out of the 38 SNPs, 17 SNPs were polymorphic in multiple isolates while 16 and four SNPs were found exclusively in CT18 and Ty2, respectively. The isolates were divided into four major clusters. All except cluster III were supported by alleles unique to or uniformly present in the cluster. Clusters I and IV were each supported by a separate SNP while four SNPs supported cluster II. However, the study relied on SNPs from comparison of only two Typhi genomes.

During the completion of the aforementioned study, a similar study by Roumagnac *et al.* (271) used 88 SNPs (referred to as biallelic polymorphisms or BiPs) as markers to differentiate 481 global Typhi isolates into 59 haplotypes. The Typhi isolates were grouped into five major clusters (Figure 5.1-1) based on the allelic patterns of four of the BiPs. These BiPs could be applied as "dichotomous keys" to differentiate each of the clusters. In the current study, these four BiPs were selected and typed in the 73 global isolates to determine if they could divide these isolates into five clusters and if our clustering is consistent with the study by Roumagnac *et al.* (271).

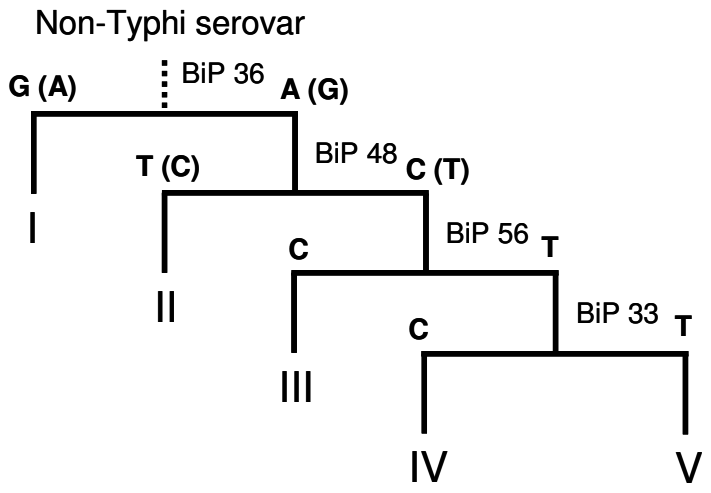


Figure 5.1-1. Four BiPs to divide 481 global Typhi isolates into five major clusters (271). These clusters, labelled as roman numerals, were differentiated by one BiP sequentially. The letters indicate the alleles of the BiP (original and corrected assignment). The letters in brackets indicate the corrected assignments of ancestral/alternative alleles for the corresponding BiPs (See results section).

There was no restriction enzyme (RE) suitable for typing these four SNPs and hence an alternative SNP typing method using a real time PCR (R-T PCR) platform was selected. This SNP detection assay relies on modifying linear primers to hairpin-shaped primers with the addition of a 5' tail complementary to the 3' end of the linear primer. The method is termed hairpin R-T PCR (HP R-T PCR) and was developed by Hazbon *et al.* (104). This HP R-T PCR assay has been shown to successfully type SNPs in bacterial species such as *Mycobacterium* (6) and *E. coli* (356), and in human DNA (29).

HP R-T PCR assay relies on the HP structures on the allele specific primers for accurate determination of the allele. The HP primers will only be linearised when they reach a certain temperature and it decreases mispriming and primer-dimer formation compared to linear primers, allowing enhanced discrimination of SNP alleles (104). Two reactions with the same DNA target are performed in parallel: one tube contains a template with the hairpin primer that will be fully complementary to one of the alleles at the SNP site, and the second tube is complementary to the target DNA sequence except for the last nucleotide at the 3' end, which is complementary to the alternative allele. The R-T PCR fluorescence curve develops more rapidly where the hairpin

sequence is fully complementary to the target DNA sequence marked by a smaller average cycle threshold (Ct value) (Figure 5.1-2).

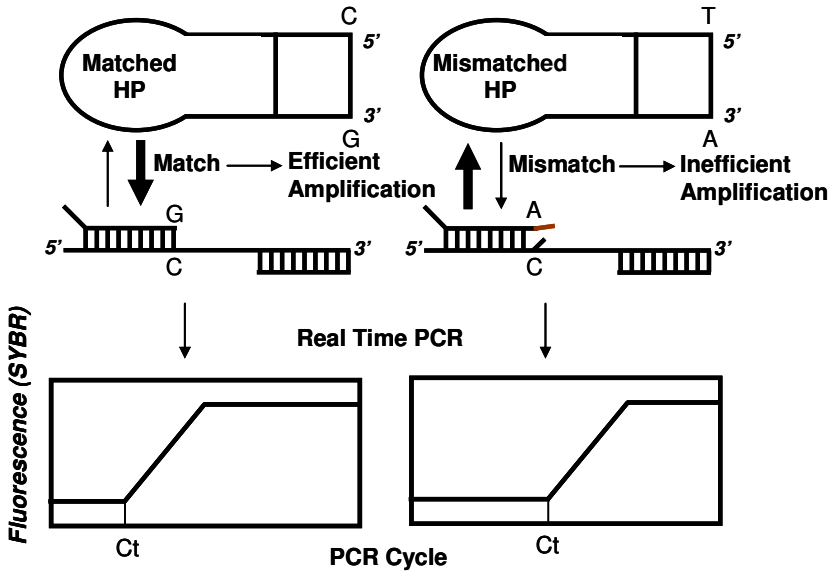


Figure 5.1-2. Principle of Hairpin (HP) R-T PCR. Adapted from Hazbon *et al.* (104). Matched hairpin primer will produce a lower Ct value as a result of more efficient PCR amplification while mismatched HP primer will have a higher Ct value.

The aim of this study was to type the four BiPs in the 73 global Typhi isolates using the HP R-T PCR assay. The Ct values obtained for matched and mismatched HP primers were compared to identify the allele for the four BiPs in each isolate. Twenty-nine of the Typhi isolates have also been typed by Roumagnac *et al.* (271), and typing of these isolates could confirm findings in their study. The BiPs were also employed to determine if the remaining isolates could be distinguished into the five clusters designated by Roumagnac *et al.* (271) and whether these clusters were consistent with the SNP-based clustering (Chapter 4).

5.2. Materials and methods

5.2.1. Bacterial strains

Seventy three global Typhi isolates (Chapter 2, Table 2.1-2), which DNA have been previously prepared, including strain CT18 and Ty2 as the positive controls, were used in this study.

5.2.2. Primer design

The primers for HP R-T PCR assay of the four BiPs were designed with the OLIGO v. 6 software (Table 5.2-1), to produce short amplicons (100 bp) and to optimally anneal at 55°C. At the 5' end of the SNP detecting primer, a tail containing five nucleotides was added to produce a stem loop with the 3' end of the primer with a Tm of 70°C and a free energey of less than -3.0. The size of the primer was 18 bp before the addition of the tail. Due to the restricted choice for the primer that covers the SNP site, the primer itself may form duplex prior to addition of the hairpin structure. A secondary mismatch was introduced to destablise the duplex and it confers more flexibility for the HP primer design. This will also decrease the affinity of the mismatched primers permiting a bigger Ct value difference between the matched and mismatched HP primers. To sequence the SNP for confirmation, another primer upstream of the hairpin primer was designed. These primers, located on 150-200 bp upstream of the SNP site (Table 5.2-1) to pair with the lower primer, were used to sequence the gene harbouring the SNP.

Table 5.2-1. The primers used for HP R-T PCR assays

BiP No.	Allele		ORF	Gene	Product	Primer's Name	Sequence 5' -> 3'[1]	Note
	Ancestral	Alternative						
33	C	T	STY2513	*glpA*	sn-glycerol-3-phosphate dehydrogenase subunit A	2513C	GCGGC'TCACGcCGGATACGCCG**C**	HP[2]
						2513T	ACGGC'TCACGcCGGATACGCCG**T**	
						2513U[3]	ATCCCAGCGGTCGTAAC	
						2513L[4]	GCGGGAAGCGAGATAATG	
36	G	A	STY2629		putative lipopolysaccharide modification acyltransferase	2629G	CAATC'CCCTGACTgAAGGATT**G**	HP
						2629A	TAATC'CCCTGACTgAAGGATT**A**	
						2629U	TTACTGCTGCCATATCCA	
						2629L	CCCACTCGCTTATTTCACTA	
48	T	C	STY3196	*lysS*	lysyl-tRNA synthetase	3196T	ATGAA'TTTCTTGAaCGCTTCA**T**	HP
						3196C	GTGAA'TTTCTTGAaCGCTTCA**C**	
						3196U	GACGTGAATGCCGATG	
						3196L	AGGATGTTCTGGGCACC	
56	C	T	STY3622	*hemD*	uroporphyrinogen-III synthase	3622C	GGTGC'TCTCGATTGGgCGCAC**C**	HP
						3622T	AGTGC'TCTCGATTGGgCGCAC**T**	
						3622U	GGTTTTTGCCCTTTCACA	
						3622L	GCAAGGCTTCGCTGATTTC	

[1] ' corresponds to addition of five nucleotides tail at the 5' of the primer to form a hairpin structure. Introduced secondary mutations are shown as small letters. Underlined in bold capital letters indicate the nucleotides that represent different alleles on the SNP site

[2] Upper primers used for HP R-T PCR with the first nucleotide at 3' end specific for a particular conformation of a SNP

[3] Upper primer used for sequencing

[4] Lower primer used for HP R-T PCR and sequencing

5.2.3. The R-T PCR reaction

All reactions were performed in a Rotor-gene™ 6000 (Corbett Life Science) sequence detector system available in the sequencing facility Ramaciotti Centre at University of New South Wales, School of Biotechnology and Biomolecular Sciences. Thermal cycling conditions were as the following: stage 1, 95°C for 2 min hold and 50°C for 2 min hold; stage 2, 10 cycles of 72°C for 30 sec, 95°C for 15 sec and 64°C for 30 sec lowering one degree in the last step for every cycle; and stage 3, 72°C for 30 sec, 95°C for 20 sec and 55°C for 30 sec repeated 40 times. Data were collected for last step of stage 3 for analysis with the Rotor-gene™ 6000 series Software v. 1.7 (Corbett Life Science). Every PCR tube contains 10 ng of chromosomal DNA, 6.25 µl of Platinum SYBR Green qPCR SuperMix-UDG with ROX (Invitrogen), 1 µl of 10 µM forward and reverse primers (Table 5.2-1) and MilliQ water to make up a final volume of 10 µl.

5.2.4. HP R-T PCR assay

In R-T PCR reaction, a fluorescent curve plotting the cycle number versus the amount of the SYBR Green signal is generated. The curve includes three phases of PCR amplification: initial amplification where the amount of PCR product is still low, the exponential amplification phase and the plateau phase where the products are saturated as a result of exhausted substrates. The Ct value also signifies the cycle number at which the level of fluorescence correlates to the number of amplicons that have accumulated in the R-T PCR reaction. The Ct value is determined when the fluorescent signal in the reaction reaches the threshold placed at the exponential amplification phase. The Ct value, however, is achieved early, when the HP primer is complementary to the SNP site. This is indicated by a more rapidly developed fluorescence curve (Figure 5.2-1).

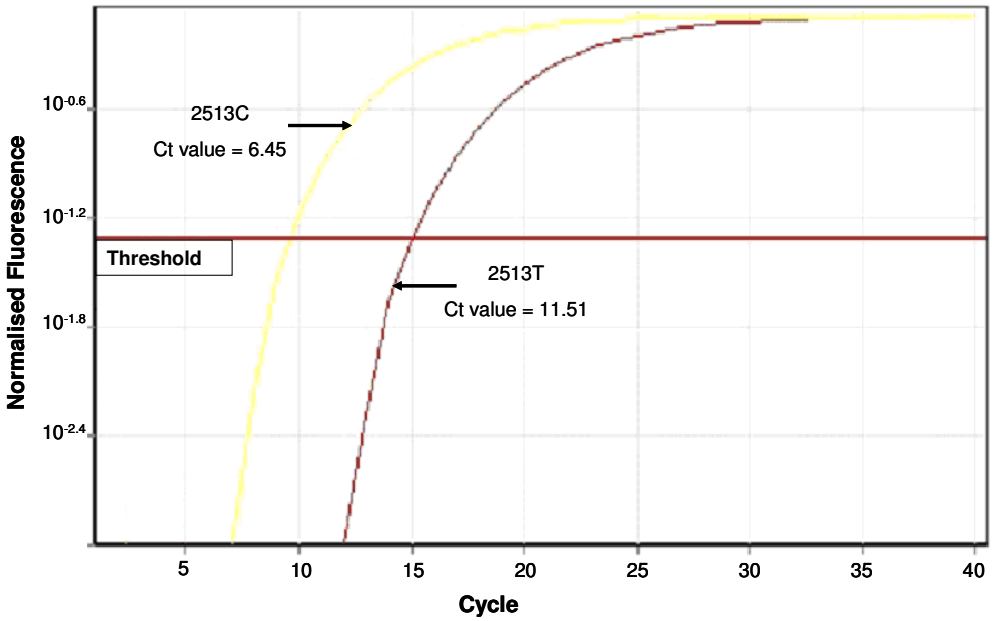


Figure 5.2-1. An example of the fluorescent curve that was generated from the HP R-T PCR assay for an isolate, which was typed for BiP 33. 2513C and 2513T indicate the two different upper primers used that correspond to two alleles, C and T, for BiP 33. The X axis corresponds to the number of cycles and the Y axis corresponds to the normalised fluorescence.

5.3. Results

5.3.1. Sensitivity of HP R-T PCR assay

The effectiveness of HP R-T PCR assay to type the four BiPs, BiP 36, BiP 48, BiP 56 and BiP 33, was only first tested using Typhi strain CT18 and Ty2 as the controls. CT18 and Ty2 differ in BiP 56, so both alleles expected for this BiP could be tested. However, the two strains are identical for the remaining three BiPs, so only one allele could be tested for the matched primer. Nevertheless, these two controls allowed the confirmation of the expected allele. Moreover, there was a clear difference in the Ct values obtained from the two HP primers set for each BiP suggesting that HP R-T PCR assay was useful for BiP typing and the remaining 72 Typhi isolates were typed for the four BiPs (Table 5.3-3). The HP assay could distinguish the SNPs with high confidence. The average differences in the Ct values for BiP 36, BiP 48, BiP 56 and BiP 33 were 6.64, 3.21, 5.48 and 5.34, respectively (Table 5.3-1).

Table 5.3-1. Summary of Ct values for all four BiPs

BiP	Allele	Ct value [Matched][1]				Ct value [Mismatched][2]				Δ Ct Value[3]			
		Average	Minimum	Maximum	STDEV	Average	Minimum	Maximum	STDEV	Average	Minimum	Maximum	STDEV
33	C	6.13	2.62	9.22	2.04	N/A[4]				9.25	6.70	18.63	2.15
	T	N/A[4]				15.38	10.92	26.67	3.67	N/A[4]			
36	A	6.38	4.76	8.92	2.23	20.18	16.02	23.75	2.02	11.68	10.90	12.53	0.82
	G	5.45	2.52	8.66	1.68	18.06	16.36	20.53	2.19	14.72	12.99	16.08	0.62
48	C	8.31	5.24	11.16	1.84	18.06	14.74	20.49	1.68	13.18	10.40	14.24	1.01
	T	9.36	6.34	11.55	1.50	21.49	19.14	24.28	1.42	8.70	6.74	9.34	0.72
56	C	6.29	2.57	9.89	1.87	15.59	12.60	18.37	2.01	8.29	5.54	9.86	1.03
	T	5.69	3.34	7.97	1.53	14.58	11.57	17.88	1.33	9.91	8.76	10.90	0.76

[1] Ct value obtained from HP R-T PCR reaction when the template is fully complementary to the upper primer

[2] Ct value obtained if the template had a mismatch at the 3' end of HP primer

[3] Δ Ct Value is the difference of the Ct values between the matched and mismatched HP primers used for R-T PCR

[4] Only strain Ty2 had a T match (Ct value=5.18) and C mismatch (Ct value=8.53), thus the data cannot be included in the table

5.3.2. Correction of the mis-assigned ancestral alleles for BiP 36 and BiP 48

Only two of the four possible alleles were observed in each of the four BiPs and one of these alleles was considered to be the ancestral allele (271). Cluster 1 was the first to emerge from non-Typhi serovars (271). An allele was considered to be the ancestral allele if it was present in isolates from cluster 1. Three out of the 73 Typhi isolates, 422Mar92, CDC1707-81 and ST60, have been typed by Roumagnac *et al.* (271) and they were grouped into cluster 1. The alleles for these isolates were expected to be G, T, C and C for BiPs 36, 48, 56 and 33, respectively.

However, based on the R-T PCR results, the isolates had alleles A, C, C and C for BiP 36, 45, 56 and 33 respectively. The allele G for BiP 36 was only observed in other isolates that were grouped into clusters 2 to 4 by Roumagnac *et al.* (271). Similarly, allele T for BiP 48 was only observed in clusters 3 and 4. Therefore, alleles A and C appeared to be the ancestral alleles for BiPs 48 and 56 instead of G and T (Table 5.3-2). We concluded from these results that ancestral alleles for two BiPs, BiP 36 and BiP 48, were mis-assigned while the ancestral alleles for the other two BiPs, BiP 56 and BiP 33, were correct.

5.3.3. Confirming the variants observed from previously typed BiPs

The 29 Typhi isolates used in this study were also typed by Roumagnac *et al.* (271). All of the alleles were identical except on five occasions (Table 5.3-2), where the alleles assigned were the opposite to what we typed. These five discrepancies were found for BiP 36 in strain ST60, BiP 48 in strains R1962 and T189, BiP 56 in strain In15 and BiP 33 in strain Ty2. The Ct value difference between the matched and mismatched HP primers was within the average, suggesting no ambiquity of HP RT-PCR typing results for these isolates. Sequencing was done for further confirmation. New upper primers were designed to pair with the lower primer. These primer pairs were used to amplify a fragment, harbouring the BiPs, of approximately 150-200 bp in length from the concerned isolates and were then sequenced. The sequencing results confirmed the HP R-T PCR results. This further showed that HP R-T PCR was a reliable method for SNP typing.

The change in allele led to a change in the cluster allocation for five isolates, which was conflicting to the result in Roumagnac *et al.* (271). These isolates have been grouped differently to the clusters reported in Roumagnac *et al.* (271). ST60 belonged to cluster 2, but R-T PCR results suggested that it belonged to cluster 1 due to mutation in BiP 36. R1962 and T189 were both in cluster 3, however they were moved in cluster 2 because of allele C for BiP 48. Interestingly, Ty2 has moved to cluster 5 as it had allele T for BiP 33. In15 had an allele T for BiP 56, which was the only allele observed when cluster 4 diverged from cluster 3. However, it also had alleles G and C for BiP 36 and BiP 48 respectively which were the characteristic of cluster 2. Based on the R-T PCR data, it was more likely that this isolate has undergone a parallel or reverse change at BiP 56 but was a sub cluster of cluster 2. This isolate was therefore designated as cluster 2a, differing from where it was previously placed by Roumagnac *et al.* (271).

Table 5.3-2. The observed alleles for each of the four BiPs in 29 Typhi isolates that have been typed by Roumagnac *et al.* (271)

Haplotype[1]	Cluster[1]		Strain Name	BiP 36	BiP 48	BiP 56	BiP 33
	Expected[2]	Observed					
H50	2	1	ST60	**A**[3]	C	C	C
H6	1	1	422Mar92	A	C	C	C
H81	1	1	CDC1707-81	A	C	C	C
H11	2	2	CDC1196-74	G	C	C	C
H14	2	2	In24	G	C	C	C
H16	2	2	ST24A	G	C	C	C
H39	2	2	SARB63	G	C	C	C
H39	2	2	SARB64	G	C	C	C
H42	3	2	R1962	G	**C**[2]	C	C
H42	3	2	T189	G	**C**[2]	C	C
H50	2	2	CDC9032-85	G	C	C	C
H50	2	2	3125	G	C	C	C
H50	2	2	CC6	G	C	C	C
H76	2	2	TYT1668	G	C	C	C
H8	2	2	PNG32	G	C	C	C
H8	2	2a	In15	G	C	**T**[2]	C
H1	3	3	CT18	G	T	C	C
H42	3	3	CDC3137-73	G	T	C	C
H59	3	3	In20	G	T	C	C
H59	3	3	414Ty	G	T	C	C
H59	3	3	416Ty	G	T	C	C
H59	3	3	417Ty	G	T	C	C
H59	3	3	420Ty	G	T	C	C
H52	4	4	CDC3434-73	G	T	T	C
H52	4	4	ST1	G	T	T	C
H52	4	4	ST1002	G	T	T	C
H52	4	4	T202	G	T	T	C
H52	4	4	TYT1669	G	T	T	C
H10	4	5	Ty2-b	G	T	T	**T**[2]

[1] The cluster and haplotype were designated Roumagnac *et al.* (271).

[2] The expected cluster number as assigned by Roumagnac *et al.* (271). for the 29 Typhi isolates that were typed in both studies.

3 Bold font indicates that the R-T PCR typing results contradicted Roumagnac *et al.* (271), however the base has been confirmed by sequencing.

5.3.4. Higher discrimination was achieved by typing the additional four SNPs and the minimum number of SNPs required for typing.

In the previous study (Chapter 4), 38 SNPs could distinguish 73 global Typhi isolates into 23 SNP profiles. Adding four BiPs resulted in the identification of four more SNP profiles. The 73 Typhi isolates were distinguished into 27 SNP profiles (Table 5.3-4). SNP profile 2 was separated into two profiles, 2a and 2b, due to the difference in BiP 48. SNP profile 2a was only represented by a single isolate, whereas SNP profile 2b had 11 isolates. SNP profile 10, which had the most isolates in the previous SNP study (Chapter 4), was further differentiated into four SNP profiles of SNP profiles 10a to 10d respectively. All of these had an allele C for BiP 48 except SNP profile 10d where it had allele T. SNP profile 10b still remained as the largest profile and contained 15 isolates. However, only three of the BiPs gave more discrimination in this study, as BiP 33 was only polymorphic in strain Ty2 that made up five unique SNPs for Ty2 out of 42 analysed SNPs (38 SNPs and 4 BiPs).

Previously, we selected 16 SNPs that were included as the minimum number of SNPs to differentiate 23 SNP profiles (Chapter 4). A Ty2 specific SNP, SNP 1, was also included into the minimum set of SNPs. However, BiP 33 has also been shown to be a Ty2 specific SNP. Since this BiP was regarded as an important BiP for cluster division by Roumagnac *et al.* (271), SNP 1 should be replaced with BiP 33 to represent the SNP specific to Ty2. A minimum of 19 SNPs, which included the three BiPs, would be required to differentiate 73 Typhi isolates into 27 SNP profiles.

Table 5.3-3. The list of observed alleles for each of the four BiPs in the 73 Typhi isolates studied

Haplotype[1]	Cluster[1]		Strain Name	BiP 36	BiP 48	BiP 56	BiP 33
	Expected[2]	Observed					
H6	1	1	422Mar92	A	C	C	C
H81	1	1	CDC1707-81	A	C	C	C
H50	2	1	ST60	**A**	C	C	C
H50	2	2	3125	G	C	C	C
		2	3126	G	C	C	C
		2	26T12	G	C	C	C
		2	26T19	G	C	C	C
		2	26T40	G	C	C	C
		2	26T50	G	C	C	C
		2	26T51	G	C	C	C
		2	26T56	G	C	C	C
		2	26T6	G	C	C	C
H50	2	2	CC6	G	C	C	C
		2	CC7	G	C	C	C
H11	2	2	CDC1196-74	G	C	C	C
		2	CDC382-82	G	C	C	C
H50	2	2	CDC9032-85	G	C	C	C
H14	2	2	In24	G	C	C	C
		2	IP.E88 353	G	C	C	C
		2	IP.E88 374	G	C	C	C
		2	PL27566	G	C	C	C
		2	PL73203	G	C	C	C
H8	2	2	PNG32	G	C	C	C
		2	R1167	G	C	C	C
H42	3	2	R1962	G	**C**	C	C
H39	2	2	SARB63	G	C	C	C
H39	2	2	SARB64	G	C	C	C
		2	ST1106	G	C	C	C
		2	ST145	G	C	C	C
H16	2	2	ST24A	G	C	C	C
		2	ST24B	G	C	C	C
H42	3	2	T189	G	**C**	C	C

H76	2	2	TYT1668	G	C	C	C
		2	TYT1677	G	C	C	C
H8	2	2a	In15	G	C	**T**	C
		3	26T30	G	T	C	C
		3	26T32	G	T	C	C
		3	26T37	G	T	C	C
H59	3	3	414Ty	G	T	C	C
		3	415Ty	G	T	C	C
H59	3	3	416Ty	G	T	C	C
H59	3	3	417Ty	G	T	C	C
		3	418Ty	G	T	C	C
		3	419Ty	G	T	C	C
H59	3	3	420Ty	G	T	C	C
		3	421Ty	G	T	C	C
		3	423Ty	G	T	C	C
		3	425Ty	G	T	C	C
		3	444Ty	G	T	C	C
		3	445Ty	G	T	C	C
		3	446Ty	G	T	C	C
		3	701Ty	G	T	C	C
		3	702Ty	G	T	C	C
H42	3	3	CDC3137-73	G	T	C	C
H1	3	3	CT18	G	T	C	C
H59	3	3	In20	G	T	C	C
		4	3123	G	T	T	C
		4	25T-36	G	T	T	C
		4	25T-40	G	T	T	C
		4	25T-44	G	T	T	C
		4	26T17	G	T	T	C
		4	26T24	G	T	T	C
		4	26T38	G	T	T	C
		4	26T49	G	T	T	C
		4	26T9	G	T	T	C
H52	4	4	CDC3434-73	G	T	T	C
		4	R1637	G	T	T	C
H52	4	4	ST1	G	T	T	C
H52	4	4	ST1002	G	T	T	C

		4	ST309	G	T	T	C
H52	4	4	T202	G	T	T	C
H52	4	4	TYT1669	G	T	T	C
H10	4	5	Ty2-b	G	T	T	**T**

[1] The cluster and haplotype have been designated by Roumagnac *et al.* (271). The clusters are labelled with numerals to prevent confusion with the clusters that we have already defined in previous study (Chapter 4).

[2] The expected cluster number as assigned by Roumagnac *et al.* (271). for the 29 Typhi isolates that were typed in both studies.

Table 5.3-4. The 27 SNP profiles arranged according to our clustering scheme

Our cluster	Rougmanac's cluster	SNP Profile	SNP No.																																			BiP No.			
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	36	48	56	33
			A	C	C	T	G	C	C	G	G	C	T	G	G	C	G	C	G	C	C	C	T	G	G	G	C	A	C	C	A	C	A	A	C	G	C	G	C	G	T	C	C
I	2	1	.	.	.	.	.	.	.	.	.	.	**G**	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	**T**	.	A	.	.	C	.	.
	2	2a	.	.	.	.	.	.	.	.	.	.	**G**	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	**T**	.	.	.	.	C	.	.
	3	2b	.	.	.	.	.	.	.	.	.	.	**G**	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	**T**	.	.	.	.	.	.	.
	3	3	.	.	.	.	.	.	.	A	.	.	**G**	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	**T**	.	.	.	.	.	.	.
	3	4	.	.	.	.	.	.	.	.	.	.	**G**	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	**T**	A	A	.	.	.	.	.
	3	5	.	.	.	.	.	.	.	.	.	.	**G**	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	**T**	A	.	.	.	.	.	.
II	2	6	.	**T**	T	C	.	.	.	**A**	.	.	.	.	.	.	.	.	.	.	.	.	.	**A**	.	.	.	.	.	.	**T**	.	.	.	.	.	.	.	.	.	C	.	.
	2	7	.	**T**	T	C	.	.	.	**A**	.	.	.	.	.	.	.	.	.	.	.	.	.	**A**	.	.	.	.	.	.	**T**	.	.	.	.	.	.	A	.	.	C	.	.
	2	8	.	**T**	.	.	A	.	.	**A**	.	.	.	.	.	.	.	.	A	.	.	.	.	**A**	.	.	.	.	.	.	**T**	.	.	.	.	.	.	A	.	.	C	.	.
	3	9	.	**T**	T	C	A	T	T	**A**	A	T	G	.	.	T	T	.	.	T	T	T	C	**A**	.	A	.	G	.	T	.	**T**	G	.	T	.	T	.	.	.	.	.	.
III	1	10a	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	C	.	.
	2	10b	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.
	2	11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	C	.	.
	2	12	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.
	2	13	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	C	.	.
	2	14	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	C	.	.
	2a	10c	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	T	.
	4	10d	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.
	4	16	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	A	A	.	.	.	T	.
	4	23	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.
IV	4	15	.	.	.	.	.	.	.	.	.	.	.	A	A	.	.	T	**A**	.	.	.	.	.	.	.	**T**	.	T	.	.	.	.	.	A	.	A	A	.	.	.	T	.
	4	17	.	.	.	.	.	.	.	.	.	.	.	A	A	.	.	.	**A**	.	.	.	.	.	.	.	**T**	.	T	.	.	.	.	.	A	.	.	A	.	.	.	T	.
	4	18	.	.	.	.	.	.	.	.	.	.	.	A	A	.	.	T	**A**	.	.	.	.	.	.	.	**T**	.	T	.	.	.	.	.	A	.	A	.	.	.	.	T	.
	4	20	.	.	.	.	.	.	.	.	.	.	.	A	A	.	.	.	**A**	.	.	.	.	.	.	.	**T**	.	T	.	.	.	.	.	A	.	A	.	.	.	.	T	.
	4	21	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	**A**	.	.	.	.	.	.	.	**T**	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.
	4	22	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	T	**A**	.	.	.	.	.	.	.	**T**	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.
	5	19	G	.	.	.	.	.	.	.	.	.	.	A	A	.	.	T	**A**	.	.	.	.	.	A	.	**T**	.	T	.	C	.	.	G	.	A	.	A	.	.	.	.	T	T
		Minimum Set[1]		*	*		*	*		*			*	*	*			*	*					*			*										*	*	*	*	*	*	*

[1] The minimal set of SNPs required to identify all 27 SNP profiles assigned in this study

5.3.5. Comparison of the clustering defined by typing 38 SNPs and four BiPs

In chapter 4, 73 Typhi isolates were distinguished into four major clusters by typing 38 SNPs. The division of most of the clusters was supported by alleles uniformly present in that particular cluster. Cluster I was supported by SNPs 11 and 35, cluster II by SNPs 2, 8, 22 and 30, and cluster IV by SNPs 17 and 25. Table 5.3-5 (A) presented only the eight of 38 SNPs supporting the clusters for ease of reading. From the four BiPs typed, two BiPs, BiP 36 and BiP 33, were unique to one SNP profile each. Allele A for BiP 36 was specific to one SNP profile 10a in cluster III while allele T for BiP 33 was specific to SNP profile 19. The allele T in BiP 56 was uniformly present in all SNP profiles of cluster IV and subset of cluster III. There was no unique distribution for either allele C or T in BiP 48. Allele C was present in clusters I to III and absent in cluster IV. The allele T of BiP 48 was present in all four clusters. This was an indication of parallel or reverse changes, where the SNP profiles from different clusters independently mutated to have the same allele. Altogether, there were five SNPs which were likely to undergo parallel or reverse changes including SNP 8, SNP 11, SNP 17, SNP 35 and BiP 48.

If the SNP profiles were alternatively arranged according to the Roumagnac *et al.* (271) clustering scheme, six different clusters were defined (Table 5.3-5, B). Five clusters have been previously defined in Roumagnac *et al.* (271) while the sixth cluster, designated as cluster 2a, was only established from the HP R-T PCR result (Table 5.3-3). Each of these major clusters was supported by a single SNP (Table 5.3-5, B). Each of clusters 1, 2a and 5 consisted of only one SNP profile, cluster 2 consisted of 10 SNP profiles, cluster 3 consisted of five SNP profiles and cluster 4 consisted of nine SNP profiles. No SNP from the 38 previously typed SNPs was found to support any of the major clusters. When the SNP profiles were arranged according to Roumagnac *et al.* (271), six SNPs were considered to undergo parallel or reverse changes. The SNPs were SNP 2, SNP 8, SNP 17, SNP 22, SNP 30 and BiP 56. These resulted in one more conflicting SNP in comparison to the SNP clustering scheme.

5.3.6. New clustering scheme

In both clustering schemes, subdivisions were seen. In our clustering scheme, there were three subdivisions in cluster I, three in cluster II, five in cluster III and two in cluster IV. The subdivisions in cluster I were resulted from SNP 8 and BiP 48; cluster II was subdivided by SNP 11, SNP 17 and BiP 48; cluster III was by SNP 25, BiP 36, BiP 48 and BiP 56; and cluster IV was by BiP 33. In Roumagnac *et al.* (271) clustering scheme, there were five subdivisions in cluster 2, three in cluster 3 and three in cluster 4. The subdivisions in cluster 2 were resulted from SNP 2, SNP 11, SNP 17 and BiP 56; cluster 3 was subdivided by SNP 2 and SNP 8; and cluster 4 was by SNP 17 and SNP 25.

Although the major clusters between the two schemes were different, the subclusters were identical between the two schemes, as defined by a combination of 12 cluster-supporting SNPs. We suggested these 13 subclusters to be renamed as an integrated scheme of 13 clusters for global epidemiology. The 13 clusters can be typed using 9 SNPs, including five SNPs from this study (Chapter 4) and four BiPs from Roumagnac *et al.* (271) (indicated in Table 5.3-5).

Table 5.3-5. Twenty seven SNP profiles were arranged according to our clustering scheme (A) and the scheme used by Roumagnac *et al.* (271) (B).

A

Our cluster	Sub-cluster	Rougmanac's cluster	SNP Profile	SNP No. 2	8	11	17	22	25	30	35	BiP No. 36	48	56	33
				C	G	T	G	G	C	C	C	G	T	C	C
I	Ia	2	1	.	.	**G**	.	.	.	.	**T**	.	C	.	.
		2	2a	.	.	**G**	.	.	.	.	**T**	.	C	.	.
	Ib	3	3	.	A	**G**	.	.	.	.	**T**	.	.	.	.
	Ic	3	2b	.	.	**G**	.	.	.	.	**T**	.	.	.	.
		3	4	.	.	**G**	.	.	.	.	**T**	.	.	.	.
		3	5	.	.	**G**	.	.	.	.	**T**	.	.	.	.
II	IIa	2	6	**T**	**A**	.	.	**A**	.	**T**	.	.	C	.	.
		2	7	**T**	**A**	.	.	**A**	.	**T**	.	.	C	.	.
	IIb	2	8	**T**	**A**	.	A	**A**	.	**T**	.	.	C	.	.
	IIc	3	9	**T**	**A**	G	.	**A**	.	**T**	T	.	.	.	.
III	IIIa	1	10a	.	.	.	.	.	.	.	.	A	C	.	.
	IIIb	2	10b	.	.	.	.	.	.	.	.	.	C	.	.
		2	11	.	.	.	.	.	.	.	.	.	C	.	.
		2	12	.	.	.	.	.	.	.	.	.	C	.	.
		2	13	.	.	.	.	.	.	.	.	.	C	.	.
		2	14	.	.	.	.	.	.	.	.	.	C	.	.
	IIIc	2a	10c	.	.	.	.	.	.	.	.	.	C	T	.
	IIId	4	10d	.	.	.	.	.	.	.	.	.	.	T	.
	IIIe	4	16	.	.	.	.	.	T	.	.	.	.	T	.
		4	23	.	.	.	.	.	T	.	.	.	.	T	.
IV	IVa	4	15	.	.	.	**A**	.	**T**	.	.	.	.	T	.
		4	17	.	.	.	**A**	.	**T**	.	.	.	.	T	.
		4	18	.	.	.	**A**	.	**T**	.	.	.	.	T	.
		4	20	.	.	.	**A**	.	**T**	.	.	.	.	T	.
		4	21	.	.	.	**A**	.	**T**	.	.	.	.	T	.
		4	22	.	.	.	**A**	.	**T**	.	.	.	.	T	.
	IVb	5	19	.	.	.	**A**	.	**T**	.	.	.	.	T	T
			Minimum Set[1]	*	*	*	*		*			*	*	*	*

B

Rougmanac's cluster	Sub-cluster	Our cluster	SNP Profile	SNP No. 2	8	11	17	22	25	30	35	BiP No. 36	48	56	33
				C	G	T	G	G	C	C	C	G	T	C	C
1		III	10a	.	.	.	.	.	.	.	.	A	C	.	.
2	2a	III	10c	.	.	.	.	.	.	.	.	.	C	T	.
		III	10b	.	.	.	.	.	.	.	.	.	C	.	.
	2b	III	11	.	.	.	.	.	.	.	.	.	C	.	.
		III	12	.	.	.	.	.	.	.	.	.	C	.	.
		III	13	.	.	.	.	.	.	.	.	.	C	.	.
		III	14	.	.	.	.	.	.	.	.	.	C	.	.
	2c	II	6	**T**	**A**	.	.	**A**	.	**T**	.	.	C	.	.
		II	7	**T**	**A**	.	.	**A**	.	**T**	.	.	C	.	.
	2d	II	8	**T**	**A**	.	A	**A**	.	**T**	.	.	C	.	.
	2e	I	1	.	.	**G**	.	.	.	.	**T**	.	C	.	.
		I	2a	.	.	**G**	.	.	.	.	**T**	.	C	.	.
3	3a	I	2b	.	.	**G**	.	.	.	.	**T**	.	.	.	.
		I	4	.	.	**G**	.	.	.	.	**T**	.	.	.	.
		I	5	.	.	**G**	.	.	.	.	**T**	.	.	.	.
	3b	I	3	.	A	**G**	.	.	.	.	**T**	.	.	.	.
	3c	II	9	**T**	**A**	G	.	**A**	.	**T**	T	.	.	.	.
4	4a	III	10d	.	.	.	.	.	.	.	.	.	.	T	.
	4b	IV	16	.	.	.	.	.	T	.	.	.	.	T	.
		IV	23	.	.	.	.	.	T	.	.	.	.	T	.
	4c	IV	21	.	.	.	**A**	.	**T**	.	.	.	.	T	.
		IV	22	.	.	.	**A**	.	**T**	.	.	.	.	T	.
		IV	20	.	.	.	**A**	.	**T**	.	.	.	.	T	.
		IV	18	.	.	.	**A**	.	**T**	.	.	.	.	T	.
		III	15	.	.	.	**A**	.	**T**	.	.	.	.	T	.
		III	17	.	.	.	**A**	.	**T**	.	.	.	.	T	.
5		IV	19	.	.	.	**A**	.	**T**	.	.	.	.	T	T
			Minimum Set[1]	*	*	*	*		*			*	*	*	*

[1] Minimum number of SNPs required to divide the isolates into subclusters

5.3.7. Phylogenetic relationship

A minimum spanning tree (MST) was constructed using the data from the 42 SNPs (38 SNPs from Chapter 4 and 4 BiPs from this study) to determine the relationships of the 27 SNP profiles. Four distinctive clusters, identical to the clusters defined in Chapter 4, were identified (Figure 5.3-1). Previous typing of 38 SNPs has suggested that SNP profile 10 was the ancestral profile connecting to the non-typhoid *S. enterica*. Four additional SNPs separated SNP profile 10 into four different SNP profiles of 10a to 10d. From comparison to the ancestral bases for the SNPs, it appeared that SNP profile 10b was the ancestral profile for cluster III, containing eight other SNP profiles. Six of these SNP profiles, 10a, 10c, 11, 12, 13 and 14, were different to SNP profile 10b by only one SNP, suggesting that they were the clones of SNP profile 10b. SNP profile 23 and SNP profile 16 differed by two and five SNPs to SNP profile 10b.

Clusters I and II seemed to arise from SNP profile 13 of cluster III. SNP profile 1 from cluster I and SNP profile 8 from cluster II were the closest to SNP profile 13. SNP profile 1 differed by two SNPs while SNP profile 8 differed by six SNPs to SNP profile 13, respectively. Cluster IV arose from SNP profile 23 of cluster III as a result of one SNP difference. The additional SNPs typed did not contradict the clustering of these profiles.

For the four BiPs (BiP 36/BiP 48/BiP 56/BiP 33) typed, there were six different patterns, each corresponding to a cluster (Figure 5.1-1): A/C/C/C, G/C/C/C, G/C/T/C, G/T/C/C, G/T/T/C and pattern G/T/T/T for cluster 1, 2, 2a, 3, 4 and 5 respectively. If clustering were designated based on these patterns, SNP profile 10a, which was the only profile that had the pattern A/C/C/C, would be the origin of the Typhi isolates. SNP profile 10a gave rise to SNP profile 10b due to a single base difference according to Roumagnac *et al.* (271) clustering scheme. SNP profile 10b which belonged to cluster 2, gave rise to SNP profile 10c of cluster 2a. Cluster 3 emerged from a single SNP difference to SNP profile 2a from cluster 2. Cluster 4 did not evolve from cluster 3 but it diverged from SNP profile 10c with a single base change. On the other hand, this reflected another conflict where previously clusters 1 to 5 have been shown to evolve in a linear fashion, and each cluster diverged from the previous cluster by differing in one BiP (271).

It has been shown in Chapter 4 that 18 isolates expressing the z66 flagellar antigen were of single origin and grouped into one cluster. Similarly, all seven Indonesian z66 positive isolates in Roumagnac *et al.* (271) emerged from a single haplotype suggesting that it has only arisen once. If the clustering was based on the BiPs, two conflicts were observed concerning the z66 positive isolates. One of the two isolates belonging to SNP profile 1 and the only isolate of SNP profile 2a were z66 positive, were located separately from the remaining 16 z66 positive isolates we studied. Both SNP profiles 1 and 2a were located in cluster 2 due to their BiPs pattern while the other z66 isolates were grouped in cluster 3. SNP profiles 1 and 2a were the only SNP profiles in cluster 2 which contained a z66 positive isolate.

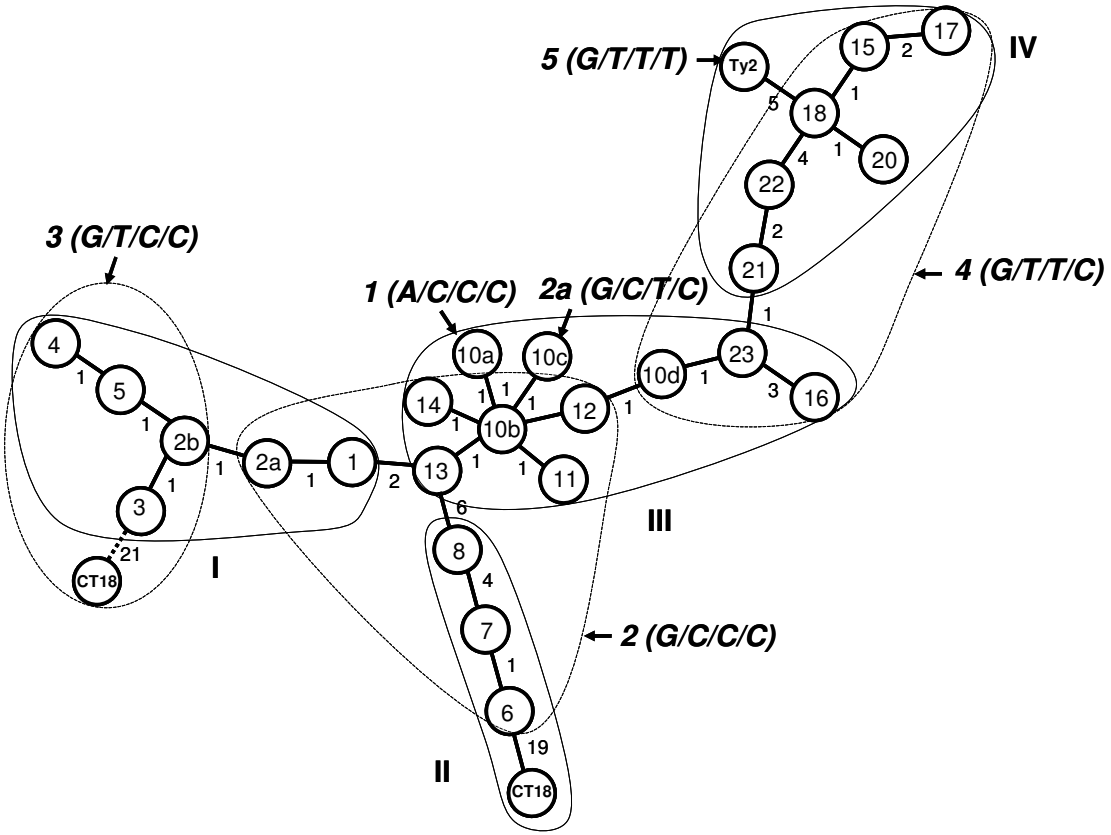


Figure 5.3-1. Minimum Spanning Tree of 27 SNP profiles. The number within the circles correspond to the SNP profile. The size of the circle is not proportional to the number of isolates having that SNP profile. The numbers on the node indicate the number of SNP differences between two profiles. Our designated clusters are indicated with roman numerals. The clustering designated according to Roumagnac *et al.* (271) are indicated by dashed lines and the cluster number with the BiP patterns for each cluster are in brackets.

5.4. Discussion

5.4.1. Hairpin Real Time PCR assay is an alternative method for SNP typing

Previously, SNP typing of Typhi isolates was achieved using restriction enzyme digestion (Chapter 4) or denaturing high performance liquid chromatography (dHPLC) as used in Roumagnac *et al.* (271). HP R-T PCR assay is an alternative method for SNP typing (104) and this was applied to type four BiPs in 73 global Typhi isolates. HP R-T PCR assay is not gel-based unlike the two aforementioned methods, and the results could be obtained directly after the completion of the PCR reactions. Therefore, the use of this method greatly reduced the time for SNP typing.

Previous SNP typing (Chapter 4) was designed to minimise the cost in consumables and it could be carried out with basic laboratory equipment. However, many of the identified SNPs from comparison of the two Typhi genomes could not be analysed due to the absence of suitable REs. The HP R-T PCR assay was more flexible, provided an easy discrimination of the SNP alleles and it did not have the limitation of the PCR-restriction enzyme digestion method. SNP detection by HP R-T PCR could be done in two assays, for which the design and setup were very straightforward. The cost of materials for each assay was low.

HP R-T PCR has been shown to be effective from the typing of four BiPs. The alleles of these BiPs were easily distinguished under a standard R-T PCR condition, allowing all BiPs to be analysed simultaneously. The results obtained from HP R-T PCR were easy to interpret since the Ct value difference was unambiguous between matched and mismatched HP primers. The average difference in Ct value for all four BiPs typed was 5.12. We have yet to test other factors that could improve the difference in Ct value. The downside was that SNP detection needed to be carried out in two separate tubes. This required considerable time for preparation if large number of isolates were to be typed simultaneously.

Based on our HP R-T PCR results, this assay appeared to be reliable and the results were repeatable. It is certainly an appealing alternative method for SNP typing if the R-T PCR platform is available. The minimum SNPs for the 38 SNPs from Chapter 4 should be further tested using HP R-T PCR assay so that they can be integrated into one assay.

5.4.2. Results conflicting with Roumagnac *et al.* suggesting errors in their typing

Twenty-nine Typhi isolates included in our investigation have been previously typed for the four BiPs by Roumagnac *et al.* (271). However, inconsistencies were observed in five isolates across the four BiPs (Table 5.3-2). One isolate each gave inconsistent results for BiPs 36, 56 and 33 respectively and two isolates were inconsistent in BiP 48 between the two studies. Our results were more likely to be correct as they have been confirmed by sequencing of the gene fragments harbouring the BiPs. One of the inconsistencies observed was in the strain Ty2. This was potentially due to the fact that our strain was different from the one used in Roumagnac *et al.* (271). Even though they were both wildtype Ty2, our Ty2 was designated as Ty2-b obtained from *Salmonella* Genetics Stock Centre (SGSC No. 2408) and was originally obtained from the study by Hone *et al.* (115). This strain has been included in several publications in the literature, the most recent being in 2004 (208). No further information was available for the Ty2 strain typed by Roumagnac *et al.* (271) except that it has been previously MLST analysed by Kidgell *et al.* (141).

Unfortunately, we could not explain the conflicts that were observed for the other four isolates. The isolates In15, ST60, T189 and R1962 were obtained from SGSC, the former three being originally isolated by Tikki Pang, while R1963 was from the Laboratory Center for Disease Control, Canada. The isolates analysed by Roumagnac *et al.* (271) were also obtained from the same sources.

According to supplementary material provided by Roumagnac *et al.* (271), the ancestral alleles for BiPs 36 and 48 were G and T, respectively. However, the HP R-T PCR results from typing of 29 common Typhi isolates suggested that the ancestral alleles for BiPs 36 and 48 were A and C instead of G and T. These inconsistencies could be due to typographical errors.

Using four of the BiPs, it was possible to differentiate 481 global Typhi isolates into five major clusters. We have identified a new BiP pattern, which has not been described previously (271). Only the isolate In15 had this BiP pattern due to a base substitution of C to T in BiP 56. This mutation was further confirmed by sequencing. Interestingly, the single base substitution in BiP 56 marked the separation of cluster 4 from cluster 3 in Roumagnac *et al.* (271). The allele T was observed in isolates belonging to clusters 4 and 5, respectively, but not in isolates from clusters 1 to 3. This was conflicting as In15 was located in cluster 2 on the constructed phylogenetic tree by Roumagnac *et al.* (271) because it was typed as allele C and not T for BiP 56. Nevertheless, HP R-T PCR results showed that In15 has the alleles that were identical to other isolates belonging to cluster 2 for the remaining three BiPs, 36, 48 and 33, respectively. The observed difference between the two studies for BiP 56 could be due either to an independent mutation or error in the dHPLC method.

5.4.3. A higher resolution was achieved by typing four more SNPs

In our study, the selection of SNPs for typing was greatly assisted by the availability of the two Typhi genomes, CT18 and Ty2**.** On the other hand, Roumagnac *et al.* (271) utilised dHPLC to discover SNPs from four Typhi strains in addition to Typhi CT18. These SNPs were screened in 105 diverse Typhi isolates to select for 88 BiPs. Strain CT18 and Ty2 had identical alleles in BiPs 36 and 48. These SNPs will never be detected by comparison of the genomes of Typhi CT18 and Ty2. These results further highlighted the existence of phylogenetic discovery bias where the number of SNPs detected depended on the number of genomes being compared (236). In Roumagnac *et al.* (271) study, five major clusters were identified using the SNPs that were discovered from comparison of four Typhi genomes to Typhi CT18. It is possible that as more SNPs are discovered from other genomes, more clusters will be identified. In order to maximise the discovery of SNPs, multiple genomes need to be compared. We have shown that by typing more SNPs, a higher resolution was achieved. Four additional SNP profiles were distinguished when three informative BiPs were included. There are more SNPs from Roumagnac *et al.* (271) study and they may be tested on our isolates for future investigations.

5.4.4. Phylogenetic rooting and conflicting phylogenetic signals

Regardless of how SNP profiles were arranged, according to either the Roumagnac *et al.* (271) clustering scheme or our own, homoplasies were observed in some of the SNPs typed. If the clustering of SNP profiles were based on our clustering scheme, six SNPs, including SNP 8, 11, 17, 35, 36 and 37, (Discussion, Chapter 4) and BiP 48 would not be uniquely present in a particular cluster. In contrast, if clustering was based on the patterns of the four BiPs in Roumagnac *et al.* (271), more homoplasies would be observed. Eleven SNPs were shown to have undergone parallel or reverse changes, including SNPs 2, 3, 4, 5, 8, 16, 17, 22, 30, 36 and 37. This contradicted by Roumagnac *et al.* (271), where all 88 BiPs typed in their study were entirely parsimonious.

The absence of homoplasy is also characterised by the linear phylogeny. This was observed in the constructed MST by Roumagnac *et al.* (271) but not in our constructed MST. According to Roumagnac *et al.* (271), the five defined clusters sequentially emerged after a base substitution in a particular BiP. Cluster 1 gave rise to cluster 2 after a mutation in BiP 36, cluster 2 to cluster 3, and so on (Figure 5.1-1). On the other hand, our four defined major clusters were not presented as a linear phylogeny. Cluster III, which was considered to be the ancestral cluster, gave rise to cluster I, II and IV due to mutations in two, four and one SNP/s, respectively. We believe that our assignment of clustering was more accurate given that there were multiple independent SNPs supported clusters I and II, as well as one SNP supporting cluster IV in comparison to only one BiP, which supports each of the five clusters. We concluded that BiP 48 has undergone parallel or reverse changes which were not observed by Roumagnac *et al.* (271).

Conflicting phylogenetic signals could have arisen from multiple substitutions at the same site or homologous recombinations. These are expected to appear less frequently in a relatively young lineage or very slowly evolving bacterial species, like *B. anthracis* and *Y. pestis* (236) for example. They do not have sufficient time to accumulate the same mutation at a particular site in multiple lineages or to experience recombination between isolates. Only one in more than 25,000 SNPs was conflicting in the species of *B. anthracis* (236) and no homoplasy was recognised when 44 SNPs were screened in 105 diverse isolates of *Y. pestis* (2). Older lineages or freely recombined species,

such as *N. meningitidis* (185)**,** have a tendency to display homoplasy due to longer time frame for accumulating mutation and and frequent recombination (236).

Typhi was predicted to have evolved 50,000 years ago from non-typhoid *S. enterica* serovar (141). It is considered to be older than *Y. pestis*, which was predicted to have evolved from *Y. pseudotuberculosis* around 1,500–20,000 years ago (3). *B. anthracis* was also predicted to have evolved from *B. cereus* around 3,277-27,245 years ago (322). Thus, it is not surprising to observe more homoplasies in Typhi. Parallel or reverse changes due either to recombination or independent mutation could play a role in shaping the polymorphisms at homoplastic SNPs and BiPs, and thus the genetic diversity among Typhi. It has recently been shown that recombination is frequent between serovars belonging to *S. enterica* subspecies I (Chapter 3). It is likely that recombination within Typhi is also frequent in contributing to the observed parallel or reverse changes.

5.5. Conclusion

Four BiPs have been typed in 73 Typhi isolates using HP R-T PCR assay. Five of the 29 Typhi isolates that have been previously typed for the four BiPs were shown to be inconsistent. There was one isolate each that showed contradictions in BiPs 36, 56 and 33, respectively, and two isolates were inconsistent in BiP 48. Three of the four BiPs have increased the discrimination of SNP typing and the isolates were differentiated into 27 SNP profiles, four more than previously observed. The relationships of the SNP profiles were illustrated in an MST as four major clusters where all clusters arose from cluster III. The clusters were not represented in a linear phylogeny, in contrast to the MST tree constructed by Roumagnac *et al.* (271).

We have shown that out of the 42 SNPs typed, seven SNPs including BiP 48 have undergone parallel or reverse changes. These could have resulted from either independent mutations or recombination. This conflicts with previous findings by Roumagnac *et al.* (271), which did not reveal any phylogenetic conflict for the four BiPs.. Furthermore, Roumagnac *et al.* (271) found no homoplasy for the other 87 BiPs typed in 481 Typhi isolates. Due to conflicting results in five of the 29 Typhi isolates typed independently between the two studies, more errors are likely to be

observed. Typing of all the BiPs corresponding to the evolutionary pathway of these isolates needs to be done for further confirmation.

A minimum of 19 SNPs (15 SNPs and four BiPs) can be used to differentiate Typhi isolates into 27 SNP profiles. We proposed nine SNPs (five SNPs and four BiPs) were sufficient to group the Typhi isolates into 13 clusters for global epidemiology.

Chapter 6: Developing a novel method for mutational discovery in *S. enterica* serovar Typhi using Surveyor™ nuclease

6.1 Introduction

Single Nucleotide Polymorphisms (SNPs) are considered to be the most valuable markers, particularly for studying the evolutionary relationships of isolates in homogeneous pathogenic clones such as *Bacillus anthracis* (236), *Mycobacterium tuberculosis* (81), *Yersinia pestis* (2) and *Salmonella enterica* serovar Typhi (Chapter 4). In our previous study, the selected SNPs were obtained from comparison of two genomes of strains CT18 and Ty2 (Chapter 4). However, the identified SNPs only revealed polymorphisms and the locations of the last common ancestor that directly connected to the evolutionary pathway of the two Typhi strains. Any polymorphisms before or after the divergence of the reference strains were undetectable. This is referred to as as the phylogenetic discovery bias (236). To avoid this phenomenon, SNPs from different lineages are needed.

The most effective strategy to directly discover SNPs is by high throughput sequencing. For example, five diverse *B. anthracis* strains were sequenced and compared to Ames *B. anthracis* reference genome (236). The whole-genome sequencing of five *B. antracis* strains allowed identification of approximately 3,500 SNPs. Twenty percent of the SNPs were selected for typing to determine the evolutionary relationships of 29 diverse *B. anthracis* strains (236).

The cost associated with genome sequencing is still high for the purpose of SNP discovery. In particular, when multiple isolates need to be sequenced and the frequencies of SNPs are relatively rare. Recent advances in microarray technology allow its use in whole genome sequencing. NimbleGen is one of the companies which provide the instruments and services for this technology (www.nimblegen.com).

NimbleGen microarray has been used for comparative genome sequencing to explore the presence of SNPs in *E. coli* strains. The microarray was first used to analyse 1199 chromosomal genes, and the large virulence plasmid (pO157) of 11 test isolates representing O157:H7 strains that were associated with human outbreaks (357). A total of putative 906 SNPs in chromosomal genes were identified in the test isolates, with Sakai used as the reference strain (357). NimbleGen was also used to compare five different *E. coli* strains that were grown in continuous logarithmic growth in glycerol minimal medium from which, a total of putative 95 SNPs were identified (108). Comparative genome sequencing using NimbleGen is an effective way to detect SNPs in an isolate relative to another. Even though it is more cost-effective than traditional capillary sequencing, it is still considerably expensive for SNP discovery from multiple isolates.

A less expensive method is the use of indirect SNP discovery, which involves the screening of mutations that depend on either the physical properties of the DNA or the recognition of a mismatch. Most of the methods involve the formation of wild type/mutant heteroduplexes. Mutational scanning could exploit the physical differences that can be detected by: electrophoretic analyses to determine the shifts in mobility due to the difference in the molecular conformation such as single-strand conformational polymorphism (SSCP) (227); denaturing gradient gel electrophoresis (DGGE) to analyse the difference in melting temperature (82); and denaturing high performance liquid chromatography (dHPLC), which is an ion-pair reverse-phase liquid chromatography on a special column matrix with partial heat denaturation of the DNA strands (224). Roumagnac *et al.* (271) utilised dHPLC to discover SNPs in *S. enterica* serovar Typhi by comparing four Typhi isolates with Typhi strain CT18 as the reference. A total of 88 SNPs were found in 66 of 200 genes tested (271).

Chemical or enzymatic mismatch cleavage methods could be used to identify DNA fragments which contain mutations. Chemical cleavage of mismatch (CCM) involves the modification of mismatched thymine residues with osmium tetroxide and cytosine residues with hydroxylamine (49). The modified bases are then cleaved by hot piperidine treatment to determine the site of the mismatch that can be detected on either denaturing acrylamide gels (49), fluorescent assay (102) or solid-phase capture of the heteroduplex (100). Enzymatic mismatch cleavage is an alternative to the chemical method, which does not require the handling of hazardous chemicals. This method also cleaves mismatches on the heteroduplexes. The enzymes used include: ribonuclease; T4 endonuclease VII (189, 355); MutS (167); MutY and thymine glycosylase (179). These enzymes

however, cannot detect all possible mutations and some could have substantial activity such as digestion of homoduplexes (222).

The recent discovery of a new nuclease has added an option for the choice of enzymes that could be used for mismatch cleavage based detection. The enzyme Surveyor™ (Transgenomics) nuclease, is a mismatch-specific DNA endonuclease commonly known as *Cel*I nuclease (225). It can recognise base substitution and deletion as short as 1-3 nucleotides (225). The advantage of this enzyme is that, it produces base overhang which allows cloning of the gene fragments carrying the SNPs. This chapter aimed to discover more SNPs in Typhi, especially in isolates that were previously shown to represent distinct phylogenetic lineages (Chapter 4 and Chapter 5), and the isolates that could not be differentiated by typing using 42 SNPs (Chapter 5). The discovery of more SNPs will provide a better insight in revealing the complexity of relationships among isolates between and within each cluster. A method was devised in an attempt to employ *Cel*I nuclease for obtaining SNPs from largescale comparisons of pairs of Typhi isolates. The principle of this method is described in the next section.

6.2. Principle of the design of the new method

*Cel*I nuclease, commercially available as Surveyor™ (Transgenomics), was chosen as an enzyme to discover more SNPs. The step by step explanations of the following figure (Figure 6.2-1) are as the following:

1. The two genomes are digested using a selected restriction enzyme, for example *BsaH*I, that leaves base overhangs at 3' end. After digestion, ligation will proceed with two sets of adaptors, one set for each reference isolate. The adaptors are designed such that once they have ligated to the digested fragments, the restriction site is removed.

2. The two ligated products will be mixed together. Denaturation and reannealing of the two DNA samples will form either matched fragments (homoduplex) or mismatched fragments (heteroduplex) if a SNP is present.

3. The hybridised DNA mixture will then be treated with *Cel*I nuclease. *Cel*I recognises a mismatch on the heteroduplex and nicks the opposite strands of the same mismatch molecule at opposite phosphodiester bonds of the mismatch, producing a one-nucleotide 3' overhang to allow non-blunt end ligation of an adaptor.

4. After *Cel*I digestion, an adaptor designed to bind to the *Cel*I single base overhang will be ligated. The adaptor has an N at the 3' end so it can bind to any of the four possible nucleotides representing an allele of the SNPs.

5. Following the ligation of *Cel*I adaptors, the sample is amplified using primers binding specifically to the adaptors. Only DNA fragments carrying *BsaH*I and *Cel*I adaptors will be exponentially amplified. The other fragments will only be amplified linearly.

6. After PCR amplification, the *BsaH*I-*Cel*I fragments will be cloned into a cloning vector.

7. Clones will be sequenced using *Cel*I adaptor primer to identify the nature of the sequence and location of the SNP.

A similar mutation detection method for large scale cloning of SNPs has been previously reported by Jiang *et al.* (129). In their study, the method involved the cleavage of chromosomal DNA by a four base pair cutter restriction enzyme, *Tsp*509I which were dephosphorylated prior to ligation with *EcoR*I adaptors. The DNA fragments were repaired by *E. coli* DNA polymerase I and T4 DNA ligase followed by hybridisation to form heteroduplex fragments. The duplex fragments were screened for mismatches by resolvase which produced staggered double-strand nicks within six nucleotides around the mismatches. Subsequently, these were treated with nuclease S1 and T4 DNA polymerase to create ligatable blunt DNA ends at the site of nicks. The fragments carrying SNPs were cloned. The major difference between this method and our design is that the use of different enzyme. *Tsp*509I is an enzyme which produces blunt end fragments while our selected enzyme will produce fragments with base overhang. Ligation of adaptors will be more specific to fragments with overhang bases. Moreover, our design is simpler to perform and requires less modification steps.

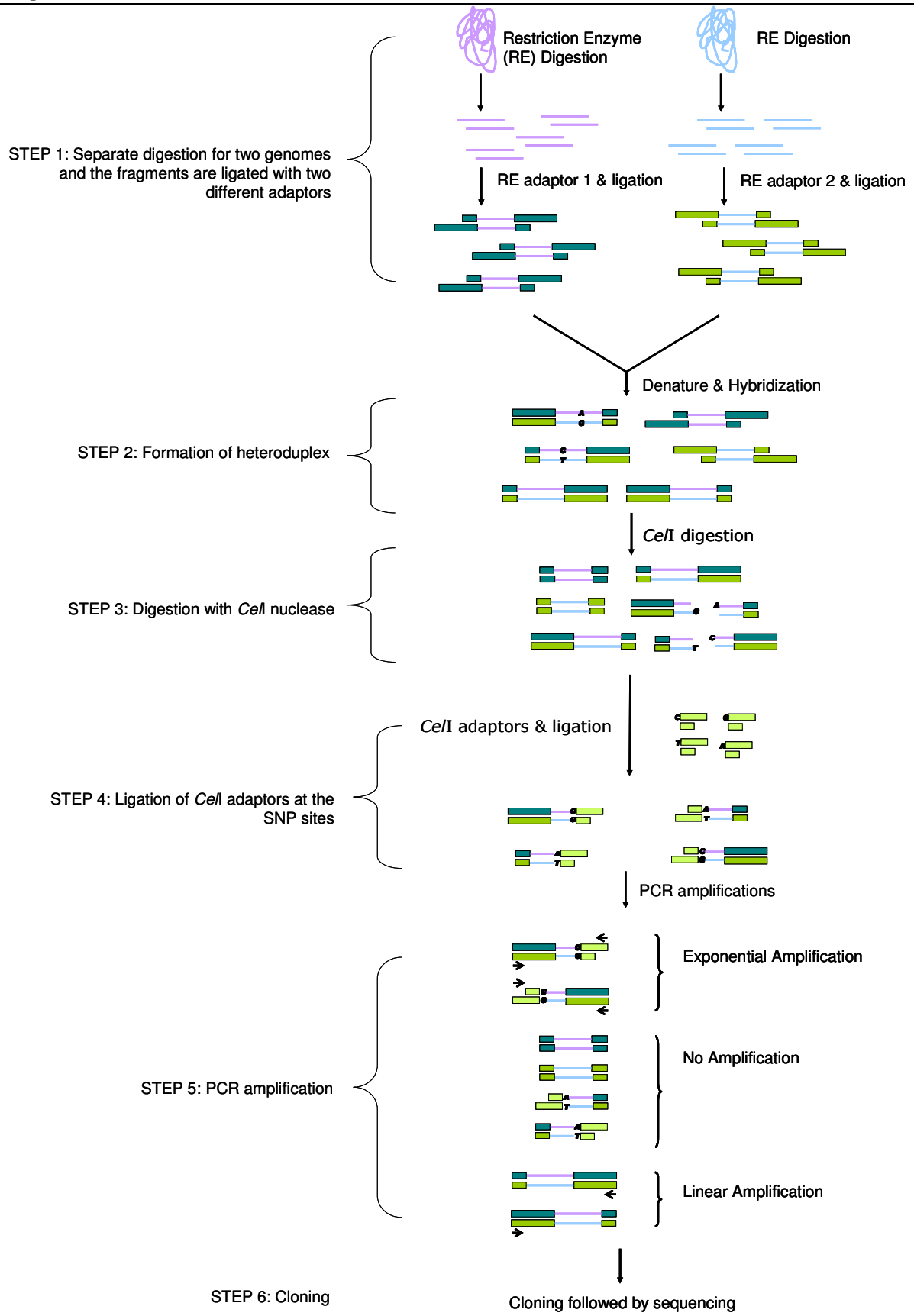


Figure 6.2-1. The principles of the proposed method using *CelI* nuclease. In this example, SNP A and C are shown.

6.3. Materials and Methods

6.3.1. Bacterial strains

Genome sequenced Typhi strains CT18 and Ty2 (Chapter 2, Table 2.1-2) were selected as the controls to test the feasibility of this method.

6.3.2. Preparation of adaptors

The adaptor consists of a pair of long and short oligonucleotides. These oligonucleotides were synthesised and purified by high-performance liquid chromatography method (Sigma-Aldrich). The oligonucleotides were diluted to 10 µM prior to use. These adaptors were prepared by mixing the pair of oligonucleotides. The mixture was heated at 96°C for 10 min which was slowly cooled at room temperature.

There are two *BsaH*I adaptors, one for each genome, and one *Cel*I adaptor. The first *BsaH*I adaptor contains complementary oligonucleotides 9281 (5'-ACTGCGTACT-3') and 9282 (5'-PCGAGTACGCAGTCTCTAGAAGTGAGGTTCATTACCATCCAGTCA-5'). The longer oligonucleotide contains a two-base overhang which complements the *BsaH*I digested fragments and the site for primer binding. The shorter fragment contains 10 nucleotides which are complementary to the 3' end of the longer fragment. The second adaptor contains oligonucleotides 9300 (5'-TGACTGGATGGTAATGAACCTCACTTCTAGAGACTGCGTACT-3') and 9301 (5'-PCGAGTACGCAGT-5'). In the second *BsaH*I adaptor, the shorter oligonucleotide contains a two-base overhang to complement the digested fragments. Ligation of the fragments requires the presence of 5' phosphate termini, hence an additional phosphate (P) is attached at the 5' end of one of the oligonicleotides on each pair.

The *Cel*I adaptor contains oligonucleotides 9288 (5'-PCTCAGGACTCATCGTTCAGTTGTAGCAATCAGGTACGT-3') and 9302 (5'-GAGTCCTGAGN-3'). The one-base overhang was designed as an N so all four possible

nucleotides, A, C, G and T, can ligate to the fragments containing SNPs, which have been cut with *Cel*I nuclease.

6.3.3. *BsaH*I digestion and adaptors ligation

Separate tubes were prepared for each sample, one tube for a different isolate, for digestion with *BsaH*I. The digested templates were ligated with *BsaH*I adaptor 1 for one and and adaptor 2 for another. Each tube contained 15 µl DNA (1 µg), 2 µl (20 U) *BsaH*I (New England Biolab), 10 µl 10 x one-phor-all buffer [100 mM Tris-acetate (pH 7.5), 100 mM magnesium acetate and 500 mM potassium acetate] (GE Healthcare), 2.5 µl T4 DNA Ligase (Promega), 5 µl 10 pmol/µl *BsaH*I adaptor (adaptor 1 or 2), 2.5 µl ATP, 5 µl BSA (New England Biolab), 2.5 µl polyethylene glycol and MilliQ water for a 50 µl reaction. The tubes were incubated overnight at room temperature. The T4 DNA ligase was inactivated by the addition of 1 µl of EDTA. The sample was purified by sodium acetate/ethanol precipitation.

6.3.4. Heteroduplex formation

Equal amounts of purified *BsaH*I digested and adaptors ligated DNA fragments (at least 200 ng total DNA) from two isolates were mixed in a 0.2 ml tube and appropriate amount of 10xTaq polymerase buffer (New England Biolab) was added to facilitate the hybridisation. The tube was placed in a thermocycler, as recommended by the manufacturer for Surveyor™ nuclease kit (Transgenomics), with the following conditions: 95ºC 2 min; 95ºC to 85ºC (decrement of 2ºC/sec); 85ºC to 25ºC (decrement of 0.1ºC/sec); and 4ºC hold.

6.3.5. *Cel*I digestion and adaptor ligation

After hybridisation, the DNA samples were treated with Surveyor™ *Cel*I nuclease (Transgenomics) in a 0.2 ml tube. Each was done in a 20 µl reaction, which contained 200-400 ng hybridised DNA, 2 µl of the reaction buffer and 1 µl each of Enhancer S and Nuclease S, supplied with the Surveyor™ Mutation Detection Kit (Transgenomics). The mixture was incubated at 42ºC

for 20 min followed by the addition of 1/10 volume of Stop Solution also provided in the kit. Subsequently, the digested samples were purified using sodium acetate/ethanol precipitation. One µl of tester DNA (120 ng) was ligated with 2 µl of *Cel*I adaptor in a tube containing 1 µl of T4 DNA ligase (4 units/µl) and 1 µl of 10X ligation buffer. The reaction was incubated at room temperature overnight. These were then loaded on the agarose gel to confirm the presence of cleavage activities. The amount of samples loaded on the gel was consistent between runs however, the contrast was maximised when using the gel viewing program, in order to see the resulting cleaved bands. Thus, the brightness of the uncleaved bands did not correspond to the DNA concentrations.

6.4. Results

6.4.1. Selection of enzyme used for genomic digestion

Digestion simulations of restriction enzymes with five to six recognition sites were done using the NIP tool available from the Australian National Genomic Information Service for the genome sequence of *S. enterica* serovar Typhi strain CT18. Among them, five are shown in Table 6.4-1. As a result, *BsaH*I was chosen because it has a two-base overhang. It also has the most number of fragments with sizes ranging between 100 bp – 3 kb.

Table 6.4-1. Total number of fragments produced by four restriction enzymes

Recognition Site	Enzyme Name	Total fragments	Number of fragments with 100 bp - 3 kb in size
Pu'AATTPy	*Apo*I	5779	4857
C'AATTG	*Mfe*I	638	212
G'AATTC	*Ecor*I	793	342
GPu'CGPyC	*BsaH*I	1727	1110

‘ – the base overhangs

Pu – purine (A, G) and Py – Pyrimidine (C, T)

6.4.2. The efficiency of BsaH*I* digestion/adaptors ligation

The gene fragment carrying SNP 24 was used to test the efficiency of *BsaH*I digestion and ligation to *BsaH*I specific adaptors. The fragment carries a *BsaH*I site (Figure 6.4-1), and primer pairs 9215/9216 (Chapter 4) were used to amplify the gene fragment for strains CT18 and Ty2. PCR was carried out in five 50 µl reactions for each strain and the products were purified by ethanol precipitation. The size of the PCR products was expected to be 475 bp. Two µl of the purified PCR products from these strains were verified by agarose gel electrophoresis (Figure 6.4-2), followed by digestion with *BsaH*I and ligation with *BsaH*I adaptors. Two fragments with sizes of 167 bp and 308 bp were expected to be observed. From the gel electrophoresis of the digested 9215/9216 gene fragments for CT18 and Ty2, the two fragments were approximately 167 and 308 bp in size respectively (Figure 6.4-2).

Following ligation with *BsaH*I adaptors, the fragment sizes were expected to be 209 bp and 350 bp respectively. The fragment with 209 bp in size was expected to contain the sequences of the lower primer, 9216, that amplified the gene fragment, while the fragment with 350 bp in size contained the sequences of the upper primer, 9215 (Figure 6.4-1). Both of the ligated fragments contained *BsaH*I adaptors. Primer 9298 which was specific to *BsaH*I adaptor was separately paired with primers 9215 and 9216 for PCR. Both *BsaH*I fragments were then amplified with expected sizes, suggesting that *BsaH*I digestion/ligation had been successful (Figure 6.4-2).

***BsaH*I digestion/ligation – CT18**

Product 1 (209 bp):

```
5′-AAACCGAGCGTGGGGCGATCTTTT————ATACCGACGAGTACGCAGTCTCTAGAAGTGAGGTTCATTACCATCCAGTCA-3′
3′-TTTGGCTCGCACCCCGCTAGAAAA————TATGGCTGCTCATGCGTCA-5′
```

Product 2 (350 bp):

```
                                  5′-ACTGCGTACTCGCCGTAGC———T———CTCACCAGCACGCCGTTTTCATAC-3′
3′-ACTGACCTACCATTACTTGGAGTGAAGATCTCTGACGCATGAGCGGCATCG———A———GAGTGGTCGTGCGGCAAAAGTATG-5′
```

***BsaH*I digestion/ligation (209/350) – Ty2**

Product 1 (209 bp):

```
5′-AAACCGAGCGTGGGGCGATCTTTT————ATACCGACGAGTACGCAGTCT-3′
3′-TTTGGCTCGCACCCCGCTAGAAAA————TATGGCTGCTCATGCGTCAGAGATCTTCACTCCAAGTAATGGTAGGTCAGT-5′
```

Product 2 (350 bp):

```
5′-TGACTGGATGGTAATGAACCTCACTTCTAGAGACTGCGTACTCGCCGTAGC———C———CTCACCAGCACGCCGTTTTCATAC-3′
                                  3′-TGACGCATGAGCGGCATCG———G———GAGTGGTCGTGCGGCAAAAGTATG-5′
```

Figure 6.4-1. Diagrammatic representation of *BsaH*I digestion of the fragments carrying SNP 24 in CT18 and Ty2. The two adaptors, adaptor I (CT18) and adaptor II (Ty2) are shown in bold. The yellow shades represent the recognition site between the adaptors and digested fragment. The grey shades represent the binding site for either primer 9215 or 9216. The green shades represent the binding site for primer 9298.

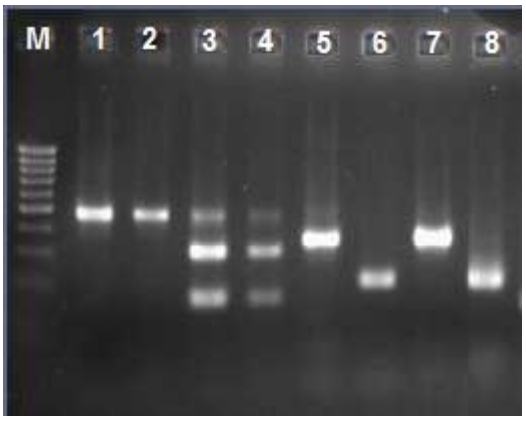


Figure 6.4-2. *BsaH*I digestion and adaptor ligation for DNA fragment amplified, using primer pairs 9215/9216 of Typhi strains CT18 and Ty2. Lane M-100 bp marker [from the largest size to the smallest size visible: 1 kb, 900 bp, 800 bp, 700 bp, 600 bp, 500 bp, 400 bp, 300 bp and 200 bp]; Lanes 1 and 2 - Undigested 9215/9216 gene fragment for CT18 and Ty2 respectively; Lanes 3 and 4 - *BsaH*I digested, and adaptors ligated for CT18 and Ty2; Lanes 5 and 7 - *BsaH*I digested fragments were ligated with the *BsaH*I adaptors and were amplified with primer pairs 9215 and 9298 for CT18 and Ty2;and Lanes 6 and 8 - *BsaH*I digested fragments were ligated with the *BsaH*I adaptors and were amplified with primer pairs 9216 and 9298 for CT18 and Ty2.

6.4.3. The effect of mismatches on the efficiency *of* Cel*I* digestion

Two controls were performed to test the efficiency of digestion. The first control was the gene fragments containing SNP 24 that were amplified in Typhi strain CT18 and Ty2. The second control was the two plamids (C and G), that were provided with the Surveyor^{TM} kit and these too were PCR amplified. The C and G controls were identical, except control C has a C instead of a G, at the 417 bp away from the start of the upper primer. The PCR products were confirmed using agarose gel electrophoresis before purification. The concentrations of these controls were adjusted to be proportionally equal (approximately 50 ng/µl), prior to hybridisation in a 10xTaq polymerase buffer. The hybridised C and G controls showed the presence of cleavage. However, for SNP 24, very faint bands of the expected cleaved fragments were observed on the agarose gel (Figure 6.4-3).

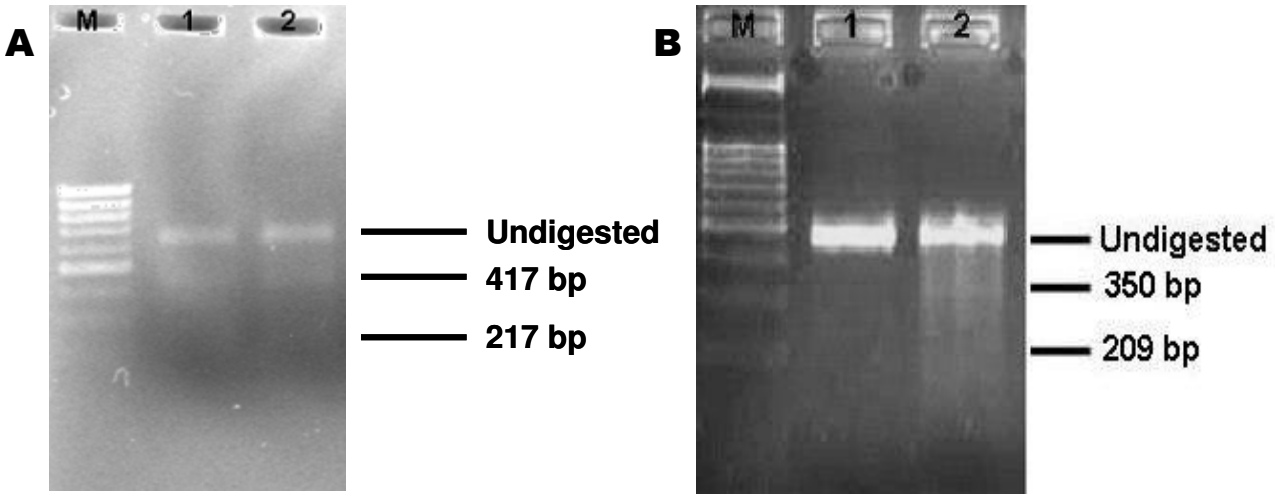


Figure 6.4-3. *Cel*I digestion. A. C/G controls provided with the kit loaded in duplicates: M-100 bp marker [from the largest size to the smallest size visible: 1 kb, 900 bp, 800 bp, 700 bp, 600 bp, 500 bp, 400 bp, 300 bp and 200 bp] and Lane 1 and 2 - Hybridised mixture of C/G control after treatment with *Cel*I nuclease. **B. SNP 24 control**: M-1 kb marker [from the largest size to the smallest size visible: 5 kb, 4 kb, 3 kb, 2 kb, 1.5 kb, 1 kb, 700 bp, 500 bp, 400 bp, 300 bp and 200 bp]; Lane 1 - Undigested gene fragment carrying SNP 24 and Lane 2 - Hybridised mixture after treatment with *Cel*I nuclease.

This could be due to the particular nature of the bases, which may affect the activity of *Cel*I. *Cel*I nuclease has been shown to have preferential substrates where the heteroduplexes resulted from

base mismatches of C/C > C/A ~ C/T > G/G > A/C ~ A/A ~ T/C > T/G ~ G/T ~ G/A ~ A/G > T/T (where > means more preferred and ~ means similar preference) (225). The controls within the kit fall into the top three preferred substrates (C/C and G/G). On the other hand, SNP 24 produced T/G and C/A heteroduplexes which were the less preferred substrates.

Four SNPs, SNP 8, SNP 9, SNP 10 and SNP14 from chapter 4, with different base mismatches were tested for this property (Table 6.4-2). SNP 10 and SNP 14 produced C/A and T/G heteroduplexes that fell into one of the most preferred substrates whilst SNP 8 and SNP 9 produced G/T and A/C heteroduplexes which were less preferred substrates for *Cel*I nuclease. The gene fragments carrying these SNPs, in Typhi strains CT18 and Ty2, were amplified using primer pairs 9137/9138, 9235/9236, 9193/9194 and 9223/9224, for SNP 8, 9, 10 and 14 respectively.

PCR products of each gene for the two strains were purified prior to hybridisation. The hybridised mixtures were incubated with 1 µl of *Cel*I nuclease for 20 min, followed by inactivation of nuclease. *Cel*I nuclease was expected to recognise and cleave the heteroduplexes into two fragments at 178 bp, 376 bp, 118 bp and 595 bp for SNP 8, 9, 10 and 14 respectively. Undigested homoduplexes fragment and two digested heteroduplexes fragments were expected to be observed for each gene carrying the SNP. The *Cel*I treated mixtures were then run and viewed on an agarose gel.

Out of the four SNPs tested, only two hybridised mixtures for gene fragments carrying SNP 8 and SNP 10, showed the presence of three fragments (Figure 6.4-4). The fragments corresponded to the expected sizes of the undigested fragment and the digested fragments. The remaining two mixtures for genes carrying SNP 14 and SNP 9 showed a single band on the agarose gel, suggesting no cleavage activity of *Cel*I nuclease. It was expected that gene fragments carrying SNP 10 and 14 would have the same cleavage activities. Both of these SNPs should be more optimally recognised and cleaved by *Cel*I nuclease than the gene fragments carrying SNP 8 and 9 based on the expected *Cel*I substrate preference. The results however showed otherwise. The substrate preference did exist but not the same as previously reported.

Table 6.4-2. Four SNPs selected to test the effect of substrate on the efficiency of *Cel*I digestion

Primer pairs	SNP No.	Heteroduplexes	Preferred substrate	Expected fragment sizes (bp)	Result[1]
9193/9194	10	C/A and T/G	Yes	118/284	Cut
9223/9224	14	C/A and T/G	Yes	595/256	Uncut
9137/9138	8	G/T and A/C	No	178/361	Uncut
9235/9236	9	G/T and A/C	No	376/152	Cut

[1] The results of the digestion could be seen on Figure 6.4-4

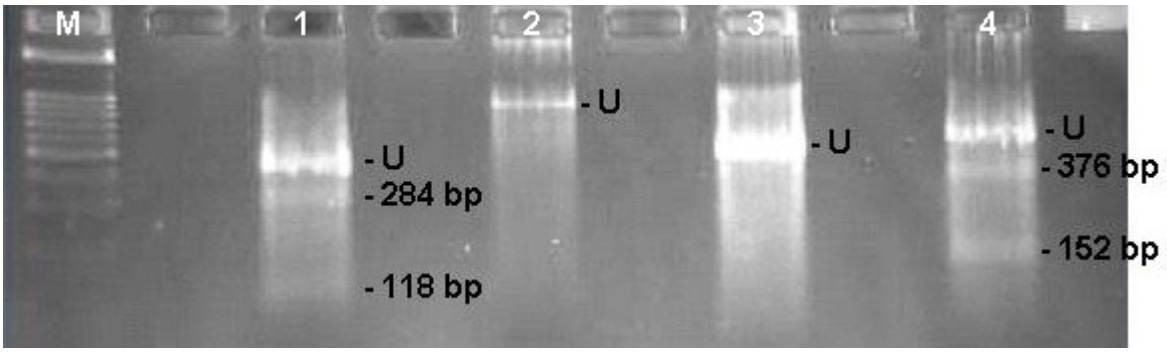


Figure 6.4-4. The specificity of *Cel*I nuclease on different heteroduplexes produced by four SNPs. Lane M – 1 kb marker [from the largest size to the smallest size visible: 5 kb, 4 kb, 3 kb, 2 kb, 1.5 kb, 1 kb, 700 bp, 500 bp, 400 bp, 300 bp and 200 bp] and Lanes 1-4 are PCR products of CT18 and Ty2 which have been hybridised and digested with *Cel*I. Lane 1- SNP 10; Lane 2 – SNP 14; Lane 3 – SNP 8; and Lane 4 – SNP 9. Undigested PCR products are indicated with a U while *Cel*I digested fragments are indicated by their expected sizes in bp.

6.4.4. Varying the composition of buffers, in particular salt concentration, results in an improved hybridisation and *CelI* activity

Although *Cel*I activity was apparent in a number of experiments, the amount of cleavage has not been satisfactory and it was not observed in all of the SNP controls. A number of factors were considered, one of which was the hybridisation buffer composition. The *Cel*I manufacturer recommendation for constituents of the buffer include: 20 mM Tris HCl, 50 mM KCl, 2 mM $MgCl_2$ and 0.1% Triton x 100. However, the recommended buffer did not work for SNP 8, 14 and 24.

Based on the recommendation from the technical support, the buffer was then changed to a 1.5xTaq polymerase buffer (15 mM KCl, 15mM $(NH_4)_2SO_4$, 30 mm Tris-HCl, 3.0 mM $MgSO_4$, and 0.15% Triton, pH 8.8). This buffer was used and it showed better cleavage for *Cel*I nuclease (Figure 6.4-5).

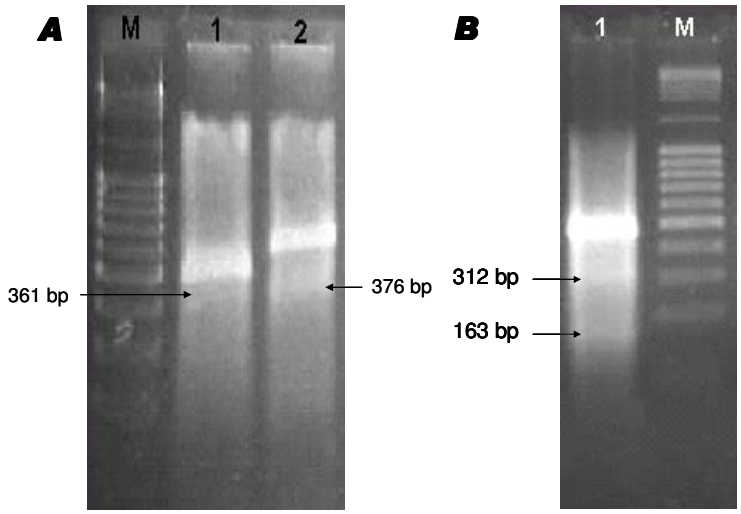


Figure 6.4-5. SNP controls digested with *Cel*I nuclease. A. SNP 8 and SNP 9. M - 1 kb marker [from the largest size to the smallest size visible: 5 kb, 4 kb, 3 kb, 2 kb, 1.5 kb, 1 kb, 700 bp, 500 bp, 400 bp, 300 bp and 200 bp]; Lane 1 – The digestion products of gene fragment carrying SNP 8 and Lane 2 – The digestion products of gene fragment carrying SNP 9. **B. SNP 24.** M - 1 kb marker [from the largest size to the smallest size visible: 5 kb, 4 kb, 3 kb, 2 kb, 1.5 kb, 1 kb, 700 bp, 500 bp, 400 bp, 300 bp and 200 bp] and Lane 1 – The digestion products of gene fragment carrying SNP 24. The arrows indicate the expected sizes of the bands that resulted from *Cel*I digestion.

6.4.5. Testing the efficiency of Cell adaptor ligation

Gene fragment, containing SNP 24, was selected to assess the effectiveness of each step. The expected sizes of the fragments obtained after *BsaH*I digestion and *BsaH*I adaptors ligation were 209 and 350 bp respectively. SNP 24 was located on the 350 bp fragment. *Cel*I nuclease will cleave this fragment into two smaller fragments with 187 bp and 163 bp in size respectively. The 187 bp also contained *BsaH*I adaptor, thus after subsequent ligation with *Cel*I adaptors, this fragment was expected to be 225 bp in size. To test whether *Cel*I digestion had occurred and whether the *Cel*I adaptor had successfully ligated, a PCR was done using primers specific to *BsaH*I and *Cel*I adaptors.

There were four bands, three of which were weak bands with sizes of approximately 350 bp, 200 bp and 180 bp respectively, as well as one bright band, which was approximately 100 bp in size (Figure 6.4-6). The band which was 350 bp in size could be the result of *Cel*I adaptors which randomly ligated to the end of the other fragments. It was likely that the 200 bp band was the

product from the *Cel*I digestion. The bright 100 bp band was most likely to be adaptor-adaptor ligated fragments. The products were then purified using ethanol precipitation for cloning.

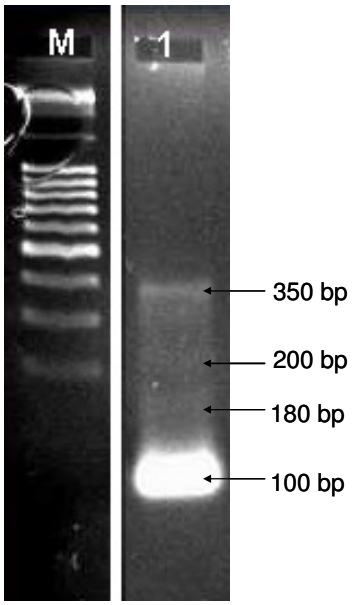


Figure 6.4-6. The PCR amplification of *Cel*I fragments using *BsaH*I adaptor specific and *Cel*I adaptor specific primers. The arrows indicate the approximate size of each band.

6.4.6. Cloning and sequencing of the transformants

The purified PCR product was cloned into *E. coli* strain DH5α using pGEMT as a vector and then plated onto the selective/differential LB plates. Forty-two white colonies were selected and the inserts were amplified using M13 primers (Figure 6.4-7). The PCR product included part of the vector sequenced (193 bp) and the expected *BsaH*I-*Cel*I fragment for a total size of approximately 400 bp. However, the size of the bands from PCR amplification of these clones was approximately 300 bp (Figure 6.4-7). One clone showed a mixture of three bands with approximately 250 bp, 300 bp and 400 bp in size, suggesting the presence of multiple inserts.

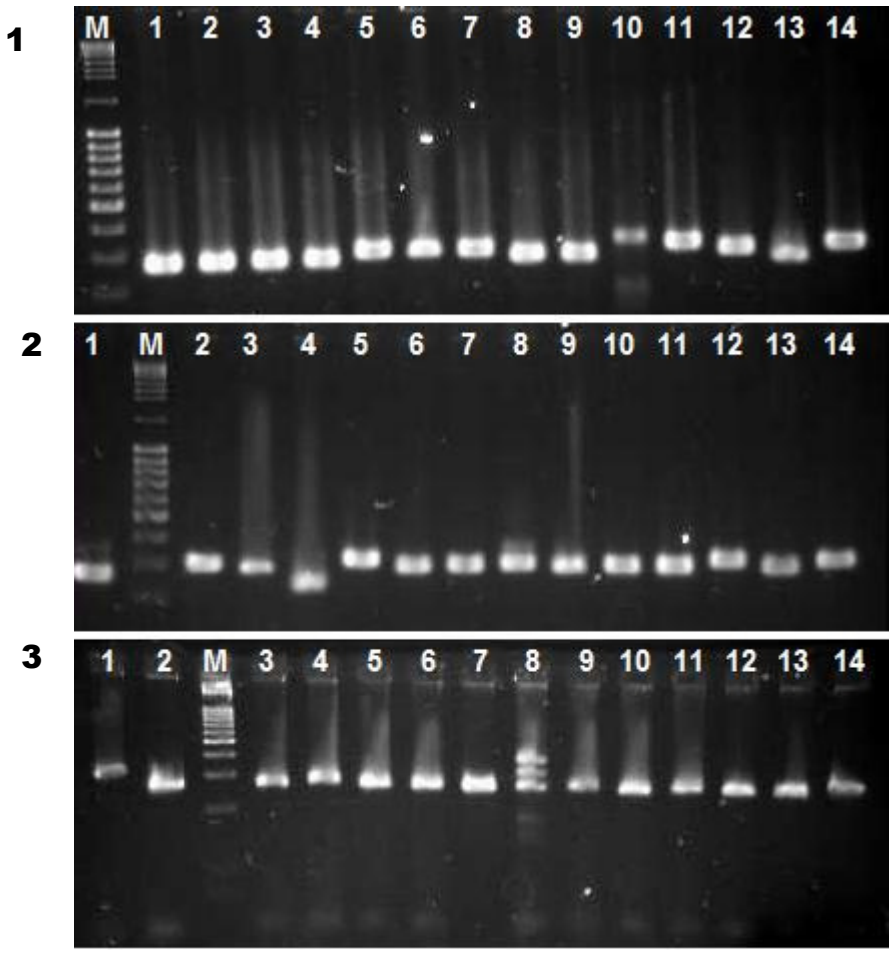


Figure 6.4-7. PCR products of the clones amplified using M13 primer sequence

To find out the nature of the inserts, a subset of PCR products were sequenced using the M13 primer. PCR products of different sizes (approximately 300-400 bp in size) from lanes 1, 5, 10 and 11 on gel one were selected. Sequencing results suggested that all the inserts were *BsaH*I adaptor sequences that were ligated with *Cel*I adaptors (Table 6.4-3). None of the sequenced products matched the *BsaH*I-*Cel*I fragments.

Table 6.4-3. The size of *BsaH*I and *Cel*I adaptors that were ligated and cloned into the vector

Sample[1]	Size (bp)[2]	
	***BsaH*I adaptor**	***Cel*I adaptor**
1	34	20
5	40	32
10	40	27
11	34	20

[1] Sample name corresponds to the selected lane from gel one

2 The expected sizes for *BsaH*I and *Cel*I adaptors were 44 and 38 respectively

6.5. Discussion

6.5.1. The ligation of *BsaH*I adaptors was successfully demonstrated

We attempted to develop an enzyme based method using the mismatch specific *Cel*I nuclease for SNP discovery. There were six steps involved: the digestion of the genomes with *BsaH*I restriction enzyme followed by ligation to two different adaptors; formation of heteroduplex; digestion with *Cel*I nuclease; ligation of *Cel*I adaptors at the SNP sites; PCR amplification using primers specific to *BsaH*I and *Cel*I adaptors; and finally cloning and identification of the SNPs. Each step was tested for feasibility by using a gene fragment that differed by a single SNP between CT18 and Ty2. The gene fragment carrying SNP 24 was selected for this purpose. There was one *BsaH*I restriction enzyme cut site on the gene fragment harbouring SNP 24 and hence two fragments were expected to be seen. The approximate sizes of the digested/ligated fragments on an agarose gel were as expected suggesting that they have been successfully digested with *BsaH*I restriction enzyme and ligated with the corresponding *BsaH*I adaptors.

6.5.2. Factors affecting the cleavage activity of CelI nuclease

We had difficulties in demonstrating the efficiency of *Cel*I nuclease. Different batches of enzymes were tested and the efficiency of *Cel*I nuclease was lower than expected. The plasmid controls C

and G provided with the kit showed a very weak cleavage activity while the SNP control using SNP 24 did not work. Several key factors were examined, including varying the concentration of *Cel*I nuclease and enhancer, and altering the length of incubation time for digestion with *Cel*I nuclease. The efficiency of digestion was not improved by increasing the amount of *Cel*I nuclease or incubation time.

The effect of the substrate on the efficiency of *Cel*I nuclease was also considered. Previously, it has been noted that the nuclease has mismatch cutting preferences of CT, AC, CC, TT, AA, GG, AG and GT (from most to least preferred) (250). The substrate preferences for *Cel*I nuclease were tested using four SNPs that produced different heteroduplexes. Two of the SNPs produced heteroduplexes which were preferred substrates for *Cel*I nuclease while the remaining two did not. Our results showed that the cleavage activity was not always observed in the heteroduplexes where base mismatches were considered to be the more preferred substrates for *Cel*I nuclease. This suggested that cleavage preference for the *Cel*I nuclease was not consistent with previous report by Qiu *et al.* (250). It was possible that the neighbouring bases within the sequence may affect the efficiency of the digestion.

6.5.3. The usefulness of *CelI* nuclease strategy for SNP discovery could not be determined due to its poor cleavage activity

SNP 24 contained a *BsaH*I site making it suitable as a control. The gene fragment containing this SNP was tested for the whole procedures. Sequencing results showed that the inserts from four clones were ligated adaptor sequences of *BsaH*I and *Cel*I, but none of these clones contained the expected *Cel*I fragments. Based on all the steps tested, the major failure was caused by the weak nuclease activity of *Cel*I. It seems that the *BsaH*I adaptor-*Cel*I adaptor ligation was amplified more efficiently than the intended SNP fragments. The adaptor-adaptor ligated fragments were amplified at a much higher rate as observed by the presence of a thick band at 100 bp. This could be due to the fact that the SNP fragments were present at much lower concentration than the *BsaH*I/*Cel*I adaptor-adaptor ligation. This *BsaH*I/*Cel*I adaptor-adaptor ligation was present at a high

concentration in the PCR products used for cloning. This may explain why we were not able to obtain an insert with the expected size, which corresponded to the SNP fragment.

6.6. Conclusion

This was the first time the enzyme *Cel*I nuclease was utilised for cloning purposes to discover more SNPs. Previous studies have shown that *Cel*I was useful for SNP detection (15, 163, 251, 291, 292). We have also demonstrated that *Cel*I can also be used for that purpose as shown for SNP 8, 9 and 24. However, the cleavage activity of *Cel*I nuclease has been inefficient for our purpose to discover new SNPs. An improved *Cel*I nuclease is required to further verify our proposed method. This project was discontinued because the difficulties in optimising the cleavage activity of *Cel*I nuclease and due to time constraints. However, we still believe that there is a potential for the mismatch cleavage method for discovering SNPs effectively and economically.

Chapter 7: Multiple locus variable number of tandem repeat analysis of *S. enterica* serovar Typhi

7.1. Introduction

The genetic homogeneity of *Salmonella enterica* serovar Typhi has significantly impeded the reconstruction of its evolutionary history. Multilocus Enzyme Electrophoresis (MLEE) (285) and Multilocus Sequence Typing (MLST) (141) have clearly shown that Typhi is highly homogoneous. MLEE analysis of 334 Typhi isolates identified two electrophoretic types (ETs). The ETs were distinguished from one another by two of the 24 metabolic enzymes analysed (285). MLST showed slightly higher discrimination. Three of the seven housekeeping genes investigated by MLST were polymorphic and four sequence types were identified amongst the 26 Typhi isolates analysed (141). These methods showed insufficient variations, for them to be useful for establishing the genetic relationships between Typhi isolates or for epidemiology studies.

Complete sequencing of the genomes of Typhi strains CT18 and Ty2 allowed us to explore the use of Single Nucleotide Polymorphisms (SNPs) as markers for typing and to determine the relationships among global Typhi isolates (Chapter 4). Seventy three isolates were divided into four major clusters using 38 SNPs. These SNPs have distinguished the isolates into 23 SNP profiles, from which 12 were represented by a single isolate while 11 others were by more than one isolates. It was demonstrated that SNP typing had a considerably higher discriminatory power in comparison to MLST. SNPs have also been utilised as a marker in a similar approach by Roumagnac *et al.* (271). Eighty eight SNPs referred to as biallelic polymorphisms (BiPs) in their study, were typed in 481 global Typhi isolates, which were differentiated into 85 haplotypes.

The addition of four BiPs into our SNP typing scheme has differentiated the 73 isolates into 27 SNP profiles, thus an increased discrimination (Chapter 5). However, SNP typing clearly has limited discriminatory power. Some of the SNP profiles still contained many isolates which could not be differentiated, For example, SNP profiles 2b and 10b that consisted of 11 and 15 isolates

respectively. This has led us to test another type of molecular marker, variable number of tandem repeats (VNTR). VNTRs are short sequence repeats which are unique DNA elements repeated in tandem. The polymorphisms in VNTRs are believed to be resulted from slippage strand misalignment (162). Therefore, isolates may contain different copy numbers for a repeat locus, allowing differentiation between isolates.

Multiple locus variable number of tandem repeat analyses (MLVA) involves determination of the number of repeats at multiple VNTR loci. MLVA is currently being investigated for its epidemiological usefulness in several species of bacteria and has been particularly effective to type homogeneous clones including *Y. pestis* (4, 144, 178, 244), *B. anthracis* (91, 126, 138, 295) and *M. tuberculosis* (88, 157, 281, 296, 302). The completed genome sequences of serovar Typhimurium strain LT2 and serovar Typhi strains CT18 and Ty2 have allowed many VNTRs to be identified. A few studies have been done to develop MLVA for typing various serovars of *S. enterica* including for serovar Typhi (164, 165, 174, 257).

Two separate MLVA studies of Typhi have been done to determine the polymorphisms of selected VNTRs loci (174, 257). Liu *et al.* (174) utilised five VNTR loci designated as TR1 to TR5, but only three VNTRs were useful including TR1, TR2 and TR3 loci. These three VNTRs have shown substantial genetic heterogeneity among the 59 Typhi strains from several Asian countries isolated between the years 2000-2001. The second study by Ramisse *et al.* (257), examined six new markers including SAL02, SAL06, SAL10, SAL15, SAL16 and SAL20. These markers could distinguish 27 Typhi strains isolated in France, between the years 1993-1999 into 25 MLVA profiles.

VNTR-based genotyping relies on the amplification of the VNTR locus using primers binding to the flanking sequences. After PCR, standard agarose gel electrophoresis is usually employed to assess the variation in tandem repeats. The method for detecting the difference in copy number has been increasingly sophisticated, including the use of capillary electrophoresis that has now been widely used in MLVA assay (165). Multicoloured fluorescent dyes allow sample pooling of multiple VNTR loci to be run simultaneously by capillary electrophoresis. Recently, MLVA assays have also been developed using real time PCR technology (247, 298).

In this study, tandem repeats from CT18 genome were surveyed for new potential polymorphisms that have not been previously identified. Together with the seven published VNTRs, including SAL02, SAL06, SAL10, SAL16, SAL20, TR1 and TR2, these were used as markers to explore its potential in studying the molecular evolution of Typhi isolates. The MLVA assay exploited universal M13-tailed primer tagged with four different fluorescent dyes to resolve the tandem repeats on capillary electrophoresis. We aimed to combine the more rapidly evolving VNTR markers with the slowly evolving SNPs to achieve an optimal resolution for typing of Typhi isolates.

7.2. Materials and Methods

7.2.1. Bacterial strains

The seventy three Typhi isolates which were typed using MLVA are listed in Chapter 2 Table 2.1-2 in the general materials and methods section.These were the same isolates used for the SNP typing (Chapter 4 and Chapter 5).

7.2.2. Identification of new VNTR markers

A publicly available database (http://minisatellites.u-psud.fr) was used to identify tandem repeats in the genomic sequence of Typhi strain CT18. The search for VNTRs was only run for Typhi strain CT18 under these two different parameters: 1) total length is more than or equal to 20 bp, unit length is more than or equal to 3 bp, copy number is more than or equal to 3 and repeats at least 80% match; and 2) total length is more than or equal to 20 bp, unit length is 2 bp, and repeats are at least 50% match. Not all copies in a repeat contain a perfectly matched motif. Repeat motifs may contain base substitutions, insertions or deletions. The percentage match of the repeat motif is calculated based on the total matched nucleotides from the total nucleotides in a repeat (Figure 7.2-1). The VNTR markers were named according to their chromosomal locations.

Figure 7.2-1. Non-perfect repeat in a potential VNTR locus. Base substitutions (indicated by stars above the bolded nucleotide) are illustrated in the two of the six copies of the tandem repeat. The % match is determined by calculating the number of matched nucleotides against the total nucleotides in the repeat. In this case, the repeat is of 83% match (25 matched nucleotides and 5 mismatched nucleotides).

7.2.3. MLVA typing

To screen for variation of the 53 potential VNTRs, PCR primers targeting the regions flanking the loci containing the repeats were designed (Table 7.2-1). Instead of direct PCR amplification of the loci containing VNTRs followed by gel electrophoresis on agarose as done by Liu *et al.* (174) and Ramisse *et al.* (257), this study utilised universal primer containing M13 sequences attached with a fluorescent dye. Each VNTR typing was run like a normal PCR reaction but three primers were included in the PCR mix: a "M13-tailed" forward primer, a reverse primer, and a dye labelled primer which is complementary to the M13 tail. Four dyes were used for pooling purpose in a single ABI 3730 capillary run at the same time, which were FAM (blue), VIC (green), NED (yellow) and PET (red). This approach allowed four VNTRs to be typed at the same time (279) (Figure 7.2-2).

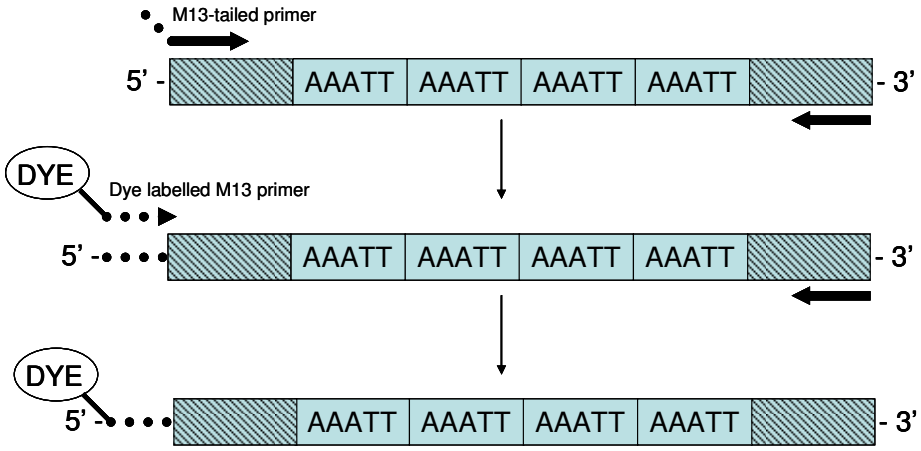
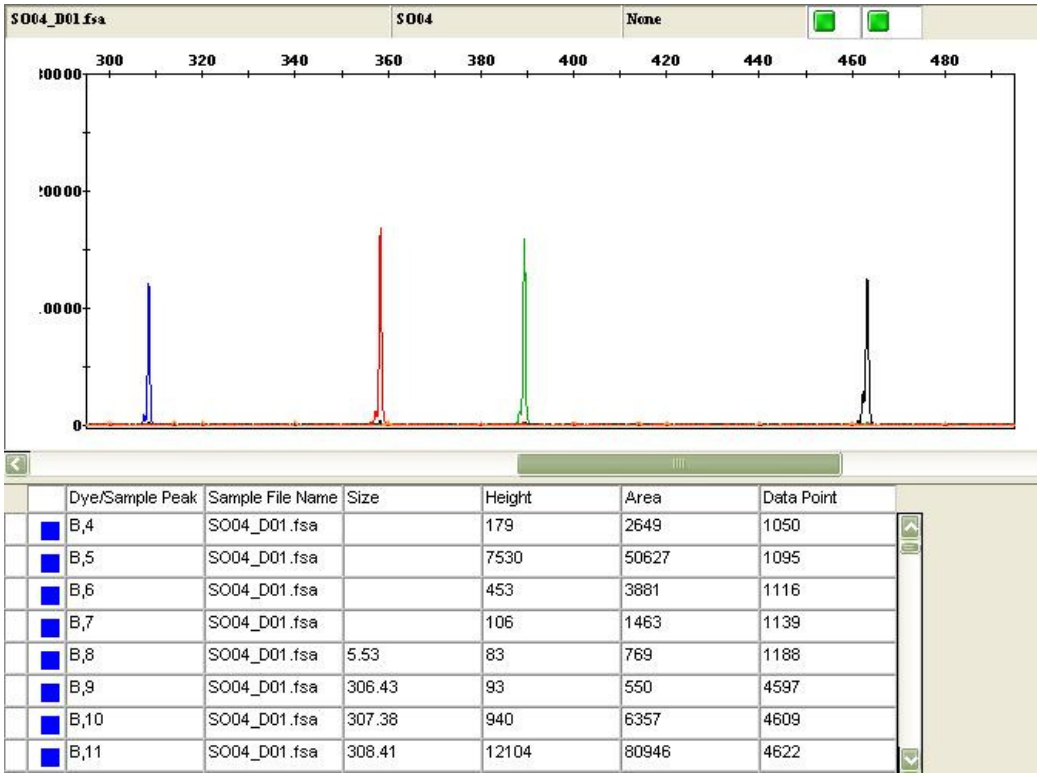


Figure 7.2-2. The diagramatic representation of the strategy used for PCR amplification of a VNTR. The striped box corresponds to the region upstream and downstream of the VNTR site. The dotted line corresponds to the M13 tail and the fluorescent dye could be either VIC, PET, FAM or NED.

The PCR was done in 20 µl reaction containing 0.2 µl DNA template (~10 ng), 50 nM, 200 nM and 250 nM for forward, M13 and reverse primers respectively, 0.2 µl 10 mM dNTPs, 2 µl 10xTaq polymerase PCR buffer (New England Biolabs), 0.125 µl (1.25 U) Taq polymerase (New England Biolabs) and MilliQ water to adjust to the final volume. The PCR conditions included a touch down cycling profile as the following: Initial denaturation 95°C for 5 min; 96°C for 1 min, 68°C for 5

min (-2°C/cycle, a decrease of 2°C after each cycle) and 72°C for 1 min for 5 cycles; 96°C for 1 min, 58°C for 2 min (-2°C/cycle) and 72°C for 1 min for 5 cycles; 96°C for 1 min, 50°C for 1 min and 72°C for 1 min for 25 cycles; and final extension at 72°C for 5 min.

The PCR products for four VNTRs of different fluorescent dyes were pooled and run as one sample on an Automated GeneScan Analyser ABI3730 (Applied Biosystem) at the sequencing facility of the School of Biotechnology and Biomolecular Sciences, the University of New South Wales. The fragment size was determined using the LIZ600 size standard (Applied Biosystem) and analysed using GeneMapper v 3.7 software (Applied Biosystem) (Figure 7.2-3). Variation in fragment size of each VNTR was scored as a new allele and each representative allele was sequenced to confirm that the size difference was genuinely due to the difference in copy number and not artifact in the assay.



Dye/Sample Peak	Sample File Name	Size	Height	Area	Data Point
B,4	SO04_D01.fsa		179	2649	1050
B,5	SO04_D01.fsa		7530	50627	1095
B,6	SO04_D01.fsa		453	3881	1116
B,7	SO04_D01.fsa		106	1463	1139
B,8	SO04_D01.fsa	5.53	83	769	1188
B,9	SO04_D01.fsa	306.43	93	550	4597
B,10	SO04_D01.fsa	307.38	940	6357	4609
B,11	SO04_D01.fsa	308.41	12104	80946	4622

Figure 7.2-3. An example of a sample generated from GeneScan run. There are four VNTRs in one sample. Each peak corresponds to a VNTR. Blue represents FAM-labelled, red for PET-labelled, green for VIC-labelled and black for NED-labelled PCR products respectively. The LIZ600 standard appears as tiny orange peaks and the peaks are 20,

40, 60, 80, 100, 114, 120, 140, 160, 180, 200, 214, 220, 240, 250, 260, 280, 300, 314, 320, 340, 360, 380, 400, 414, 420, 440, 460, 480, 500, 514, 520, 540, 560, 580 and 600 bps (Applied Biosystems). Each peak gives a certain size which corresponds to the size of the PCR product in bp (indicated on the X-axis) and the height corresponds to the amount of signal of the corresponding peak (indicated on the Y-axis).

Table 7.2-1. Primers used for VNTR typing

VNTR Name	Positions on the chromosome[1]			Primer Sequence 5' -> 3'
	Start	Finish		
33	33883	33902	U	cacgacgttgtaaaacgac[2]TGCTGGATACTGTGTTGATG
			L	TAAGGCGTACTGCTCTTGA
88	88000	88020	U	cacgacgttgtaaaacgacATCTCTTTTTGGCATAGGG
			L	GAACTGGCACGCTAAATAA
89	89849	89874	U	cacgacgttgtaaaacgacTGAAGCAGGACACGGTCGTT
			L	AATACCCGTCTTCTCTTCCA
131	131320	131340	U	cacgacgttgtaaaacgacTTACGGTTGCTGGTGGA
			L	GCTGGCTTATATGGGCTT
322	322437	322457	U	cacgacgttgtaaaacgacTGGACGGTGATGTGAAAG
			L	TGTCGCTGTTCCCCAGA
325	325814	325835	U	cacgacgttgtaaaacgacCTGGTGGCTATGATTGGT
			L	CCTGTTTTCGGACTGACG
392	392259	392288	U	cacgacgttgtaaaacgacGAATGTCACGTTAAGCGG
			L	TTAGGCGGCGGTTGGCGG
853	853288	853307	U	cacgacgttgtaaaacgacCGGGCGAGGTAAAAATCA
			L	CTATCTGCGGGCGGTGTC
1070	1070419	1070443	U	cacgacgttgtaaaacgacAGTGAAGTGCTTGGTCTGT
			L	ATCGCAACCCGTCTGGAA
1305	1305593	1305614	U	cacgacgttgtaaaacgacTGGCGGCTAATGAAAACG
			L	TGGCAATCACCGTAGCAA
1400	1400723	1400742	U	cacgacgttgtaaaacgacCTCGGTTTCCCAGATACA
			L	ATGCGGTGATGTTCTCCA
1736	1736093	1736113	U	cacgacgttgtaaaacgacATAACACGCCCTGATAGC
			L	GTCCTTCCGTCACCAAAC
1754	1754938	1754957	U	cacgacgttgtaaaacgacGCAGCCCAATGATACGA
			L	GCCAGAATAGCAAGAACG
1773	1773334	1773355	U	cacgacgttgtaaaacgacCTACCGTGGATTTCTGGC
			L	AGTCTGGACCCGATGGAA

1794	1794548	1794570	U	cacgacgttgtaaaacgacTCATGTCGAACTGCTGCC
			L	CAAAACCTGTCGTATTCT
1882	1882677	1882696	U	cacgacgttgtaaaacgacTGTCGGTTGGGTATTTCT
			L	ACTGGTTGTTGCGTGATT
2086	2086054	2086073	U	cacgacgttgtaaaacgacGTCGGCGTAGTCGTCAA
			L	GATGTCGTGGGCAGAGAT
2126	2126315	2126342	U	cacgacgttgtaaaacgacAACCCTGATGAACGTGATAG
			L	TCATCGGCGTTGGGAGTGTG
2156	2156354	2156378	U	cacgacgttgtaaaacgacTCGCCATACTCCACCACC
			L	GCCAACGACTATGTGTGC
2305	2305346	2305369	U	cacgacgttgtaaaacgacGCTGTTGTTCTGGCATTT
			L	CCTTCCTCATCGACGGCA
2521	2521851	2521870	U	cacgacgttgtaaaacgacATCCGTATATTCCCACCA
			L	AGCCGTTACCGACCGCCT
2625	2625517	2625537	U	cacgacgttgtaaaacgacATCAATAAAGCGAATACC
			L	TTGCACTGATGTTTGTCC
2634	2634594	2634613	U	cacgacgttgtaaaacgacCACCACAGCGGACCATC
			L	TCAGCAGCGTTCGTTTCT
3035	3035585	3035605	U	cacgacgttgtaaaacgacGTTCATTACGCATCACAA
			L	CGGATGTACTGGAAACCT
3311	3111715	3111738	U	cacgacgttgtaaaacgacTTTGATGACCTGAGAACC
			L	AAGCGGAAGAAGAAGCAC
3325	3325781	3325811	U	cacgacgttgtaaaacgacATCAGCGTATTCAGGTTT
			L	CATTGGTCATCGTATTTG
3338	3338374	3338397	U	cacgacgttgtaaaacgacTTGTCGGGCTGCTCGGTC
			L	CAGGTTATCAAATACGCT
3369	3369113	3369135	U	cacgacgttgtaaaacgacTATCCGTTCATTCAAGGT
			L	CGACTTTTACGCTGACGA
3404	3404661	3404681	U	cacgacgttgtaaaacgacGCGTGCCGACCGTACCAG
			L	CGATGTGGCGTGGATTTT
3463	3463849	3463868	U	cacgacgttgtaaaacgacGAGTATGGCGAGTTGTTT
			L	GGGCGTTTTTAGTTTTCA

3509	3509161	3509182	U	cacgacgttgtaaaacgacACAGGCGGAAGGTTTCGT
			L	CGGATATTAAAGTCGATG
3666	3666066	3666089	U	cacgacgttgtaaaacgacGCCTGCGGTGTGAGTTGC
			L	CACACGTTTCGCGGTATT
3699	3699824	3699843	U	cacgacgttgtaaaacgacGTCTGGTCTCGGCAACA
			L	CGTCGGCATCGTCTTTA
3835	3835379	3835402	U	cacgacgttgtaaaacgacTGCGGGCGATTTGAATGC
			L	TCATACACCCTGCGGCG
4030	4030409	4030432	U	cacgacgttgtaaaacgacTACTTCGGCGGTCTGGTC
			L	ATAATCGCGGTTCACCTC
4067	4067718	4067738	U	cacgacgttgtaaaacgacGGGGTCAGGGCGTCAGA
			L	CACACTTTGCCACTCAGGT
4107	4107596	4107616	U	cacgacgttgtaaaacgacCGGCAAAATCTTCTTCAG
			L	AGAGTTGAATCGGGAAGT
4207	4207838	4207864	U	cacgacgttgtaaaacgacTGCTTTATTTCTGCTCTCTT
			L	ATTCGGTTTCTTCTGCTG
4295	4295722	4295743	U	cacgacgttgtaaaacgacATCATCGTCGTAAGTGCG
			L	TGCCCTTCTCCCTTCGTT
4353	4353030	4353052	U	cacgacgttgtaaaacgacCTTCCTTTACCCCTTATG
			L	ATTCCCACAACCGTTACC
4500	4500182	4500203	U	cacgacgttgtaaaacgacCGTTGCTGCTCCGAAAT
			L	GCGGTGAAGTGGAAAAAG
4546	4546242	4546261	U	cacgacgttgtaaaacgacGGAGCAGATAGCCAGAAC
			L	CGGGTAGCGTCATCCAG
4578	4578239	4578259	U	cacgacgttgtaaaacgacCTGGCGAAGTGGTGGAA
			L	GCGAGACCGAACTGATAC
4585	4585683	4585709	U	cacgacgttgtaaaacgacCGGTGTTGTGCTGTCTC
			L	CGTGCCACTTTATCTTCA
4694	4694785	4694816	U	cacgacgttgtaaaacgacCACCAACTCATCACCACT
			L	CGGACTCAAAACACAACAT
4699	4699223	4699322	U	cacgacgttgtaaaacgacTATTCTACTTCAGTCCCCCC
			L	AACCTCCCTGTATCTACCAA

TR1[3]	2017168	2017251	U	cacgacgttgtaaaacgacAGAACCAGCAATGCGCCAACGA
			L	CAAGAAGTGCGCATACTACACC
TR2[3]	2557000	2557222	U	cacgacgttgtaaaacgacCCCTGTTTTTCGTGCTGATACG
			L	CAGAGGATATCGCAACAATCGG
SAL02[3]	666037	666093	U	cacgacgttgtaaaacgacGGAAAGACTGGCGAACAAAT
			L	TCGCCAATACCATGAGTACG
SAL06[3]	764489	764518	U	cacgacgttgtaaaacgacTTGGTCGCGGAACTATAACTG
			L	CTTCGTCTGATTGCCACTCC
SAL10[3]	2016742	2016765	U	cacgacgttgtaaaacgacAAGCGACGTTCTTCTGCAAC
			L	TGGAATATGATGGCATGACG
SAL16[3]	3041486	3041568	U	cacgacgttgtaaaacgacCCATGGCTGCAGTTAATTTCT
			L	TGATACGCTTTTGACGTTGC
SAL20[3]	3643794	3643841	U	cacgacgttgtaaaacgacCAGCCGACACAACTTAACGA
			L	ACTGTACCGTGCGCGTTT

[1] The location of the VNTR on the chromosome of *S. enterica* serovar Typhi strain CT18

[2] cacgacgttgtaaaacgac is the M13 tail attached at the 5' end of the upper primers

[3] Primers designed from previously published literatures with additional of 18-mers M13 tail at 5' end of the upper primer

7.2.4. Sequencing for confirmation of predicted copy numbers in VNTR loci

The DNA fragments harbouring the repeats were amplified in isolates representing different sizes according to GeneScan analysis. The primers used for PCR amplification were identical to the pairs used for MLVA typing without the M13 dye labelled forward primer added in the PCR reaction. The PCR products were viewed on 2% agarose gel in TBE buffer and purified by sodium acetate/ethanol precipitation. One µl of the PCR product was used in sequencing reaction and amplified using unlabelled M13 forward primer (5'-CACGACGTTGTAAAACGAC-3'). The copy numbers were determined by counting the presence of repeat unit in the sequence trace file.

7.2.5. Bioinformatic analysis

Dendogram to represent the genetic relationships of the MLVA profiles was done using unweighted pair group method with arithmetic means (UPGMA) which was available in PHYLIP package. This package was accessible through the Australian National Genomic Information Service (ANGIS). A minimum spanning tree (MST) was also generated using Arlequin v. 3.1 available from http://cmpg.unibe.ch/software/arlequin3. The Simpson's index of diversity (D value) was calculated for each VNTR locus, using an in-house program, MLEECOMP package (249).

7.3. Results

7.3.1. VNTR markers selected for typing

The VNTR database resource accessible from http://minisatellites.u-psud.fr was used to explore the complete genome sequence of serovar Typhi strain CT18. A total of 174 potential VNTRs were found which have a total length of more than or equal to 20 bp, unit length of more than or equal to 3 bp, copy number of more than or equal to 3 and repeats of at least 80% match. The proportion of the tandem repeats for each categories: the % match; number of copies; and the unit length, is presented as pie charts (Figure 7.3-1). The majority of the repeats was only 80% match (Figure 7.3-1, A). Out of the 174 possible VNTRs, the copy numbers of the repeats ranged from three to 28 copies with most of the repeats were only three copies (Figure 7.3-1, B). The unit length for the repeat ranged from three to 273 bp and 30% of the 174 potential VNTRs had unit length of 6 bp (Figure 7.3-1, C). Forty three of 174 repeats were chosen because they had the highest copy numbers, highest overall % match and at least two perfect copies.

Fifty seven additional VNTRs with a total length of more than or equal to 20 bp, unit length of 2 bp, repeats were at least 50% match and spread around the CT18 chromosome. The % match was reduced to 50% as no repeat was identified if the parameter was set to 80%. The distribution of these repeats for each criteria is summarised in Figure 7.3-2. The majority of these repeats had 12 copies (Figure 7.3-2, A). These were further refined to include only those with five continuous perfect copies. Only three repeats were selected resulting in a total of 46 VNTRs surveyed for possible polymorphisms in Typhi.

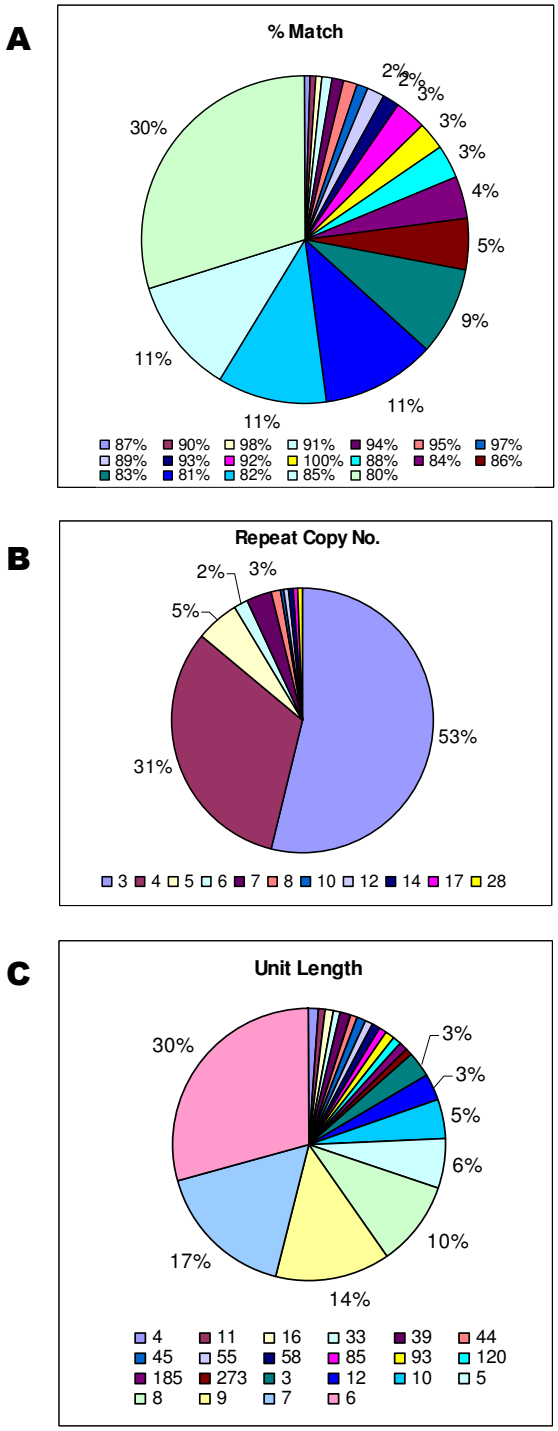


Figure 7.3-1. The proportion of repeats for different categories. **A.** % match, **B.** copy numbers and **C.** unit length, presented as percentage from the 174 possible VNTRs identified with the parameter of: total length more than or equal to 20 bp; unit length more than or equal to 3 bp; copy number of more than or equal to 3; and repeats of at least 80%. Percentage of 1 was not displayed on the pie charts.

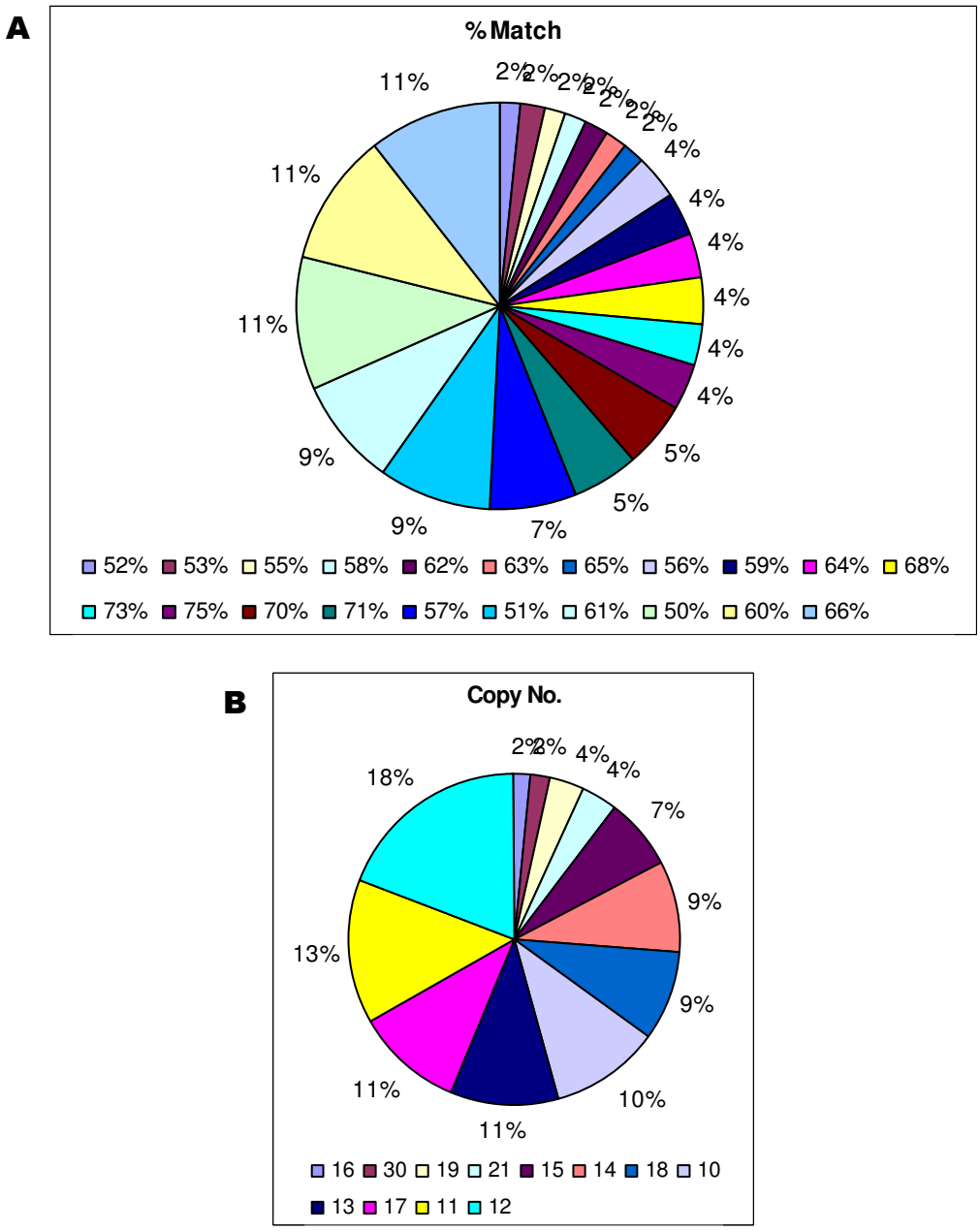


Figure 7.3-2. The proportion of repeats for different categories. A. % match and **B.** copy numbers, presented as percentage from the 57 possible VNTRs identified with the parameter of: total length more than or equal to 20 bp; unit length is 2 bp; and the repeats are at least 50% match.

7.3.2. Optimisation of VNTR typing

7.3.2.1. Determination of ratio of each M13 dye-labelled PCR product when pooling for GeneScan analysis

It is known that different fluorescent dyes have different signal strength. FAM and VIC have the strongest fluorescent signals followed by NED and PET according to GeneScan® Reference Guide (1). The recommendation is a ratio of 3:3:4:6 for FAM:VIC:NED:PET. We initially used this ratio for our sample pooling but found out that VIC was far too strong (Figure 7.3-3) while FAM was too weak.

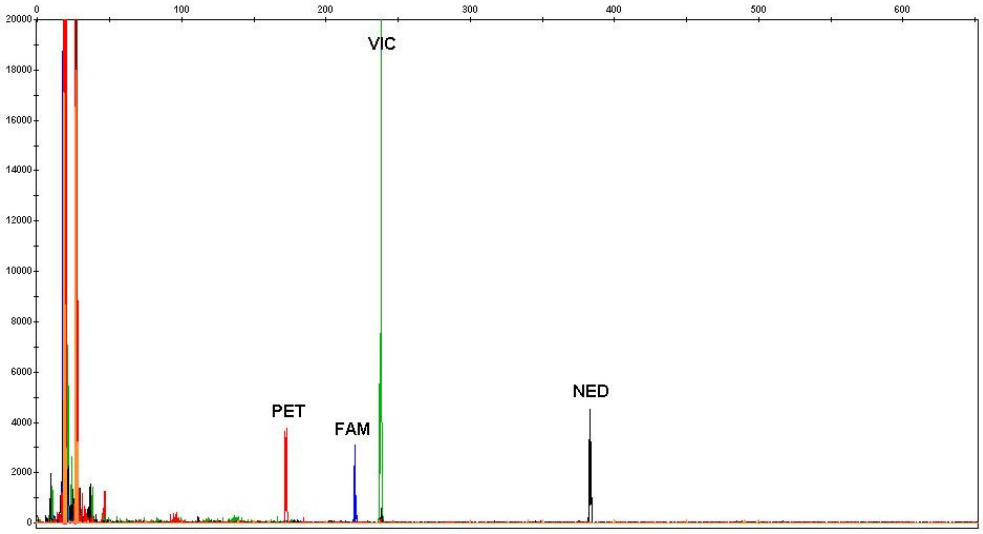


Figure 7.3-3. An electrophoretic diagram illustrating different intensity in fluorescent signals for four dyes.

We used four VNTRs, VNTR 392, VNTR 2156, VNTR 3338 and TR2 and four Typhi isolates to test the ratios. The tests showed that the signal strength was greatest for VIC followed by NED and the least were for FAM and PET. To compensate for the weaker signal, the PCR products from the all VNTRs corresponding to each of the four dyes were mixed in the ratio 2:6:6:3 for VIC, PET, FAM and NED respectively, in all of the subsequent pooled samples. This ratio was applied to all the VNTRs typed.

7.3.2.2. Adjustment of size calling for different dyes used

In most cases the same fluorescent dye-labelled M13 primers were used to type a VNTR. However, there were cases where we needed to use a different dye for pooling purpose to maximise the use of multi-colour analyses. As the dye could affect the relative electrophoretic mobilities of a fragment, we tested the effect of the dyes on size calling using four VNTRs, VNTR 4500, SAL10, SAL16 and TR2 respectively. These VNTRs were amplified from strain CT18 and the sizes are shown in Table 7.3-1. VIC and NED labelled products ran at the same rate and consistently gave the same size for the four VNTRs typed. FAM was 1 bp smaller and PET was 3 bp larger. Therefore, the product size was corrected for the dye size effect if different dyes were used.

Table 7.3-1. The effect of different dyes on the predicted size of VNTR

VNTR Name	Dye label used	Predicted Size (bp)	Difference[1]
4500	FAM	312	-1
	NED	313	0
	PET	316	+3
	VIC	313	0
SAL10	FAM	208	-1
	NED	209	0
	PET	212	+3
	VIC	209	0
SAL16	FAM	239	-1
	NED	240	0
	PET	243	+3
	VIC	240	0
TR2	FAM	524	-1
	NED	525	0
	PET	528	+3
	VIC	525	0

[1] The difference between the predicted size and the expected actual size

7.3.3. Preliminary typing of selected repeats for identification of VNTRs

The 46 potential VNTRs from section 3.1 were assessed, using a selection of 12 Typhi isolates (Table 7.3-2). These isolates were selected based on SNP typing data (Chapter 4). Out of the 12 selected isolates, one was from cluster I, two from cluster II, five from cluster III and four from cluster IV. There were smaller numbers of isolates to represent cluster I and II than cluster III and IV, because the former two were less divergent than the latter. Three isolates from SNP profile 10 were included in the panel as this SNP profile was the profile with the largest numbers of isolates. From the 46 VNTRs, five failed to produce a PCR product at annealing temperature ranging from 48 to 60^{o}C (Table 7.3-3). Thirty nine VNTRs were monomoprhic which showed no variation amongst these isolates (Table 7.3-4). Only two VNTRs were polymorphic and showed size differences amongst the 12 isolates. The seven published VNTRs have also shown variations between the 12 Typhi isolates.

Table 7.3-2. The panel of 12 isolates used to asses the 46 VNTRs selected for typing

Cluster[1]	SNP Profile[1]	Strain Name
i	3	CDC3137-73
ii	7	Tp1
ii	8	Tp2
iii	10	CDC3434-73
iii	10	CDC1707-81
iii	13	CDC1196-74
iii	13	CDC9032-85
iii	10	CDC382-82
iv	19	Ty2-b
iv	20	25T-36
iv	18	25T-40
iv	22	25T-44

[1] The cluster and SNP profile are defined from typing of 42 SNPs (**Chapters 4 and 5**)

Table 7.3-3. List of VNTRs which failed to produce PCR products and were excluded from further analyses

VNTR	Unit Length (bp)[1]	Repeat No.[2]	Total (bp)	% Match[3]	Consensus[4]	Expected PCR product size (bp)[5]
89	7	4	28	80%	AAAAGCA	216
131	6	4	21	80%	CTTCCG	328
1754	3	6	20	82%	AAC	266
2126	2	14	28	66%	AT	279
4067	6	4	21	80%	CTCCGG	264

[1] The size of the VNTR

[2] The total number of repeats in Typhi strain CT18

[3] The % match of the consensus pattern amongst all the copies

[4] Sequence of the repeat unit

[5] Based on Typhi CT18 sequence

Table 7.3-4. List of VNTRs which showed no variation amongst a panel of 12 Typhi isolates

VNTR	Unit Length (bp)	Repeat No.	Total (bp)	% Match	Consensus	Product Size (bp)[1]
33	6	3	18	85%	TGTTTG	347
88	7	3	21	85%	TTTTCAT	299
322	3	7	21	83%	CTG	401
325	7	3	21	86%	TAAAAAA	389
392	2	15	30	70%	GC	240
853	6	3	18	92%	AAGGCG	235
1070	6	4	24	89%	CCGGAG	201
1305	3	7	21	80%	CGC	329
1400	5	4	20	81%	CGATG	358
1736	6	4	24	86%	CGCTCG	361
1773	6	3	18	82%	CCATGC	308
1794	6	4	24	83%	CAGCGC	191
1882	7	3	21	85%	ATTACTG	407
2086	7	3	21	92%	GACCGTT	413
2156	8	3	24	100%	GGCAGGCT	239
2305	6	4	24	83%	CAGCAT	294
2521	5	4	20	81%	GCACC	350
2625	4	5	20	82%	ACAG	261
2634	7	3	21	80%	ACCGGCG	218
3035	5	4	20	81%	CATGG	217
3311	7	4	28	80%	GCGATGC	463
3325	9	3	27	86%	CGTCGCCGC	274
3338	2	12	24	75%	CG	220

3369	5	4	20	80%	CCAGG	224
3404	3	7	21	88%	CGG	266
3463	7	3	21	85%	TTCCCGC	299
3509	3	7	21	80%	CCG	435
3666	5	5	25	85%	GCGGC	252
3699	6	4	24	93%	TGGGCA	384
3835	6	4	24	83%	ATGCTG	310
4030	6	4	24	83%	CTGGTG	224
4107	3	8	24	84%	CGT	263
4207	9	3	27	83%	GCTCTTTCT	332
4295	6	4	24	82%	CCCTAC	437
4353	6	4	24	88%	GCTGGC	236
4546	6	3	18	85%	CAGCGC	322
4578	6	4	24	93%	GGCGCT	288
4585	9	3	27	83%	GATGAAAGA	222
4694	10	3	30	81%	ACTCATCACCA	308

[1] The PCR product size detected by the GeneMapper software

Therefore, a total of 9 VNTRs were polymorphic including two found in this study and seven from previously published studies. The nature of these VNTRs is summarised in Table 7.3-5. Five VNTRs were found on genic regions, three on the intergenic regions and one on a pseudogene. Typhi strains CT18 and Ty2 have different copy numbers for these VNTRs, except for SAL10 where both strains have two copies. The unit length for five VNTRs was 6 bp and the remaining four VNTRs consisted of 3, 7, 8 and 12 bp units respectively.

Table 7.3-5. VNTRs, including two found from this study and seven from published literatures, which showed variation among a panel of 12 Typhi isolates

VNTR Name	Gene	Product	Consensus	Unit Length	No. of Copies	
					CT18	Ty2
4500	*STY4635*	hypothetical protein	GGACTC	6	4	3
4699	*sefC*	outer membrane fimbrial usher protein	TGTTGG	6	17	23
TR1		intergenic region between *yedD* and *yedE*	AGAAGAA	7	12	11
TR2		intergenic region between *acrD* and *yffB*	CCAGTTCC	8	28	11
SAL02	*citT*	citrate carrier	TACCAG	6	10	15
SAL06	*STY0765*	Pseudogene	CTCAAT	6	5	6
SAL10	*yedD*		ACGCCGCTGCCG	12	2	2
SAL16		Intergenic region between *STY3169* (pseudogene) and *STY3172*	ACCATG	6	14	16
SAL20	*ftsN*	cell division protein	CAG	3	16	17

7.3.4. Locus comparison and polymorphism in repeat numbers among Typhi isolates

The remaining 61 Typhi isolates were typed for the nine polymorphic VNTR loci. Four isolates; ST1, 26T19, T202 and 415Ty failed to give any PCR products for VNTR 4500 and two strains; 26T17 and 26T19 failed to amplify VNTR 4699. Therefore, these five strains were removed from subsequent analyses. Thus, the total number of isolates analysed was 68 Typhi isolates, including the 12 isolates that were analysed in the previous section. The size of the PCR products varied from 146 bp in VNTR 4699 to 620 bp in TR2. Different size PCR amplicons corresponded to variation in the number of copies of repeats and thus were assigned different allelic numbers. The genotype data for nine VNTR loci and for all 68 Typhi isolates are presented as allelic numbers in Table 7.3-6.

Table 7.3-6. The MLVA profiles that were distinguished by 9 VNTR loci for 73 Typhi isolates

Strain Name	MLVA profile	Allelic No.								
		4500	4699	SAL02	SAL06	SAL10	SAL16	SAL20	TR1	TR2
Tp1	1	2	6	9	3	2	1	10	4	15
IP.E88 374	1	2	6	9	3	2	1	10	4	15
Tp2	2	2	6	11	3	2	1	8	6	15
Ty2	3	3	17	11	4	2	7	9	7	9
CDC3434-73	4	3	23	5	4	2	5	10	11	20
CDC3137-73	5	3	8	9	3	2	3	8	4	23
CDC1707-81	6	3	1	3	3	2	4	7	5	17
CDC1196-74	7	3	10	6	3	2	5	6	4	18
CDC9032-85	8	2	22	11	4	2	10	9	8	19
CDC382-82	9	3	9	11	3	2	3	2	5	8
25T-36	10	2	8	8	3	2	7	10	7	22
25T-40	11	2	12	11	3	2	4	9	6	10
25T-44	12	1	10	16	3	1	8	9	1	27
R1637	13	3	8	19	3	2	6	8	7	16
R1962	14	3	14	4	3	2	2	5	8	10
R1167	15	3	1	3	3	2	5	4	6	2
ST60	16	3	13	7	3	2	10	4	9	12
ST24A	17	4	6	6	3	2	10	7	9	5
ST24B	18	4	13	7	3	2	10	6	9	5
ST145	19	4	3	7	3	2	4	4	11	12
ST1002	20	3	8	13	3	2	7	10	6	10
ST309	21	4	9	1	3	2	5	7	7	9
ST1106	22	4	23	2	3	2	4	3	10	25
In15	23	2	3	13	3	2	9	4	8	9
PL27566	24	2	18	15	3	2	3	6	7	14
PL73203	25	3	16	5	3	2	7	7	8	16
26T6	26	3	1	13	4	2	6	5	6	26
26T9	27	3	20	7	3	2	5	7	5	7
26T12	28	3	6	20	3	2	3	8	1	13
26T24	29	1	7	15	3	2	8	8	1	8
26T30	30	3	16	5	3	2	3	8	4	16
26T32	31	4	17	5	3	2	3	8	4	16
26T37	32	3	3	10	3	2	3	8	12	18
26T38	33	5	1	14	3	2	4	1	7	21

26T40	34	3	19	12	3	2	4	9	5	4
26T49	35	3	21	12	3	2	5	7	9	5
26T50	36	3	13	18	3	2	6	10	5	9
26T51	37	3	22	9	3	2	4	5	7	8
26T56	38	3	4	6	3	2	9	10	9	21
3123	39	5	1	13	3	2	4	10	6	21
3125	40	3	23	5	3	2	9	7	2	14
3126	41	3	23	6	3	2	9	11	2	14
T189	42	3	10	8	3	2	1	4	7	7
In20	43	4	4	14	4	2	4	6	3	22
In24	44	3	16	17	4	2	8	4	8	1
PNG32	45	2	7	14	3	2	9	7	7	20
CC6	46	3	15	8	3	2	3	7	8	6
CC7	46	3	15	8	3	2	3	7	8	6
414Ty	47	4	4	7	3	2	4	7	11	7
416Ty	48	3	4	6	3	2	4	7	12	9
417Ty	49	4	5	7	3	2	4	7	12	3
418Ty	50	4	5	6	3	2	4	7	12	8
419Ty	51	4	4	6	3	2	4	7	12	10
420Ty	52	4	4	6	3	2	4	5	12	21
421Ty	53	4	3	5	3	2	4	5	12	17
423Ty	54	4	4	9	3	2	4	6	12	20
425Ty	55	4	6	6	3	2	4	8	12	24
444Ty	56	4	5	6	3	2	4	8	12	15
702Ty	56	4	5	6	3	2	4	8	12	15
445Ty	57	4	4	6	2	2	4	8	12	11
446Ty	58	4	5	6	3	2	4	8	12	8
701Ty	59	4	4	1	3	2	4	8	12	28
TYT1668	60	3	12	4	3	2	2	2	4	10
TYT1669	61	4	18	8	3	2	3	7	6	10
TYT1677	62	3	2	6	3	2	7	5	9	6
IP.E88 353	63	4	6	9	4	2	4	8	6	15
CT18	64	4	11	4	3	2	5	8	8	25
422Mar92	65	4	7	5	3	2	3	8	8	12
26T17	N/A[1]	2	-	11	3	2	3	6	6	9
26T19	N/A	-[2]	-	15	3	2	1	5	9	2
415Ty	N/A	-	4	6	3	2	4	7	12	21

ST1	N/A	-	14	18	1	2	3	9	7	22
T202	N/A	-	16	15	3	2	7	4	7	22

[1] Due to incomplete dataset, these strains were omitted from further analyses

[2] No GeneScan data was available

The allelic variability at nine VNTR loci among the 68 isolates is summarised in Table 7.3-7. Continuous range of repeat units was only observed in TR1, from 5 to 16 copies. Overall, two to 28 alleles were observed and the repeat units ranged from a single repeat, to as many as 40 repeats.

Table 7.3-7. Features of the nine polymorphic VNTRs observed in the 68 Typhi isolates

VNTR Name	No. of alleles	PCR Product Size Range (bp)	No. of Copies
4500	5	294-366	1, 2, 3, 4, 13
4699	23	163-307	7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19, 20, 21, 22, 23, 24, 25, 26, 29, 30, 31, 34
TR1	12	227-304	5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16
TR2	28	284-620	1, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 26, 27, 28, 32, 37, 40
SAL02	20	146-278	6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 23, 24, 26, 28
SAL06	3	180-198	3, 5, 6
SAL10	2	197-209	1, 2
SAL16	10	216-276	4, 5, 6, 7, 8, 9, 10, 11, 13, 14
SAL20	11	183-219	8, 9, 11, 12, 13, 14, 15, 16, 17, 18, 20

In each VNTR locus, each different number of copies was considered as a new allele. The frequency for each allele in different VNTR locus was different (Figure 7.3-4). The VNTR with the highest number of alleles was VNTR 4699 followed by TR2, SAL02, TR1, SAL20 and SAL16. All of them had 10 or more alleles. The remaining three VNTRs, VNTR 4500, SAL06 and SAL10 had less than 10 alleles. There were five, three and two alleles for VNTR 4500, SAL06 and SAL10 respectively.

Neither high nor low copy numbers were always dominant in all VNTR loci. The majority isolates could have less copy for one VNTR but more in another. It was easier to recognise a particular copy number that appeared at the highest frequency in the VNTR locus which consisted lesser number of alleles. The majority of isolates had three, five and two copies for VNTR 4500, SAL06 and SAL10 respectively. In the VNTR with

higher number of alleles, the number of isolates was more distributed for each observed number of copies. A large number of isolates had 3, 10, 11 and 7 copies for VNTR 4500, VNTR 4699, SAL02 and SAL16 respectively. Most of the isolates had 16 copies for both SAL20 and TR1. In TR2, the copy numbers of 12 and 17 were equally shared by six isolates each.

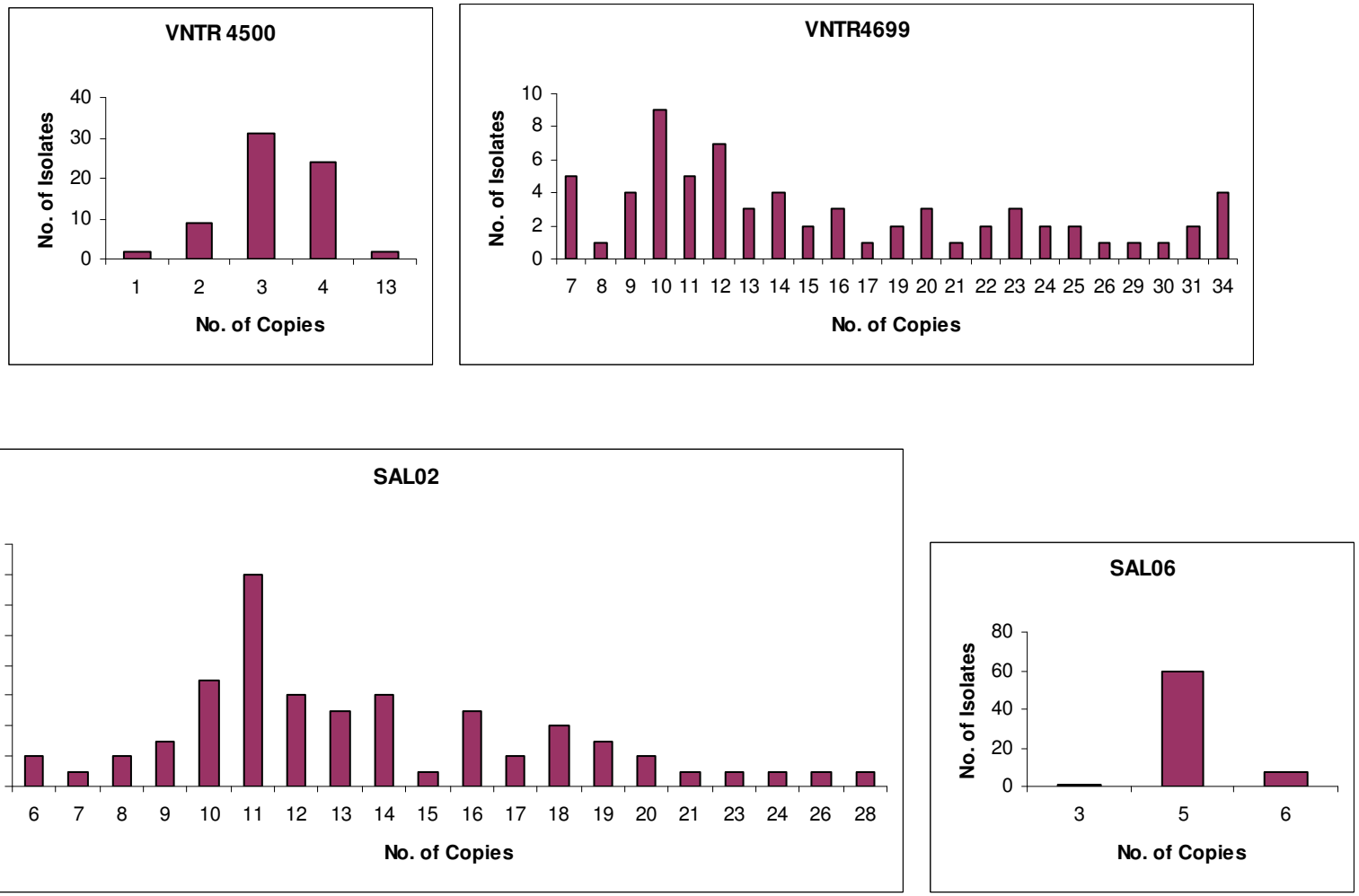
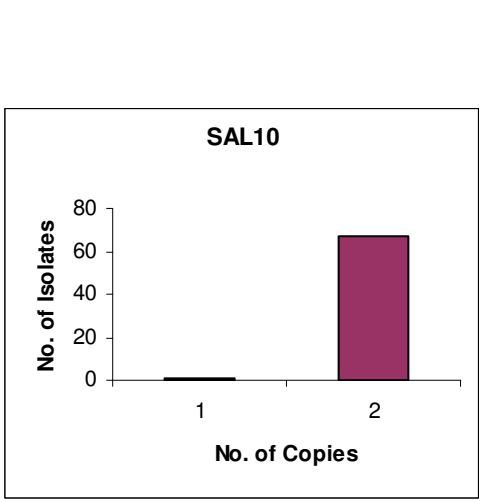


Figure 7.3-4. The distribution of copy numbers in 68 Typhi isolates



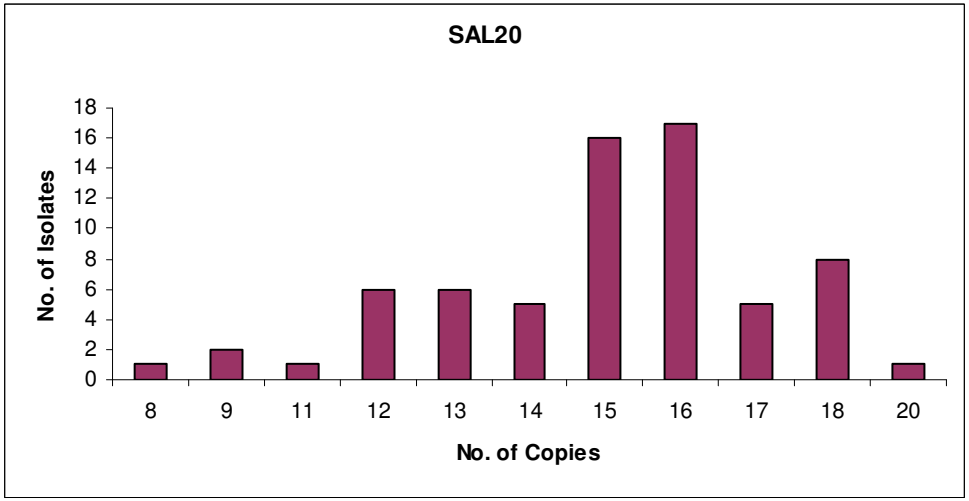
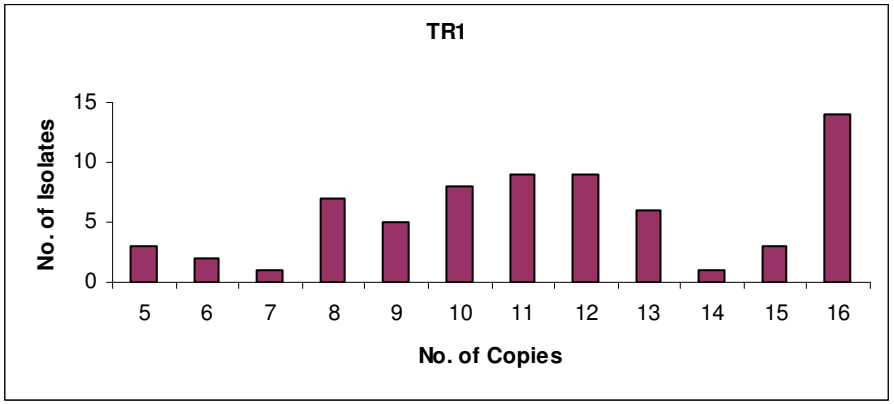


Figure 7.3-4 (Cont.). The distribution of copy numbers in 68 Typhi isolates



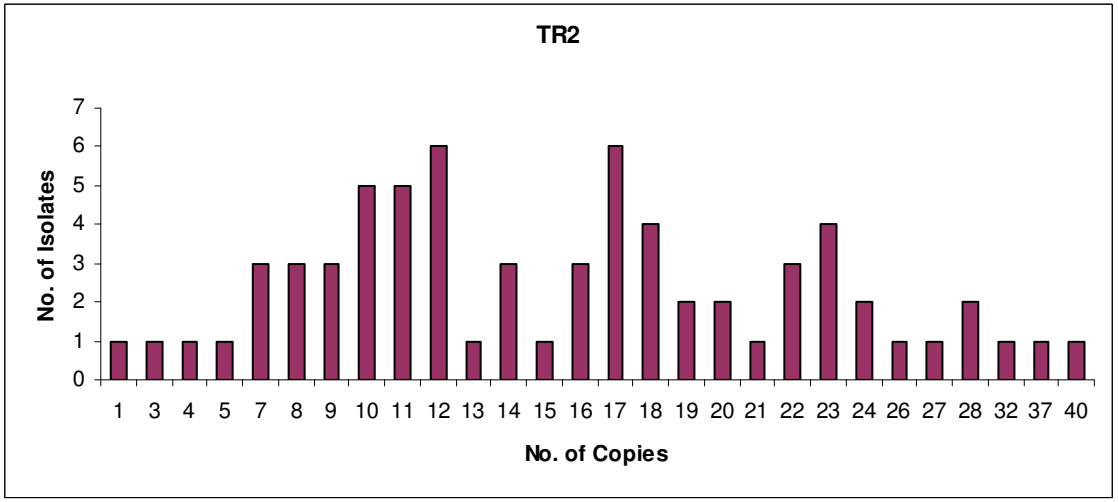


Figure 7.3-4 (Cont.). The distribution of copy numbers in 68 Typhi isolates

7.3.5. Discriminatory power in each VNTR locus

Not all loci had equal discriminatory power. The discriminatory power was measured by Simpson's Index of diversity (D value) (124). The D value could also be used to determine the usefulness of these VNTRs loci for typing purposes. The closer the value is to one, the better the marker is to differentiate the isolates for epidemiological purpose. The D values ranged from 0.044 to 0.964 and averaged at 0.706 (Table 7.3-8).

Table 7.3-8. The diversity for each VNTR locus represented as the D value

VNTR	D
4500	0.663
4699	0.951
TR1	0.894
TR2	0.964
SAL02	0.923
SAL06	0.225
SAL10	0.044
SAL16	0.831
SAL20	0.855
All	0.999

The three most variable VNTR loci, TR2, VNTR 4699 and SAL02 exhibited D values of 0.964, 0.951 and 0.923 respectively The 68 Typhi isolates were distinguished into 65 VNTR types where only three pairs of identical strains were observed for all VNTR loci. These are, SARB63 with IP.E88 374; CC6 with CC7; and 444Ty with 702Ty. Typing of all nine VNTR loci gave a D value of 0.999. This D value could also be achieved by only typing four of the nine VNTR loci, VNTR 4699, SAL02, SAL20 and TR2. All except SAL20 were the VNTRs with the three highest D values. Even though TR1 has the fourth highest D value, it could be replaced by VNTR 4699 that gave the same differentiation for the 68 Typhi isolates. MLVA was more superior in comparison to SNP typing of 42 SNPs, ribotyping and MLST (Figure 7.3-5).

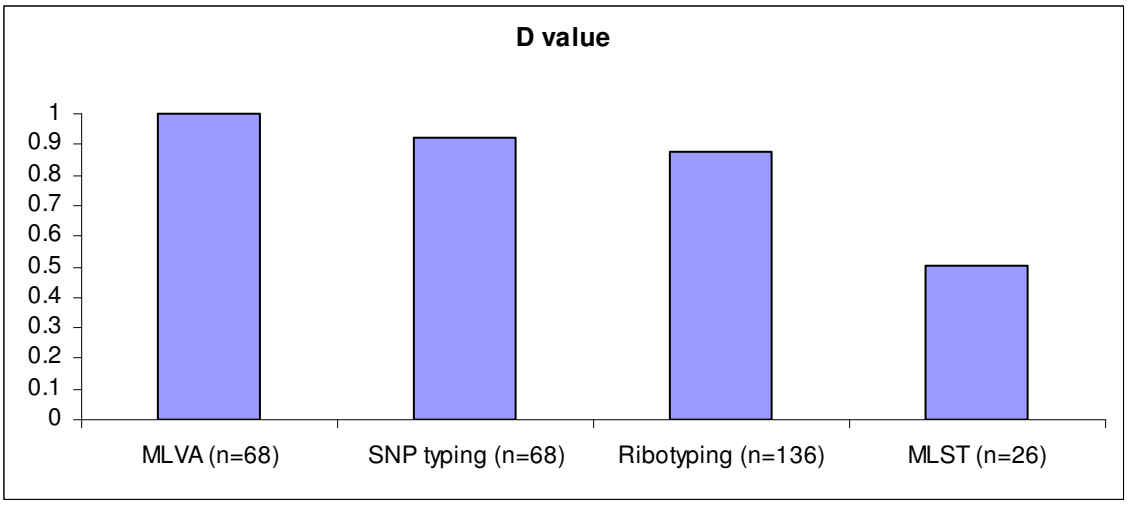


Figure 7.3-5. Comparison of discriminatory powers for different typing methods. n corresponds to the number of isolates typed. The D value of MLVA was calculated based on nine VNTR loci typed in 68 Typhi isolates and for SNP typing it was calculated based on 42 SNPs typed in 68 isolates using the data obtained from Chapter 4 and 5. The D values for ribotyping and MLST were obtained from Chapter 4.

7.3.6. Inconsistencies between genome data and VNTR analyses were observed for CT18 and Ty2

Two discrepancies were observed between theoretical and observed alleles for SAL02 and SAL16 for CT18. According to the genome sequence of CT18, SAL02 and SAL16 were predicted to consist of 10 and 14 copies respectively. However, our data suggested that the SAL02 and SAL16 in CT18 were of nine and eight copies respectively, confirmed by sequencing. Similar inconsistencies were also observed between the published genome data of Ty2 and our Ty2 for these two VNTRs in addition to VNTR 4699. According to genome data, Ty2 has copy numbers of 23, 15 and 16 while our results suggested that the copy numbers are 24, 16 and 10 for VNTR 4699, SAL02 and SAL16 respectively. Nonetheless, all of these repeats were also observed in other Typhi isolates.

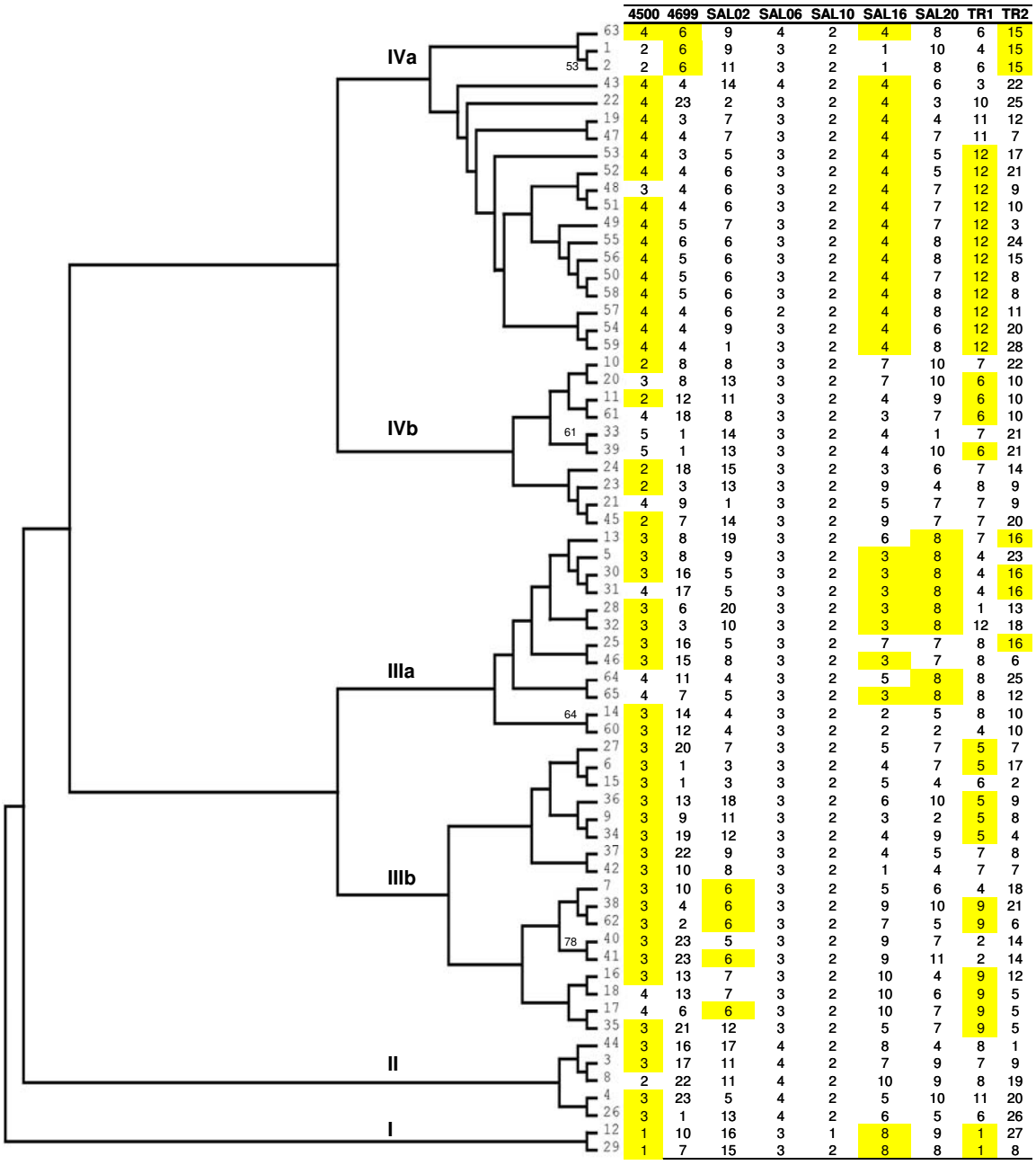
7.3.7. Relationships determined by MLVA

MLVA is designed to target genomic regions, which are prone to rapid mutation marked by changes in copy number of tandem repeat sequences. Initially, a dendogram was generated using UPGMA to illustrate the genetic relationships among the MLVA profiles. The tree gave the same weight for all alleles that were present at each locus. The UPGMA clustering of the MLVA data showed that there were four major clusters (Figure 7.3-6).

The allele of VNTR 4500 could divide the major clusters. Only allele 1 was found exclusively in one cluster, cluster I. The other alleles can be observed in MLVA profiles belonging to multiple different clusters, indicating random slippage of these copies. The majority of isolates had allele 2 in cluster IVb, allele 3 in cluster II and III, and allele 4 in cluster IVa. Other than the allele from VNTR 4500, there were also alleles from other VNTR loci which supported the cluster.

Cluster I which contained two MLVA profiles, was supported by allele 1 for VNTR 4500, allele 8 for SAL 16 and allele 1 for TR1. Only allele 1 for VNTR 4500 was unique within this cluster while the other alleles were also present in other clusters. Allele 8 for SAL16 was observed in an isolate belonging to cluster II while allele 1 for TR1 was observed in an isolate belonging to cluster IIIa. Cluster IIIa was supported by allele 8 of SAL20 and allele 16 of TR2. Cluster IIIb was supported by alleles 5 and 9 of TR1, which divided them into two subclusters. Cluster IVa was supported by allele 12 of TR1 and cluster IVb was supported by allele 6 of TR1.

None of the major cluster was supported by bootstrap value. Only four branches were supported by bootstrap values greater than 50% suggesting that this tree may not be suitable to represent the relationships of the Typhi isolates based on the VNTR data.



	4500	4699	SAL02	SAL06	SAL10	SAL16	SAL20	TR1	TR2
63	4	6	9	4	2	4	8	6	15
1	2	6	9	3	2	1	10	4	15
2	2	6	11	3	2	1	8	6	15
43	4	4	14	4	2	4	6	3	22
22	4	23	2	3	2	4	3	10	25
19	4	3	7	3	2	4	4	11	12
47	4	4	7	3	2	4	7	11	7
53	4	3	5	3	2	4	5	12	17
52	4	4	6	3	2	4	5	12	21
48	3	4	6	3	2	4	7	12	9
51	4	4	6	3	2	4	7	12	10
49	4	5	7	3	2	4	7	12	3
55	4	6	6	3	2	4	8	12	24
56	4	5	6	3	2	4	8	12	15
50	4	5	6	3	2	4	7	12	8
58	4	5	6	3	2	4	8	12	8
57	4	4	6	2	2	4	8	12	11
54	4	4	9	3	2	4	6	12	20
59	4	4	1	3	2	4	8	12	28
10	2	8	8	3	2	7	10	7	22
20	3	8	13	3	2	7	10	6	10
11	2	12	11	3	2	4	9	6	10
61	4	18	8	3	2	3	7	6	10
33	5	1	14	3	2	4	1	7	21
39	5	1	13	3	2	4	10	6	21
24	2	18	15	3	2	3	6	7	14
23	2	3	13	3	2	9	4	8	9
21	4	9	1	3	2	5	7	7	9
45	2	7	14	3	2	9	7	7	20
13	3	8	19	3	2	6	8	7	16
5	3	8	9	3	2	3	8	4	23
30	3	16	5	3	2	3	8	4	16
31	4	17	5	3	2	3	8	4	16
28	3	6	20	3	2	3	8	1	13
32	3	3	10	3	2	3	8	12	18
25	3	16	5	3	2	7	7	8	16
46	3	15	8	3	2	3	7	8	6
64	4	11	4	3	2	5	8	8	25
65	4	7	5	3	2	3	8	8	12
14	3	14	4	3	2	2	5	8	10
60	3	12	4	3	2	2	2	4	10
27	3	20	7	3	2	5	7	5	7
6	3	1	3	3	2	4	7	5	17
15	3	1	3	3	2	5	4	6	2
36	3	13	18	3	2	6	10	5	9
9	3	9	11	3	2	3	2	5	8
34	3	19	12	3	2	4	9	5	4
37	3	22	9	3	2	4	5	7	8
42	3	10	8	3	2	1	4	7	7
7	3	10	6	3	2	5	6	4	18
38	3	4	6	3	2	9	10	9	21
62	3	2	6	3	2	7	5	9	6
40	3	23	5	3	2	9	7	2	14
41	3	23	6	3	2	9	11	2	14
16	3	13	7	3	2	10	4	9	12
18	4	13	7	3	2	10	6	9	5
17	4	6	6	3	2	10	7	9	5
35	3	21	12	3	2	5	7	9	5
44	3	16	17	4	2	8	4	8	1
3	3	17	11	4	2	7	9	7	9
8	2	22	11	4	2	10	9	8	19
4	3	23	5	4	2	5	10	11	20
26	3	1	13	4	2	6	5	6	26
12	1	10	16	3	1	8	9	1	27
29	1	7	15	3	2	8	8	1	8

Figure 7.3-6. Unweighted pair group method with arithmetic means (UPGMA) dendogram of 65 MLVA profiles. The roman numerals indicate the assigned cluster numbers. Bootstrap values, if greater than 50%, are presented at the nodes of the tree. The table corresponds to the allelic pattern for each MLVA profile. The shaded cells correspond to the alleles that supported the clustering of the MLVA profile.

A minimum-spanning tree (MST) was also generated to calculate the distance between isolates by scoring the number of VNTR loci that differed between two MLVA profiles. If two profiles differed by a single VNTR locus, the distance between them will be 1; if two loci were different, the distance will be 2 and so on (Figure 7.3-7). The majority of the MLVA profiles differed by four VNTR loci, with the minimum difference of one VNTR locus and the maximum of five VNTRs loci. Only two pairs, MLVA profiles 50 and 58; and MLVA profiles 56 and 58, differed by one VNTR locus, SAL20 and TR2 respectively. The connections between MLVA profiles that differed by more than a single locus may not represent an evolutionary relationship, nevertheless they illustrated the connections of the most similar MLVA profiles. There was no apparent major cluster in MST. The four major clusters that were observed in UPGMA tree were not observed in the MST.

Similar to previous SNP typing (Chapter 4 and Chapter 5), there was no observed epidemiological connection on the Typhi isolates typed. It was evident that the isolates were distributed throughout the MST tree. No branching was completely dominated by isolates from the same country, year of isolation and/or localities of isolation, suggesting an absence of geographical association as observed in SNP typing (Chapter 4).

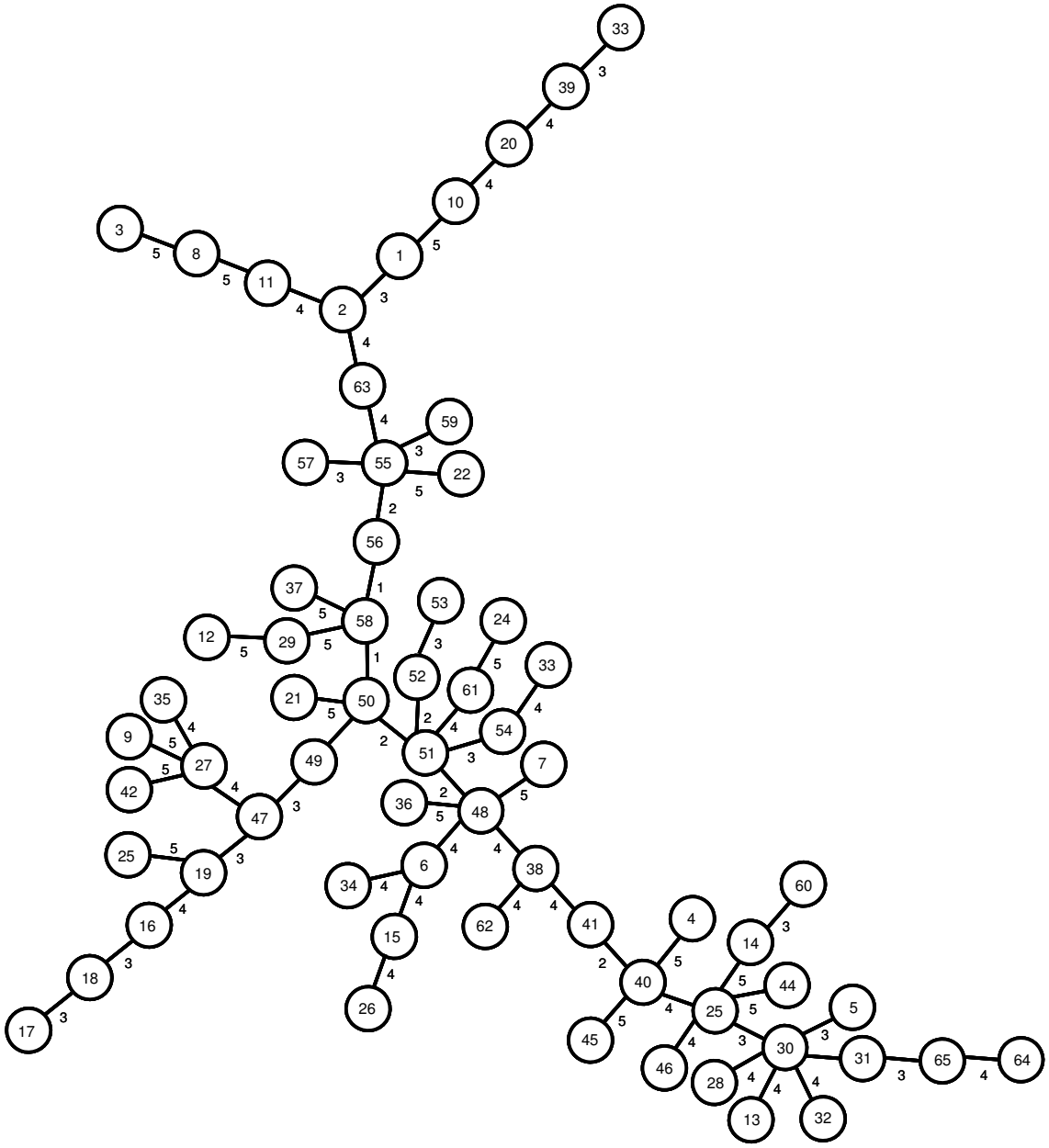


Figure 7.3-7. The Minimum Spanning Tree of the 68 Typhi isolates, generated using the nine VNTR data. The number within the circle corresponds to the MLVA profile in Table 7.3-6. The size of the circle does not reflect the size (the number of isolates) of that particular MLVA profile. The number on the branch represents the number of VNTR difference between the connecting profiles.

7.3.8. Comparison between SNP typing and VNTR typing

Previously, we have typed the 73 global Typhi isolates using 42 SNPs. The present study addressed whether MLVA could provide higher resolution than SNP typing and can be used to determine the relationships of the 73 Typhi isolates. The isolates were separated according to the corresponding clusters defined by SNPs. To allow easier differentiation between isolates, different numbers of copies for each VNTR locus were presented as allelic numbers as shown in Table 7.3-6. The allele numbers were used to generate MST. The MST generated from the VNTR data was compared to the MST generated using data of 42 SNPs (Chapter 5). The MLVA profiles were divided according to the clusters that have been identified by the SNP data. For each of the clusters, a new MST was generated using VNTR data, and it was compared cluster by cluster to the MST generated by SNP data.

None of the allele on any VNTR locus was unique to only a single cluster. Out of the nine polymorphic VNTRs, the differences between each isolate ranged from one to six. From the four distinguished clusters, only cluster I had the least VNTR differences while the isolates in the remaining clusters had large number of differences. This was an indication that the relationships between these isolates were not strongly supported and the trees showed homoplasies.

Generally, the relationships of the isolates represented by MST were conflicting with the MST generated from SNPs data if they differed by more than three VNTRs. This was indicated by the straight lines that connected one MLVA profile to another on the MST dendograms (except cluster IV) but dashed lines otherwise (Figure 7.3-8 to Figure 7.3-11).

7.3.8.1. Cluster I

We know from previous studies that all SNP profiles from cluster I were differentiated by only one SNP suggesting a clonal complex. Twenty one MLVA profiles differed by one to five VNTRs were identified (Table 7.3-9 and Figure 7.3-8). VNTR typing differentiated SNP profile 1 into two MLVA profiles, profile 2b into 10 profiles and SNP profiles 3 and 5 into three MLVA profiles respectively. MLVA profiles with SNP profile 1 and 3 were still grouped together. Although the MST showed that two MLVA profiles belonging to SNP profile 1 were connected to each other, they differed in five of the nine VNTR loci.

The majority of MLVA profiles belonging to SNP profile 2b were connected to each other, except MLVA profile 43 and 52 which were connected to SNP profile 4 and 5 respectively. MLVA profile 43 was connected to MLVA profile 54 of SNP profile 4 by difference in four VNTR loci. MLVA profile 52 was connected to MLVA profile 51 and 53 that belonged to SNP profile 5 by differences in two and three VNTR loci respectively. Similarly, MLVA profile 53 was separated from the remaining two profiles belonging to SNP profile 5.

From the MST generated using SNP data, SNP profile 1 gave rise to SNP profile 2a, which was consistent to the the MST using VNTR data. Based on VNTR data, SNP profile 2b appeared to arise directly from SNP profile 1 however, according to the SNP data profile 2b seemed to arise from 2a.

Table 7.3-9. The VNTR data presented as the allele number for cluster I

SNP.MLVA profile[1]	Allelic No.								
	4500	4699	SAL02	SAL06	SAL10	SAL16	SAL20	TR1	TR2
1.22	4	23	2	3	2	4	3	10	25
1.47	4	4	7	3	2	4	7	11	7
2a.19	4	3	7	3	2	4	4	11	12
2b.32	3	3	10	3	2	3	8	12	18
2b.43	4	4	14	4	2	4	6	3	22
2b.49	4	5	7	3	2	4	7	12	3
2b.50	4	5	6	3	2	4	7	12	8
2b.52	4	4	6	3	2	4	5	12	21
2b.55	4	6	6	3	2	4	8	12	24
2b.56	4	5	6	3	2	4	8	12	15
2b.57	4	4	6	2	2	4	8	12	11
2b.58	4	5	6	3	2	4	8	12	8
2b.59	4	4	1	3	2	4	8	12	28
3.30	3	16	5	3	2	3	8	4	16
3.31	4	17	5	3	2	3	8	4	16
3.5	3	8	9	3	2	3	8	4	23
4.54	4	4	9	3	2	4	6	12	20
5.48	3	4	6	3	2	4	7	12	9
5.51	4	4	6	3	2	4	7	12	10
5.53	4	3	5	3	2	4	5	12	17

[1] For easier identification, SNP profile was added as the prefix. The first number in the column SNP.MLVA profile represents the SNP profile and the second number represents the MLVA profile

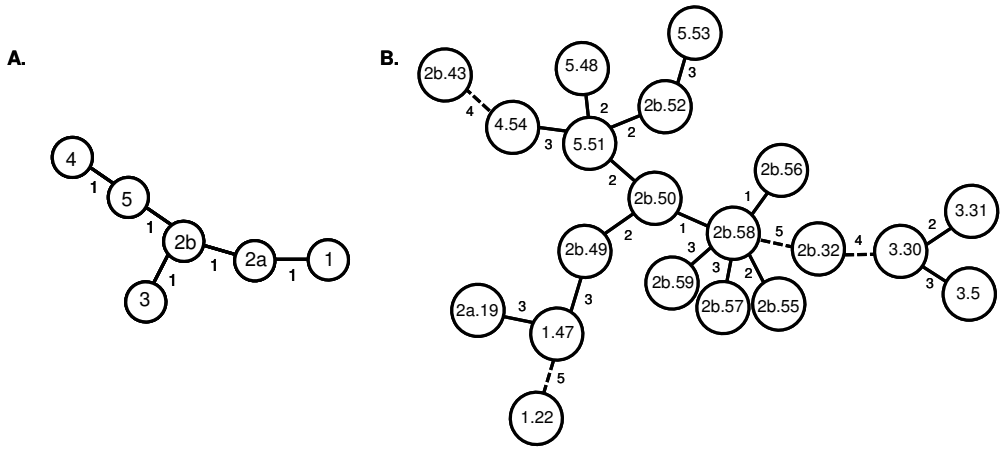


Figure 7.3-8. Comparison of Minimum Spanning Trees (MST) for the profiles from cluster I based on (A) SNP data and (B) VNTR data. The size of the circle does not correspond to the number of isolates with that particular SNP profile or MLVA profile. The numbers on the node correspond to the number of differences in either number of SNPs (for MST generated from SNP profile) or VNTRs (for MST generated from VNTR data). The dashed lines indicate that the differences obtained from VNTR typing were more than three and the relationships between isolates are in conflict with the the MST generated from SNP data**.**

7.3.8.2. Cluster II

In cluster II, SNP typing was able to differentiate the isolates into four SNP profiles, namely SNP profile 6, 7, 8 and 9 respectively. Typing of nine VNTR loci further differentiated SNP profile 7 into two MLVA profiles (Table 7.3-10 and Figure 7.3-9). The two MLVA profiles differed by six loci.

From the two MSTs that were generated by SNP and VNTR data respectively, only SNP profile 9 was inconsistent. Even though the MST based on VNTR data suggested that SNP profile 9 was closer to SNP profile 7, SNP profile 9 differed by six out of seven informative VNTRs, to both SNP profile 7 and 8, respectively (Table 7.3-10). According to the SNP data, SNP profile 9 was closer to SNP profile 6 but these profiles did not share identical allele in any of the nine VNTRs typed.

The least alleles difference in the VNTR loci was observed between SNP profile 7 and SNP profile 8, which corresponded to MLVA profile 1 and 2 respectively. These MLVA profiles differed by three VNTR loci while according to SNP data, these profiles differed by four SNPs.

Table 7.3-10. The VNTR data presented as the allele number for cluster II

SNP.MLVA profile[1]	Allelic No.								
	4500	4699	SAL02	SAL06	SAL10	SAL16	SAL20	TR1	TR2
6.40	3	23	5	3	2	9	7	2	14
7.1	2	6	9	3	2	1	10	4	15
7.17	4	6	6	3	2	10	7	9	5
8.2	2	6	11	3	2	1	8	6	15
9.64	4	11	4	3	2	5	8	8	25

[1] For easier identification, SNP profile was added as the prefix. The first number in the column SNP.MLVA profile represents the SNP profile and the second number represents the MLVA profile

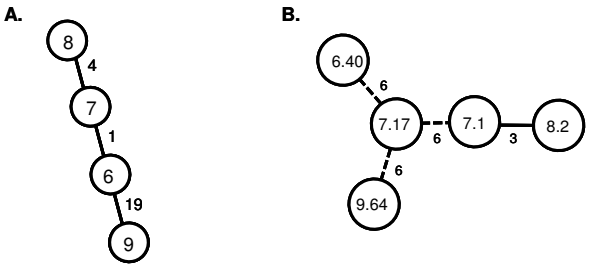


Figure 7.3-9. Comparison of Minimum Spanning Trees (MST) for the profiles from cluster II based on (A) SNP data and (B) VNTR data. The size of the circle does not correspond to the number of isolates with that particular SNP profile or MLVA profile. The numbers on the node correspond to the number of differences in either number of SNPs (for MST generated from SNP profile) or VNTRs (for MST generated from VNTR data). The dashed lines indicate that the differences obtained from VNTR typing were more than three and the relationships between isolates are in conflict with the the MST generated from SNP data**.**

7.3.8.3. Cluster III

There were 10 SNP profiles and 34 MLVA profiles belonging to cluster III. All SNP profiles which contained multiple isolates had different MLVA profiles, excluding SNP profile 11. Two isolates belonging to SNP profile 11 were found to have the same MLVA profile and one of these isolates was excluded in the construction of MST.

The relationships illustrated by the two MSTs were different in most cases. SNP profiles belonging to this cluster appeared as a clonal complex where each SNP profile differed from each other by one SNP. According to the VNTR data, the MLVA profiles were very different that no clonal complex was observed. The MLVA profiles differed from each other by three to six VNTR loci (Table 7.3-11 and Figure 7.3-10). The MLVA profiles with the smallest difference (three VNTR loci) were only found between three pairs of MLVA profiles. These included: MLVA profile 16 to MLVA profile 18; MLVA profile 14 to MLVA profile 60; and MLVA profile 33 to MLVA profile 39. These MLVA profiles were not from the same SNP profile, except MLVA profiles 33 and 39 which were from SNP profile 23.

Multiple MLVA profiles belonging to the same SNP profile were not always grouped together, except for those of SNP profile 23. For example, four MLVA profiles (MLVA profile 7, 8, 42 and 44) from SNP profile 13 were separately grouped. Only MLVA profiles 8 and 44 were connected to each other but they differed by six VNTR loci. MLVA profile 42 was connected to MLVA profile 16 from SNP profile 10a, and MLVA profile 7 was connected to MLVA profile 15 from SNP profile 10b.

From SNP data, SNP profile 10b was shown to be the ancestral SNP profile. It was connected to six other SNP profiles, including SNP profile 10a, 10c, 11, 12, 13 and 14, by one SNP difference. Typing of nine VNTR has divided the isolates in SNP profile 10b into 14 MLVA profiles. These MLVA profiles were scattered around and were independently connected to the six aforementioned SNP profiles. One of the MLVA profile from SNP profile 10b was connected to SNP profile 10d by six VNTR

differences. SNP typing showed that SNP profile 10d was connected to SNP profile 10b through SNP profile 10c. However, according to the VNTR typing, three MLVA profiles from SNP profile 10d were not connected to SNP profile 10c. Further inconsistencies between the two MSTs were observed for SNP profile 23 and SNP profile 16. SNP typing has shown that SNP profile 23 was connected to SNP profile 10d by one SNP difference and to SNP profile 16 by three SNPs differences. In contrast, VNTR typing has shown that SNP profile 23 was connected to SNP profile 10a. The only MLVA profile from SNP profile 16 was not connected to any MLVA profile from SNP profile 23, but it was connected to MLVA profile 24 from SNP profile 10b and MLVA profile 46 from SNP profile 11.

VNTR typing data suggested that this cluster was much more diverse than the other three clusters. Most inconsistencies between the two MSTs were observed within this cluster than in other clusters. It was not surprising considering that the resolution of SNP typing was low for this cluster. VNTR typing definitely has increased the discrimination of the SNP profiles. However, the majority of the MLVA profiles differed by five loci, and it is uncertain whether the relationships illustrated by MST generated by VNTR data were reliable.

Table 7.3-11. The VNTR data presented as the allele number for cluster III

SNP.MLVA profile[1]	Allelic No.								
	4500	4699	SAL02	SAL06	SAL10	SAL16	SAL20	TR1	TR2
10a.16	3	13	7	3	2	10	4	9	12
10a.6	3	1	3	3	2	4	7	5	17
10a.65	4	7	5	3	2	3	8	8	12
10b.15	3	1	3	3	2	5	4	6	2
10b.18	4	13	7	3	2	10	6	9	5
10b.24	2	18	15	3	2	3	6	7	14
10b.25	3	16	5	3	2	7	7	8	16
10b.26	3	1	13	4	2	6	5	6	26
10b.34	3	19	12	3	2	4	9	5	4
10b.36	3	13	18	3	2	6	10	5	9
10b.37	3	22	9	3	2	4	5	7	8
10b.38	3	4	6	3	2	9	10	9	21
10b.41	3	23	6	3	2	9	11	2	14
10b.45	2	7	14	3	2	9	7	7	20
10b.60	3	12	4	3	2	2	2	4	10
10b.62	3	2	6	3	2	7	5	9	6
10b.9	3	9	11	3	2	3	2	5	8
10c.23	2	3	13	3	2	9	4	8	9
10d.27	3	20	7	3	2	5	7	5	7
10d.35	3	21	12	3	2	5	7	9	5
10d.4	3	23	5	4	2	5	10	11	20
11.46	3	15	8	3	2	3	7	8	6
12.28	3	6	20	3	2	3	8	1	13
13.42	3	10	8	3	2	1	4	7	7
13.44	3	16	17	4	2	8	4	8	1
13.7	3	10	6	3	2	5	6	4	18
13.8	2	22	11	4	2	10	9	8	19
14.14	3	14	4	3	2	2	5	8	10
14.63	4	6	9	4	2	4	8	6	15
16.61	4	18	8	3	2	3	7	6	10
23.33	5	1	14	3	2	4	1	7	21
23.39	5	1	13	3	2	4	10	6	21

[1] For easier identification, SNP profile was added as the prefix. The first number in the column SNP.MLVA profile represents the SNP profile and the second number represents the MLVA profile

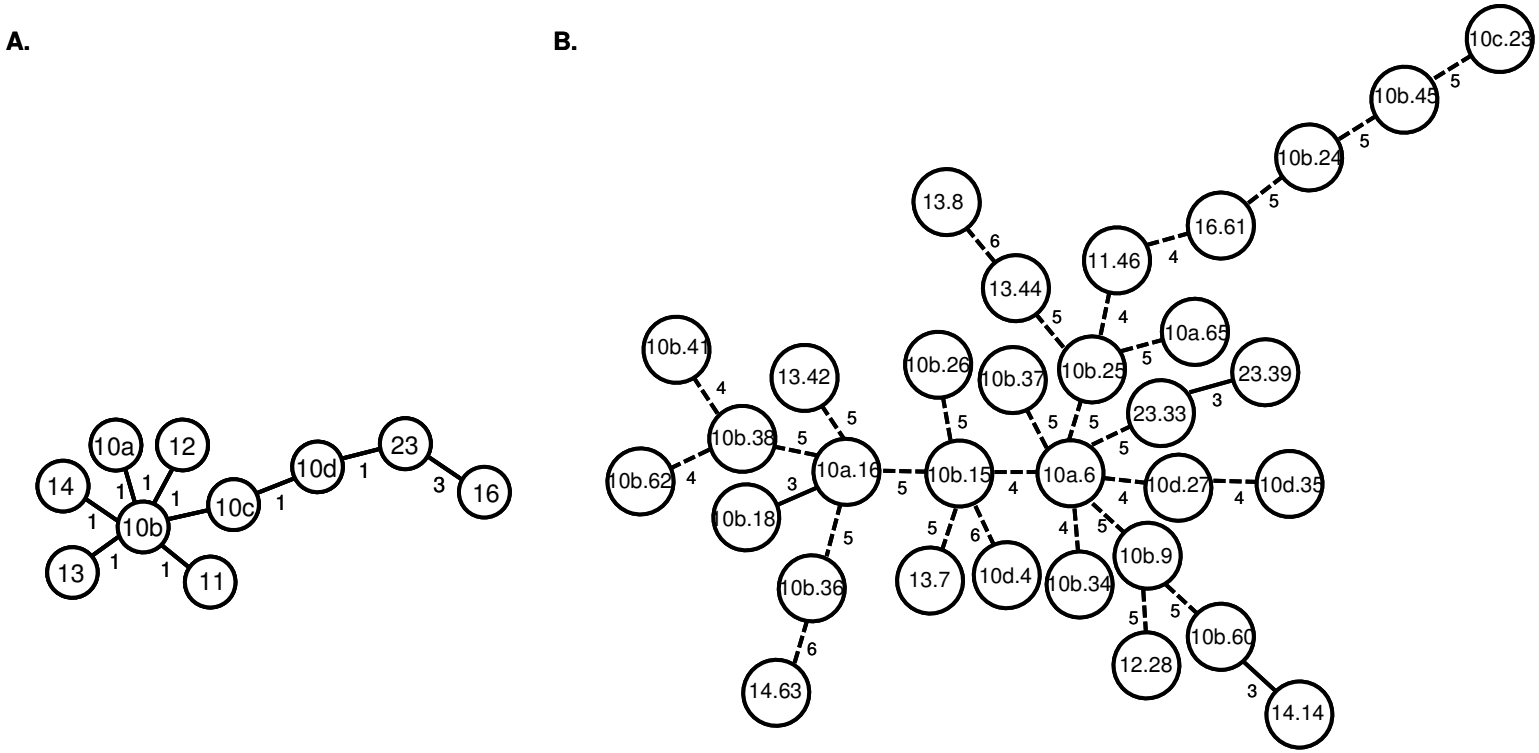


Figure 7.3-10. Comparison of Minimum Spanning Trees (MST) for the profiles from cluster III based on (A) SNP data and (B) VNTR data. The size of the circle does not correspond to the number of isolates with that particular SNP profile or MLVA profile. The numbers on the node correspond to the number of differences in either number of SNPs (for MST generated from SNP profile) or VNTRs (for MST generated from VNTR data). The dashed lines indicate that the differences obtained from VNTR typing were more than three and the relationships between isolates are in conflict with the the MST generated from SNP data**.**

7.3.8.4. Cluster IV

Eight MLVA profiles belonged to cluster IV (Table 7.3-12 and Figure 7.3-11). The isolates representing SNP profiles 15 (ST1) and 17 (T202) that belonged to cluster IV were excluded, due to failures in amplifying the region of VNTR 4500. Thus, the relationships of these SNP profiles could not be determined. Four to six differences in VNTRs were observed in isolates from this cluster. A higher resolution was achieved by typing the VNTR loci. Isolates of SNP profile 18 were differentiated into four MLVA profiles. MLVA profile 21 belonging to SNP profile 18 was not clustered together with the other three MLVA profiles of the same SNP profile. Even though the three MLVA profiles were connected, they differed in five VNTR loci.

Most of the relationships on the MST generated from MLVA data in this cluster were consistent with the MST from SNP data. MLVA profile 10 representing SNP profile 20 was linked to MLVA profile 20 of SNP profile 18 and MLVA profile 3 representing SNP profile 19 was connected to MLVA profile 11 of SNP profile 18. These were consistent with SNP data where SNP profiles 19 and 20 arose from SNP profile 18. SNP profile 21, represented by MLVA profile 29 and SNP profile 22, represented by MLVA profile 12 differed by five VNTR loci. They were located adjacent to one another, which was consistent with the SNP data. However, VNTR data suggested that SNP profile 21 was closer to SNP profile 18, than SNP profile 22 to SNP profile 18, which contradicted the SNP data.

Table 7.3-12. The VNTR data presented as the allele number for cluster IV

SNP.MLVA profile[1]	Allelic No.								
	4500	4699	SAL02	SAL06	SAL10	SAL16	SAL20	TR1	TR2
18.11	2	12	11	3	2	4	9	6	10
18.13	3	8	19	3	2	6	8	7	16
18.20	3	8	13	3	2	7	10	6	10
18.21	4	9	1	3	2	5	7	7	9
19.3	3	17	11	4	2	7	9	7	9
20.10	2	8	8	3	2	7	10	7	22
21.29	1	7	15	3	2	8	8	1	8
22.12	1	10	16	3	1	8	9	1	27

[1] For easier identification, SNP profile was added as the prefix. The first number in the column SNP.MLVA profile represents the SNP profile and the second number represents the MLVA profile

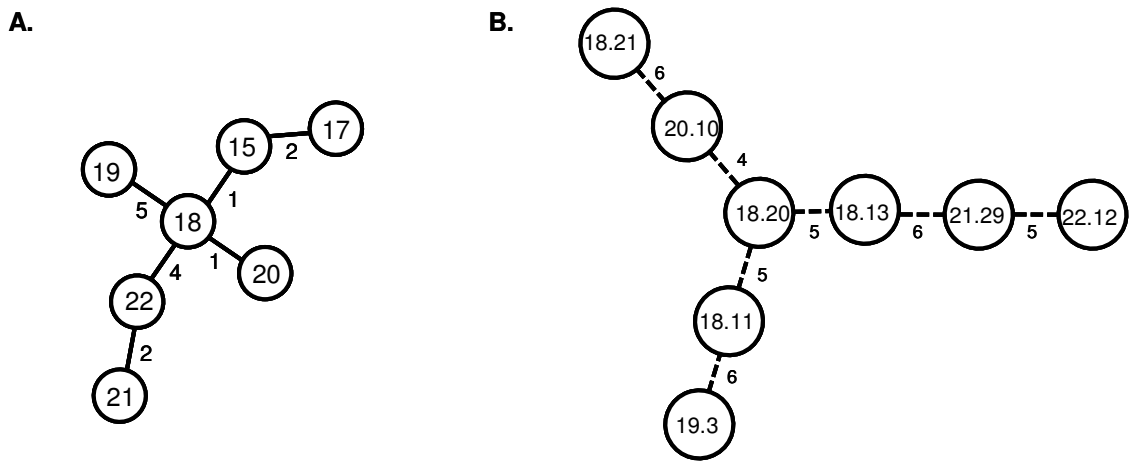


Figure 7.3-11. Comparison of Minimum Spanning Trees (MST) for the profiles from cluster IV based on (A) SNP data and (B) VNTR data. The size of the circle does not correspond to the number of isolates with that particular SNP profile or MLVA profile. The numbers on the node correspond to the number of differences in either number of SNPs (for MST generated from SNP profile) or VNTRs (for MST generated from VNTR data). The dashed lines indicate that the differences obtained from VNTR typing were more than three and the relationships between isolates are in conflict with the the MST generated from SNP data**.**

7.3.9. Genetic relationships of 68 Typhi isolates by combining VNTR and SNP markers

The data from VNTR typing and SNP typing were combined to establish the overall relationships of 73 global Typhi isolates (Figure 7.3-12). VNTRs and SNPs provide differing level of diversities. VNTR is a more rapidly evolving marker (139) and was more appropriate in determining relationships of very closely related isolates while SNP is more slowly evolving and was used to establish the major phylogenetic relationships between isolates. Therefore, combining these two markers would enable high discrimination between isolates while minimising the homoplasy, which arose due to the highly polymorphic VNTR loci. The SNP data was converted into numerical data of 1 and 2 to represent the two alleles, 1 for the most common allele and 2 for the alternative allele. There were only two alleles for all the SNPs typed. Data from SNP typing and MLVA were combined to form large composite data and a new MST was constructed based on the combined data.

By combining the two markers, more unique genotypes were identified. A total of 42 SNPs and nine VNTRs could distinguish the 68 Typhi isolates into 66 different profiles, one more profile than that were distinguished by VNTR typing alone (Table 7.3-13). Two pairs of isolates, CC6 and CC7 and 444Ty and 702Ty, were identical for the SNPs and VNTRs tested. Both CC6 and CC7 were isolated in 1995 in Thailand and had the same phage type and genotype (Chapter 2, Table 2.1-2). 444Ty and 702Ty also had the same genome type, however the locality and year of isolation were not known for these two isolates.

Each of the profiles was connected to one another with a minimum of one difference to a maximum of 26 differences. From the MST, it could be seen that cluster III, which were defined from the SNPs typing was no longer visible as a clonal complex. This cluster was the largest and it radiated to give rise to the remaining clusters. The SNP profiles, which were considered to be the ancestral profiles for cluster I and II respectively by typing of 42 SNPs, were different based on the combined data. SNP profile 10a of cluster III gave rise to cluster I while SNP profile 10b also from cluster III gave rise to cluster II. This was dissimilar to the findings from SNP typing where both cluster I and II emerged from SNP profile 13.

Further dissimilarities were observed where cluster I arose from SNP profile 2a instead of SNP profile 1 and cluster II arose from SNP profile 6 instead of SNP profile 8. Nevertheless, they were connected to SNP profile 2b but of different MLVA profiles. Cluster I still encompassed all the isolates expressing z66 flagellar antigens. Typhi strain CT18 that represented the SNP profile 9 was no longer belonged to cluster II and was relocated to cluster I with a total of 26 differences. Two SNP profiles, SNP profile 21 and 22 respectively, which were assigned to cluster IV have now been directly linked to SNP profile 23 that was assigned to cluster III.

Table 7.3-13. The combined data of SNPs and VNTRs for the 68 Typhi isolates typed

Cluster	SNP.VNTR Profile	SNP No.																																						BiP No.				VNTR								
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	36	48	56	33	4500	4699	SAL02	SAL06	SAL10	SAL16	SAL20	TR1	TR2
	1.22	1	2	2	2	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	2	1	1	1	1	1	1	1	4	23	2	3	2	4	3	10	25
	1.47	1	2	2	2	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	2	1	1	1	1	2	1	4	4	7	3	2	4	7	11	7
	2a.19	1	2	2	2	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	2	1	2	1	1	1	1	1	4	3	7	3	2	4	4	11	12
	2b.32	1	2	2	2	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	2	1	2	1	1	1	2	1	3	3	10	3	2	3	8	12	18
	2b.43	1	2	2	2	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	2	1	2	1	1	1	2	1	4	4	14	4	2	4	6	3	22
	2b.49	1	2	2	2	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	2	1	2	1	1	1	2	1	4	5	7	3	2	4	7	12	3
	2b.50	1	2	2	2	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	2	1	2	1	1	1	2	1	4	5	6	3	2	4	7	12	8
	2b.52	1	2	2	2	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	2	1	2	1	1	1	2	1	4	4	6	3	2	4	5	12	21
	2b.55	1	2	2	2	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	2	1	2	1	1	1	2	1	4	6	6	3	2	4	8	12	24
I	2b.56	1	2	2	2	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	2	1	2	1	1	1	2	1	4	5	6	3	2	4	8	12	15
	2b.57	1	2	2	2	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	2	1	2	1	1	1	2	1	4	4	6	2	2	4	8	12	11
	2b.58	1	2	2	2	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	2	1	2	1	1	1	2	1	4	5	6	3	2	4	8	12	8
	2b.59	1	2	2	2	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	2	1	2	1	1	1	2	1	4	4	1	3	2	4	8	12	28
	3.30	1	2	2	2	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	2	1	2	1	1	1	2	1	3	16	5	3	2	3	8	4	16
	3.31	1	2	2	2	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	2	1	2	1	1	1	2	1	4	17	5	3	2	3	8	4	16
	3.5	1	2	2	2	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	2	1	2	1	1	1	2	1	3	8	9	3	2	3	8	4	23
	4.54	1	2	2	2	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	2	2	1	1	1	1	2	1	4	4	9	3	2	4	6	12	20
	5.48	1	2	2	2	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	2	2	2	1	1	1	2	1	3	4	6	3	2	4	7	12	9
	5.51	1	2	2	2	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	2	2	2	1	1	1	2	1	4	4	6	3	2	4	7	12	10
	5.53	1	2	2	2	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	2	2	2	1	1	1	2	1	4	3	5	3	2	4	5	12	17
	6.40	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	3	23	5	3	2	9	7	2	14
	7.1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	6	9	3	2	1	10	4	15
II	7.17	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	4	6	6	3	2	10	7	9	5
	8.2	1	1	2	2	2	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	6	11	3	2	1	8	6	15
	9.64	1	1	1	1	2	2	2	1	2	2	2	1	1	2	2	1	1	2	2	2	2	1	1	2	1	2	1	2	1	1	2	1	2	1	2	1	2	1	1	1	2	1	4	11	4	3	2	5	8	8	25
	10a.16	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	2	1	1	3	13	7	3	2	10	4	9	12
	10a.6	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	2	1	1	3	1	3	3	2	4	7	5	17
	10a.65	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	2	1	1	4	7	5	3	2	3	8	8	12
	10b.15	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	3	1	3	3	2	5	4	6	2
	10b.18	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	4	13	7	3	2	10	6	9	5
	10b.24	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	2	18	15	3	2	3	6	7	14
	10b.25	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	3	16	5	3	2	7	7	8	16
	10b.26	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	3	1	13	4	2	6	5	6	26
	10b.34	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	3	19	12	3	2	4	9	5	4
	10b.36	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	3	13	18	3	2	6	10	5	9
	10b.37	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	3	22	9	3	2	4	5	7	8
	10b.38	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	3	4	6	3	2	9	10	9	21
	10b.41	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	3	23	6	3	2	9	11	2	14
	10b.45	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	2	7	14	3	2	9	7	7	20
	10b.60	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	3	12	4	3	2	2	2	4	10
	10b.62	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	3	2	6	3	2	7	5	9	6
III	10b.9	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	3	9	11	3	2	3	2	5	8
	10c.23	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	2	2	3	13	3	2	9	4	8	9
	10d.27	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	2	2	3	20	7	3	2	5	7	5	7
	10d.35	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	2	2	3	21	12	3	2	5	7	9	5
	10d.4	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	2	2	3	23	5	4	2	5	10	11	20
	11.46	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	2	2	1	1	1	1	3	15	8	3	2	3	7	8	6
	12.28	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	3	6	20	3	2	3	8	1	13
	13.42	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	3	10	8	3	2	1	4	7	7
	13.44	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	3	16	17	4	2	8	4	8	1
	13.7	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	3	10	6	3	2	5	6	4	18
	13.8	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	2	22	11	4	2	10	9	8	19
	14.1	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	2	2	1	1	1	1	1	1	2	6	9	3	2	1	10	4	15
	14.14	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	2	2	1	1	1	1	1	1	3	14	4	3	2	2	5	8	10
	14.63	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	2	2	1	1	1	1	1	1	4	6	9	4	2	4	8	6	15
	16.61	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	2	1	1	2	1	1	1	1	2	1	1	1	1	2	1	1	1	1	2	2	1	4	18	8	3	2	3	7	6	10
	23.33	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	2	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	2	2	5	1	14	3	2	4	1	7	21
	23.39	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	2	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	2	2	5	1	13	3	2	4	10	6	21
	18.11	1	2	2	2	1	1	1	2	1	1	1	2	2	1	1	2	2	1	1	1	1	2	1	1	2	1	2	1	1	2	1	1	1	2	1	2	2	1	1	1	2	2	2	12	11	3	2	4	9	6	10
	18.13	1	2	2	2	1	1	1	2	1	1	1	2	2	1	1	2	2	1	1	1	1	2	1	1	2	1	2	1	1	2	1	1	1	2	1	2	2	1	1	1	2	2	3	8	19	3	2	6	8	7	16
	18.20	1	2	2	2	1	1	1	2	1	1	1	2	2	1	1	2	2	1	1	1	1	2	1	1	2	1	2	1	1	2	1	1	1	2	1	2	2	1	1	1	2	2	3	8	13	3	2	7	10	6	10
IV	18.21	1	2	2	2	1	1	1	2	1	1	1	2	2	1	1	2	2	1	1	1	1	2	1	1	2	1	2	1	1	2	1	1	1	2	1	2	2	1	1	1	2	2	4	9	1	3	2	5	7	7	9
	19.3	2	2	2	2	1	1	1	2	1	1	1	2	2	1	1	2	2	1	1	1	1	2	2	1	2	1	2	1	2	2	1	2	1	2	1	2	2	1	2	1	2	2	3	17	11	4	2	7	9	7	9
	20.10	1	2	2	2	1	1	1	2	1	1	1	2	2	1	1	1	2	1	1	1	1	2	1	1	2	1	2	1	1	2	1	1	1	2	1	2	2	1	1	1	2	2	2	8	8	3	2	7	10	7	22
	21.29	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	2	1	1	1	1	2	1	1	2	1	1	1	1	2	1	1	1	1	1	1	2	2	1	1	2	2	1	7	15	3	2	8	8	1	8
	22.12	1	2	2	2	1	1	1	2	1	1	1	1	2	1	1	2	2	1	1	1	1	2	1	1	2	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	2	2	1	10	16	3	1	8	9	1	27

The SNP data were converted into numerical data to generate a Minimum Spanning Tree. Only two possible alleles were observed in the SNP data and therefore for each SNP, the major allele was converted into number 1 while the minor one was designated as 2.

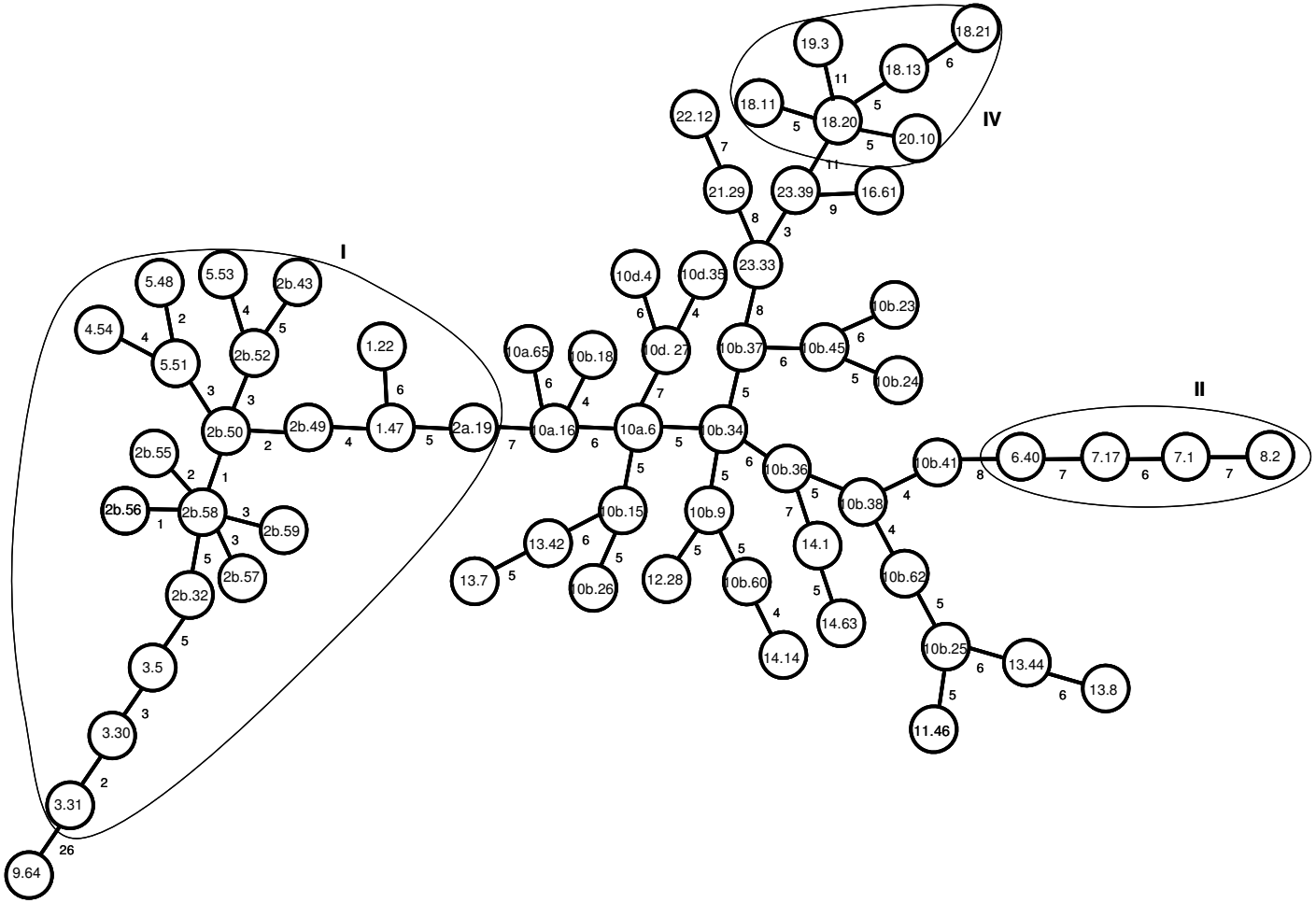


Figure 7.3-12. The minimum spanning tree to represent the relationship of the 66 profiles of 68 Typhi isolates, typed with 42 SNPs and 9 VNTRs markers. The numbers within the circle, which are separated by a dot, correspond to the SNP and MLVA profiles respectively. The size of the circle does not correspond to the number of isolates with that particular SNP and MLVA profile. The number on the node connecting two circles corresponds to the total difference between two profiles out of 51 markers. Clusters I, II and IV are labelled with roman numerals while the unlabelled profiles belonged to cluster III.

7.4. Discussion

7.4.1. Design and factors considered in current MLVA typing

In the present study, we have explored VNTR loci to differentiate 73 global Typhi isolates for typing purpose. We have only explored the possible polymorphism in VNTRs with repeat units of two or more. Mononucleotide repeat was not investigated because of the tendency of Taq polymerase to introduce an extra A at the 3' end of the amplified product (121). However, the addition of this extra A is inconsistent, making it difficult to accurately determine the number of repeat unit.

Forty-six potential VNTRs were identified in Typhi strain CT18 using the web-based resource tool. Five failed to produce any PCR product. PCR failure could be due to point mutations in flanking regions resulting in change of one or both primer sites. This was not investigated further. Only two of the 39 VNTRs with successful PCR amplification were polymorphic. These, together with the seven previously described VNTR markers, were used to type the Typhi isolates.

There was no correlation of either the repeat unit length, copy number or the % match with the level of polymorphisms. Typhi CT18 had more copies of SAL16 than SAL02 and TR1 respectively, but SAL02 and TR1 were found to be more discriminating than SAL16. TR2 had a unit length of 8 bp, which was the third longest amongst the nine VNTRs typed, yet it was the most polymorphic VNTR. Furthermore, not all polymorphic VNTRs had 100% match in all repeats. This was in contrast to previous findings that compared completed genomes from multiple strains of three bacterial species using the web-based tandem repeat database, including six strains of *Staphylococcus aureus*, four strains of *E. coli* and four strains of *Streptococcus pyogenes* (58). Polymorphic tandem repeats that differed in at least two of the strains were shown to have a higher percentage match (more than 80%), indicating a higher internal conservation between repeat units and a higher number of copies (total length of the repeat is longer than 80 bp). Similarly, in *Y. pestis*, higher polymorphism in VNTR was achieved with higher conservation of the repeat motif

(158) while in *B. anthracis*, a highly significant association was observed between polymorphism, total length and %GC content (158).

When performing fluorescent based MLVA typing, variability in fluorescent intensities needs to be considered. This knowledge is valuable especially if the four dyes are utilised simultaneously for typing VNTR loci. We have shown that the intensity was greatest for VIC followed by NED and least for both FAM and PET. This was in contrast to the manufacturer's recommendation (Applied Biosystem) where VIC and FAM have the same fluorescent intensity and their signal strengths are better than NED followed by PET. This discrepancy could be due to the quality of our M13 dye-labelled primers. The synthesised primers could have a different proportion of dye labelled primers in comparison to the unlabelled ones. Prolonged storage or frequent thawing/freezing could also contribute to the observed difference of fluorescent intensity. Newly synthesised M13 primers need to be tested to confidently explain the variation observed in the intensity of fluorescent signals.

Moreover, we have also established that each dye produced variable electrophoretic mobilities, so that the VNTRs ran as if they were shorter or longer by 1 or 2 bp, than their actual sizes. We have derived a formula to correct the size obtained when different dyes are used. VIC and NED produced identical size calling while FAM was 1 bp shorter and PET was 2 bp longer in comparison to VIC and NED respectively. Care must be taken when different dyes are utilised for analysing the same VNTR. One to 2 bp differences may not affect the VNTRs which consist of longer repeat unit as they could be considered as an artefact or what is usually referred as a stutter product. The longer the length of the repeat unit, the lesser stutter is produced. In VNTR of the same copy numbers, the percentage of stuttering is greater in VNTR consisting of two units than of three units and so on (332). However, in the case of shorter repeat units, this is significant as the results may be ambiguous. Mobility shift corrections must therefore be applied as above to the collected sizing data for VNTRs run using different dyes.

7.4.2. Comparison to previous MLVA schemes for Typhi and the advantage of typing VNTR using an automated DNA sequencer

7.4.2.1. The distribution of number of copies for VNTR loci was different between studies

Previously, two separate MLVA studies of Typhi have been done to determine the polymorphisms of selected VNTRs loci (174, 257). The first study utilised three VNTR loci designated as TR1 to TR3. There were 14, 29 and 4 distinct alleles detected for TR1, TR2 and TR3 respectively. A combination of the three TRs was able to discriminate 59 strains isolated from several Asian countries, including Singapore, Indonesia, India, Bangladesh, Malaysia, and Nepal between year 2000-2001 into 49 distinct MLVA profiles (174). Ramisse *et al.* (257) also typed TR1 to discriminate 27 Typhi strains isolated from blood and stool samples of symptomatic patients in France between the years 1993-1999. Together with six other VNTR loci, SAL02, SAL06, SAL10, SAL15, SAL16 and SAL20, they distinguished 27 Typhi isolates into 25 MLVA profiles. Altogether, seven VNTR loci with considerable D values, SAL02, SAL06, SAL10, SAL16, SAL20, TR1 and TR2, were selected to type 73 global Typhi isolates. TR3 was not included in our typing even though it has been shown to be discriminating in Liu *et al.* (174). This is because the length polymorphisms that were observed in TR3 could be due to the variations in the number of copies as well as deletions in the nucleotides at the 5' upstream region (174).

SAL10 and SAL06 have been shown to have the least discriminatory powers, consistent with previous study by Ramisse *et al.* (257). However, the D values were 0.14 and 0.43 in SAL10 and SAL06 respectively for 27 Typhi isolates, both of which were higher then when they were used to type 73 Typhi isolates (D values=0.04 and 0.23 for SAL10 and SAL06 respectively). The discriminatory power of SAL16 and SAL20 were shown to be higher in our study than the study by Ramisse *et al.* (257). TR1 and SAL02 were shown to have comparable discriminatory power (D value=0.87) for both VNTRs. However in our study SAL02 (D value=0.92) was more superior than TR1 (D value=0.89).

The alleles and the length polymorphisms were resulted from variations in the number of copies, and to confirm the accuracy, the PCR amplicon was sequenced. This was done in the study by Liu *et al.* (174) but not in Ramisse *et al.* (257). In both our study and the study by Ramisse *et al.* (257), the majority of isolates had five copies of SAL06 and two copies of SAL10. However, the distribution of number of copies in the remaining VNTR loci was different between studies. The majority of French isolates had either 14 or 17 copies, 14 and 15 copies for SAL02, SAL16 and SAL20 respectively (257) while the majority of our global isolates had 11, 7 and 16 copies for SAL02, SAL16 and SAL20 respectively. The majority of Asian isolates had 28 copies for TR2 (174) whilst most of the isolates in our study had either 12 or 17 copies of TR2. Interestingly, the distribution of number of copies for TR1 was different in all three studies, Ramisse *et al.* (257) observed 17 copies in most of their French isolates and we observed 16 copies in the majority of our isolates while 12 copies were observed in Typhi strains isolated from different Asian countries (174). This suggested that isolates with a particular number of copies for these VNTRs are dominant in a certain continent. Unfortunately, we could not validate this assumption as we only have a limited number of isolates from the same localities.

VNTRs have been increasingly used as a molecular marker to type homogeneous pathogens. To date, there are three studies, including our current study that utilised VNTRs to discriminate Typhi isolates from different localities. The collected sizing data from all three studies could be stored in a database accessible globally. Independent results could be shared between laboratories despite the different techniques utilised for VNTR typing. This also allows rapid computerised identification and classification of any Typhi isolates for epidemiological study.

7.4.2.2. The advantage of using multicoloured capillary over standard gel electrophoresis

Standard gel electrophoresis was used in previous MLVA studies of Typhi (174, 257). Although a multiplex scheme was done on three VNTRs in one of the studies (174), the resolution was too low and comigrating bands would influence the interpretation of results. Lane to lane and run to run

variabilities that result in varying electrophoretic mobilities need to be considered when using standard gel electrophoresis.

We could not mix the PCR products from our selected VNTRs to be run on 2% agarose gel as their size determinations would be difficult. In our study, the PCR products amplifying the VNTR loci were analysed with an automated DNA sequencer allowing the discrimination of fragments differing in size by four different fluorescent dyes. The incorporation of an internal standard sizing ladder in the same lane as an unknown VNTRs mixture minimises sizing errors due to those variations. Each of the four VNTRs, which were pooled in a sample, was run together with a size standard. Capillary electrophoresis makes the fragment size determination far more accurate than any standard gel based system.

Previously, Lindstedt *et al.* (165) has multiplexed two VNTR loci for serovar Typhimurium, using one fluorescent dye and they were run in a single cycling condition. This could be also applied for Typhi. Multiple VNTR loci can be amplified, using their respective primer pairs together with only one colour fluoresecent dye-labelled M13 primer, in a single PCR cycling condition. All primers are designed to amplify the regions containing VNTR loci. However, the sizes of the PCR products for each multiplexed VNTR locus need to be differed by at least 150 bp to prevent overlapping between loci and difficulty in interpretation of alleles. Some of the VNTRs had a wider range of sizes that corresponded to the different alleles. This has to be taken into consideration when creating spacing for co-typing different VNTR loci. Unfortunately multiplexing of all nine VNTRs from this study, by using only one type of fluorescent dye in a single PCR reaction will not be possible, until a longer size standard is available. Morever, designing primers for this purpose will be a challenge, especially to achieve a common annealing temperature for amplifying all VNTR loci. A more realistic multiplexing could be attained by using four types of fluorescent dyes. Three different dye-labelled M13 primers can be used to amplify two VNTR loci and one dye can be used to amplify three different VNTR loci. The VNTR loci that use the same fluorescent dye labelled primer will be run on the same PCR reaction. These different dyes will be pooled and run in a single capillary electrophoresis for each isolate.

7.4.3. VNTR variation observed in different stocks of CT18 and Ty2

Two discrepancies were observed between the theoretical predicted number of repeat unit and the observed one for SAL02 and SAL16 in strain CT18. Four gene sequences (Chapter 3) and forty one SNPs (Chapter 4) have shown that our CT18 strain was identical to the sequenced CT18 suggesting that the inconsistency was not because of a mix up in strain CT18. Possible explanations for this observation are sequencing errors or strain divergence between our CT18 and the sequenced CT18.

Another discrepancy was also observed in strain Ty2. The strain Ty2 we included in the analysis was designated as Ty2-b while the genome sequenced was LCDC TY2, both of which were wildtype Ty2 strains. These two strains have been shown to have different VNTR alleles in VNTR 4699 and both SAL02 and SAL16. Our Ty2 strain has also been shown to have different allele in one of the BiPs typed (Chapter 5) but not in 38 SNPs typed when compared to the genome data. These inconsistencies were not observed in the most diverse VNTRs. VNTR 4699 was the second most varied while SAL02 and SAL16 were the 3rd and 6th most varied respectively. Nonetheless, it was interesting to observe two identical pairs of strains, CC6 and CC7 and 444Ty and 702Ty, despite the rapidly evolving VNTR loci were typed.

A similar inconsistency has also been previously reported in a published literature concerning SAL16 for serovar Typhimurium (257). A discrepancy was observed between theoretical and observed number of copies for strain LT2, where 13 copies (instead of 15) were detected. It was not known whether freezing and passaging isolates multiple times as well as prolonged storage, would affect the stability of our nine VNTR loci. The observed differences in the copy numbers could have been due to this reason. MLVA typing of *S. enterica* serovar Enteritidis revealed different stability for the VNTRs typed (30). Two of the 12 isolates that were isolated from an individual over the course of 36 days had different copy number in one VNTR locus. However, MLVA profiles remained identical for isolates that have been passaged several times in laboratory settings.

7.4.4. Possible role of VNTR on the lifestyle of Typhi

Of the nine VNTR loci, only TR1 marker had a continuous range of copy numbers indicating a balanced distribution in the level of diversity in the Typhi isolates analysed while other VNTR loci with discontinuous range suggested the possibility of sampling bias or relatively high mutation rates for these loci. Another likely explanation is that a particular number of copies may be disadvantageous for the survival of these isolates. One of the polymorphic VNTRs was located on a pseudogene, two were on intergenic regions and the remaining six VNTRs were on the genic regions.

SAL06 was located on a pseudogene *STY0765*. Since pseudogenes are less likely to be under any selection pressure, it was expected that this VNTR was more likely to be variable. However, SAL06 was shown to be the VNTR with the second lowest discrimination out of nine VNTR loci typed. SAL16, TR1 and TR2 were located on the intergenic regions. Variations of the tandem repeats in intergenic regions have been shown to be useful for maintaining the secondary structure of the neighbouring genes (188). The remaining five VNTRs were all in genic regions. Two of the VNTRs, 4500 and SAL10, were located on *STY4635* and *yedD*, both encoding for hypothetical proteins. Both of these VNTRs were multiples of 6 bp and 12 bp respectively suggesting that variation in their copy numbers did not affect the translation of the reading frame. VNTR 4500 was 2.5 times more variable than SAL10 when these two VNTRs were typed for 73 global Typhi isolates. SAL10 was the least discriminating amongst all VNTR loci typed. Interestingly, the flanking region upstream of gene *yedD* contained the VNTR TR1, which was six times more polymorphic than SAL10. The significance of these hypothetical proteins is not yet known.

VNTR 4699 encoded for outer membrane proteins while two VNTRs, SAL02 and SAL20, encoded proteins which may be involved in metabolic activity of *S. enterica* serovar Typhi. VNTR 4699 was located on the gene *sefC* of the *sef* fimbrial operon, which encoded for outer membrane fimbrial usher protein. Two of the four genes within this operon, *sefA* and *sefD*, contained a stop codon in strain Typhi CT18 suggesting no selection pressure to maintain the genes within this operon (231). It was highly likely that the polymorphism in VNTR 4699 did not affect the overall expression of *sef* fimbrial operon as the operon itself is possibly non functional (315).

The repeat SAL02 was located on the *citT* gene encoding the citrate carrier. It appears that the protein, CitT, is a novel family of eubacterial transporters involving in the transport of di- and tricarboxylic acid (243). In Typhi CT18, this gene was located on an operon containing *citCDEFXGT* genes that coded for proteins and were not pseudogenes. The function of this operon has not been characterised in *S. enterica* serovars, but *citT* was present and was polymorphic in serovars Typhi strains CT18 and Ty2, Typhimurium strain LT2 and Paratyphi A strain SARB42. This suggested that the gene was not under stringent control and may not be essential for metabolism of *S. enterica* as citrate could not be utilised as a carbon source.

SAL02 was located on the *ftsN* gene and it was the third most polymorphic VNTRs that were typed in 73 global Typhi isolates. *ftsN* gene encodes for a membrane protein for localisation of cell division however, not much is known about its role in *S. enterica* serovars. In *E. coli*, FtsN protein is one of 10 other proteins that are involved in septum formation which occurs at midcell. FtsN is not essential and its absence could be replaced by a mutation that mimics the FtsN-induced conformational change in FtsA. A single amino acid change in FtsA could readily compensate the lack of functional FtsN (23). This suggests that the presence of SAL20 on this gene may not affect the cell division process.

A microarray study, which compared the genome contents of eight diverse Typhi isolates to Typhi strain CT18, revealed 13 regions of absent or divergent gene. However, the genes where VNTRs were located did not fall into any of these regions, suggesting that these genes were present and not divergent in all eight Typhi isolates examined. All six VNTRs in the genic regions consisted of repeat units which were multiples of three bases. Variation in the repeat units may not disrupt the reading frame. The VNTRs were located on hypothetical proteins, operon containing pseudogenes and genes which were not required for survival, suggesting that these VNTRs may not be associated with any biological function. Polymorphisms in these VNTRs may be random as they were not under selection pressures but may provide advantages in maintaining the bacterial diversities.

7.4.5. MLVA is a highly discriminating method to type the global Typhi isolates

A high level of discrimination among the 73 global Typhi isolates was achieved by MLVA. Nine VNTRs were able to differentiate 68 Typhi isolates into 65 distinct profiles with the discriminatory power of 0.999 suggesting that these VNTRs are highly discriminating. This means that if two isolates are sampled randomly from the global population, they will be classified as having different MLVA type on 99.9% of isolated strains. This was much higher than SNP typing (Chapter 4) and ribotyping (148) where the the discriminatory power was only 0.87 for both of them. The same level of discrimination could also be achieved by typing only four of the nine VNTR loci.

The individual VNTR markers were either not informative enough as in the case of SAL06 and SAL10 or too highly variable and showing a high level of homoplasies as shown in the remaining markers. We have used multilocus VNTRs to detect and quantify the similarities between Typhi isolates and it was expected that these loci would show different rates of evolution. Fast evolving loci allow diffentiation among closely related isolates while slow loci are valuable to discriminate distant Typhi isolates. Nevertheless, all of the VNTR loci appeared to mutate so fast that no particular allele became fixed in any group of isolates. For example, the five alleles of VNTR 4500 were not exclusively present in any of the clusters. Combination of independent loci was able to reveal phylogenetic story. However, MLVA alone did not retain any phylogenetic information which could be used to trace the evolutionary histories and relationships of the 68 global Typhi isolates.

High level of diversity within VNTR loci has enabled differentiation of these closely related strains, however they could not be used to determine the true evolutionary relationships due to homoplasy, as seen in the constructed MST. There were clear differences in the distributions of various strains on the MST created with both SNP typing and MLVA. The congruence between the two trees was low suggesting very different evolutionary events. We attempted to circumvent this problem by combining the VNTR data with previously collected SNPs data to determine whether the overall genetic relationships of Typhi isolates can be better resolved. No clonal complex was observed however, three of the four major clusters were still evident.

Noller *et al.* (223) suggests that for isolates to be considered belonging to the same outbreak, the intralocus differences should only be contributed by one repeat at a time among them. The majority of the Typhi isolates typed differed by more than one VNTR between one another. Only isolate 445Ty differed by one VNTR to 417Ty and 425Ty VNTR respectively. Unfortunately, only the location of the strain 417Ty was known while there was a limited information on the other two strains. Therefore, we were unable to confirm if these isolates were involved in or isolated from a local outbreak.

7.5. Conclusion

Typing of tandem repeats were explored in the genome sequence data of Typhi strain CT18 for typing. Forty six potential VNTRs were selected and examined for possible polymorphisms in a panel of 12 Typhi isolates. Only two tandem repeats were found to be polymorphic. Together with seven previously published VNTRs they were typed in a collection of 73 global Typhi isolates. Only 68 Typhi isolates were selected for VNTR typing because five isolates were unable to amplify at least one of the VNTR loci. Nine VNTR loci could differentiate 68 Typhi isolates into 65 MLVA profiles and the same discrimination could be achieved by typing only four of the nine VNTR loci. MLVA was much more discriminating than SNP typing, ribotyping and MLST. Our data suggested that VNTR was a more rapidly evolving marker in comparison to SNP. However, VNTR alone was unable to resolve the relationships of most of the Typhi isolates analysed. A combination of both SNP and VNTR markers was unable to determine any clonal complexes that were identified using 42 SNPs as the markers (Chapter 4 and Chapter 5).

Although no causal relationship was established between the isolates typed, MLVA could be helpful to indicate the presence of clonal expansions, especially in epidemiologically linked isolates. We have shown that MLVA could be used for global epidemiology of Typhi, however we were unable to determine if the same level of high discrimination could also be achieved when typing local outbreaks. A limitation of the present study was the restricted number of isolates investigated. Testing of a larger number of isolates from various geographic areas will ultimately

lead to the development of an international Typhi fingerprint database accessible for epidemiological use.

Furthermore, we did not examine the stability of these VNTR markers. Future studies are required to investigate whether MLVA profiles remain constant for Typhi isolates that have undergone series of laboratory passages and strains which are isolated from the same patient over a period of time. This is significant to determine the long term epidemiological value of these VNTRs as markers for molecular typing.

8. General discussion and conclusion

8.1. The importance of this study

This thesis focused on investigating the evolution and diversity of *Salmonella enterica*, in particular serovar Typhi. *S. enterica* serovars show different host specificity and disease severity. The serovars causing 99% of *Salmonella*-related infection (Salmonellosis) in humans belong to subspecies I. The most common disease caused is acute enterocolitis but enteric fever is a significant disease in developing countries. Several serovars cause enteric fever. Serovar Typhi causes typhoid fever while serovars Paratyphi A, Paratyphi B clone b1, Paratyphi C and Sendai, cause the milder form of typhoid-like enteric fevers. Little is known about the relationship among the different serovars which cause enteric fever; whether these serovars are closely related or whether they evolved from multiple phylogenetic descents. Sequence data of six genes were used to establish the extent of the genetic diversity and the evolutionary relationship between enteric fever causing serovars and other serovars belonging to subspecies I.

As a relatively young pathogen, Typhi is highly homogeneous (141). The lack of genetic diversity is a major challenge to develop suitable typing methods to differentiate Typhi isolates, for both phylogenetic and epidemiological purposes. In this study, two molecular markers were used to differentiate and to establish the evolutionary relationships of Typhi isolates. These markers were genome wide single nucleotide polymorphisms (SNPs) and variable number of tandem repeats (VNTRs). They exhibited different levels of discriminatory power. The methods developed could be used for surveillance and epidemiological investigations. The data obtained, provided a better insight into the diversity and evolution of Typhi.

8.2. Recombination is a major factor contributing to the diversification of S. enterica subspecies I and the relationships of enteric fever causing serovars remain unresolved

Multilocus enzyme electrophoresis has shown that there is no close evolutionary relationship between enteric fever causing serovars, except for serovar Paratyphi A and Sendai (285). In contrast, microarray study has shown that Typhi is closely related to serovars Paratyphi A and Sendai by the gene content difference (43). Clearly, the relationships of enteric fever causing serovars need to be resolved using DNA sequence of housekeeping genes. Knowing the relationships will enable us to understand the evolution, emergence and development of enteric fever clones.

The genetic relationships of serovars from subspecies I were analysed using nucleotide sequences of six genes (*mdh*, *mglA*, *mutS*, *proV*, *torC* and *speC*). In particular, we focused on the serovar closest to serovar Typhi, and the relative contributions of recombination and point mutation to clone diversification in subspecies I. We analysed 15 strains from 13 serovars obtained from the *Salmonella* Reference Collection B (SARB) representing the genetic diversity of the subspecies I.

Phylogenetic analysis revealed that there was a lack of congruence among the six gene trees. In general, only three serovars broadly fell within the same cluster when compared to MLEE tree. Split decomposition analysis only resolved five strains with a network structure while others showed a star phylogeny. The level of recombination in *S. enterica* subspecies I was determined by calculating the compatibility values for within and between the six genes. The values were compared with those obtained from strains representing different subspecies of *S. enterica* and with the pathogenic *E. coli*. *torC*, *speC* and *mdh* were the most compatible within and between gene loci while *mutS*, *proV* and *mglA* were the least compatible. The average compatibility values were lower for the SARB strains in comparison to SARC strains and *E. coli*.

The degree of incongruence among the six genes trees was established using Maximum Likelihood (ML) analysis. Gene trees with the largest number of congruences were of *mdh* and *mutS*, which were congruent to three other gene trees. *proV* and *torC* were only congruent to two gene trees while *speC* was only to another gene tree. The *mglA* tree was not congruent to any of the gene trees. Overall, only 37% congruencies were detected among the SARB strains while all the gene trees were congruent to each other among the strains representing different subspecies of *S. enterica*. High rate of recombination was evident from the low compatibility values and incongruencies between gene trees in the SARB strains studied. These results showed that the genes studied have undergone frequent recombination, suggesting a low level of clonality within subspecies I of *S. enterica*.

This was in contrast with previous findings where *S. enterica* has been generally regarded as the species with the highest level of clonality (192). *S. enterica* was viewed as a species with a low rate of recombination and any genetic variations between clones were considered to be largely resulted from mutations (192). MLEE studies of *S. enterica* natural populations have shown that *S. enterica* has a strong linkage disequilibrium between alleles in the 24 metabolic enzyme loci. It has a higher or equal index of association (I_A) for individual serovars than that as a species (192). Unfortunately, this conclusion was reinforced by the misinterpretation of the I_A analysis from the subspecies I MLEE data. Heterogeneous serovars, such as serovar Paratyphi C and Choleraesuis were treated as a single population and this resulted in an artificially high I_A.

Sequence data of five housekeeping genes from 16 strains of the *Salmonella* Reference Collection C (SARC), representing each of the eight subspecies provided a strong support of the MLEE data that *S. enterica* is highly clonal. The topologies of phylogeny trees constructed from these sequence data are generally congruent, suggesting a low level of recombination (34). However, the housekeeping gene studies have only used two isolates to represent a subspecies, which would not allow identification of recombination events within a subspecies.

Our data suggested that recombination is one of the key factors which lead to the evolution and diversification of *S. enterica* subspecies I. The evolutionary origins and the genetic relationships of enteric fever causing serovars remain unresolved. Currently, there are only 12 genomes for serovars belonging to subspecies I that are either fully or almost completely sequenced. Considering the high levels of recombination, the entire genome sequences will be required to obtain the best analysis to determine the evolutionary relationships. This will be achievable in the near future with the lowering cost of full genome sequencing.

8.3. Single base mutations are valuable markers to establish the evolutionary relationships of global Typhi isolates

Population structure studies have revealed that Typhi is highly homogeneous; the clone was estimated to be about 50,000 years old suggesting that Typhi only had a relatively short time frame to accumulate variation (141). Due to its homogeneity, it is often challenging to distinguish isolates responsible for outbreaks or sporadic cases of typhoid fever and even harder to establish the evolutionary relationships between these isolates. DNA-based molecular typing methods such as pulse field gel electrophoresis and ribotyping have demonstrated their effectiveness to differentiate isolates for epidemiology studies of Typhi. However, these methods have limited capacity to define the phylogenetic relationships. Therefore, alternative methods with a higher discriminatory power at nucleotide level are needed.

We have developed a typing strategy using genome-wide Single Nucleotide Polymorphisms (SNPs) as the markers. The principle of the method is to characterize a SNP by the presence or absence of the restriction enzyme recognition site. Two complete Typhi genomes of strains CT18 and Ty2 allowed us to screen for candidate SNPs that could be subjected to typing. Seven economical 4-bp cutter restriction enzymes were utilised to type 38 SNPs on a collection of 73 worldwide Typhi isolates.

The 73 isolates can be grouped into 23 SNP profiles, 12 profiles were represented by a single isolate and 11 profiles were shared by more than one isolate. The overall relationships of these SNP profiles were visualised using a minimum spanning network. We could resolve the relationships of most SNP profiles with the exception of three profiles (4, 16 and 17) which had equal distance to more than one SNP profile, shown as network structures on the tree. The 23 SNP profiles were divided into four major clusters, I to IV. Cluster III was directly connected to the non-Typhi serovars as the outgroup, and gave rise to the other 3 clusters. SNP profile 10 within this cluster appeared to be the ancestral profile. Correlations between phylogenetic groupings with the phage types and/or genome types were only observed in two clusters. This study has demonstrated the usefulness of genome wide SNPs for molecular typing and determining relationships among the Typhi isolates.

8.4. z66 flagellar antigen and the origin of Typhi

The majority of serovars in this subspecies are diphasic as they express two flagellar antigens encoded by *fliC* and *fljB* genes at H1 and H2 locus respectively. In contrast, serovar Typhi is mostly monophasic and can only express the antigen at H1 locus. However, some Indonesian isolates can also express an additional antigen, named as the z66 antigen, thought to be encoded by *fljB* at the H2 locus (96). Based on this observation, it has been proposed that serovar Typhi was originated in Indonesia and was initially diphasic carrying a H2-like-operon which was subsequently lost to become monophasic before they spread globally.

The 18 Typhi isolates expressing the unique flagellar antigen z66 included in our study were divided into four SNP profiles. They were found to be clustered together and branched off from the same ancestral group, suggesting a single origin. Our new data suggested that serovar Typhi only had an H1 antigen and then gained a new phase-2-flagellin operon only recently during the divergence of cluster I. The gene encoding z66

antigen was more similar to H27 *fliC* of *E. coli* than to the other H antigen genes *of S. enterica* (122) and was located on a linear plasmid (13), which further supports the hypothesis that z66 was obtained only recently by lateral transfer and it appears that serovar Typhi did not originate in Indonesia.

8.5. HP R-T PCR is an alternative method for SNP typing

In chapter 4, the method used for SNP typing was restriction enzyme digestion. This method has been shown to be simple, economical and reliable. However, one major setback of this method was the availability of suitable and cheap restriction enzyme to recognise the SNP while maintaining the low cost for typing. The study by Roumagnac *et al.* (271) utilised denaturing liquid performance chromatography (dHPLC) for SNP typing. However, this was more laborious. An alternative to these SNP typing methods was hairpin real time PCR assay (HP R-T PCR assay) (104). This method was not gel-based like the other two methods and the results could be obtained directly at the completion of real time PCR. This method relied on the hairpin (HP) structure on the allele specific primers. By far, this was the best choice of methods for SNP typing provided that real time platform was available.

The method has been tested to type four SNPs (Chapter 5) and it has been shown to be reliable. The identity of the allele for each SNP could be confidently identified based on the difference of the Ct values, with an average of 5.16, obtained between matched and mismatched HP primers. This method was relatively simple. The only disadvantage was that each SNP needed to be typed in separate reactions for each allele and each SNP was typed individually. Thus, considerable preparations will be required to type a large number of isolates if many SNPs are used.

There are other high throughput methods for SNP typing, including microarrays (44); mass spectrometry based techniques (276); TaqMan probes (175); and molecular beacons (316). However, these methods are very costly and may require special equipments and

laboratory settings. These methods are impractical to be applied, especially in developing countries where typhoid fever is prevalent and cost-efficient surveillance is most needed.

8.6. Typing of four additional SNP using HP R-T PCR assay improved the resolution our SNP typing

A study by Roumagnac *et al.* (271) used 88 SNPs, which they referred to as biallelic polymorphisms (BiPs), as markers to determine the evolutionary relationships between 481 global Typhi isolates. Four BiPs, BiP 36, 48, 56 and 33, were found to divide these isolates into five major clusters. These BiPs were selected for typing of our 73 Typhi isolates using HP R-T PCR.

Previously, we have shown that 73 Typhi isolates could be distinguished into 23 SNP profiles (Chapter 4). Typing an additional of four BiPs differentiated two SNP profiles into two and four subprofiles respetively. SNP profile 2 was differentiated into two subprofiles, 2a and 2b and SNP profile 10 was differentiated into four SNP subprofiles, 10a to 10d respectively. A new cluster was also found, based on Roumagnac *et al.* (271) clustering scheme, therefore there were six clusters in total. The combination of our cluster supporting SNPs and Roumagnac *et al.* (271) cluster supporting BiPs allowed subdivision of the clusters into 13 subclusters. These can be used for global epidemiological study of Typhi.

8.7. Parallel or reverse changes were observed in Typhi isolates, suggesting recombination within a clone

In chapter 4, six of the 38 selected SNPs including SNP 8, 11, 17, 35, 36 and 37 were found be present in multiple lineages. This was an indication of conflicting phylogenetic signal or homoplasy. Homoplasy was resulted from parallel or reverse changes that could be either due to multiple independent mutations or recombination. This was in contrast

with the study by Roumagnac *et al.* (271), in which no homoplasy was observed in all 88 BiPs analysed.

However, based on our clustering scheme, one of the four BiPs from Roumagnac *et al.* (271), BiP 48 has also been shown to undergo parallel or reverse changes in our isolates (Chapter 5). It was shown that SNP profiles from different clusters carried the same allele. If our SNP profiles were arranged according to Roumagnac *et al.* (271) scheme, more SNPs were found to be conflicting. Therefore, in both situations, parallel or reverse changes must have occurred. Homoplasy was present and this was likely to have resulted from recombination within the clone.

8.8. SNP-base typing method for global epidemiology

We have shown that SNPs were useful for molecular typing and inferring the phylogenetic relationships of global Typhi isolates. A total of 42 SNPs were typed in the 73 global Typhi isolates, including 38 SNPs from Chapter 4 and four BiPs from Roumagnac *et al.* (271). From this data, we selected a set of minimum number of SNPs for identification at three levels: 1) the major clusters, 2) 13 subclusters and 3) individual SNP profiles. A minimum of three cluster supporting SNPs (SNP 2, SNP 11 and SNP 17) or four BiPs (BiP 36, BiP 48, BiP 56 and BiP 33) were sufficient to group Typhi isolates into major clusters according to either clustering schemes. Nine SNPs (five SNPs from our study and four BiPs from Roumagnac *et al.* (271)) could be used to distinguish Typhi isolates into 13 subclusters (Table 5.3-5). Lastly, a minimum of 19 SNPs (15 SNPs and four BiPs) could differentiate Typhi isolates into 27 SNP profiles (Table 5.3-4). This subset of SNPs could be used to perform significantly larger studies to distinguish any isolates at different evolutionary scales, and to determine the origin and the global distribution of Typhi clones. The SNP typing method developed in this study (Chapter 4 and Chapter 5) will be a valuable tool for global epidemiology of Typhi.

8.9. A new strategy to discover novel SNPs using enzymatic based method

The 38 SNPs selected for typing were identified from the comparison of two complete genomes of Typhi, strains CT18 (231) and Ty2 (57). This resulted in the phylogenetic bias where only the path of the last common ancestors connecting strains CT18 and Ty2 was revealed. No branch was evident in any other cluster despite the node locations for remaining clusters being well established. Roumagnac *et al.* (271) has demonstrated the use of dHPLC for discovering more SNPs. Their study has shown the presence new SNPs, which could not be identified solely by comparison of the two Typhi genomes. This further highlights the need to discover more SNPs, especially from isolates that represent distinct phylogenetic lineages. The discovery of more SNPs will provide a better insight into the complexity of relationships among isolates between and within each cluster. To achieve this, we attempted to develop an enzymatic-based approach, which was simpler and less labourious than dHPLC, to discover more SNPs. The single stranded nuclease, *Cel*I available in a kit, Surveyor^{TM} (Transgenomics) could specifically recognise and digest fragments containing single base mutations. Previous studies have shown that *Cel*I was useful for SNP detection (15, 163, 251, 291, 292). The potential use of this *Cel*I nuclease for SNP discovery was investigated (Chapter 6). However, we have shown that *Cel*I has a poor cleavage activity and we were unable to demonstrate the usefulness of our proposed method for SNP discovery. The project was discontinued due to time constraints. We have demonstrated that our method was valid for most of the steps. This method still has its potential provided that the cleavage activity of *Cel*I nuclease can be optimised.

8.10. Two new VNTRs were identified and included in MLVA for Typhi

Variable number of tandem repeats (VNTRs) appear to be the most variable marker. Multiple VNTRs have been increasingly investigated for typing of bacterial pathogens in

an assay called multiple-locus VNTR analysis (MLVA). Since its recent development, MLVA has proven to be valuable for discriminating isolates from homogeneous species such as *B. antracis* and *Y. pestis*. Two MLVA studies have also been performed for *S. enterica* serovar Typhi. In the first MLVA study for Typhi, three VNTRs were found to be polymorphic in 59 Typhi strains that were isolated from several Asian countries (174). One of these VNTRs was used in the second MLVA study along with six other polymorphic VNTRs to type 27 Typhi strains isolated from France (257). We explored the Typhi CT18 genome to find more polymorphic VNTRs to be included in the MLVA assay.

There were 46 potential VNTRs and these were typed on a panel of 12 diverse Typhi isolates. Five failed to produce a PCR product at various annealing temperature and 39 VNTRs showed no variation among the 12 isolates. Only two were found to be polymorphic, and together with the seven of the most polymorphic VNTRs from the previous two MLVA studies (174, 257), they were included in our MLVA typing of global Typhi isolates.

8.11. Method advancement for VNTR typing

Previous MLVA studies for Typhi (174, 257) used conventional gel electrophoresis to differentiate the copy numbers in each VNTR locus. Our current MLVA assay employed universal M13 primers that have been attached with dye and a hemi-nested touch down PCR was developed, as based on the principle by Schuelke *et al.* (279), to differentiate the VNTRs on capillary electrophoresis. Having fluorescent dyes only on the universal M13 primer has significantly reduced the cost for our typing. Furthermore, four VNTR loci were simultaneously typed by using four different fluorescent dyes (VIC, PET, FAM and NED) attached to the universal M13 primer. These VNTRs were pooled in a sample and run together with a size standard. The use of capillary electrophoresis has made the fragment size determination more accurate than the standard gel based system. We have also shown that the dyes have different fluorescent intensity. This knowledge is valuable

especially if the four dyes are to be utilised simultaneously for typing VNTR loci. Furthermore, we have also established that different dyes have different electrophoretic mobility. We have derived a formula to correct the size obtained when different dyes are used. VIC and NED produced identical size calling while FAM was 1 bp shorter and PET was 2 bp longer in comparison to VIC and NED respectively.

8.12. VNTR typing offers higher discriminatory power than SNP typing for molecular typing of Typhi

The nine VNTR loci were typed in 73 Typhi isolates to determine the discriminatory power. One VNTR (VNTR 4500) was found to be untypeable in four isolates and an additional VNTR (VNTR 4699) was untypeable in two isolates. The discriminatory power of each VNTR differed. It ranged from 0.044 to 0.964 and averaged at 0.706. Typing of nine VNTR loci gave a discriminatory power of 0.999. The nine VNTR loci could differentiate 68 fully typeable Typhi isolates into 65 MLVA profiles. The same discrimination can also be achieved by typing only four of the nine VNTR loci, including VNTR 4699, SAL02, SAL20 and TR2. VNTR typing offered a much higher discriminatory power than SNP typing of 42 SNPs, ribotyping and MLST (Chapter 7, Figure 7.3-5) and is expected to be highly valuable for local epidemiology.

8.13. Comparison of VNTRs and SNPs as molecular markers to establish the genetic relationships of global Typhi isolates

An unweighted pair group method with arithmetic means (UPGMA) dendogram divided the 65 MLVA profiles into four major clusters while a minimum spanning tree (MST) showed no clustering of MLVA profiles. The four major clusters identified by SNP typing were not apparent in either the UPGMA dendogram or MST generated from

VNTR data. This suggested that VNTRs could not be used to determine relationships between the Typhi isolates analysed as the VNTRs may have been evolving too fast.

To further investigate this problem, the MLVA profiles were divided according to the four major clusters that have been identified by SNP typing. For each individual cluster, a MST was generated using the VNTR data and it was compared to the MST generated by the SNP data. The MSTs from the two datasets were congruent for the majority of profiles from cluster I, II and IV. The relationships of the profiles within these clusters were generally consistent. However, in cluster III, the relationships of the profiles were conflicting between the two types of data. This could be due to a higher level of diversities of isolates within this cluster in comparison to the other three clusters. This was also reflected by the absence of cluster supporting SNP for cluster III. This suggested that VNTRs could be used to determine the relationships of closely related isolates. However, in more divergent Typhi isolates, VNTRs were only useful to reveal the extent of genetic diversity but they could not be solely used for determining the evolutionary relationships of Typhi. SNPs are required to determine the phylogenetic relationships of these diverse isolates and for cluster III, more SNPs are needed for a better resolution.

In conclusion, the two typing methods offer different levels of sensitivity and reveal phylogenetic signals at different levels. A combination of SNP typing and VNTR typing allows the best resolution to establish the evolutionary relationships and highly discriminating for local and global epidemiology.

8.14. Concluding remarks

S. enterica has traditionally been thought as the species with high level of clonality (192, 283). Our study has challenged the clonal paradigm as evidenced by substantial recombination between serovars of subspecies I. This suggested that the relative contributions of point mutation and recombination to the divergence of *S. enterica* are different between and within subspecies. We were unable to determine the relationship

between enteric fever causing serovars due to frequent recombinations within subspecies I. Sequencing a large number of genes or full genome data may allow this to be resolved in the future.

Typhoid fever remains as a devastating disease in developing countries. For better public health control and prevention, epidemiological monitoring of the spread of Typhi clones, locally and globally, is essential. Current DNA based typing methods include pulse field gel electrophoresis and ribotyping. Unfortunately, these methods could not be used to establish the evolutionary relationships of Typhi isolates. This thesis represents one of the first studies to explore the use of SNPs and VNTRs for molecular typing of global Typhi isolates. SNP typing was more discriminating than ribotyping and MLST, and it could be used to resolve the evolutionary relationships for most of the isolates analysed. We have also shown that MLVA of nine VNTRs has provided an excellent discrimination of the 73 Typhi isolates used. However, due to a limited number of isolates, we are yet to demonstrate the usefulness of both SNPs and VNTRs for local epidemiological studies.

Both SNP typing and VNTR typing will be valuable for global surveillance of Typhi. These molecular typing methods are useful in resolving the evolutionary relationships of Typhi isolates and revealing the extent of the genetic diversity within Typhi. The information gained provides a better insight into the genome dynamics of Typhi as a host restricted pathogen. An understanding in the biological correlation of this genetic diversity will have important implications on the factors that are involved in the epidemiology of Typhi. Ultimately, more effective strategies can be developed to control the persistence of Typhi and prevent the emergence and spread of multidrug resistant Typhi strains.

References

1. GeneScan® Reference Guide: Chemistry Reference for ABI Prism® 310 Genetic Analyzer. Applied Biosystems.
2. **Achtman, M., G. Morelli, P. Zhu, et al.** 2004. Microevolution and history of the plague bacillus, *Yersinia pestis*. Proc. Natl. Acad. Sci. U. S. A. **101:**17837–17842.
3. **Achtman, M., K. Zurth, G. Morelli, et al.** 1999. *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*.[erratum appears in Proc Natl Acad Sci U S A 2000 Jul 5;97(14):8192]. Proc. Natl. Acad. Sci. U. S. A. **96:**14043-8.
4. **Adair, D. M., P. L. Worsham, K. K. Hill, et al.** 2000. Diversity in a variable-number tandem repeat from *Yersinia pestis*. J. Clin. Microbiol. **38:**1516-9.
5. **Ali, S., A. M. Vollaard, S. Widjaja, et al.** 2006. PARK2/PACRG polymorphisms and susceptibility to typhoid and paratyphoid fever. Clin. Exp. Immunol. **144:**425-431.
6. **Alland, D., D. W. Lacher, M. H. Hazbo?n, et al.** 2007. Role of large sequence polymorphisms (LSPs) in generating genomic diversity among clinical isolates of Mycobacterium tuberculosis and the utility of LSPs in phylogenetic analysis. J. Clin. Microbiol. **45:**39-46.
7. **Amavisit, P., D. Lightfoot, G. F. Browning, et al.** 2003. Variation between pathogenic serovars within *Salmonella* pathogenicity islands. J. Bacteriol. **185:**3624-35.
8. **Andersson, J. O., and S. G. Andersson.** 2001. Pseudogenes, junk DNA, and the dynamics of *Rickettsia* genomes. Mol. Biol. Evol. **18:**829-39.
9. **Ansong, C., H. Yoon, A. D. Norbeck, et al.** 2008. Proteomics Analysis of the Causative Agent of Typhoid Fever. J. Proteome Res. **7:**546-557.
10. **Arricau, N., D. Hermant, H. Waxin, et al.** 1998. The RcsB-RcsC regulatory system of *Salmonella typhi* differentially modulates the expression of invasion proteins, flagellin and Vi antigen in response to osmolarity. Mol. Microbiol. **29:**835-50.

11. **Arya, S. C.** 1999. Efficacy of Vi polysaccharide vaccine against *Salmonella typhi*.[comment]. Vaccine **17:**1015-6.

12. **Baehr, W., E. C. Gotschlich, and P. J. Hitchcock.** 1989. The virulence-associated gonococcal H.8 gene encodes 14 tandemly repeated pentapeptides. Mol. Microbiol. **3:**49-55.

13. **Baker, S., J. Hardy, K. E. Sanderson, et al.** 2007. A novel linear plasmid mediates flagellar variation in *Salmonella typhi*. PLoS Path. **3:**605-610.

14. **Bandelt, H. J., and A. W. Dress.** 1992. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. Mol. Phylogenet. Evol. **1:**242-252.

15. **Bannwarth, S., V. Procaccio, and V. Paquis-Flucklinger.** 2005. Surveyor Nuclease: a new strategy for a rapid identification of heteroplasmic mitochondrial DNA mutations in patients with respiratory chain defects. Hum. Mutat. **25:**575-82.

16. **Barrett, T. J., J. D. Snyder, P. A. Blake, et al.** 1982. Enzyme-linked immunosorbent assay for detection of *Salmonella typhi* Vi antigen in urine from typhoid patients. J. Clin. Microbiol. **15:**235-7.

17. **Bastin, D. A., L. K. Romana, and P. R. Reeves.** 1991. Molecular cloning and expression in *Escherichia coli* K-12 of the *rfb* gene cluster determining the O antigen of an *E. coli* O111 strain. Mol. Microbiol. **5:**2223-2231.

18. **Baumler, A., R. Tsolis, T. Ficht, et al.** 1998. Evolution of Host Adaptation in *Salmonella enterica*. Infect. Immun. **66:**4579-4587.

19. **Baumler, A. J., R. M. Tsolis, and F. Heffron.** 1996. The *lpf* fimbrial operon mediates adhesion of *Salmonella typhimurium* to murine Peyer's patches. Proc. Natl. Acad. Sci. U. S. A. **93:**279-83.

20. **Bayliss, C. D., M. J. Callaghan, and E. R. Moxon.** 2006. High allelic diversity in the methyltransferase gene of a phase variable type III restriction-modification system has implications for the fitness of *Haemophilus influenzae*. Nucleic Acids Res. **34:**4046-59.

21. **Beltran, P., J. M. Musser, R. Helmuth, et al.** 1988. Toward a population genetic analysis of *Salmonella*: genetic diversity and relationships among strains of

serotypes *S. choleraesuis*, *S. derby*, *S. dublin*, *S. enteritidis*, *S. heidelberg*, *S. infantis*, *S. newport*, and *S. typhimurium*. Proceedings of the National Academy of Sciences USA **85:**7753-7757.

22. **Beltran, P., S. A. Plock, N. H. Smith, et al.** 1991. Reference collection of strains of the *Salmonella typhimurium* complex from natural populations. J. Gen. Microbiol. **137:**601-606.
23. **Bernard, C. S., M. Sadasivam, D. Shiomi, et al.** 2007. An altered FtsA can compensate for the loss of essential cell division protein FtsN in *Escherichia coli*. Mol. Microbiol. **64:**1289-305.
24. **Bernstein, A., and E. M. Wilson.** 1963. An Analysis of the Vi-Phage Typing Scheme for *Salmonella typhi*. J. Gen. Microbiol. **32:**349-73.
25. **Best, E. L., B. A. Lindstedt, A. Cook, et al.** 2007. Multiple-locus variable-number tandem repeat analysis of *Salmonella enterica* subsp. *enterica* serovar Typhimurium: comparison of isolates from pigs, poultry and cases of human gastroenteritis. J. Appl. Microbiol. **103:**562-572.
26. **Blanc-Potard, A. B., and E. A. Groisman.** 1997. The *Salmonella selC* locus contains a pathogenicity island mediating intramacrophage survival. EMBO J. **16:**5376-85.
27. **Blanc-Potard, A. B., F. Solomon, J. Kayser, et al.** 1999. The SPI-3 pathogenicity island of *Salmonella enterica*. J. Bacteriol. **181:**998-1004.
28. **Bonfield, J., K. Smith, and R. Staden.** 1995. A new DNA sequence assembly program. Nucleic Acids Res. **23:**4992-4999.
29. **Boniotto, M., M. H. Hazbo?n, W. J. Jordan, et al.** 2004. Novel hairpin-shaped primer assay to study the association of the -44 single-nucleotide polymorphism of the DEFB1 gene with early-onset periodontal disease. Clin. Diagn. Lab. Immunol. **11:**766-769.
30. **Boxrud, D., K. Pederson-Gulrud, J. Wotton, et al.** 2007. Comparison of multiple-locus variable-number tandem repeat analysis, pulsed-field gel electrophoresis, and phage typing for subtype analysis of *Salmonella enterica* serotype Enteritidis. J. Clin. Microbiol. **45:**536-43.

31. **Boyd, E. F., K. Nelson, F.-S. Wang, et al.** 1994. Molecular genetic basis of allelic polymorphism in malate dehydrogenase (*mdh*) in natural populations of *Escherichia coli* and *Salmonella enterica*. Proc. Natl. Acad. Sci. U. S. A. **91:**1280-1284.

32. **Boyd, E. F., S. Porwollik, F. Blackmer, et al.** 2003. Differences in gene content among *Salmonella enterica* serovar Typhi isolates. J. Clin. Microbiol. **41:**3823-8.

33. **Boyd, E. F., F.-S. Wang, P. Beltran, et al.** 1993. *Salmonella* reference collection B (SARB): strains of 37 serovars of subspecies 1. J. Gen. Microbiol. **139:**1125-1132.

34. **Boyd, E. F., F. S. Wang, T. S. Whittam, et al.** 1996. Molecular genetic relationships of the *Salmonellae*. Appl. Environ. Microbiol. **62:**804-808.

35. **Brenner, F. W., R. G. Villar, F. J. Angulo, et al.** 2000. *Salmonella* Nomenclature. J. Clin. Microbiol. **38:**2465-2467.

36. **Brinig, M. M., C. A. Cummings, G. N. Sanden, et al.** 2006. Significant gene order and expression differences in *Bordetella pertussis* despite limited gene content variation. J. Bacteriol. **188:**2375-2382.

37. **Brown, E. W., M. L. Kotewicz, and T. A. Cebula.** 2002. Detection of recombination among *Salmonella enterica* strains using the incongruence length difference test. Mol. Phylogenet. Evol. **24:**102-120.

38. **Brown, E. W., M. K. Mammel, J. E. LeClerc, et al.** 2003. Limited boundaries for extensive horizontal gene transfer among *Salmonella* pathogens. Proc. Natl. Acad. Sci. U. S. A. **100:**15676-15681.

39. **Bueno, S. M., C. A. Santiviago, A. A. Murillo, et al.** 2004. Precise excision of the large pathogenicity island, SPI7, in *Salmonella enterica* serovar Typhi. J. Bacteriol. **186:**3202-13.

40. **Campo, N., M. J. Dias, M. L. Daveran-Mingot, et al.** 2004. Chromosomal constraints in Gram-positive bacteria revealed by artificial inversions. Mol. Microbiol. **51:**511-22.

41. **Chabaud, M., J. M. Durand, N. Buchs, et al.** 1999. Human interleukin-17: A T cell-derived proinflammatory cytokine produced by the rheumatoid synovium. Arthritis Rheum. **42:**963-70.

42. **Chain, P. S., E. Carniel, F. W. Larimer, et al.** 2004. Insights into the evolution of *Yersinia pestis* through whole-genome comparison with *Yersinia pseudotuberculosis*. Proc. Natl. Acad. Sci. U. S. A. **101:**13826-13831.

43. **Chan, K., S. Baker, C. C. Kim, et al.** 2003. Genomic comparison of *Salmonella* enterica serovars and *Salmonella bongori* by use of an *S. enterica* serovar Typhimurium DNA microarray. J. Bacteriol. **185:**553-63.

44. **Chee, M., R. Yang, E. Hubbell, et al.** 1996. Accessing genetic information with high-density DNA arrays. Science **274:**610-4.

45. **Cho, S., D. J. Boxrud, J. M. Bartkus, et al.** 2007. Multiple-locus variable-number tandem repeat analysis of *Salmonella enteritidis* isolates from human and non-human sources using a single multiplex PCR. FEMS Microbiol. Lett. **275:**16-23.

46. **Collier-Hyams, L. S., H. Zeng, J. Sun, et al.** 2002. Cutting edge: *Salmonella* AvrA effector inhibits the key proinflammatory, anti-apoptotic NF-kappa B pathway.[see comment]. J. Immunol. **169:**2846-50.

47. **Contreras, I., Toro, C. S., Troncoso, G. and Mora, G. C.** 1997. *Salmonella typhi* mutants defective in anaerobic respiration are impaired in their ability to replicate within epithelial cells. Microbiology **143:**2665-2672.

48. **Cooper, V. S., and R. E. Lenski.** 2000. The population genetics of ecological specialization in evolving *Escherichia coli* populations. Nature **407:**736-739.

49. **Cotton, R. G., N. R. Rodrigues, and R. D. Campbell.** 1988. Reactivity of cytosine and thymine in single-base-pair mismatches with hydroxylamine and osmium tetroxide and its application to the study of mutations. Proc. Natl. Acad. Sci. U. S. A. **85:**4397-401.

50. **Craigie, J., and A. Felix.** 1947. Typing of typhoid bacilli with Vi bacteriophage. Lancet **252:**823-827.

51. **Craigie, J., and C. Yen.** 1938. The demonstration of types of *B. typhosus* by means of preparations of type II Vi phage. 1. Principles and technique and 2. The stability and epidemiological significance of V form types of *B. typhosus*. Can. J. Public Health **29:**448-496.

52. **Crosa, J. H., D. J. Brenner, W. H. Ewing, et al.** 1973. Molecular relationships among the *Salmonellae*. J. Bacteriol. **115:**307-315.

53. **Dawid, S., S. J. Barenkamp, and J. W. St Geme, 3rd.** 1999. Variation in expression of the *Haemophilus influenzae* HMW adhesins: a prokaryotic system reminiscent of eukaryotes. Proc. Natl. Acad. Sci. U. S. A. **96:**1077-82.

54. **Day, W. A., Jr., R. E. Fernandez, and A. T. Maurelli.** 2001. Pathoadaptive mutations that enhance virulence: genetic organization of the *cadA* regions of *Shigella* spp. Infect. Immun. **69:**7471-80.

55. **De Bolle, X., C. D. Bayliss, D. Field, et al.** 2000. The length of a tetranucleotide repeat tract in *Haemophilus influenzae* determines the phase variation rate of a gene with homology to type III DNA methyltransferases.[erratum appears in Mol Microbiol 2002 Oct;46(1):293]. Mol. Microbiol. **35:**211-22.

56. **Deng, W., V. Burland, G. Plunkett, 3rd, et al.** 2002. Genome sequence of *Yersinia pestis* KIM. J. Bacteriol. **184:**4601-4611.

57. **Deng, W., S. R. Liou, G. Plunkett, 3rd, et al.** 2003. Comparative genomics of *Salmonella enterica* serovar Typhi strains Ty2 and CT18. J. Bacteriol. **185:**2330-2337.

58. **Denoeud, F., and G. Vergnaud.** 2004. Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains : a web-based resource. BMC Bioinformatics **5:**4.

59. **Didelot, X., M. Achtman, J. Parkhill, et al.** 2007. A bimodal pattern of relatedness between the *Salmonella* Paratyphi A and Typhi genomes: convergence or divergence by homologous recombination? Genome Res. **17:**61-8.

60. **Dolz, R.** 1994. GCG: Comparison of sequences. Methods Mol. Biol. **24:**64-82.

61. **Drancourt, M., and D. Raoult.** 2002. Molecular insights into the history of plague. Microbes Infect. **4:**105-9.

62. **Echeita, M. A., and M. A. Usera.** 1998. Chromosomal rearrangements in *Salmonella enterica* serotype typhi affecting molecular typing in outbreak investigations. J. Clin. Microbiol. **36:**2123-6.

63. **Elsinghorst, E. A., L. S. Baron, and D. J. Kopecko.** 1989. Penetration of human intestinal epithelial cells by *Salmonella*: molecular cloning and expression of

Salmonella typhi invasion determinants in *Escherichia coli*. Proc. Natl. Acad. Sci. U. S. A. **86:**5173-7.

64. **Ernst, R. K., T. Guina, and S. I. Miller.** 2001. *Salmonella typhimurium* outer membrane remodeling: role in resistance to host innate immunity. Microbes Infect. **3:**1327-34.
65. **Everest, P., Wain, J., Roberts, M., Rook, G. and Dougan, G.** 2001. The molecular mechanisms of severe typhoid fever. Trends Microbiol. **9:**316-320.
66. **Excoffier, L., G. Laval, and S. Schneider.** 2005. Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evol Bioinform Online **1:**47-50.
67. **Falush, D., M. Torpdahl, X. Didelot, et al.** 2006. Mismatch induced speciation in *Salmonella*: Model and data. Philos Trans R Soc Lond B Biol Sci **361:**2045-2053.
68. **Fang, F. C., M. A. DeGroote, J. W. Foster, et al.** 1999. Virulent *Salmonella typhimurium* has two periplasmic Cu, Zn-superoxide dismutases. Proc. Natl. Acad. Sci. U. S. A. **96:**7502-7.
69. **Fasanella, A., M. Van Ert, S. A. Altamura, et al.** 2005. Molecular diversity of *Bacillus anthracis* in Italy. J. Clin. Microbiol. **43:**3398-401.
70. **Faucher, S. P., R. Curtiss, 3rd, and F. Daigle.** 2005. Selective capture of *Salmonella enterica* serovar Typhi genes expressed in macrophages that are absent from the *Salmonella enterica* serovar Typhimurium genome. Infect. Immun. **73:**5217-21.
71. **Faucher, S. P., S. Porwollik, C. M. Dozois, et al.** 2006. Transcriptome of *Salmonella enterica* serovar Typhi within macrophages revealed through the selective capture of transcribed sequences. Proc. Natl. Acad. Sci. U. S. A. **103:**1906-11.
72. **Feil, E. J., E. C. Holmes, D. E. Bessen, et al.** 2001. Recombination within natural populations of pathogenic bacteria: Short-term empirical estimates and long-term phylogenetic consequences. Proc. Natl. Acad. Sci. U. S. A. **98:**182-187.

73. **Feil, E. J., B. C. Li, D. M. Aanensen, et al.** 2004. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J. Bacteriol. **186:**1518-30.

74. **Feil, E. J., M. C. Maiden, M. Achtman, et al.** 1999. The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. Mol. Biol. Evol. **16:**1496-502.

75. **Feil, E. J., J. M. Smith, M. C. Enright, et al.** 2000. Estimating recombinational parameters in *Streptococcus pneumoniae* from multilocus sequence typing data. Genetics **154:**1439-1450.

76. **Felix, A., and R. M. Pitt.** 1935. Virulence and immunogenic activities of *B. typhosus* in relation to its antigenic constituents. J Hyg (Lond) **35:**428-436.

77. **Felsenstein, J.** 1983. Parsimony in systematics: biological and statistical issues. Ann. Rev. Ecol. Syst. **14:**313-333.

78. **Felsenstein, J.** 1989. PHYLIP-phylogeny inference package. Cladistics **5:**164-166.

79. **Fica, A. E., S. Prat-Miranda, A. Fernandez-Ricci, et al.** 1996. Epidemic typhoid in Chile: analysis by molecular and conventional methods of *Salmonella typhi* strain diversity in epidemic (1977 and 1981) and nonepidemic (1990) years. J. Clin. Microbiol. **34:**1701-7.

80. **Field, D., and C. Wills.** 1998. Abundant microsatellite polymorphism in *Saccharomyces cerevisiae*, and the different distributions of microsatellites in eight prokaryotes and *S. cerevisiae*, result from strong mutation pressures and a variety of selective forces. Proc. Natl. Acad. Sci. U. S. A. **95:**1647-52.

81. **Filliol, I., A. Motiwala, M. Cavatore, et al.** 2006. Global Phylogeny of *Mycobacterium tuberculosis* Based on Single Nucleotide Polymorphism (SNP) Analysis: Insights into Tuberculosis Evolution, Phylogenetic Accuracy of Other DNA Fingerprinting Systems, and Recommendations for a Minimal Standard SNP Set. J. Bacteriol. **188:**759–772.

82. **Fischer, S. G., and L. S. Lerman.** 1983. DNA fragments differing by single base-pair substitutions are separated in denaturing gradient gels: correspondence with melting theory. Proc. Natl. Acad. Sci. U. S. A. **80:**1579-1583.

83. **Folkesson, A., A. Advani, S. Sukupolvi, et al.** 1999. Multiple insertions of fimbrial operons correlate with the evolution of *Salmonella* serovars responsible for human disease. Mol. Microbiol. **33:**612-22.

84. **Folkesson, A., S. Lofdahl, and S. Normark.** 2002. The *Salmonella enterica* subspecies I specific centisome 7 genomic island encodes novel protein families present in bacteria living in close contact with eukaryotic cells. Res. Microbiol. **153:**537-45.

85. **Forest, C., S. P. Faucher, K. Poirier, et al.** 2007. Contribution of the stg fimbrial operon of *Salmonella enterica* serovar Typhi during interaction with human cells. Infect. Immun.

86. **Fouet, A., K. L. Smith, C. Keys, et al.** 2002. Diversity among French *Bacillus anthracis* isolates. J. Clin. Microbiol. **40:**4732-4.

87. **Frankel, G., S. M. Newton, G. K. Schoolnik, et al.** 1989. Intragenic recombination in a flagellin gene: characterization of the H1-j gene of *Salmonella typhi*. EMBO J. **8:**3149-3152.

88. **Frothingham, R., and W. A. Meeker-O'Connell.** 1998. Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats. Microbiology **144:**1189-96.

89. **Fu, Y., and J. E. Galan.** 1999. A *Salmonella* protein antagonizes Rac-1 and Cdc42 to mediate host-cell recovery after bacterial invasion.[see comment]. Nature **401:**293-7.

90. **Gauthier, A., M. Turmel, and C. Lemieux.** 1991. A group I intron in the chloroplast large subunit rRNA gene of *Chlamydomonas eugametos* encodes a double-strand endonuclease that cleaves the homing site of this intron. Curr. Genet. **19:**43-7.

91. **Gierczyski, R., S. Kaluzewski, A. Rakin, et al.** 2004. Intriguing diversity of *Bacillus anthracis* in eastern Poland--the molecular echoes of the past outbreaks. FEMS Microbiol. Lett. **239:**235-40.

92. **Gilson, E., S. Bachellier, S. Perrin, et al.** 1990. Palindromic unit highly repetitive DNA sequences exhibit species specificity within Enterobacteriaceae. Res. Microbiol. **141:**1103-16.

93. **Goh, S. H., S. K. Byrne, J. L. Zhang, et al.** 1992. Molecular typing of *Staphylococcus aureus* on the basis of coagulase gene polymorphisms. J. Clin. Microbiol. **30:**1642-5.

94. **Gordon, D., C. Abajian, and P. Green.** 1998. CONSED - A graphical tool for sequence finishing. Genome Res. **8:**195-202.

95. **Groisman, E. A., A. Blanc-Potard, and K. Uchiya.** 1999. Pathogenicity islands and the evolution of *Salmonella* virulence, p. 127-150. *In* J. B. Kaper and J. Hacker (ed.), Pathogenicity Islands and other mobile virulence elements. ASM press, Washington, D.C.

96. **Guinee, P. A., W. H. Jansen, H. M. Maas, et al.** 1981. An unusual H antigen (z66) in strains of *Salmonella typhi*. Ann. Microbiol. (Paris). **132:**331-4.

97. **Gutacker, M. M., J. C. Smoot, C. A. Lux Miglicaccio, et al.** 2002. Genome-wide analysis of snynonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms: Resolution of genetic relationships among closely related microbial strains. Genetics **162:**1533-1543.

98. **Guttman, D. S., and D. E. Dykhuizen.** 1994. Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. Science **266 (5189):**1380-1383.

99. **Haghjoo, E., and J. E. Gala?n.** 2004. *Salmonella typh*i encodes a functional cytolethal distending toxin that is delivered into host cells by a bacterial-internalization pathway. Proc. Natl. Acad. Sci. U. S. A. **101:**4614-4619.

100. **Hansen, L. L., J. Justesen, and T. A. Kruse.** 1996. Sensitive and fast mutation detection by solid phase chemical cleavage. Hum. Mutat. **7:**256-63.

101. **Hardy, K. J., D. W. Ussery, B. A. Oppenheim, et al.** 2004. Distribution and characterization of staphylococcal interspersed repeat units (SIRUs) and potential use for strain differentiation. Microbiology **150:**4045-52.

102. **Haris, I. I., P. M. Green, D. R. Bentley, et al.** 1994. Mutation detection by fluorescent chemical cleavage: application to hemophilia B. PCR Methods Appl. **3:**268-71.

103. **Harris, J. B., A. Baresch-Bernal, S. M. Rollins, et al.** 2006. Identification of in vivo-induced bacterial protein antigens during human infection with *Salmonella enterica* serovar Typhi. Infect. Immun. **74:**5161-8.

104. **Hazbon, M. H., and D. Alland.** 2004. Hairpin primers for simplified single-nucleotide polymorphism analysis of *Mycobacterium tuberculosis* and other organisms. J. Clin. Microbiol. **42:**1236-42.

105. **Heinrich, P. C., I. Behrmann, S. Haan, et al.** 2003. Principles of interleukin (IL)-6-type cytokine signalling and its regulation. Biochem. J. **374:**1-20.

106. **Hermans, P. W., S. K. Saha, W. J. van Leeuwen, et al.** 1996. Molecular typing of Salmonella typhi strains from Dhaka (Bangladesh) and development of DNA probes identifying plasmid-encoded multidrug-resistant isolates. J. Clin. Microbiol. **34:**1373-9.

107. **Hermans, P. W., D. van Soolingen, E. M. Bik, et al.** 1991. Insertion element IS*987* from *Mycobacterium bovis* BCG is located in a hot-spot integration region for insertion elements in *Mycobacterium tuberculosis* complex strains. Infect. Immun. **59:**2695-705.

108. **Herring, C. D., A. Raghunathan, C. Honisch, et al.** 2006. Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. Nat. Genet. **38:**1406-1412.

109. **Hickman-Brenner, F., and J. Farmer III.** 1983. Bacteriophage types of *Salmonella typhi* in the United States from 1974 through 1981. J. Clin. Microbiol. **17:**172-174.

110. **Hill, C. W., and J. A. Gray.** 1988. Effects of chromosomal inversion on cell fitness in *Escherichia coli* K-12. Genetics **119:**771-8.

111. **Hirose, K., T. Ezaki, M. Miyake, et al.** 1997. Survival of Vi-capsulated and Vi-deleted *Salmonella typhi* strains in cultured macrophage expressing different levels of CD14 antigen. FEMS Microbiol. Lett. **147:**259-65.

112. **Hoare, A., M. Bittner, J. Carter, et al.** 2006. The outer core lipopolysaccharide of *Salmonella enterica* serovar Typhi is required for bacterial entry into epithelial cells. Infect. Immun. **74:**1555-1564.

113. **Hoffmaster, A. R., C. C. Fitzgerald, E. Ribot, et al.** 2002. Molecular subtyping of *Bacillus anthracis* and the 2001 bioterrorism-associated anthrax outbreak, United States. Emerg. Infect. Dis. **8:**1111-6.

114. **Holt, K. E., J. Parkhill, C. J. Mazzoni, et al.** 2008. High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi. Nat. Genet. **40:**987-993.

115. **Hone, D. M., A. M. Harris, S. Chatfield, et al.** 1991. Construction of genetically defined double aro mutants of *Salmonella typhi*. Vaccine **9:**810-816.

116. **Hood, D. W., M. E. Deadman, M. P. Jennings, et al.** 1996. DNA repeats identify novel virulence genes in *Haemophilus influenzae*. Proc. Natl. Acad. Sci. U. S. A. **93:**11121-5.

117. **Hoorfar, J., D. L. Baggesen, and P. H. Porting.** 1999. A PCR-base strategy for simple and rapid identification of rough presumptive *Salmonella* isolates. J. Microbiol. Methods **35:**77-84.

118. **Hornick, R. B., S. E. Greisman, T. E. Woodward, et al.** 1970. Typhoid fever: pathogenesis and immunologic control. N Engl J Med **283:**686-746.

119. **Hosoglu, S., M. Loeb, M. F. Geyik, et al.** 2003. Molecular epidemiology of invasive *Salmonella typhi* in southeast Turkey. Clin Microbiol Infect **9:**727-30.

120. **House, D., A. Bishop, C. Parry, et al.** 2001. Typhoid fever: pathogenesis and disease. Curr. Opin. Infect. Dis. **14:**573-578.

121. **Hu, G.** 1993. DNA polymerase-catalyzed addition of nontemplated extra nucleotides to the 3' end of a DNA fragment. DNA Cell Biol. **12:**763-70.

122. **Huang, X., V. Phung le, S. Dejsirilert, et al.** 2004. Cloning and characterization of the gene encoding the z66 antigen of *Salmonella enterica* serovar Typhi. FEMS Microbiol. Lett. **234:**239-46.

123. **Hulton, C. S., C. F. Higgins, and P. M. Sharp.** 1991. ERIC sequences: a novel family of repetitive elements in the genomes of *Escherichia coli*, *Salmonella typhimurium* and other enterobacteria. Mol. Microbiol. **5:**825-34.

124. **Hunter, P. R., and M. A. Gaston.** 1988. Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J. Clin. Microbiol. **26:**2465-2466.

125. **Irino, K., F. Grimont, I. Casin, et al.** 1988. rRNA gene restriction patterns of *Haemophilus influenzae* biogroup aegyptius strains associated with Brazilian purpuric fever. J. Clin. Microbiol. **26:**1535-8.

126. **Jackson, P. J., E. A. Walthers, A. S. Kalif, et al.** 1997. Characterization of the variable-number tandem repeats in *vrrA* from different *Bacillus anthracis* isolates. Appl. Environ. Microbiol. **63:**1400-5.

127. **Jakobsen, I. B., and S. Easteal.** 1996. A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. CABIOS **12:**291-295.

128. **Jarosik, G. P., and E. J. Hansen.** 1994. Identification of a new locus involved in expression of *Haemophilus influenzae* type b lipooligosaccharide. Infect. Immun. **62:**4861-7.

129. **Jiang, M. C., P. C. Jiang, C. F. Liao, et al.** 2005. A modified mutation detection method for large-scale cloning of the possible single nucleotide polymorphism sequences. J Biochem Mol Biol **38:**191-197.

130. **Jin, Q., Z. Yuan, J. Xu, et al.** 2002. Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. Nucleic Acids Res. **30:**4432-4441.

131. **Joiner, K. A.** 1988. Complement evasion by bacteria and parasites. Annu. Rev. Microbiol. **42:**201-30.

132. **Jones, B. D.** 1996. Salmonellosis: host immune responses and bacterial virulence determinants. Annu. Rev. Immunol. **14:**533-561.

133. **Jonsson, K., C. Signas, H. P. Muller, et al.** 1991. Two different genes encode fibronectin binding proteins in *Staphylococcus aureus*. The complete nucleotide sequence and characterization of the second gene. Eur. J. Biochem. **202:**1041-8.

134. **Josenhans, C., K. A. Eaton, T. Thevenot, et al.** 2000. Switching of flagellar motility in *Helicobacter pylori* by reversible length variation of a short homopolymeric sequence repeat in *fliP*, a gene encoding a basal body protein. Infect. Immun. **68:**4598-603.

135. **Kapoor, S., Singh, R. D., Sharma, P. C. and Khular, M.** 2002. Anaerobiosis induced virulence of *Salmonella typhi*. Indian J. Med. Res. **115:**184-188.

136. **Kariuki, S., G. Revathi, J. Muyodi, et al.** 2004. Characterization of multidrug-resistant typhoid outbreaks in Kenya. J. Clin. Microbiol. **42:**1477-82.

137. **Keddy, K. H., K. P. Klugman, and J. B. Robbins.** 1998. Efficacy of Vi polysaccharide vaccine against strains of *Salmonella typhi*: reply.[see comment]. Vaccine **16:**871-2.

138. **Keim, P., L. B. Price, A. M. Klevytska, et al.** 2000. Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within *Bacillus anthracis*. J. Bacteriol. **182:**2928-36.

139. **Keim, P., M. N. Van Ert, T. Pearson, et al.** 2004. Anthrax molecular epidemiology and forensics: Using the appropriate marker for different evolutionary scales. Infect., Genet. Evol. **4:**205-213.

140. **Kidgell, C., D. Pickard, J. Wain, et al.** 2002. Characterisation and distribution of a cryptic *Salmonella typhi* plasmid pHCM2. Plasmid **47:**159-171.

141. **Kidgell, C., U. Reichard, J. Wain, et al.** 2002. *Salmonella typhi*, the causative agent of typhoid fever, is approximately 50,000 years old. Infect., Genet. Evol. **2:**39-45.

142. **Kim, S. R., and T. Komano.** 1992. Nucleotide sequence of the R721 shufflon. J. Bacteriol. **174:**7053-8.

143. **Kiss, T., E. Morgan, and G. Nagy.** 2007. Contribution of SPI-4 genes to the virulence of *Salmonella enterica*. FEMS Microbiol. Lett. **275:**153-9.

144. **Klevytska, A. M., L. B. Price, J. M. Schupp, et al.** 2001. Identification and characterization of variable-number tandem repeats in the *Yersinia pestis* genome. J. Clin. Microbiol. **39:**3179-85.

145. **Koay, A. S., M. Jegathesan, M. Y. Rohani, et al.** 1997. Pulsed-field gel electrophoresis as an epidemiologic tool in the investigation of laboratory acquired *Salmonella typhi* infection. Southeast Asian J. Trop. Med. Public Health **28:**82-4.

146. **Komano, T., A. Kubo, and T. Nisioka.** 1987. Shufflon: multi-inversion of four contiguous DNA segments of plasmid R64 creates seven different open reading frames. Nucleic Acids Res. **15:**1165-72.

147. **Kops, S. K., D. K. Lowe, W. M. Bement, et al.** 1996. Migration of *Salmonella typhi* through intestinal epithelial monolayers: an in vitro study. Microbiol. Immunol. **40:**799-811.

148. **Kothapalli, S., S. Nair, S. Alokam, et al.** 2005. Diversity of genome structure in *Salmonella enterica* serovar Typhi populations. J. Bacteriol. **187:**2638-50.

149. **Krawiec, S., and M. Riley.** 1990. Organization of the bacterial chromosome. Microbiol. Rev. **54:**502-39.

150. **Kubori, T., and J. E. Galan.** 2003. Temporal regulation of *Salmonella* virulence effector function by proteasome-dependent protein degradation. Cell **115:**333-42.

151. **Kubota, K., T. J. Barrett, M. L. Ackers, et al.** 2005. Analysis of *Salmonella enterica* serotype Typhi pulsed-field gel electrophoresis patterns associated with international travel. J. Clin. Microbiol. **43:**1205-9.

152. **Kuhle, V., and M. Hensel.** 2004. Cellular microbiology of intracellular *Salmonella enterica*: functions of the type III secretion system encoded by Salmonella pathogenicity island 2. Cell. Mol. Life Sci. **61:**2812-26.

153. **Kumar, S., S. Kumar, and S. Kumar.** 2006. Infection as a risk factor for gallbladder cancer. J. Surg. Oncol. **93:**633-9.

154. **Lam, S., and J. R. Roth.** 1983. IS*200*: a *Salmonella*-specific insertion sequence. Cell **34:**951-60.

155. **Lan, R., M. C. Alles, K. Donohoe, et al.** 2004. Molecular evolutionary relationships of enteroinvasive *Escherichia coli* and *Shigella* spp. Infect. Immun. **72:**5080-5088.

156. **Lan, R., and P. R. Reeves.** 2006. Evolution of enteric pathogens, p. 273-299. *In* H. S. Seifert and V. J. DiRita (ed.), Evolution of microbial pathogens ASM Press, Washington, D.C.

157. **Le Fleche, P., M. Fabre, F. Denoeud, et al.** 2002. High resolution, on-line identification of strains from the *Mycobacterium tuberculosis* complex based on tandem repeat typing. BMC Microbiol. **2:**37.

158. **Le Fleche, P., Y. Hauck, L. Onteniente, et al.** 2001. A tandem repeats database for bacterial genomes: application to the genotyping of *Yersinia pestis* and *Bacillus anthracis*. BMC Microbiol. **1:**2.

159. **Le Minor, L., and M. Y. Popoff.** 1987. Designation of *Salmonella enterica* sp. nov., nom. rev., as the type and only species of the genus *Salmonella*. Int. J. Syst. Bacteriol. **37:**465-468.

160. **Le Minor, L., M. Y. Popoff, B. Laurent, et al.** 1986. Individualisation D'une septieme sous-espece de *Salmonella*: *S. choleraesuis* subsp. *indica* subsp. nov. Annales de l'Institut Pasteur - Microbiologie **137B:**211-217.

161. **Le, T. A., M. Lejay-Collin, P. A. Grimont, et al.** 2004. Endemic, epidemic clone of Salmonella enterica serovar typhi harboring a single multidrug-resistant plasmid in Vietnam between 1995 and 2002. J. Clin. Microbiol. **42:**3094-9.

162. **Levinson, G., and G. A. Gutman.** 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution.[see comment]. Mol. Biol. Evol. **4:**203-21.

163. **Li, J., R. Berbeco, R. J. Distel, et al.** 2007. s-RT-MELT for rapid mutation scanning using enzymatic selection and real time DNA-melting: new potential for multiplex genetic analysis. Nucleic Acids Res. **35:**e84.

164. **Lindstedt, B.-A., E. Heir, E. Gjernes, et al.** 2003. DNA fingerprinting of *Salmonella enterica* subsp. *enterica* serovar typhimurium with emphasis on phage type DT104 based on variable number of tandem repeat loci. J. Clin. Microbiol. **41:**1469-79.

165. **Lindstedt, B.-A., T. Vardund, L. Aas, et al.** 2004. Multiple-locus variable-number tandem-repeats analysis of *Salmonella enterica* subsp. *enterica* serovar Typhimurium using PCR multiplexing and multicolor capillary electrophoresis. J. Microbiol. Methods **59:**163-72.

166. **Ling, J. M., N. W. Lo, Y. M. Ho, et al.** 2000. Molecular methods for the epidemiological typing of *Salmonella enterica* serotype Typhi from Hong Kong and Vietnam. J. Clin. Microbiol. **38:**292-300.

167. **Lishanski, A., E. A. Ostrander, and J. Rine.** 1994. Mutation detection by mismatch binding protein, MutS, in amplified DNA: Application to the cystic fibrosis gene. Proc. Natl. Acad. Sci. U. S. A. **91:**2674-2678.

168. **Liu, G.-R., A. Rahn, W.-Q. Liu, et al.** 2002. The evolving genome of *Salmonella enterica* serovar Pullorum. J. Bacteriol. **184:**2626-33.

169. **Liu, S.-L., and K. E. Sanderson.** 1996. Highly plastic chromosomal organization in *Salmonella typhi*. Proc. Natl. Acad. Sci. USA **93:**10303-10308.

170. **Liu, S.-L., and K. E. Sanderson.** 1995. Rearrangements in the genome of the bacterium *Salmonella typhi*. Proc. Natl. Acad. Sci. USA **92:**1018-1022.

171. **Liu, S. L., and K. E. Sanderson.** 1995. I-*Ceu*I reveals conservation of the genome of independent strains of *Salmonella typhimurium*. J. Bacteriol. **177:**3355-7.

172. **Liu, S. L., A. B. Schryvers, K. E. Sanderson, et al.** 1999. Bacterial phylogenetic clusters revealed by genome structure. J. Bacteriol. **181:**6747-55.

173. **Liu, W.-Q., G.-R. Liu, J.-Q. Li, et al.** 2007. Diverse genome structures of *Salmonella paratyphi* C. BMC Genomics **8:**290.

174. **Liu, Y., M.-A. Lee, E.-E. Ooi, et al.** 2003. Molecular typing of *Salmonella enterica* serovar Typhi isolates from various countries in Asia by a multiplex PCR assay on variable-number tandem repeats. J. Clin. Microbiol. **41:**4388-94.

175. **Livak, K. J.** 1999. Allelic discrimination using fluorogenic probes and the 5' nuclease assay. Genet. Anal. **14:**143-9.

176. **Looney, R. J., and R. T. Steigbigel.** 1986. Role of the Vi antigen of *Salmonella typhi* in resistance to host defense in vitro. J. Lab. Clin. Med. **108:**506-16.

177. **Lostroh, C. P., and C. A. Lee.** 2001. The *Salmonella* pathogenicity island-1 type III secretion system. Microbes Infect. **3:**1281-91.

178. **Lowell, J. L., D. M. Wagner, B. Atshabar, et al.** 2005. Identifying sources of human exposure to plague. J. Clin. Microbiol. **43:**650-6.

179. **Lu, A. L., and I. C. Hsu.** 1992. Detection of single DNA base mutations with mismatch repair enzymes. Genomics **14:**249-255.

180. **Luk, J. M. C., U. Kongmuang, P. R. Reeves, et al.** 1993. Selective amplification of abequose and paratose synthase genes (*rfb*) by polymerase chain reaction for identification of *Salmonella* major serogroups (A, B, C2, and D). J. Clin. Microbiol. **31:**2118-2123.

181. **Lyczak, J. B.** 2003. Commensal Bacteria Increase Invasion of Intestinal Epithelium by *Salmonella enterica* serovar Typhi. Infect. Immun. **71:**6610-6614.

182. **Lyczak, J. B., and G. B. Pier.** 2002. *Salmonella enterica* serovar Typhi modulates cell surface expression of its receptor, the cystic fibrosis

transmembrane conductance regulator, on the intestinal epithelium. Infect. Immun. **70:**6416-23.

183. **Lyczak, J. B., T. S. Zaidi, M. Grout, et al.** 2001. Epithelial cell contact-induced alterations in *Salmonella enterica* serovar Typhi lipopolysaccharide are critical for bacterial internalization. Cell. Microbiol. **3:**763-772.

184. **Maho, A., A. Rossano, H. Hachler, et al.** 2006. Antibiotic susceptibility and molecular diversity of *Bacillus anthracis* strains in Chad: detection of a new phylogenetic subgroup. J. Clin. Microbiol. **44:**3422-5.

185. **Maiden, M. C., J. A. Bygraves, E. Feil, et al.** 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proceedings of the National Academy of Sciences USA **95:**3140-3145.

186. **Malik, A. S.** 2002. Complications of bacteriologically confirmed typhoid fever in children. J. Trop. Pediatr. **48:**102-8.

187. **Manfredi, R., C. Donzelli, S. Talo, et al.** 1998. Typhoid fever and HIV infection: a rare disease association in industrialized countries. Int. J. Infect. Dis. **3:**105-8.

188. **Martin, B., O. Humbert, M. Camara, et al.** 1992. A highly conserved repeated DNA element located in the chromosome of *Streptococcus pneumoniae*. Nucleic Acids Res. **20:**3479-3483.

189. **Mashal, R. D., J. Koontz, and J. Sklar.** 1995. Detection of mutations by cleavage of DNA heteroduplexes with bacteriophage resolvases. Nat. Genet. **9:**177-183.

190. **Matic, I., F. Taddei, and M. Radman.** 1996. Genetic barriers among bacteria. Trends Microbiol. **4:**69-73.

191. **Matusevicius, D., P. Kivisakk, B. He, et al.** 1999. Interleukin-17 mRNA expression in blood and CSF mononuclear cells is augmented in multiple sclerosis. Mult. Scler. **5:**101-4.

192. **Maynard Smith, J., N. H. Smith, M. O'Rourke, et al.** 1993. How clonal are bacteria? Proc. Natl. Acad. Sci. U. S. A. **90:**4384-4388.

193. **McClelland, M., K. E. Sanderson, S. W. Clifton, et al.** 2004. Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of *Salmonella enterica* that cause typhoid. **36:**1268-1274.

194. **McClelland, M., K. E. Sanderson, J. Spieth, et al.** 2001. Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. Nature **413:**852-856.

195. **McCormick, B. A., S. I. Miller, D. Carnes, et al.** 1995. Transepithelial signaling to neutrophils by *Salmonellae*: a novel virulence mechanism for gastroenteritis. Infect. Immun. **63:**2302-9.

196. **McDevitt, D., and T. J. Foster.** 1995. Variation in the size of the repeat region of the fibrinogen receptor (clumping factor) of *Staphylococcus aureus* strains. Microbiology **141:**937-43.

197. **McPeek, F. D., Jr., J. F. Coyle-Morris, and R. M. Gemmill.** 1986. Separation of large DNA molecules by modified pulsed field gradient gel electrophoresis. Anal. Biochem. **156:**274-85.

198. **Mehta, G., and S. C. Arya.** 2002. Capsular Vi polysaccharide antigen in *Salmonella enterica* serovar typhi isolates. J. Clin. Microbiol. **40:**1127-8.

199. **Merabishvili, M., M. Natidze, S. Rigvava, et al.** 2006. Diversity of *Bacillus anthracis* strains in Georgia and of vaccine strains from the former Soviet Union. Appl. Environ. Microbiol. **72:**5631-6.

200. **Mills, S. D., and B. B. Finlay.** 1994. Comparison of *Salmonella typhi* and *Salmonella typhimurium* invasion, intracellular growth and localization in cultured human epithelial cells. Microb. Pathog. **17:**409-23.

201. **Miyake, M., L. Zhao, T. Ezaki, et al.** 1998. Vi-deficient and nonfimbriated mutants of *Salmonella typhi* agglutinate human blood type antigens and are hyperinvasive. FEMS Microbiol. Lett. **161:**75-82.

202. **Monot, M., N. Honore, T. Garnier, et al.** 2005. On the origin of leprosy.[see comment]. Science **308:**1040-2.

203. **Morris, C., C. M. C. Yip, I. S. M. Tsui, et al.** 2003. The Shufflon of *Salmonella enterica* Serovar Typhi Regulates Type IVB Pilus-Mediated Bacterial Self-Association. Infect. Immun. **71:**1141-1146.

204. **Moshitch, S., L. Doll, B. Z. Rubinfeld, et al.** 1992. Mono- and bi-phasic *Salmonella typhi*: genetic homogeneity and distinguishing characteristics. Mol. Microbiol. **6:**2589-2597.

205. **Mrazek, J., X. Guo, and A. Shah.** 2007. Simple sequence repeats in prokaryotic genomes. Proc. Natl. Acad. Sci. U. S. A. **104:**8472-7.

206. **Mroczenski-Wildey, M. J., J. L. Di Fabio, and F. C. Cabello.** 1989. Invasion and lysis of HeLa cell monolayers by *Salmonella typhi*: the role of lipopolysaccharide.[erratum appears in Microb Pathog 1989 Apr;6(4):precedi]. Microb. Pathog. **6:**143-52.

207. **Muralidharan, K., A. Stern, and T. F. Meyer.** 1987. The control mechanism of opacity protein expression in the pathogenic *Neisseriae*. Antonie Van Leeuwenhoek **53:**435-40.

208. **Nair, S., S. Alokam, S. Kothapalli, et al.** 2004. *Salmonella enterica* serovar Typhi strains from which SPI7, a 134-kilobase island with genes for Vi exopolysaccharide and other functions, has been deleted. J. Bacteriol. **186:**3214-23.

209. **Nair, S., C. L. Poh, Y. S. Lim, et al.** 1994. Genome fingerprinting of *Salmonella typhi* by pulsed-field gel electrophoresis for subtyping common phage types. Epidemiol. Infect. **113:**391-402.

210. **Nair, S., E. Schreiber, K. L. Thong, et al.** 2000. Genotypic characterization of *Salmonella typhi* by amplified fragment length polymorphism fingerprinting provides increased discrimination as compared to pulsed-field gel electrophoresis and ribotyping. J. Microbiol. Methods **41:**35-43.

211. **Nakata, N., T. Tobe, I. Fukuda, et al.** 1993. The absence of a surface protease, OmpT, determines the intercellular spreading ability of *Shigella*: the relationship between the *ompT* and *kcpA* loci. Mol. Microbiol. **9:**459-68.

212. **Nastasi, A., C. Mammina, and M. R. Villafrate.** 1991. rDNA fingerprinting as a tool in epidemiological analysis of *Salmonella typhi* infections. Epidemiol. Infect. **107:**565-76.

213. **Navarro, F., T. Llovet, M. A. Echeita, et al.** 1996. Molecular typing of Salmonella enterica serovar typhi. J. Clin. Microbiol. **34:**2831-4.

214. **Nelson, K., and R. K. Selander.** 1992. Evolutionary genetics of the proline permease gene (*putP*) and the control region of the proline utilization operon in populations of *Salmonella* and *Escherichia coli*. J. Bacteriol. **174:**6886-6895.

215. **Nelson, K., and R. K. Selander.** 1994. Intergenic transfer and recombination of the 6-phosphogluconate dehydrogenase gene (*gnd*) in enteric bacteria. Proc. Natl. Acad. Sci. U. S. A. **91:**10227-10231.

216. **Nelson, K., F. S. Wang, E. F. Boyd, et al.** 1997. Size and sequence polymorphism in the isocitrate dehydrogenase kinase/phosphatase gene (*aceK*) and flanking regions in *Salmonella enterica* and *Escherichia coli.* Genetics **147:**1509-1520.

217. **Nelson, K., T. S. Whittam, and R. K. Selander.** 1991. Nucleotide polymorphism and evolution in the glyceraldehyde-3-phosphate dehydrogenase gene (*gapA*) in natural populations of *Salmonella* and *Escherichia coli*. Proc. Natl. Acad. Sci. U. S. A. **88:**6667-6671.

218. **Nevius, P., G. Controni, and W. J. Rodriguez.** 1980. Meningitis in typhoid fever: an unusual complication. South. Med. J. **73:**269-70.

219. **Ng, I., S. L. Liu, and K. E. Sanderson.** 1999. Role of genomic rearrangements in producing new ribotypes of *Salmonella typhi*. J. Bacteriol. **181:**3536-3541.

220. **Nie, H., F. Yang, X. Zhang, et al.** 2006. Complete genome sequence of *Shigella flexneri* 5b and comparison with *Shigella flexneri* 2a. BMC Genomics **7**.

221. **Nolan, C. M., P. C. White, Jr., J. C. Feeley, et al.** 1981. Vi serology in the detection of typhoid carriers. Lancet **1:**583-5.

222. **Nollau, P., and C. Wagener.** 1997. Methods for detection of point mutations: performance and quality assessment. IFCC Scientific Division, Committee on Molecular Biology Techniques. Clin. Chem. **43:**1114-28.

223. **Noller, A. C., M. C. McEllistrem, O. C. Stine, et al.** 2003. Multilocus sequence typing reveals a lack of diversity among *Escherichia coli* O157:H7 isolates that are distinct by pulsed-field gel electrophoresis. J. Clin. Microbiol. **41:**675-679.

224. **Oefner, P. J., C. G. Huber, F. Umlauft, et al.** 1994. High-resolution liquid chromatography of fluorescent dye-labeled nucleic acids. Anal. Biochem. **223:**39-46.

225. **Oleykowski, C. A., C. R. Bronson Mullins, A. K. Godwin, et al.** 1998. Mutation detection using a novel plant endonuclease. Nucleic Acids Res. **26:**4597-602.

226. **Olsen, S. J., B. Kafoa, N. S. Win, et al.** 2001. Restaurant-associated outbreak of *Salmonella typhi* in Nauru: an epidemiological and cost analysis. Epidemiol. Infect. **127:**405-12.

227. **Orita, M., H. Iwahana, H. Kanazawa, et al.** 1989. Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. Proc. Natl. Acad. Sci. U. S. A. **86:**2766-70.

228. **Oscarsson, J., M. Westermark, S. Lo?fdahl, et al.** 2002. Characterization of a pore-forming cytotoxin expressed by *Salmonella enterica* serovars Typhi and Paratyphi A. Infect. Immun. **70:**5759-5769.

229. **Pabbaraju, K., W. Miller, and K. Sanderson.** 2000. Distribution of intervening sequences in the genes for 23S rRNA and rRNA fragmentation among strains of the *Salmonella* reference collection B (SARB) and SARC sets. J. Bacteriol. **182:**1923-9.

230. **Pai, H., J. H. Byeon, S. Yu, et al.** 2006. *Salmonella enterica* serovar typhi strains isolated in Korea containing a multidrug resistance class 1 integron. Antimicrob. Agents Chemother. **47:**2006-8.

231. **Parkhill, J., G. Dougan, K. D. James, et al.** 2001. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. Nature **413:**848-852.

232. **Parkhill, J., M. Sebaihia, A. Preston, et al.** 2003. Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. Nat. Genet. **35:**32-40.

233. **Parkhill, J., B. W. Wren, N. R. Thomson, et al.** 2001. Genome sequence of *Yersinia pestis*, the causative agent of plague. Nature **413:**523-527.

234. **Parry, C. M., T. T. Hien, G. Dougan, et al.** 2002. Typhoid fever. N. Engl. J. Med. **22:**1770-1782.

235. **Pascopella, L., B. Raupach, N. Ghori, et al.** 1995. Host restriction phenotypes of *Salmonella typhi* and *Salmonella gallinarum*. Infect. Immun. **63:**4329-35.

236. **Pearson, T., J. Busch, J. Ravel, et al.** 2004. Phylogenetic discovery bias in *Bacillus anthracis* using single-nucleotide polymorphisms from whole-genome sequencing. Proc. Natl. Acad. Sci. U. S. A. **101:**13536–13541.

237. **Pickard, D., J. Li, M. Roberts, et al.** 1994. Characterization of defined *ompR* mutants of *Salmonella typhi*: *ompR* is involved in the regulation of Vi polysaccharide expression. Infect. Immun. **62:**3984-93.

238. **Pickard, D., J. Wain, S. Baker, et al.** 2003. Composition, acquisition, and distribution of the Vi exopolysaccharide-encoding *Salmonella enterica* pathogenicity island SPI-7. J. Bacteriol. **185:**5055-65.

239. **Pier, G. B., M. Grout, T. Zaidi, et al.** 1998. *Salmonella typhi* uses CFTR to enter intestinal epithelial cells. Nature **393:**79-82.

240. **Popoff, M. Y.** 2001. Antigenic formulas of the *Salmonella* serovars, 8th edition. WHO Collaborating Centre for Reference and Research on *Salmonella*. Institut Pasteur, Paris, France.

241. **Popoff, M. Y., and L. L. Minor.** 1997. Antigenic formulas of the *Salmonella* serovars, 7th revision. WHO Collaborating Centre for Reference and Research on *Salmonella*. Institut Pasteur, Paris, France.

242. **Porwollik, S., E. F. Boyd, C. Choy, et al.** 2004. Characterization of *Salmonella enterica* subspecies I genovars by use of microarrays. J. Bacteriol. **186:**5883-5898.

243. **Pos, K. M., P. Dimroth, and M. Bott.** 1998. The *Escherichia coli* citrate carrier CitT: a member of a novel eubacterial transporter family related to the 2-oxoglutarate/malate translocator from spinach chloroplasts. J. Bacteriol. **180:**4160-5.

244. **Pourcel, C., F. Andre-Mazeaud, H. Neubauer, et al.** 2004. Tandem repeats analysis for the high resolution phylogenetic analysis of *Yersinia pestis*. BMC Microbiol. **4:**22.

245. **Pradier, C., O. Keita-Perse, E. Bernard, et al.** 2000. Outbreak of typhoid fever on the French Riviera. Eur. J. Clin. Microbiol. Infect. Dis. **19:**464-7.

246. **Prentice, M. B., James, K. D., Parkhill, J., Baker, S. G., Stevens, K., Simmonds, M. N., Mungall, K. L., Churcher, C., Oyston, P. C. F., Titball, R.**

W., Wren, B. W., Wain, J., Pickar, D., Hien, T. T., Farrar, J. J. and Dougan, G. 2002. *Yersinia pestis* pFra shows biovar-specific differences and recent common ancestry with a *Salmonella enterica* serovar Typhi plasmid. J. Bacteriol. **183:**2586-2594.

247. **Price, E. P., H. Smith, F. Huygens, et al.** 2007. High-resolution DNA melt curve analysis of the clustered, regularly interspaced short-palindromic-repeat locus of *Campylobacter jejuni*. Appl. Environ. Microbiol. **73:**3431-6.

248. **Pupo, G. M., R. Lan, and P. R. Reeves.** 2000. Multiple independent origins of Shigella clones of *Escherichia coli* and convergent evolution of many of their characteristics. Proc. Natl. Acad. Sci. U. S. A. **97:**10567-10572.

249. **Pupo, G. M., R. Lan, P. R. Reeves, et al.** 2000. Population Genetics of *Escherichia coli* in a Natural Population of Native Australian Rats. Environ. Microbiol. **2:**594-610.

250. **Qiu, P., H. Shandilya, J. M. D'Alessio, et al.** 2004. Mutation detection using Surveyor? nuclease. Biotechniques **36:**702-707.

251. **Qiu, P., H. Shandilya, and G. F. Gerard.** 2005. A method for clone sequence confirmation using a mismatch-specific DNA endonuclease. Mol. Biotechnol. **29:**11-8.

252. **Quintaes, B. R., N. C. Leal, E. M. Reis, et al.** 2004. Optimization of randomly amplified polymorphic DNA-polymerase chain reaction for molecular typing of Salmonella enterica serovar Typhi. Rev. Soc. Bras. Med. Trop. **37:**143-7.

253. **Radman, M., I. Matic, and F. Taddei.** 1999. Evolution of evolvability. Ann. N. Y. Acad. Sci. **870:**146-155.

254. **Radnedge, L., P. G. Agron, P. L. Worsham, et al.** 2002. Genome plasticity in *Yersinia pestis*. Microbiology **148:**1687-1698.

255. **Raffatellu, M., D. Chessa, R. P. Wilson, et al.** 2005. The Vi capsular antigen of *Salmonella enterica* serotype Typhi reduces Toll-Like Receptor-dependent Interleukin-8 expression in the intestinal mucosa. Infect. Immun. **73:**3367-3374.

256. **Raffatellu, M., R. L. Santos, D. Chessa, et al.** 2007. The capsule encoding the *viaB* locus reduces interleukin-17 expression and mucosal innate responses in the

bovine intestinal mucosa during infection with *Salmonella enterica* serotype Typhi. Infect. Immun. **75:**4342-50.

257. **Ramisse, V., P. Houssu, E. Hernandez, et al.** 2004. Variable number of tandem repeats in *Salmonella enterica* subsp. *enterica* for typing purposes. J. Clin. Microbiol. **42:**5722-30.

258. **Rasmussen, M. A., S. A. Carlson, S. K. Franklin, et al.** 2005. Exposure to rumen protozoa leads to enhancement of pathogenicity of and invasion by multiple-antibiotic-resistant *Salmonella enterica* bearing SGI1. Infect. Immun. **73:**4668-75.

259. **Rayssiguier, C., D. S. Thaler, and M. Radman.** 1989. The barrier to recombination between *Escherichia coli* and *Salmonella typhimurium* is disrupted in mismatch-repair mutants. Nature **342:**396-400.

260. **Read, T. D., S. L. Salzberg, M. Pop, et al.** 2002. Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*.[see comment]. Science **296:**2028-33.

261. **Reeves, M. W., G. M. Evins, A. A. Heiba, et al.** 1989. Clonal nature of *Salmonella typhi* and its genetic relatedness to other Salmonellae as shown by multilocus enzyme electrophoresis, and proposal of *Salmonella bongori.* J. Clin. Microbiol. **27:**313-320.

262. **Reeves, P.** 1993. Evolution of *Salmonella* O antigen variation by interspecific gene transfer on a large scale. Trends Genet. **9:**17-22.

263. **Reeves, P. R., L. Farnell, and R. Lan.** 1994. MULTICOMP: a program for preparing sequence data for phylogenetic analysis. Computer applications in the biosciences: CABIOS **10:**281-284.

264. **Reid, S. D., C. J. Herbelin, A. C. Bumnaugh, et al.** 2000. Parallel evolution of virulence in pathogenic *Escherichia coli.* Nature **406:**64-67.

265. **Ren, Z., H. Jin, P. W. Whitby, et al.** 1999. Role of CCAA nucleotide repeats in regulation of hemoglobin and hemoglobin-haptoglobin binding protein genes of *Haemophilus influenzae*. J. Bacteriol. **181:**5865-70.

266. **Ribot, E. M., R. K. Wierzba, F. J. Angulo, et al.** 2002. *Salmonella enterica* serotype Typhimurium DT104 isolated from humans, United States, 1985, 1990, and 1995. Emerg. Infect. Dis. **8:**387-91.

267. **Rich, R. L., B. Demeler, K. Ashby, et al.** 1998. Domain structure of the *Staphylococcus aureus* collagen adhesin. Biochemistry (Mosc). **37:**15423-33.

268. **Rocha, E. P. C., J. Maynard Smith, L. D. Hurst, et al.** 2006. Comparisons of d*N*/d*S* are time dependent for closely related bacterial genomnes. J. Theor. Biol. **239:**226-235.

269. **Rojas, G., S. Saldias, M. Bittner, et al.** 2001. The *rfaH* gene, which affects lipopolysaccharide synthesis in *Salmonella enterica* serovar Typhi, is differentially expressed during the bacterial growth phase. FEMS Microbiol. Lett. **204:**123-128.

270. **Ross, I. L., and M. W. Heuzenroeder.** 2005. Use of AFLP and PFGE to discriminate between *Salmonella enterica* serovar Typhimurium DT126 isolates from separate food-related outbreaks in Australia. Epidemiol. Infect. **133:**635-644.

271. **Roumagnac, P., F.-X. Weill, C. Dolecek, et al.** 2006. Evolutionary history of *Salmonella typhi*. Science **314:**1301-1304.

272. **Ryu, C., K. Lee, H.-J. Hawng, et al.** 2005. Molecular characterization of Korean *Bacillus anthracis* isolates by amplified fragment length polymorphism analysis and multilocus variable-number tandem repeat analysis. Appl. Environ. Microbiol. **71:**4664-71.

273. **Sabat, A., J. Krzyszton-Russjan, W. Strzalka, et al.** 2003. New method for typing *Staphylococcus aureus* strains: multiple-locus variable-number tandem repeat analysis of polymorphism and genetic relationships of clinical isolates. J. Clin. Microbiol. **41:**1801-4.

274. **Sambrook, J., and D. Russell.** 2001. Molecular cloning: A laboratory manual, 3rd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

275. **Santander, J., S.-Y. Wanda, C. A. Nickerson, et al.** 2007. Role of RpoS in fine-tuning the synthesis of Vi capsular polysaccharide in *Salmonella enterica* serotype Typhi. Infect. Immun. **75:**1382-92.

276. **Sauer, S., H. Lehrach, and R. Reinhardt.** 2003. MALDI mass spectrometry analysis of single nucleotide polymorphisms by photocleavage and charge-tagging. Nucleic Acids Res **31:**e63.

277. **Saunders, N. J., J. F. Peden, D. W. Hood, et al.** 1998. Simple sequence repeats in the *Helicobacter pylori* genome. Mol. Microbiol. **27:**1091-8.

278. **Schnaitman, C. A., and J. D. Klena.** 1993. Genetics of lipopolysaccharide biosynthesis in enteric bacteria. Microbiol. Rev. **57:**655-82.

279. **Schuelke, M.** 2000. An economic method for the fluorescent labeling of PCR fragments. Nat. Biotechnol. **18:**233-4.

280. **Schwartz, D. C., and C. R. Cantor.** 1984. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. Cell **37:**67-75.

281. **Scott, A. N., D. Menzies, T.-N. Tannenbaum, et al.** 2005. Sensitivities and specificities of spoligotyping and mycobacterial interspersed repetitive unit-variable-number tandem repeat typing methods for studying molecular epidemiology of tuberculosis. J. Clin. Microbiol. **43:**89-94.

282. **Segall, A., M. J. Mahan, and J. R. Roth.** 1988. Rearrangement of the bacterial chromosome: forbidden inversions. Science **241:**1314-8.

283. **Selander, R. K., P. Beltran, and N. H. Smith.** 1991. Evolutionary genetics of *Salmonella*, p. 25-27. *In* R. K. Selander, A. G. Clark, and T. S. Whittam (ed.), Evolution at the molecular level. Sinauer Associates, Sunderland, Massecheusetts.

284. **Selander, R. K., P. Beltran, N. H. Smith, et al.** 1990. Genetic population structure, clonal phylogeny and pathogenicity of *Salmonella paratyphi* B. Infect. Immun. **58:**1891-1901.

285. **Selander, R. K., P. Beltran, N. H. Smith, et al.** 1990. Evolutionary genetic relationships of clones of *Salmonella* serovars that cause human typhoid and other enteric fevers. Infect. Immun. **58:**2262-2275.

286. **Selander, R. K., J. Li, and K. Nelson.** 1996. Evolutionary genetics of *Salmonella enterica*, p. 2691-2707. *In* F. C. Neidhardt, R. C. III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella* Cellular

and Molecular Biology 2nd edition, vol. 2. American Society for Microbiology, Washington, D. C.

287. **Shangkuan, Y. H., and H. C. Lin.** 1998. Application of random amplified polymorphic DNA analysis to differentiate strains of *Salmonella typhi* and other Salmonella species. J. Appl. Microbiol. **85:**693-702.

288. **Sharma, A., and A. Qadri.** 2004. Vi polysaccharide of *Salmonella typhi* targets the prohibitin family of molecules in intestinal epithelial cells and suppresses early inflammatory responses. Proc. Natl. Acad. Sci. U. S. A. **101:**17492-7.

289. **Sharma, K. B., and S. C. Arya.** 1995. Detection of *Salmonella typhi* by nested PCR based on the ViaB sequence.[comment]. J. Clin. Microbiol. **33:**3361.

290. **Sherburne, C. K., T. D. Lawley, M. W. Gilmour, et al.** 2000. The complete DNA sequence and analysis of R27, a large IncHI plasmid from *Salmonella typhi* that is temperature sensitive for transfer. Nucl. Acids Res. **28:**2177-2186.

291. **Shi, R., K. Otomo, H. Yamada, et al.** 2006. Temperature-mediated heteroduplex analysis for the detection of drug-resistant gene mutations in clinical isolates of *Mycobacterium tuberculosis* by denaturing HPLC, SURVEYOR nuclease. Microbes Infect. **8:**128-35.

292. **Shi, Y., S. F. Terry, P. F. Terry, et al.** 2007. Development of a rapid, reliable genetic test for *Pseudoxanthoma elasticum*. J Mol Diagn **9:**105-12.

293. **Shukla, V. K., H. Singh, M. Pandey, et al.** 2000. Carcinoma of the gallbladder--is it a sequel of typhoid? Dig. Dis. Sci. **45:**900-3.

294. **Silverman, M., and M. Simon.** 1980. Phase variation: genetic analysis of switching mutants. Cell **19:**845-854.

295. **Smith, K. L., V. DeVos, H. Bryden, et al.** 2000. *Bacillus anthracis* diversity in Kruger National Park. J. Clin. Microbiol. **38:**3780-4.

296. **Spurgiesz, R. S., T. N. Quitugua, K. L. Smith, et al.** 2003. Molecular typing of *Mycobacterium tuberculosis* by using nine novel variable-number tandem repeats across the Beijing family and low-copy-number IS6110 isolates. J. Clin. Microbiol. **41:**4224-30.

297. **Sreenu, V. B., P. Kumar, J. Nagaraju, et al.** 2006. Microsatellite polymorphism across the *M. tuberculosis* and *M. bovis* genomes: implications on genome evolution and plasticity. BMC Genomics **7:**78.

298. **Stephens, A. J., J. Inman-Bamber, P. M. Giffard, et al.** 2008. High-Resolution Melting Analysis of the *spa* Repeat Region of *Staphylococcus aureus*. Clin. Chem. **54:**432-6.

299. **Stern, A., and T. F. Meyer.** 1987. Common mechanism controlling phase and antigenic variation in pathogenic *Neisseriae*. Mol. Microbiol. **1:**5-12.

300. **Stull, T. L., J. J. LiPuma, and T. D. Edlind.** 1988. A broad-spectrum probe for molecularepidemiology of bacteria: Ribosomal RNA. J. Infect. Dis. **157:**280-286.

301. **Suerbaum, S., J. M. Smith, K. Bapumia, et al.** 1998. Free recombination within *Helicobacter pylori*. Proceedings of the National Academy of Sciences USA **95:**12619-12624.

302. **Supply, P., S. Lesjean, E. Savine, et al.** 2001. Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units. J. Clin. Microbiol. **39:**3563-71.

303. **Supply, P., E. Mazars, S. Lesjean, et al.** 2000. Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome. Mol. Microbiol. **36:**762-71.

304. **Swofford, D. L.** 1998. PAUP: phylogenetic analysis using parsimony, 4.0 beta ed. Sinauer Associates, Sunderland.

305. **Szu, S. C., and S. Bystricky.** 2003. Physical, chemical, antigenic, and immunologic characterization of polygalacturonan, its derivatives, and Vi antigen from *Salmonella typhi*. Methods Enzymol. **363:**552-67.

306. **Szu, S. C., X. R. Li, A. L. Stone, et al.** 1991. Relation between structure and immunologic properties of the Vi capsular polysaccharide. Infect. Immun. **59:**4555-61.

307. **Thomson, N., S. Baker, D. Pickard, et al.** 2004. The role of prophage-like elements in the diversity of *Salmonella enterica* serovars. J. Mol. Biol. **339:**279-300.

308. **Thong, K. L., Z. A. Bhutta, and T. Pang.** 2000. Multidrug-resistant strains of Salmonella enterica serotype typhi are genetically homogenous and coexist with antibiotic-sensitive strains as distinct, independent clones. Int. J. Infect. Dis. **4:**194-7.

309. **Thong, K. L., M. Passey, A. Clegg, et al.** 1996. Molecular analysis of isolates of *Salmonella typhi* obtained from patients with fatal and nonfatal typhoid fever. J. Clin. Microbiol. **34:**1029-33.

310. **Thong, K. L., S. Puthucheary, R. M. Yassin, et al.** 1995. Analysis of *Salmonella typhi* isolates from Southeast Asia by pulsed-field gel electrophoresis. J. Clin. Microbiol. **33:**1938-41.

311. **Threlfall, E. J., E. Torre, L. R. Ward, et al.** 1993. Insertion sequence IS200 can differentiate drug-resistant and drug-sensitive *Salmonella typhi* of Vi-phage types E1 and M1. J. Med. Microbiol. **39:**454-8.

312. **Tomb, J. F., O. White, A. R. Kerlavage, et al.** 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*.[see comment][erratum appears in Nature 1997 Sep 25;389(6649):412]. Nature **388:**539-47.

313. **Torpdahl, M., and P. Ahrens.** 2004. Population structure of Salmonella investigated by amplified fragment length polymorphism. J. Appl. Microbiol. **97:**566-73.

314. **Torpdahl, M., G. Sorensen, B.-A. Lindstedt, et al.** 2007. Tandem repeat analysis for surveillance of human *Salmonella typhimurium* infections. Emerg. Infect. Dis. **13:**388-95.

315. **Townsend, S. M., N. E. Kramer, R. Edwards, et al.** 2001. *Salmonella enterica* Serovar Typhi Possesses a Unique Repertoire of Fimbrial Gene Sequences. Infect. Immun. **69:**2894-2901.

316. **Tyagi, S., D. P. Bratu, and F. R. Kramer.** 1998. Multicolor molecular beacons for allele discrimination. Nat. Biotechnol. **16:**49-53.

317. **Umemura, M., A. Yahagi, S. Hamada, et al.** 2007. IL-17-mediated regulation of innate and acquired immune response against pulmonary *Mycobacterium bovis* bacille Calmette-Guerin infection. J. Immunol. **178:**3786-96.

318. **Upadhye, V., S. Gujral, A. Maheshwari, et al.** 2005. Benign cystic teratoma of ovary perforating into small intestine with co-existent typhoid fever. Indian J. Gastroenterol. **24:**216-7.

319. **Van Belkum, A., S. Scherer, L. Van Alphen, et al.** 1998. Short-sequence DNA repeats in prokaryotic genomes. Microbiol. Mol. Biol. Rev. **62:**275-293.

320. **van de Verg, L. L., C. P. Mallett, H. H. Collins, et al.** 1995. Antibody and cytokine responses in a mouse pulmonary model of *Shigella flexneri* serotype 2a infection. Infect. Immun. **63:**1947-1954.

321. **van der Velden, A. W., A. J. Baumler, R. M. Tsolis, et al.** 1998. Multiple fimbrial adhesins are required for full virulence of *Salmonella typhimurium* in mice. Infect. Immun. **66:**2803-8.

322. **Van Ert, M. N., W. R. Easterday, L. Y. Huynh, et al.** 2007. Global Genetic Population Structure of *Bacillus anthracis*. PLoS ONE **2:**e461.

323. **Vernikos, G. S., N. R. Thomson, and J. Parkhill.** 2007. Genetic flux over time in the *Salmonella* lineage. Genome Biology **8:**R100-R100.16.

324. **Vieu, J. F., H. Binette, and M. Leherissey.** 1986. Absence of the antigen H:z66 in 2355 strains of *Salmonella typhi* from Madagascar and several countries of tropical Africa. Bull. Soc. Pathol. Exot. Filiales **79:**22-26.

325. **Vieu, J. F., and M. Leherissey.** 1988. The antigen H:z66 in 1,000 strains of *Salmonella typhi* from the Antilles, Central America and South America. Bull. Soc. Pathol. Exot. Filiales **81:**198-201.

326. **Vinogradov, E. V., Y. A. Knirel, N. K. Kochetkov, et al.** 1994. The structure of the O-specific polysaccharidee of Salmonella arizonae O62. Carbohydr. Res. **253:**101-110.

327. **Virlogeux, I., H. Waxin, C. Ecobichon, et al.** 1995. Role of the *viaB* locus in synthesis, transport and expression of *Salmonella typhi* Vi antigen. Microbiology **141:**3039-47.

328. **Vladoianu, I. R., H. R. Chang, and J. C. Pechere.** 1990. Expression of host resistance to *Salmonella typhi* and *Salmonella typhimurium*: bacterial survival within macrophages of murine and human origin. Microb. Pathog. **8:**83-90.

329. **Vos, P., R. Hogers, M. Bleeker, et al.** 1995. AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. **23:**4407-4414.

330. **Vulic, M., F. Dionisio, F. Taddei, et al.** 1997. Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. Proc. Natl. Acad. Sci. U. S. A. **94:**9763-9767.

331. **Wain, J., D. House, A. Zafar, et al.** 2005. Vi antigen expression in *Salmonella enterica* serovar Typhi clinical isolates from Pakistan. J. Clin. Microbiol. **43:**1158-65.

332. **Walsh, P. S., N. J. Fildes, and R. Reynolds.** 1996. Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA. Nucleic Acids Res. **24:**2807-2812.

333. **Wang, F., T. Whittam, and R. Selander.** 1997. Evolutionary genetics of the isocitrate dehydrogenase gene (*icd*) in *Escherichia coli* and *Salmonella enterica*. J. Bacteriol. **179:**6551-6559.

334. **Wang, L., D. Rothemund, H. Curd, et al.** 2003. Species-wide variation in the *Escherichia coli* flagellin (H antigen) gene. J. Bacteriol. **185:**2936-2943.

335. **Wani, T., D. K. Kakru, R. Shaheen, et al.** 2004. Infective endocarditis due to *Salmonella typhi*--a case report. Indian J. Pathol. Microbiol. **47:**76-7.

336. **Wei, J., M. B. Goldberg, V. Burland, et al.** 2003. Complete genome sequence and comparative genomics of Shigella flexneri serotype 2a strain 2457T. Infect. Immun. **71:**2775-2786.

337. **Weinstein, D. L., B. L. O'Neill, D. M. Hone, et al.** 1998. Differential early interactions between *Salmonella enterica* serovar typhi and two other pathogenic *Salmonella* serovars with intestinal epithelial cells. Infect. Immun. **66:**2310-2318.

338. **Weinstein, D. L., B. L. O'Neill, and E. S. Metcalf.** 1997. *Salmonella typhi* Stimulation of Human Intestinal Epithelial Cells Induces Secretion of Epithelial Cell-Derived Interleukin-6. Infect. Immun. **65:**395-404.

339. **Weiser, J. N., J. M. Love, and E. R. Moxon.** 1989. The molecular mechanism of phase variation of *H. influenzae* lipopolysaccharide. Cell **59:**657-65.

340. **WHO.** 2005. Typhoid fever, Democratic Republic of the Congo. Wkly. Epidemiol. Rec. **80:**1-8.

341. **Wiehler, S., and D. Proud.** 2007. Interleukin-17A modulates human airway epithelial responses to human rhinovirus infection. Am J Physiol Lung Cell Mol Physiol **293:**L505-15.

342. **Williams, J. G. K., A. R. Kubelik, K. J. Livak, et al.** 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res. **18:**6531-6535.

343. **Winter, S. E., M. Raffatellu, P. R. Wilson, et al.** 2008. The *Salmonella enterica* serotype Typhi regulator TviA reduces interleukin-8 production in intestinal epithelial cells by repressing flagellin secretion. Cell. Microbiol. **10:**247-261.

344. **Witonski, D., R. Stefanova, A. Ranganathan, et al.** 2006. Variable-number tandem repeats that are useful in genotyping isolates of *Salmonella enterica* subsp. *enterica* serovars Typhimurium and Newport. J. Clin. Microbiol. **44:**3849-54.

345. **Wong, C. K., C. Y. Ho, F. W. Ko, et al.** 2001. Proinflammatory cytokines (IL-17, IL-6, IL-18 and IL-12) and Th cytokines (IFN-gamma, IL-4, IL-10 and IL-13) in patients with allergic asthma. Clin. Exp. Immunol. **125:**177-83.

346. **Wong, K. K., M. McClelland, L. C. Stillwell, et al.** 1998. Identification and sequence analysis of a 27-kilobase chromosomal fragment containing a *Salmonella* pathogenicity island located at 92 minutes on the chromosome map of *Salmonella enterica* serovar Typhimurium LT2. Infect. Immun. **66:**3365-71.

347. **Wood, M. W., M. A. Jones, P. R. Watson, et al.** 1998. Identification of a pathogenicity island required for *Salmonella* enteropathogenicity. Mol. Microbiol. **29:**883-91.

348. **Woodward, T. E., J. E. Smadel, H. L. Ley, et al.** 1948. Preliminary report on the beneficial effect of choloromycetin in the treatment of typhoid fever. Ann. Intern. Med. **29:**131-134.

349. **Worth Jr., L., S. E. Clark, M. Radman, et al.** 1994. Mismatch repair proteins MutS and MutL inhibit RecA-catalyzed strand transfer between diverged DNAs. Proc. Natl. Acad. Sci. U. S. A. **91:**3238-3241.

350. **Wu, K.-Y., G.-R. Liu, W.-Q. Liu, et al.** 2005. The genome of *Salmonella enterica* serovar Gallinarum: distinct insertions/deletions and rare rearrangements. J. Bacteriol. **187:**4720-7.

351. **Yamaoka, Y., T. Kodama, K. Kashima, et al.** 1998. Variants of the 3' region of the *cagA* gene in *Helicobacter pylori* isolates from patients with different *H. pylori*-associated diseases.[see comment]. J. Clin. Microbiol. **36:**2258-63.

352. **Yang, F., J. Yang, X. Zhang, et al.** 2005. Genome dynamics and diversity of *Shigella* species, the etiologic agents of bacillary dysentery. Nucleic Acids Res. **33:**6445-58.

353. **Yang, H. H., C. G. Wu, G. Z. Xie, et al.** 2001. Efficacy trial of Vi polysaccharide vaccine against typhoid fever in south-western China. Bull. World Health Organ. **79:**625-31.

354. **Ye, P., P. B. Garvey, P. Zhang, et al.** 2001. Interleukin-17 and lung host defense against *Klebsiella pneumoniae* infection. Am. J. Respir. Cell Mol. Biol. **25:**335-40.

355. **Youil, R., B. W. Kemper, and R. G. H. Cotton.** 1995. Screening for mutations by enzyme mismatch cleavage with T4 endonuclease VII. Proc. Natl. Acad. Sci. U. S. A. **92:**87-91.

356. **Zhang, W., W. Qi, T. Albert, et al.** 2006. Probing genomic diversity and evolution of *Eschericia coli* O157 by single nucleotide polymorphisms. Genome Res. **16:**757-767.

357. **Zhang, W., W. Qi, T. J. Albert, et al.** 2006. Probing genomic diversity and evolution of *Escherichia coli* O157 by single nucleotide polymorphisms. Genome Res. **16:**757-767.

358. **Zhang, X.-L., I. S. M. Tsui, C. M. C. Yip, et al.** 2000. *Salmonella enterica* Serovar Typhi Uses Type IVB Pili To Enter Human Intestinal Epithelial Cells. Infect. Immun. **68:**3067-3073.

359. **Zhao, L., T. Ezak, Z. Y. Li, et al.** 2001. Vi-Suppressed wild strain *Salmonella typhi* cultured in high osmolarity is hyperinvasive toward epithelial cells and destructive of Peyer's patches. Microbiol. Immunol. **45:**149-58.