Frequent recombination and low level of clonality within *Salmonella enterica* subspecies I

Sophie Octavia and Ruiting Lan

Correspondence
Ruiting Lan
r.lan@unsw.edu.au

School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW 2052, Australia

The genetic relationship and population structure of *Salmonella enterica* subspecies I strains were analysed using nucleotide sequences of four genes (*mgIA*, *proV*, *torC* and *speC*). Fifteen strains from the *Salmonella* reference collection B (SARB), belonging to 13 serovars, were analysed. Sequence data of two housekeeping genes, *mdh* and *mutS*, of the same 15 strains reported by Brown *et al.* (2003) (*Proc Natl Acad Sci U S A* **100**, 15676–15681) were also included in the analyses. Phylogenetic analysis revealed that there was a lack of congruence among the six gene trees. Split decomposition analysis resolved only five strains with a network structure, while others showed a star phylogeny. Compatibility values for the SARB strains were the lowest in comparison to those for strains representing different subspecies of *S. enterica*. These results showed that the genes studied have undergone frequent recombination, suggesting a low level of clonality within subspecies I of *S. enterica*.

Received 2 September 2005
Revised 21 November 2005
Accepted 2 December 2005

INTRODUCTION

Salmonella has been assigned to more than 2400 different serovars, based on a serotyping scheme that accounts for the differences in antigenic properties of the LPS (O antigen) and the flagellin (H antigen) (Popoff, 2001). These serovars were originally designated by Latin binomial species names, but, because of their close relatedness, the species names were subsequently retained as the serovar names of the single *Salmonella* species known as *Salmonella enterica* (Le Minor & Popoff, 1987; Brenner *et al.*, 2000). For example, the name *Salmonella typhi* refers to *S. enterica* serovar Typhi, or simply Typhi (the latter convention is used in this paper). Based on DNA hybridization and biotyping studies, the *Salmonella* serovars have been classified into seven subspecies: I, II, IIIa, IIIb, IV, V and VI (Crosa *et al.*, 1973; Le Minor *et al.*, 1986). Multilocus enzyme electrophoresis (MLEE) has defined an eighth group, designated subspecies VII, which consists of five isolates of two serovars that were initially allocated to subspecies IV on the basis of biochemical characteristics (Boyd *et al.*, 1996).

Most subspecies of *S. enterica* are not commonly associated with disease, and they may behave like commensals in cold-blooded animals (Baumler *et al.*, 1998). However, subspecies I strains cause intestinal infections in warm-blooded animals, and are responsible for 99 % of *Salmonella*-related

infections in humans (Selander *et al.*, 1996; Popoff, 2001). The widely prevalent serovar Typhimurium causes gastroenteritis in humans, but mainly asymptomatic chronic infection in chickens. A number of serovars have a restricted host range; for example, Typhi exclusively infects humans, causing typhoid fever.

MLEE has been used extensively to study the extent of genetic diversity within *S. enterica* natural populations. The technique has shown that many serovars vary genetically, and are represented by multiple electrophoretic types (ETs) (Beltran *et al.*, 1988, 1991; Reeves *et al.*, 1989; Selander *et al.*, 1990a, b). Some serovars are genotypically heterogeneous; for example, Derby and Newport (Beltran *et al.*, 1988) include divergent isolates, with ETs clustered distantly in MLEE trees, while other serovars can be confined within a single cluster of closely related ETs, in which each serovar has a predominant widely distributed ET (Beltran *et al.*, 1988, 1991; Reeves *et al.*, 1989; Selander *et al.*, 1990a, b). From 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 (Beltran *et al.*, 1991); *Salmonella* reference collection B (SARB), which consists of 72 strains of 37 subspecies I serovars (Boyd *et al.*, 1993); and *Salmonella* reference collection C (SARC), which consists of 16 strains representing the eight subspecies (Boyd *et al.*, 1996).

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 (Beltran *et al.*,

Abbreviations: ET, electrophoretic type; I_A , index of association; ML, maximum likelihood; MLEE, multilocus enzyme electrophoresis; MLST, multilocus sequence typing; NJ, neighbour joining.

The GenBank/EMBL/DDBJ accession numbers for the nucleotide sequences reported in this paper are DQ285482–DQ285541.

1988, 1991; Reeves *et al.*, 1989; Selander *et al.*, 1990a, b). 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 (Nelson *et al.*, 1991, 1997; Nelson & Selander, 1992, 1994; Boyd *et al.*, 1994; Selander *et al.*, 1996; Wang *et al.*, 1997). These findings led to the conclusion that *S. enterica* has one of the highest levels of clonality among bacterial species.

In this study, we sequenced four genes from a selected number of SARB strains in order to determine the genetic relationships of strains belonging to subspecies I, looking in particular for the existence of 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 most of the isolates studied.

METHODS

Bacterial isolates. Fifteen SARB strains were chosen (Table 1). The strains were obtained from the *Salmonella* Genetic Stock Centre (SGSC), University of Calgary, Canada. For convenience, the strain names designated by Boyd *et al.* (1993) have been used instead of the SARB numbers. SARB contains two Typhi strains, Tp1 and Tp2; Tp2 alone was selected for this study, since Tp1 is identical to genome sequence strain CT18, based on multilocus sequence typing (MLST) by Kidgell *et al.* (2002). We further confirmed the identity of Tp2 by sequencing three MLST genes (*hemD*, *hisD* and *thrA*), which were shown to vary among the Typhi isolates studied by Kidgell *et al.* (2002). The other SARB strains were selected because they have the least number of allelic differences to Tp2, according to MLEE data (Boyd *et al.*, 1993), or because they cause enteric fever in humans. The identity of all other strains used in this study was confirmed by PCR serogrouping (Luk *et al.*, 1993; Hoorfar *et al.*, 1999), 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, as described by Bastin *et al.* (1991).

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: *mgIA* (galactoside transport ATP-binding protein MglA), *proV* (glycine betaine L-proline transport ATP-binding protein), *speC* (ornithine decarboxylase) and *torC* (cytochrome-*c*-type protein) (Parkhill *et al.*, 2001). These genes are functional in Typhimurium LT2, but they are pseudogenes in Typhi CT18. The primer pairs used in this study were designed based on the Typhi CT18 genome sequence (Table 2), and synthesized commercially (Sigma). The sequences reported in this paper have been deposited in the GenBank database (accession nos DQ285482–DQ285541).

PCR assay and DNA sequencing. Each PCR reaction included 2.5 µl DNA template (approx. 20 ng), 0.5 µl (30 pmol µl⁻¹) of each forward and reverse primer, 0.5 µl 10 mM dNTPs, 5 µl 10 × PCR buffer (500 mM KCl, 100 mM Tris/HCl, pH 9.0, 1% Triton X-100 and 15 mM MgCl₂), 0.25 µl (1.25 U) *Taq* polymerase (Promega), and MilliQ water to a total volume of 50 µl. PCR cycles were performed in a Hybaid PCR Sprint Thermocycler (Hybaid): initial DNA denaturation for 2 min at 94 °C, followed by DNA denaturation for 15 s at 94 °C, primer annealing for 30 s at 50 °C, and polymerization for 90 s at 72 °C for 35 cycles, with a final extension of 5 min at 72 °C. PCR products were verified on ethidium-bromide-stained agarose gels, before purification using sodium acetate/ethanol precipitation. The PCR sequencing reactions contained BigDye, and were done as recommended by the manufacturer (Applied Biosystems). We sequenced both forward and reverse directions. Unincorporated dye terminators were removed by ethanol precipitation. The reaction products were separated and detected by gel electrophoresis, using the Automated DNA Sequence Analyser ABI377 or ABI3730 (Applied Biosystems) at the sequencing facility of the School of Biotechnology and Biomolecular Sciences, University of New South Wales, Australia.

Bioinformatic analysis. The CONSED version 8.0 (Gordon *et al.*, 1998) program package, accessed through the Australian National Genomic Information Service, was used for sequence editing. PILEUP

Table 1. *S. enterica* strains used in this study

Strain	SARB no.	Serovar	Source	Locality (year)	MLEE difference*
Cs6	SARB5	Choleraesuis	Unknown	Switzerland	16
De1	SARB9	Derby	Avian	Oklahoma (1986)	17
In1	SARB26	Infantis	Human	North Carolina	11
Mo1	SARB30	Montevideo	Human	Georgia	10
Np8	SARB36	Newport	Human	North Carolina	10
Pa1	SARB42	Paratyphi A	Human	Laboratory strain	12
Pb1	SARB43	Paratyphi B	Human	France (1976)	14
Pb7	SARB47	Paratyphi B	Human	Africa (1981)	11
Pc2	SARB49	Paratyphi C	Human	France (1988)	13
Pc4	SARB50	Paratyphi C	Human	France (1977)	9
Pn1	SARB39	Panama	Human	Italy	10
Sf1	SARB59	Seftenberg	Chicken	Maryland (1987)	11
Sw1	SARB57	Schwarzengrund	Unknown	Scotland (1988)	10
Tm1	SARB65	Typhimurium	Human	Mexico	14
Tp2	SARB64	Typhi	Human	Senegal (1988)	—

*Based on MLEE data from Boyd *et al.* (1993); the value given for allelic difference is between the strain and Typhi Tp2.

Table 2. Genes and primers used in this study

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

*According to the Typhi CT18 genome (accession no. NC_003198).

†Relative to the first base of the initiation codon.

from the Genetics Computer Group package (Dolz, 1994), and MULTICOMP (Reeves *et al.*, 1994), were used for multiple sequence alignment and comparison. PHYLIP (Felsenstein, 1989) was used to generate phylogenetic trees and bootstrap values. SPLITTREE version 3.2 (Bandelt & Dress, 1992) was used to create network structures using the distance method. Overall compatibility of informative sites was measured by using the RETICULATE program (Jakobsen & Easteal, 1996), 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.* (2000, 2001), was done using PAUP version 4.0 beta (Swofford, 1998), with the parameters of the HKY85 model of DNA substitutions, estimation of the transition/transversion (Ti/Tv) ratio, and α parameter, assuming gamma distribution. ML generates scores for comparison of one gene tree against another based on the 99th 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 ($\Delta\text{-lnL}$) is lower than that of any of the 200 random trees, since the second gene tree should be of better fit to the data from the first gene than the 200 random trees (Feil *et al.*, 2000, 2001). Calculation of the linkage disequilibrium index (I_A) (Maynard Smith *et al.*, 1993) from MLEE data was done using an in-house program, MLEECOMP.

RESULTS

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, 818, 820 and 604 bp for *mglA*, *proV*, *speC* and *torC*, respectively. The mean pairwise percentage difference for all genes and strains was 1.06 (Table 3). A total of 133 sites were polymorphic (sites at which more than one type of nucleotide 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.* (2003) study for the same SARB strains used in this study, were included for comparison (Table 3) and subsequent analyses.

Comparison of Typhi Tp2 with the two genome sequence strains CT18 (Parkhill *et al.*, 2001) and Ty2 (Deng *et al.*, 2003) revealed that Tp2 was identical to Ty2 in all four genes sequenced, but differed from CT18 by one base in

Table 3. Pairwise nucleotide difference

Gene	Size (bp)	Pairwise percentage difference			No. polymorphic sites	No. informative sites
		Mean	Minimum	Maximum		
<i>mglA</i>	743	1.35	0.27	2.02	37	19
<i>proV</i>	818	0.78	0.12	1.59	26	12
<i>speC</i>	820	1.27	0.00	2.32	49	24
<i>torC</i>	604	0.83	0.17	1.82	21	11
<i>mdh</i> *	831	0.81	0.00	1.56	24	13
<i>mutS</i> *	1098	0.81	0.09	1.28	35	21

*Genes sequenced by Brown *et al.* (2003).

torC. While most SARB strains have functionally intact sequences, four cases of gene inactivation were observed. Two strains, Pc2 and Pc4, had the same deletion as the Typhi strain, i.e. deletion of a CG repeat in *mgIA*. The changes in Pc2 and Pc4 must be independent, as their sequences are very different from each other. Strain Pc2 also had a C-to-T substitution, forming a stop codon in *torC*. Strain Pa1 had a C-to-T substitution, leading to a stop codon in *mgIA*.

The sequence alignment for informative sites is shown in Fig. 1. It is clear that none of the strains is consistently similar in all six genes, suggesting the presence of intergenic recombination. Only two pairs of strains, Pn1 and Mo1, and Sw1 and Pc4, shared similarity in two or more genes. Pn1 and Mo1 shared similarity over the entire sequence in three genes (*proV*, *torC* and *mdh*), but only parts of the sequences in two other genes (*speC* and *mutS*). Sw1 and Pc4 had almost identical sequences in two genes (*proV* and *mdh*), but only some segments similar in three other genes (*mgIA*, *speC* and *mutS*).

Phylogenetic relationships

Evolutionary trees were constructed by the neighbour-joining (NJ) method for each of the six genes, and also for the concatenated sequences of all six genes (Fig. 2). The individual gene trees did not resemble one another in their topology, and inconsistent clustering of strains was observed. However, there were three cases where strains were grouped closely together in two or more genes: In1 with Np8 in *mgIA* 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, since bootstrap values, including those for the combined six-gene tree, were low. We believe that this is due to conflicting signals resulting from recombination, which will be discussed later, and not because of low phylogenetic signal in our data. The combined six-gene tree was then compared with an MLEE tree using data from Boyd *et al.* (1993), which was reconstructed to include only the strains used in this study. Except for three strains, Sw1, Pn1 and Pc4, which broadly fell within the same cluster, other strains were inconsistently clustered in the two trees.

Split decomposition (Bandelt & Dress, 1992) was then used to visualize the relationship of the strains. The method displays conflicting phylogenetic signals resulting from recombination as network structures. As shown in Fig. 2, 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 suggests that recombination is extensive, and that the strain relationships are not well represented by a splits graph.

Congruence analysis

To establish the degree of incongruence among the six gene trees, ML analysis (Feil *et al.*, 2000, 2001) was carried out, and the results are summarized in Table 4. None of the six gene trees was congruent to all the other gene trees. The gene trees with the largest number of congruencies were those of

	<i>mgIA</i>	<i>proV</i>	<i>speC</i>	<i>torC</i>	<i>mdh</i>	<i>mutS</i>
			111111111111111111111111111111			1111111111111111222222222
	23334444445556677888	223444558889	02222233444555556666677	22333555666	234446666677	456778888889902333444
	9033022472730656008	665258275581	711125281581156635778929	03135012689	7170461346759	625171118898953079567
	4316889145803392488	288774340936	125845633552572356149881	71678510461	8928728059265	149632351769885497768
Consensus	CACAGCGGCATCCCTTCT	TCGCGTGGTCCT	TCCGTACGGGTCCTGGGGCCAGGG	CCCGCGCCCA	CCTTTATATAATC	GTGGATCGGGCCTCGGAACGT
Del1	T.T.A.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

Fig. 1. Informative sites of the four genes sequenced in this study, and two additional genes (*mdh* and *mutS*) from the study of Brown *et al.* (2003). The numbers above the alignment, reading vertically, are base positions.

Fig. 2. Phylogenetic trees. Shown are NJ trees of the individual genes, the concatenated sequences of six genes, and the MLEE data; and the split tree of the concatenated sequences of six genes. Bootstrap values, if greater than 50 %, are presented at nodes of the NJ trees.

mdh and *mutS*, which were congruent to three other gene trees. *proV* and *torC* trees were congruent to two other gene trees, while the *speC* tree was congruent to one other gene tree. The *mgIA* tree was not congruent to any 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 housekeeping-gene trees (*aceK*, *gapA*, *icd*, *mdh*, *putP* and *gnd*) of SARC strains sequenced by Selander's group (Nelson *et al.*, 1991, 1997; Nelson & Selander, 1992, 1994; Boyd *et al.*,

1994; Wang *et al.*, 1997); all the gene trees were congruent to each other (data not shown).

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 & Eastal (1996). We calculated compatibility values both within a gene, and between genes (Fig. 3). *mgIA* had the lowest

Table 4. ML analysis for congruence between each gene tree of the SARB strains analysed in this study

Gene	α^*	Ti/Tv ratio†	-lnL	-lnL (99th)‡	Δ -lnL score of gene trees§					
					<i>mgla</i>	<i>proV</i>	<i>speC</i>	<i>torC</i>	<i>mdh</i>	<i>mutS</i>
<i>mgla</i>	0.61	4.42	1416	28	—	46	44	38	45	42
<i>proV</i>	0.02	5.09	1389	25	40	—	35	25	25	30
<i>speC</i>	0.69	4.80	1482	64	110	72	—	89	44	77
<i>torC</i>	0.29	14.03	1016	30	48	27	34	—	28	31
<i>mdh</i>	0.49	5.63	1386	39	62	36	23	47	—	38
<i>mutS</i>	0.02	5.74	1876	46	65	33	36	48	34	—

*Nucleotide substitution rate variation between sites, with gamma distribution as the parameter.
†Estimated transition/transversion ratio.
‡Difference in -lnL score from -lnL column (reference data) and the 99th percentile from random topology.
§Difference in -lnL score from reference data to each calculated score from other genes. Trees are deemed to be congruent if Δ -lnL is equal to or lower than the 99th percentile of random trees, when compared with reference data (Feil *et al.*, 2000, 2001). Congruent gene trees are in bold type.

within-locus compatibility, at 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 %, were 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 *aceK*, *gapA*, *icd*, *mdh*, *putP* and *gnd* from the 16 SARC strains sequenced by Selander’s group (Nelson *et al.*, 1991, 1997; Nelson & Selander, 1992, 1994; Boyd *et al.*, 1994; Wang *et al.*, 1997). As shown in Fig. 3, 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 *Escherichia coli*, using data from Reid *et al.* (2000) 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* (Fig. 3).

DISCUSSION

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 observed could be a result of low sequence variation, but this is less likely, since ML analysis is relatively insensitive to sequence variation, and has been applied to other species with a comparable low level of variation (Feil *et al.*, 2001). Compatibility analysis was consistent with ML analysis, showing that recombination occurs frequently within *S. enterica* subspecies I. Altogether, the results suggest that the level of clonality within subspecies I is low.

It is apparent from our analysis that there are different levels of clonality within *S. enterica* subspecies. Comparison of our

Fig. 3. Comparison of mean compatibility values within and between loci of *S. enterica* and *E. coli* strains. Selander’s set refers to the sequence data for six housekeeping genes of 16 *S. enterica* strains representing different subspecies, obtained from the studies of Selander’s group (Nelson *et al.*, 1991, 1997; Nelson & Selander, 1992, 1994; Boyd *et al.*, 1994; Wang *et al.*, 1997). Reid’s set refers to the sequence data for seven housekeeping genes of 14 common pathogenic *E. coli* strains, obtained from Reid *et al.* (2000).

data with those from the SARC set, which represents the eight different subspecies, by both ML and compatibility analyses, showed that recombination occurs far more frequently within *S. enterica* subspecies I than between *S. enterica* subspecies. This situation is rather similar to the case of *Rhizobium meliloti*; there are two major divisions of the species, and recombination is rare between the divisions, but common within the divisions (Maynard Smith *et al.*, 1993).

The level of recombination in *S. enterica* subspecies I can be compared with that in other species. By compatibility analysis, we showed that the frequency of recombination in subspecies I was higher than in *E. coli*. The ML analysis allowed comparisons with a number of species to which the method has been applied (Feil *et al.*, 2000, 2001). The percentages of gene-tree comparisons which are congruent are 88, 75, 55 and 7 % for *E. coli* (Feil *et al.*, 2000), *Haemophilus influenzae* (Feil *et al.*, 2000), *Staphylococcus aureus* (Feil *et al.*, 2001) and *Neisseria meningitidis* (Feil *et al.*, 2000), respectively. In our study, 37 % of the comparisons were congruent among the SARB strains (Table 4). Therefore, the level of clonality of subspecies I is at the lower end of the spectrum, in comparison with *E. coli*, *H. influenzae*, *S. aureus* and *N. meningitidis*.

It is interesting to note that Brown *et al.* (2003) recently reported that *mutS*, a gene involved in mismatch repair and a strong mutator, undergoes frequent recombination in SARB strains, in comparison with SARC strains. That study used *mdh* for comparison, and attributed incongruence of the two gene trees to recombination in *mutS*. Using two genes only, one cannot determine which gene has undergone recombination; *mdh* was regarded as 'non-recombinant', on the assumption that it does not undergo recombination in the 'highly clonal' species *S. enterica*. However, a comparison of the six gene trees shows that *mdh* undergoes frequent recombination, as does *mutS*. Interestingly, despite the potential mutator properties of *mutS*, the *mutS* tree is not the most incongruent to the other gene trees. It seems that *mutS* is not more recombinogenic than the other genes, although the implications of this are not yet clear.

Re-examination of the MLEE data challenges the myth of high clonality at all levels in *S. enterica*

The results from this study are in sharp contrast to the long-held view that *S. enterica* has a highly clonal population structure (Selander *et al.*, 1991; Maynard Smith *et al.*, 1993). In the landmark paper on bacterial population structures by Maynard Smith *et al.* (1993), 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. That study used I_A to measure the extent of linkage disequilibrium from MLEE data; I_A values for *S. enterica* were

significantly greater than zero. Our sequence data seem 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 I_A as a relative measure of clonality. The MLEE data for *S. enterica* used by Maynard Smith *et al.* (1993) were, in fact, for 14 serovars of subspecies I, originating from Selander *et al.* (1990b). The dataset thus represents subspecies I rather than the whole species. We obtained the MLEE data from Boyd *et al.* (1996) for 80 ETs that represent all eight subspecies. The I_A for the 80 ETs is 3.219 ± 0.156 , which is almost two and half times greater 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 this effect on the scale of the I_A . The difference in I_A for data between subspecies I and the whole species seems to reflect their difference in the level of clonality, and is consistent with our sequence data.

We further examined the subspecies I MLEE data, reported by Selander *et al.* (1990b). We tested whether removing closely related ETs, which potentially correspond to clonal complexes, affects the I_A . We used eBURST (Feil *et al.*, 2004) to identify closely related ETs, and one ET was selected to represent each cluster. When ETs differing by one, two and three 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 disappears when closely related ETs are treated as a unit. This change in I_A resembles that which occurs in an organism with an epidemic population structure, such as *N. meningitidis* (Maynard Smith *et al.*, 1993). 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.* (1993) that *S. enterica* is clonal at all levels was based on the finding that I_A values for individual serovars are equal to or higher than those for the whole dataset (see Table 1 of Maynard Smith *et al.*, 1993). Their interpretation assumed that a serovar represents a real genetic group. However, a number of serovars, including Paratyphi C and Choleraesuis, used in Maynard Smith *et al.* (1993), are known to be not clustered together with a single origin (Selander *et al.*, 1990b). We suspected that this may have contributed to the high I_A values, and examined the data for Paratyphi C and Choleraesuis from Selander *et al.* (1990b). 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 our I_A calculations. Similarly, when we took out the two divergent isolates (Cs6 and Cs13) from the Choleraesuis 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 erroneous interpretation of high clonality at the serovar level, at least for Paratyphi C and Choleraesuis.

In a series of studies of the SARC set using six housekeeping genes (Nelson *et al.*, 1991; Nelson & Selander, 1992, 1994; Boyd *et al.*, 1994, 1996; Wang *et al.*, 1997), 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 (Brown *et al.*, 2002). Additionally, in most cases, the branching patterns among the subspecies are also consistent, suggesting only low levels of recombination. The sequence data were interpreted as a strong support of the interpretation from the MLEE data that *S. enterica* is highly clonal, but, as one can now see, wrongly reinforced the erroneous interpretation of the I_A analysis from the subspecies I MLEE data. Further, the housekeeping gene studies used only two isolates to represent a subspecies, which would not allow identification of recombination events within a subspecies.

Predominance of intra-subspecies recombinational exchange

Based on the level of variation in the six genes, it seems that recombinational exchange occurs 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 (Nelson *et al.*, 1991, 1997; Nelson & Selander, 1992, 1994; Boyd *et al.*, 1994; Wang *et al.*, 1997) 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 *mutS* between SARC and SARB strains sequenced by Brown *et al.* (2003). For *mdh*, the difference between strains of subspecies I and the other subspecies ranged from 2.31 to 8.66%. In contrast, *mdh* from the 15 SARB strains of this study had a mean of 0.81%, and a maximum of 1.56%. Similarly for *mutS*, no strain within subspecies I had a level of difference equal to or higher than that between subspecies. The predominance of intra-subspecies recombination may be the result of a number of factors. *MutS* creates a barrier to the recombination of divergent DNA (Rayssiguier *et al.*, 1989; Worth *et al.*, 1994; Vulic *et al.*, 1997; Radman *et al.*, 1999; Brown *et al.*, 2003) that may block recombination with other subspecies. There could also be a niche barrier (Matic *et al.*, 1996). *S. enterica* strains of subspecies I usually share a common niche, 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*.

Relationships of subspecies I isolates

Typhi has been shown to be a homogeneous clone, and has been suggested to have arisen about 50 000 years ago (Kidgell *et al.*, 2002). Using sequence data, we initially wished to determine which SARB strain is the closest relative of Typhi. In the MLEE tree of 72 SARB strains, the two Typhi strains Tp1 and Tp2 are grouped together, and clustered with Derby De1 (Boyd *et al.*, 1993). However, Typhi Tp2

differs from the other serovars by at least 9 of the 24 enzyme loci studied by Boyd *et al.* (1993), with the least allelic difference to Pc4, rather than to De1, of 17 differences. The NJ trees (Fig. 2) showed that Typhi was placed inconsistently in the six gene trees. Typhi was clustered together with Pb7 in *mglA* and *mutS*, and with Pa1 in *speC*; these relationships were also reflected in the near-identical sequences in *mglA* and *speC* (Fig. 1). 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 is no clear indication of the closest relative of the Typhi clone in the 15 SARB strains analysed.

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 split tree and the combined six-gene NJ tree showed that the four strains formed one group, in which Sw1 and Pc4 appeared to be more closely related. A high level of recombination appears 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 (Pabbaraju *et al.*, 2000; Torpdahl & Ahrens, 2004). Although we can only make a comparison of 15 of the 72 strains, no consistency of clustering of strains was observed between the MLEE tree and the combined sequence tree (Fig. 2), 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.

Concluding comments

Our study has contributed to a better understanding of the population structure of *S. enterica*. Previous sequence studies, using strains of different subspecies (Selander *et al.*, 1996), have shown 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 suggests that recombination has occurred at a frequency sufficiently high to have eliminated many of the phylogenetic signals. Statistical analyses using compatibility, split decomposition and ML provide further evidence that recombination is frequent in *S. enterica* subspecies I. These findings reveal that the clonality of *S. enterica* varies within the species. Further studies are required to quantify recombination and mutation parameters in subspecies I, and to ascertain these parameters in the other subspecies.

Our observation of a high level of recombination within subspecies I indicates the need for further work on the evolution of *S. enterica* clones. Nearly 1500 serovars in subspecies I have been reported, comprising 60% of known *S. enterica* serovars (Popoff, 2001). Only 2% of subspecies I serovars have been studied at population genetic level, largely by MLEE, which has provided a limited picture of

evolutionary origins of the specialized (e.g. host-adapted) clones, and the diversity of the subspecies. The findings of frequent recombination from this study have now blurred that picture, since the MLEE relationships between more distantly related ETs can no longer be considered reliable. There is a great need to determine the relationships at sequence level, using multilocus sequence data, of clones encompassing the whole subspecies, in addition to those frequently encountered in human and domestic animal infections, and this will provide a better framework within which to study the evolution of pathogenicity and host adaptation.

ACKNOWLEDGEMENTS

This work was supported by a Faculty Research Grant from the University of New South Wales. We thank Alfred Tay for technical assistance, Dr Ken Sanderson, The University of Calgary, for strains, Dr Ed Feil for advice on ML analysis, and Dr Fidelma Boyd for kindly providing us with the MLEE data of the whole species. We also thank Dr Mark Tanaka for critical reading of the manuscript, and the anonymous referees for suggestions.

REFERENCES

- Bandelt, H. J. & Dress, A. W. (1992). Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Mol Phylogenet Evol* **1**, 242–252.
- Bastin, D. A., Romana, L. K. & Reeves, P. R. (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.
- Baumler, A., Tsolis, R., Ficht, T. & Adams, L. (1998). Evolution of host adaptation in *Salmonella enterica*. *Infect Immun* **66**, 4579–4587.
- Beltran, P., Musser, J. M., Helmuth, R. & 8 other authors (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*. *Proc Natl Acad Sci U S A* **85**, 7753–7757.
- Beltran, P., Plock, S. A., Smith, N. H., Whittam, T. S., Old, D. C. & Selander, R. K. (1991). Reference collection of strains of the *Salmonella typhimurium* complex from natural populations. *J Gen Microbiol* **137**, 601–606.
- Boyd, E. F., Wang, F.-S., Beltran, P., Plock, S. A., Nelson, K. & Selander, R. K. (1993). *Salmonella* reference collection B (SARB): strains of 37 serovars of subspecies I. *J Gen Microbiol* **139**, 1125–1132.
- Boyd, E. F., Nelson, K., Wang, F.-S., Whittam, T. S. & Selander, R. K. (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.
- Boyd, E. F., Wang, F. S., Whittam, T. S. & Selander, R. K. (1996). Molecular genetic relationships of the salmonellae. *Appl Environ Microbiol* **62**, 804–808.
- Brenner, F. W., Villar, R. G., Angulo, F. J., Tauxe, R. & Swaminathan, B. (2000). *Salmonella* nomenclature. *J Clin Microbiol* **38**, 2465–2467.
- Brown, E. W., Kotewicz, M. L. & Cebula, T. A. (2002). Detection of recombination among *Salmonella enterica* strains using the incongruence length difference test. *Mol Phylogenet Evol* **24**, 102–120.
- Brown, E. W., Mammel, M. K., LeClerc, J. E. & Cebula, T. A. (2003). Limited boundaries for extensive horizontal gene transfer among *Salmonella* pathogens. *Proc Natl Acad Sci U S A* **100**, 15676–15681.
- Crosa, J. H., Brenner, D. J., Ewing, W. H. & Falkow, S. (1973). Molecular relationships among the *Salmonellae*. *J Bacteriol* **115**, 307–315.
- Deng, W., Liou, S. R., Plunkett, G., 3rd, Mayhew, G. F., Rose, D. J., Burland, V., Kodoyianni, V., Schwartz, D. C. & Blattner, F. R. (2003). Comparative genomics of *Salmonella enterica* serovar Typhi strains Ty2 and CT18. *J Bacteriol* **185**, 2330–2337.
- Dolz, R. (1994). GCG: comparison of sequences. *Methods Mol Biol* **24**, 64–82.
- Feil, E. J., Smith, J. M., Enright, M. C. & Spratt, B. G. (2000). Estimating recombinational parameters in *Streptococcus pneumoniae* from multilocus sequence typing data. *Genetics* **154**, 1439–1450.
- Feil, E. J., Holmes, E. C., Bessen, D. E. & 9 other authors (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.
- Feil, E. J., Li, B. C., Aanensen, D. M., Hanage, W. P. & Spratt, B. G. (2004). eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J Bacteriol* **186**, 1518–1530.
- Felsenstein, J. (1989). PHYLIP – phylogeny inference package. *Cladistics* **5**, 164–166.
- Gordon, D., Abajian, C. & Green, P. (1998). CONSED – a graphical tool for sequence finishing. *Genome Res* **8**, 195–202.
- Hoorfar, J., Baggesen, D. L. & Porting, P. H. (1999). A PCR-based strategy for simple and rapid identification of rough presumptive *Salmonella* isolates. *J Microbiol Methods* **35**, 77–84.
- Jakobsen, I. B. & Easteal, S. (1996). A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *CABIOS* **12**, 291–295.
- Kidgell, C., Reichard, U., Wain, J., Linz, B., Torpdahl, M., Dougan, G. & Achtman, M. (2002). *Salmonella typhi*, the causative agent of typhoid fever, is approximately 50 000 years old. *Infect Genet Evol* **2**, 39–45.
- Le Minor, L. & Popoff, M. Y. (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.
- Le Minor, L., Popoff, M. Y., Laurent, B. & Hermant, D. (1986). Individualisation D'une septieme sous-espece de *Salmonella*: *S. choleraesuis* subsp. *indica* subsp. nov. *Ann Inst Pasteur Microbiol* **137B**, 211–217.
- Luk, J. M. C., Kongmuang, U., Reeves, P. R. & Lindberg, A. A. (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.
- Matic, I., Taddei, F. & Radman, M. (1996). Genetic barriers among bacteria. *Trends Microbiol* **4**, 69–73.
- Maynard Smith, J., Smith, N. H., O'Rourke, M. & Spratt, B. G. (1993). How clonal are bacteria? *Proc Natl Acad Sci U S A* **90**, 4384–4388.
- Nelson, K. & Selander, R. K. (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.
- Nelson, K. & Selander, R. K. (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.
- Nelson, K., Whittam, T. S. & Selander, R. K. (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.

Nelson, K., Wang, F. S., Boyd, E. F. & Selander, R. K. (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.

Pabbaraju, K., Miller, W. & Sanderson, K. (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**, 1923–1929.

Parkhill, J., Dougan, G., James, K. D. & 38 other authors (2001). Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature* **413**, 848–852.

Popoff, M. Y. (2001). *Antigenic Formulas of the Salmonella Serovars*, 8th edn. Paris, France: WHO Collaborating Centre for Reference and Research on *Salmonella*, Institut Pasteur.

Radman, M., Matic, I. & Taddei, F. (1999). Evolution of evolvability. *Ann N Y Acad Sci* **870**, 146–155.

Rayssiguier, C., Thaler, D. S. & Radman, M. (1989). The barrier to recombination between *Escherichia coli* and *Salmonella typhimurium* is disrupted in mismatch-repair mutants. *Nature* **342**, 396–400.

Reeves, M. W., Evins, G. M., Heiba, A. A., Plikaytis, B. D. & Farmer, J. J., III (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.

Reeves, P. R., Farnell, L. & Lan, R. (1994). MULTICOMP: a program for preparing sequence data for phylogenetic analysis. *CABIOS* **10**, 281–284.

Reid, S. D., Herbelin, C. J., Burnnough, A. C., Selander, R. K. & Whittam, T. S. (2000). Parallel evolution of virulence in pathogenic *Escherichia coli*. *Nature* **406**, 64–67.

Selander, R. K., Beltran, P., Smith, N. H., Barker, R. M., Crichton, P. B., Old, D. C., Musser, J. M. & Whittam, T. S. (1990a). Genetic population structure, clonal phylogeny and pathogenicity of *Salmonella paratyphi* B. *Infect Immun* **58**, 1891–1901.

Selander, R. K., Beltran, P., Smith, N. H. & 7 other authors (1990b). Evolutionary genetic relationships of clones of *Salmonella* serovars that cause human typhoid and other enteric fevers. *Infect Immun* **58**, 2262–2275.

Selander, R. K., Beltran, P. & Smith, N. H. (1991). Evolutionary genetics of *Salmonella*. In *Evolution at the Molecular Level*, pp. 25–27. Edited by R. K. Selander, A. G. Clark & T. S. Whittam. Sunderland, MA: Sinauer Associates.

Selander, R. K., Li, J. & Nelson, K. (1996). Evolutionary genetics of *Salmonella enterica*. In *Escherichia coli and Salmonella: Cellular and Molecular Biology*, 2nd edn, pp. 2691–2707. Edited by F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter & H. E. Umbarger. Washington, DC: American Society for Microbiology.

Swofford, D. L. (1998). PAUP – phylogenetic analysis using parsimony. Sunderland, MA: Sinauer Associates.

Torpdahl, M. & Ahrens, P. (2004). Population structure of *Salmonella* investigated by amplified fragment length polymorphism. *J Appl Microbiol* **97**, 566–573.

Vulic, M., Dionisio, F., Taddei, F. & Radman, M. (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.

Wang, F., Whittam, T. & Selander, R. (1997). Evolutionary genetics of the isocitrate dehydrogenase gene (*icd*) in *Escherichia coli* and *Salmonella enterica*. *J Bacteriol* **179**, 6551–6559.

Worth, L., Jr, Clark, S. E., Radman, M. & Modrich, P. (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.