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

Sophie Octavia and Ruiting Lan

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 (*mglA, 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*.

**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.,...
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: mglA (galactoside transport ATP-binding protein MglA), proV (glycine betaine l-proline transport ATP-binding protein), speC (ornithine decarboxylase) and forC (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

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

*Based on MLEE data from Boyd et al. (1993); the value given for allelic difference is between the strain and Typhi Tp2.
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 a 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 two gene trees 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 ($I_D$) (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 2. Genes and primers used in this study

<table>
<thead>
<tr>
<th>Gene</th>
<th>Chromosomal location (bp)*</th>
<th>Direction</th>
<th>Size (bp)</th>
<th>Sequence (5’–3’)</th>
<th>Position†</th>
</tr>
</thead>
<tbody>
<tr>
<td>torC</td>
<td>3 824 187</td>
<td>Forward</td>
<td>18</td>
<td>GGTCAATGTCGGGATTG</td>
<td>60–77</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reverse</td>
<td>18</td>
<td>TGGCTCCAGCGCTCGAT</td>
<td>728–745</td>
</tr>
<tr>
<td>speC</td>
<td>3 129 168</td>
<td>Forward</td>
<td>17</td>
<td>AAAATCGGGCATCTCTG</td>
<td>892–909</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reverse</td>
<td>17</td>
<td>CGGCTGCGGTGATAGCA</td>
<td>1876–1893</td>
</tr>
<tr>
<td>proV</td>
<td>2 808 847</td>
<td>Forward</td>
<td>18</td>
<td>GGCTCGGGTGAAATCCAC</td>
<td>190–207</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reverse</td>
<td>18</td>
<td>TCGTGACACCGCGCAC</td>
<td>1090–1107</td>
</tr>
<tr>
<td>mglA</td>
<td>2 251 323</td>
<td>Forward</td>
<td>20</td>
<td>GTCTTTGCGATTTATCA</td>
<td>173–190</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reverse</td>
<td>18</td>
<td>AATAGCCACGGACAAAT</td>
<td>1229–1246</td>
</tr>
</tbody>
</table>

*According to the Typhi CT18 genome (accession no. NC_003198).
†Relative to the first base of the initiation codon.

### Table 3. Pairwise nucleotide difference

<table>
<thead>
<tr>
<th>Gene</th>
<th>Size (bp)</th>
<th>Pairwise percentage difference</th>
<th>No. polymorphic sites</th>
<th>No. informative sites</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>Minimum</td>
<td>Maximum</td>
</tr>
<tr>
<td>mglA</td>
<td>743</td>
<td>1.35</td>
<td>0.27</td>
<td>2.02</td>
</tr>
<tr>
<td>proV</td>
<td>818</td>
<td>0.78</td>
<td>0.12</td>
<td>1.59</td>
</tr>
<tr>
<td>speC</td>
<td>820</td>
<td>1.27</td>
<td>0.00</td>
<td>2.32</td>
</tr>
<tr>
<td>torC</td>
<td>604</td>
<td>0.83</td>
<td>0.17</td>
<td>1.82</td>
</tr>
<tr>
<td>mdh*</td>
<td>831</td>
<td>0.81</td>
<td>0.00</td>
<td>1.56</td>
</tr>
<tr>
<td>mutS*</td>
<td>1098</td>
<td>0.81</td>
<td>0.09</td>
<td>1.28</td>
</tr>
</tbody>
</table>

*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 mglA. 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 mglA.

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, had almost identical sequences in two genes (proV and mglA), but only some segments similar in three other genes (mglA, 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 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, 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 mglA, proV, speC, torC, mdh and mutS.
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 mglA 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). mglA had the lowest

---

**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.
within-locus compatibility, at 52%, followed by proV (53%), mutS (65%), torC (67%) 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

---

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

<table>
<thead>
<tr>
<th>Gene</th>
<th>α*</th>
<th>Ti/Tv ratio†</th>
<th>−lnL</th>
<th>−lnL (99th)‡</th>
<th>Δ−lnL score of gene tree§</th>
</tr>
</thead>
<tbody>
<tr>
<td>mglA</td>
<td>0.61</td>
<td>0.42</td>
<td>1416</td>
<td>28</td>
<td>46 44 38 45 42</td>
</tr>
<tr>
<td>proV</td>
<td>0.02</td>
<td>0.09</td>
<td>1389</td>
<td>25</td>
<td>40 35 25 25 30</td>
</tr>
<tr>
<td>speC</td>
<td>0.69</td>
<td>0.80</td>
<td>1482</td>
<td>64</td>
<td>110 72 89 44 77</td>
</tr>
<tr>
<td>torC</td>
<td>0.29</td>
<td>0.03</td>
<td>1016</td>
<td>30</td>
<td>48 27 34 28 31</td>
</tr>
<tr>
<td>mdh</td>
<td>0.49</td>
<td>0.63</td>
<td>1386</td>
<td>39</td>
<td>62 36 23 47 — 38</td>
</tr>
<tr>
<td>mutS</td>
<td>0.02</td>
<td>0.74</td>
<td>1876</td>
<td>46</td>
<td>65 33 36 48 34 —</td>
</tr>
</tbody>
</table>

* 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.

---

**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 ± 0.549. 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 4.32 ± 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 Iα 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 (Raysiguier 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

S. Octavia and R. Lan
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


