Phylogenetic relationships and species differentiation of 39 Legionella species by sequence determination of the RNase P RNA gene \textit{rnpB}

Carl-Johan Rubin,\textsuperscript{1} Mikael Thollesson,\textsuperscript{2} Leif A. Kirsebom\textsuperscript{3} and Björn Herrmann\textsuperscript{1}

\textsuperscript{1}Department of Clinical Microbiology, University Hospital, SE-751 85 Uppsala, Sweden  
\textsuperscript{2}Department of Molecular Evolution, EBC, Uppsala University, Norbyvägen 18C, SE-19530 Uppsala, Sweden  
\textsuperscript{3}Department of Cell and Molecular Biology, Box 596, Biomedical Centre, SE-75124 Uppsala, Sweden

The \textit{rnpB} gene is ubiquitous in \textit{Bacteria}, \textit{Archaea} and \textit{Eucarya} and encodes the RNA component of RNase P, an endoribonuclease P that consists of one RNA and one protein subunit (C5). In this study, partial \textit{rnpB} genes were sequenced from 39 type strains and 16 additional strains of the genus \textit{Legionella}. Models of the putative secondary structures of the RNase P RNA in the genus \textit{Legionella} are proposed and possible interactions between RNase P RNA and C5 are discussed. The phylogenetic relationships within the genus \textit{Legionella} were examined and \textit{rnpB} sequences indicated six main clades that together comprised 27 of the 39 species examined. The phylogenetic relationships were further inferred by analysing combined datasets of sequences from the \textit{rnpB}, \textit{mip}, 16S rRNA and \textit{rpoB} genes. It is concluded that \textit{rnpB} is suitable for use in phylogenetic studies of closely related species and that it exhibits the potential to discriminate between \textit{Legionella} species.

\textbf{INTRODUCTION}

The genus \textit{Legionella} comprises more than 50 species (Park et al., 2004) and new species are frequently described (http://www.dsmz.de/bactnom/bactname.htm). Some \textit{Legionella} species cause human disease, while others have been detected only in the environment (Fields et al., 2002). The diseases caused by \textit{Legionella} include the pneumonic form, Legionnaires’ disease, and the extrapulmonary flu-like form, Pontiac fever. \textit{Legionella pneumophila} serogroup (sg) 1 is the causative agent in up to 84\% of disease cases that are due to \textit{Legionella} infection (Yu et al., 2002). There may, however, be a bias towards detecting \textit{L. pneumophila} sg 1, since commonly used urinary antigen tests mainly detect this serogroup (Helbig et al., 2003). Other \textit{L. pneumophila} serogroups, as well as \textit{Legionella longbeachae}, \textit{Legionella micdadei}, \textit{Legionella dumoffii} and \textit{Legionella feeleii}, are also often associated with human disease (O’Connell et al., 1996). In Australia, \textit{L. longbeachae} is the leading cause of Legionnaires’ disease (Alli et al., 2003). Immuno-compromised individuals are especially susceptible to infection and may be colonized by species not normally associated with disease. Therefore, it is important for epidemiological investigations to identify species within the genus accurately.

In clinical diagnostics, \textit{Legionella} bacteria are commonly identified by culture since they have very specific growth requirements. If bacteria grow on cysteine-enriched buffered charcoal-yeast extract (BCYE) agar, but not on BCYE agar without cysteine, the bacteria presumably belong to the genus \textit{Legionella} (Murray et al., 2002). To discriminate between species, phenotypic tests such as growth characteristics, auto-fluorescence and serological methods targeting membrane proteins are often used. These methods provide low resolution and antigen cross-reactivity limits the specificity of antibody tests (Verissimo et al., 1996).
Several DNA-based classification systems have been described for *Legionella*, some of which target the sequence variation of specific genes such as *mip* (Ratcliff et al., 1998), 16S rRNA (Fry et al., 1991), *rpoB* (Ko et al., 2002) and *gyrA* (Feddersen et al., 2000). Analysis of transfer DNA intergenic spacer length polymorphism has been shown to be highly discriminatory in the identification of *Legionella* species (De Gheldre et al., 2001). Amplified fragment length polymorphism (AFLP) and multilocus sequence typing (MLST) are very discriminatory techniques for subtyping within the species *L. pneumophila* (Gaia et al., 2003), where they also distinguish within serogroups (Jonas et al., 2000; Valsangiacomo et al., 1995). Thus, DNA-based analyses have greatly improved *Legionella* phylogenetics and the ability to discriminate between bacteria within the *Legionella* genus.

The *rnpB* gene encodes the catalytic RNA moiety of endoribonuclease P (RNase P), the enzyme that removes the 5'-leader of precursor tRNAs. In *Bacteria* and some *Archaea*, the RNA component alone exhibits catalytic activity *in vitro*; however, the protein subunit C5 is essential for activity *in vivo* and is encoded by the *rnpA* gene (Altman & Kirsebom, 1999; Pannucci et al., 1999). Based on the secondary structure of RNase P RNA, two types of bacterial RNase P RNA have been identified, type A and type B. RNase P RNA derived from *Legionella* belongs to type A, which is suggested to be the ancestral form of RNase P RNA found in

![Fig. 1. Minimum consensus secondary structure of bacterial type A RNase P RNA and deduced secondary structures of RNase P RNA from *Legionella pneumophila* ATCC 33152\(^1\) sg 1, *Legionella moravica* ATCC 42877\(^1\) and *Legionella israelensis* ATCC 43119\(^1\). Lower case letters indicate nucleotides in primer sequences and nucleotides denoted N are tentative nucleotides flanking the primers. Boxed nucleotides show positions proposed to interact with the C5 protein.](image-url)
bacteria. Type B RNase P RNA may have emerged later within the low G+C Gram-positive lineage (Haas & Brown, 1998).

In bacteria, *rnpB* comprises about 400 nt (Pace & Brown, 1995) and several conserved regions essential for the functioning of RNase P RNA have been identified. Certain regions with high variability have also been identified, for example, P3, P12 and P17 (see Fig. 1). The sequence variation of *rnpB* has been used previously to differentiate closely related species of *Chlamydiaceae* (Herrmann et al., 2000) and *Streptococcus* (Tapp et al., 2003).

In this report, we investigate the phylogeny within the genus *Legionella*, discuss the interaction between RNase P RNA and the C5 protein and evaluate *rnpB* as a tool for discriminating between *Legionella* species.

### METHODS

#### Bacterial strains.

The 55 *Legionella* strains (39 type strains and 16 additional reference strains) included in this study are listed in Table 1. Furthermore, 12 clinical and three environmental isolates (1777/97, 1780/97 and 1829/00) are detailed (see Fig. 5).

#### DNA preparation.

DNA for PCR amplification was extracted either by using the Qiagen DNA mini kit or by using a phenol/chloroform-based method.

#### PCR amplification.

The *rnpB* gene was amplified using the primer pair LP3 (5'CA(INOSINE)AGTYGGTCAGGCAAT-3') and BM1-2 (5'TGTTAAACGACCGGCAARTAAGCCGGGTCTCTGT-3'). The reaction mixture consisted of 0.8 μM of each primer, 200 μM dNTPs, 2 mM MgCl₂, 2 μL HotStar Taq DNA polymerase (Qiagen) and 20–100 ng template DNA, as measured from semiquantitative measurement on ethidium bromide-stained agarose gels. The reaction mixture was subjected to enzyme activation for 15 min at 95 °C followed by 37 cycles of amplification. Each amplification cycle consisted of denaturation for 30 s at 95 °C, primer annealing for 40 s at 55 °C and elongation for 40 s at 72 °C. A final step of 7 min at 72 °C was performed to ensure complete extension.

The 16S rRNA and *mip* genes of the environmental isolate 1829/00 were sequenced as described elsewhere (Johansson et al., 1995; Ratcliff et al., 1998).

#### Sequence analyses.

Initially, partial *rnpB* gene sequences were PCR-amplified with primers binding to the conserved regions defined by P4. The products generated by these primers, however, did not include the hypervariable P3 loop.

To facilitate the design of the LP3 primer, the sequence between the promoter region and the P3 loop was determined for ten *Legionella* species by amplifying the 5’-flanking DNA of the P4 loop in a PCR with biotinylated primers and streptavidin-coated beads, as described elsewhere (Sorensen et al., 1993). Sequencing of PCR products defined by LP3 and BM1-2 was performed on both DNA strands of all strains using polymer POP6 in an ABI 310 Genetic Analyser (Applied Biosystems). BM1-2 and LP3 were also used as primers in the sequencing PCR, in which the BigDye terminator-labelled cycle sequencing chemistry kit version 2.0 (Applied Biosystems) was used.

#### Phylogenetic analysis.

The *rnpB* gene sequences were aligned using CLUSTAL W, but required some manual editing to align homologous sites according to the secondary structures of RNase P RNA. The 16S rRNA, *mip* and *rpoB* gene sequences were obtained from GenBank and their accession numbers are presented in Table 1. The 16S rRNA and *mip* gene sequences varied in size, and sequences from some species were therefore trimmed at the ends to generate sequences with homologous sites in the CLUSTAL W alignments. The *rnpB* fragments included were 304–354 nt long. 16S rRNA genes were approximately 1350 nt in length. All *rpoB* fragments were 300 nt long and approximately 650 nt were included from *mip* sequences, including the hypervariable insert adjacent to the signal sequence. The combined dataset, consisting of *rnpB*, 16S rRNA, *mip* and *rpoB* gene sequences, was constructed by concatenating the alignments for the individual genes and included 39 *Legionella* species, of which 37 had sequences available for all four genes.

The alignments obtained (individual genes and concatenated alignments) were used for phylogenetic inference using a Bayesian approach as implemented in MrBayes 3.0B4 (Huelsenbeck & Ronquist, 2001). MrBayes uses Metropolis-coupled Markov chain Monte Carlo methods to calculate the posterior probabilities for the parameters of interest. Each analysis was run for 1 x 10⁷ generations with four differentially heated chains; generations before convergence (as monitored on tree likelihood and total tree length) were discarded as burn-in.

To select an adequate model for the Bayesian analysis, as well as pairwise distances, we used a hierarchical likelihood ratio test approach (Huelsenbeck & Crandall, 1997). To do this, we used *PAUP* + 4.0b8-10 Linux and Macintosh versions (Swofford, 2000) and the same test hierarchy (and thus model selection) as implemented in the programs Posada & Crandall, 1998 at P<0.01. Neighbour-joining (NJ) trees under the Jukes–Cantor model were produced for each of the datasets/partitions separately as well as in combination and the parameters for each model were estimated using these trees. The same model was used for the pairwise distances (e.g. gamma shape, proportion invariant sites) and the parameters assigned were based on the maximum-likelihood estimate.

In addition to the Bayesian analysis, we performed bootstrap analyses as implemented in *PAUP* + 4.0b8-10, using maximum parsimony (MP) as optimality criterion and also using NJ. For the optimality criterion method, 1000 bootstrap replicates were performed and heuristic search algorithms were used, namely simple stepwise addition and tree bisection reconnection branch swapping.

To compare the variation in the different genes, we used the Shannon–Wiener information index, *H*, defined as

\[
H = - \sum_{i=1}^{4} p_i \log p_i
\]

where \(p_i\) is the proportion of A, T, C and G (Shannon & Weaver, 1949; Wiener, 1949). The mean value for all sites in the same set of taxa was calculated for each dataset.

### RESULTS AND DISCUSSION

#### Sequence comparison of *rnpB* in the genus *Legionella*

PCR amplicons between 304 and 354 bp (primers excluded) were obtained, constituting 89–91% of the total *rnpB* gene sequences. Partial *rnpB* gene sequences of 39 type strains were compared and the pairwise similarities ranged between 75% and 100%, with *Legionella micdadei* ATCC 33218 and *Legionella maccachernii* ATCC 35300 as the only species having identical *rnpB* gene sequences. These two species are difficult to distinguish from each other phenotypically and...
<table>
<thead>
<tr>
<th>Strain</th>
<th>Serogroup</th>
<th>GenBank accession no.</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Coxiella burnetii</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L. adelaidensis ATCC 49625&lt;sup&gt;T&lt;/sup&gt; = CCUG 31231 A&lt;sup&gt;T&lt;/sup&gt;</td>
<td></td>
<td>AE016960</td>
<td>GenBank</td>
</tr>
<tr>
<td>L. anisa ATCC 35292&lt;sup&gt;T&lt;/sup&gt; = CCUG 29665&lt;sup&gt;T&lt;/sup&gt;</td>
<td></td>
<td>AJ781468</td>
<td>SMI</td>
</tr>
<tr>
<td>L. birminghamensis ATCC 43702&lt;sup&gt;T&lt;/sup&gt; = CCUG 31233 A&lt;sup&gt;T&lt;/sup&gt;</td>
<td></td>
<td>Z49717</td>
<td>SMI</td>
</tr>
<tr>
<td>L. bozemanae ATCC 33217&lt;sup&gt;T&lt;/sup&gt; = CCUG 11880&lt;sup&gt;T&lt;/sup&gt;</td>
<td>1</td>
<td>Z49719</td>
<td>SMI</td>
</tr>
<tr>
<td>L. bozemanae ATCC 35545&lt;sup&gt;T&lt;/sup&gt; = CCUG 16416</td>
<td>2</td>
<td>Z49718</td>
<td>SMI</td>
</tr>
<tr>
<td>L. brunensis ATCC 43878&lt;sup&gt;T&lt;/sup&gt; = CCUG 31114&lt;sup&gt;T&lt;/sup&gt;</td>
<td></td>
<td>Z49718</td>
<td>SMI</td>
</tr>
<tr>
<td>L. cherriei ATCC 35252&lt;sup&gt;T&lt;/sup&gt; = CCUG 29665&lt;sup&gt;T&lt;/sup&gt;</td>
<td></td>
<td>Z49720</td>
<td>SMI</td>
</tr>
<tr>
<td>L. cincinnatiensis ATCC 43753&lt;sup&gt;T&lt;/sup&gt; = CCUG 31230 A&lt;sup&gt;T&lt;/sup&gt;</td>
<td></td>
<td>Z49721</td>
<td>SMI</td>
</tr>
<tr>
<td>L. dimitoffii ATCC 33279&lt;sup&gt;T&lt;/sup&gt; = CCUG 11881&lt;sup&gt;T&lt;/sup&gt;</td>
<td></td>
<td>Z49722</td>
<td>SMI</td>
</tr>
<tr>
<td>L. erythra ATCC 35303&lt;sup&gt;T&lt;/sup&gt; = CCUG 29667&lt;sup&gt;T&lt;/sup&gt;</td>
<td></td>
<td>Z32636</td>
<td>CCUG</td>
</tr>
<tr>
<td>L. fairfieldensis ATCC 49588&lt;sup&gt;T&lt;/sup&gt; = CCUG 31235 A&lt;sup&gt;T&lt;/sup&gt;</td>
<td></td>
<td>Z32639</td>
<td>SMI</td>
</tr>
<tr>
<td>L. fellutii ATCC 35072&lt;sup&gt;T&lt;/sup&gt; = CCUG 16417&lt;sup&gt;T&lt;/sup&gt;</td>
<td>1</td>
<td>Z32639</td>
<td>SMI</td>
</tr>
<tr>
<td>L. geestiana ATCC 49504&lt;sup&gt;T&lt;/sup&gt; = CCUG 44893&lt;sup&gt;T&lt;/sup&gt;</td>
<td></td>
<td>Z32639</td>
<td>SMI</td>
</tr>
<tr>
<td>L. gormanii ATCC 33297&lt;sup&gt;T&lt;/sup&gt; = CCUG 12267&lt;sup&gt;T&lt;/sup&gt;</td>
<td></td>
<td>Z32639</td>
<td>SMI</td>
</tr>
<tr>
<td>L. grattiana ATCC 49413&lt;sup&gt;T&lt;/sup&gt; = CCUG 44894&lt;sup&gt;T&lt;/sup&gt;</td>
<td></td>
<td>Z49725</td>
<td>SMI</td>
</tr>
<tr>
<td>L. hackelae ATCC 35250&lt;sup&gt;T&lt;/sup&gt; = CCUG 31232 A&lt;sup&gt;T&lt;/sup&gt;</td>
<td></td>
<td>Z49725</td>
<td>SMI</td>
</tr>
<tr>
<td>L. israelensis ATCC 43119&lt;sup&gt;T&lt;/sup&gt; = CCUG 31115&lt;sup&gt;T&lt;/sup&gt;</td>
<td></td>
<td>Z32640</td>
<td>SMI</td>
</tr>
<tr>
<td>L. jametowniensis ATCC 35298&lt;sup&gt;T&lt;/sup&gt; = CCUG 29669&lt;sup&gt;T&lt;/sup&gt;</td>
<td></td>
<td>Z49726</td>
<td>SMI</td>
</tr>
<tr>
<td>L. jordani ATCC 33623&lt;sup&gt;T&lt;/sup&gt; = CCUG 16413&lt;sup&gt;T&lt;/sup&gt;</td>
<td></td>
<td>Z49726</td>
<td>SMI</td>
</tr>
<tr>
<td>L. lansingensis ATCC 49751&lt;sup&gt;T&lt;/sup&gt; = CCUG 31227 B&lt;sup&gt;T&lt;/sup&gt;</td>
<td></td>
<td>Z49727</td>
<td>SMI</td>
</tr>
<tr>
<td>L. londinensis ATCC 49505&lt;sup&gt;T&lt;/sup&gt; = CCUG 44895&lt;sup&gt;T&lt;/sup&gt;</td>
<td></td>
<td>Z49728</td>
<td>SMI</td>
</tr>
<tr>
<td>L. longbeachae ATCC 33462&lt;sup&gt;T&lt;/sup&gt; = CCUG 46622&lt;sup&gt;T&lt;/sup&gt;</td>
<td>1</td>
<td>Z49728</td>
<td>SMI</td>
</tr>
<tr>
<td>L. longbeachae ATCC 33484 = CCUG 46623</td>
<td>2</td>
<td>Z49728</td>
<td>SMI</td>
</tr>
<tr>
<td>L. macachernii ATCC 35300&lt;sup&gt;T&lt;/sup&gt; = CCUG 31116&lt;sup&gt;T&lt;/sup&gt;</td>
<td></td>
<td>X60081</td>
<td>SMI</td>
</tr>
<tr>
<td>L. micdadei ATCC 33218&lt;sup&gt;T&lt;/sup&gt; = CCUG 31229 A&lt;sup&gt;T&lt;/sup&gt;</td>
<td>1 subsp. pneumophila</td>
<td>X73395</td>
<td>SMI</td>
</tr>
<tr>
<td>L. moravaica ATCC 43877&lt;sup&gt;T&lt;/sup&gt; = CCUG 31117&lt;sup&gt;T&lt;/sup&gt;</td>
<td>2 subsp. pneumophila</td>
<td>X227162</td>
<td>SMI</td>
</tr>
<tr>
<td>L. nautarum ATCC 49506&lt;sup&gt;T&lt;/sup&gt; = CCUG 44900&lt;sup&gt;T&lt;/sup&gt;</td>
<td>3 subsp. pneumophila</td>
<td>Z32639</td>
<td>SMI</td>
</tr>
<tr>
<td>L. oakridgensis ATCC 33761&lt;sup&gt;T&lt;/sup&gt; = CCUG 16414&lt;sup&gt;T&lt;/sup&gt;</td>
<td>4 subsp. fraseri</td>
<td>X73397</td>
<td>SMI</td>
</tr>
<tr>
<td>L. parisiensis ATCC 35299&lt;sup&gt;T&lt;/sup&gt; = CCUG 29670&lt;sup&gt;T&lt;/sup&gt;</td>
<td>5 subsp. pneumophila</td>
<td>X227162</td>
<td>SMI</td>
</tr>
<tr>
<td>L. pneumophila ATCC 33152&lt;sup&gt;T&lt;/sup&gt; = CCUG 9568&lt;sup&gt;T&lt;/sup&gt;</td>
<td>6 subsp. pneumophila</td>
<td>X227162</td>
<td>SMI</td>
</tr>
<tr>
<td>L. pneumophila ATCC 33154 = CCUG 13396</td>
<td>7 subsp. pneumophila</td>
<td>X227162</td>
<td>SMI</td>
</tr>
<tr>
<td>L. pneumophila ATCC 33155 = CCUG 13397</td>
<td>8 subsp. pneumophila</td>
<td>X227162</td>
<td>SMI</td>
</tr>
<tr>
<td>L. pneumophila ATCC 33156&lt;sup&gt;T&lt;/sup&gt; = CCUG 13398&lt;sup&gt;T&lt;/sup&gt;</td>
<td>9 subsp. pneumophila</td>
<td>X227162</td>
<td>SMI</td>
</tr>
<tr>
<td>L. pneumophila ATCC 33215</td>
<td>6 subsp. pneumophila</td>
<td>X227162</td>
<td>SMI</td>
</tr>
<tr>
<td>L. pneumophila ATCC 33216 = CCUG 13399</td>
<td>5 subsp. fraseri</td>
<td>X227162</td>
<td>SMI</td>
</tr>
<tr>
<td>L. pneumophila ATCC 33823 = CCUG 16411</td>
<td>7 subsp. pneumophila</td>
<td>X227162</td>
<td>SMI</td>
</tr>
<tr>
<td>L. pneumophila ATCC 35096 = CCUG 16412</td>
<td>8 subsp. pneumophila</td>
<td>X227162</td>
<td>SMI</td>
</tr>
<tr>
<td>L. pneumophila ATCC 35251 = CCUG 30657 A&lt;sup&gt;T&lt;/sup&gt;</td>
<td>9 subsp. pneumophila</td>
<td>X227162</td>
<td>SMI</td>
</tr>
<tr>
<td>L. pneumophila ATCC 35289 = CCUG 30704</td>
<td>11 subsp. pneumophila</td>
<td>X227162</td>
<td>SMI</td>
</tr>
<tr>
<td>L. pneumophila ATCC 43130 = CCUG 30660</td>
<td>10 subsp. pneumophila</td>
<td>X227162</td>
<td>SMI</td>
</tr>
<tr>
<td>L. pneumophila ATCC 43283 = CCUG 30705</td>
<td>12 subsp. pneumophila</td>
<td>X227162</td>
<td>SMI</td>
</tr>
<tr>
<td>L. pneumophila ATCC 43290 = CCUG 30661</td>
<td>14 subsp. pneumophila</td>
<td>X227162</td>
<td>SMI</td>
</tr>
<tr>
<td>L. pneumophila ATCC 43703 = CCUG 44898</td>
<td>13 subsp. pneumophila</td>
<td>X227162</td>
<td>SMI</td>
</tr>
<tr>
<td>L. pneumophila ATCC 43736 = CCUG 30662</td>
<td></td>
<td>X227162</td>
<td>SMI</td>
</tr>
<tr>
<td>L. quaierezensis ATCC 49507&lt;sup&gt;T&lt;/sup&gt; = CCUG 44899&lt;sup&gt;T&lt;/sup&gt;</td>
<td></td>
<td>Z49732</td>
<td>SMI</td>
</tr>
<tr>
<td>L. quinlivanii ATCC 43830&lt;sup&gt;T&lt;/sup&gt; = CCUG 31234 A&lt;sup&gt;T&lt;/sup&gt;</td>
<td></td>
<td>Z49733</td>
<td>SMI</td>
</tr>
<tr>
<td>L. rubrulucens ATCC 35304&lt;sup&gt;T&lt;/sup&gt; = CCUG 29671&lt;sup&gt;T&lt;/sup&gt;</td>
<td>1</td>
<td>Z32643</td>
<td>SMI</td>
</tr>
<tr>
<td>L. saithelensi ATCC 35248&lt;sup&gt;T&lt;/sup&gt; = CCUG 29672&lt;sup&gt;T&lt;/sup&gt;</td>
<td>1</td>
<td>Z49734</td>
<td>SMI</td>
</tr>
<tr>
<td>L. sanitcricus ATCC 35301&lt;sup&gt;T&lt;/sup&gt; = CCUG 29673&lt;sup&gt;T&lt;/sup&gt;</td>
<td>1</td>
<td>Z49735</td>
<td>SMI</td>
</tr>
</tbody>
</table>
were previously referred to as the Tatlockia species (Fox et al., 1991); they are however distinguishable using mip gene sequencing.

Differentiation of Legionella bozemanae and Legionella anisa has been proven difficult by transfer DNA intergenic spacer length polymorphisms (De Gheldre et al., 2001). A close species relationship was found when comparing rnpB gene sequences derived from these two Legionella species. In sequence analysis, L. anisa ATCC 35292\(^T\) differs in four nucleotide positions compared with L. bozemanae ATCC 35545 sg 2, while the corresponding differences in comparisons between L. anisa ATCC 35292\(^T\) and L. bozemanae ATCC 33217\(^T\) sg 1, as well as between the two serogroups of L. bozemanae, were seven nucleotide positions. The finding that rnpB gene sequences from L. bozemanae sg 2 and L. anisa are more similar than those from the two serogroups of L. bozemanae was surprising since sequence data from previous studies examining other genes, such as rpoB, have indicated a more distant relationship between the two species (Ko et al., 2002; Ratcliff et al., 1998).

The L. pneumophila serogroups 4, 5 and 15, together constituting L. pneumophila subsp. fraseri, differed in 14–15 nt from the rnpB gene sequence of L. pneumophila ATCC 33152\(^T\) sg 1, while the sequence variation within subspecies fraseri was 1 nt. Similarly, the sequence variation within the subspecies of L. pneumophila subsp. pneumophila varied between 0 and 4 nt.

### Secondary structures and intermolecular interactions of the RNase P RNA

The RNase P RNA gene has been sequenced and characterized from a large number of different bacteria (Brown, 1999). This information has been incorporated into models of the three-dimensional structures of both types of RNase P RNA, as well as in a complex with the protein subunit (Massire et al., 1998; Tsai et al., 2003). The majority of these previous studies on bacterial RNase P RNA have, however, focused on structural and functional differences between members of different bacterial genera (Haas & Brown, 1998). This study examined 39 species within one genus.

With the guidance of a type A RNase P RNA minimum consensus structure model (Brown, 1999), secondary structures of RNase P RNAs from three Legionella type strains were generated. As shown in Fig. 1, the secondary structures are very similar to those of some \(\gamma\)-Proteobacteria, e.g. Pseudomonas aeruginosa and Escherichia coli. Within the Legionella genus, RNase P RNA showed structural variations, particularly in P3 and P12 but also in the structural element between P16 and P17.

Two P3 variants were observed where the major variant was an 8 bp stem structure with a loop varying in size between three and six residues. In one species, Legionella israelensis ATCC 43119\(^T\), the sequence suggested the existence of a P3 helix composed of 20 bp with a U-rich internal bulge and a loop with eight residues (Fig. 1). Moreover, we noted that the first 6 bp of P3 are highly conserved irrespective of Legionella species (boxed residues in Fig. 1; see also sequence alignment data in Supplementary Material in IJSEM Online). Likewise, comparison of RNase P RNA structures derived from closely related bacterial species, e.g. Mycoplasma species (Svard et al., 1994), also reveals that the length of P3 varies and that residues in the boxed region (Fig. 1) in P3 are conserved. In contrast, a comparison of the length and sequence variation of the P3 domain derived from different bacterial branches indicates that the structure of the P3 is highly variable (Haas & Brown, 1998). The three-dimensional structural model of E. coli RNase P RNA (type A) in complex with the RNase P protein (C5) suggests that part of P3 interacts with the C5 protein (Tsai et al., 2003). In this model, the nucleotides in P3 that are suggested to interact with C5 are positioned closer to the P3 loop (the size of P3 in E. coli RNase P RNA is similar to that of Legionella ATCC 43119\(^T\)). However, if P3 serves as a binding region for C5, then the available structural data (i.e. comparison of the P3 structure derived from closely related bacterial species, see above) suggest that the interaction between P3 and C5 should not depend on the length of P3. We therefore argue that it is possible that C5 binds to a specific sequence motif in P3. We suggest that it is likely that the ‘boxed’ residues in Fig. 1 serve as a binding motif for the C5 protein. The extra P3 residues in L. israelensis ATCC 43119\(^T\) (and in other RNase P RNAs with long P3 helices)

---

**Table 1.** cont.

<table>
<thead>
<tr>
<th>Strain</th>
<th>Serogroup</th>
<th>GenBank accession no.</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>L. shakespearei ATCC 49655(^T) = CCUG 31228 A(^T)</td>
<td></td>
<td>AJ781435</td>
<td>Z49736</td>
</tr>
<tr>
<td>L. spiritensis ATCC 35249(^T) = CCUG 31118(^T)</td>
<td>1</td>
<td>AJ781434</td>
<td>M36030</td>
</tr>
<tr>
<td>L. steigerwaltii ATCC 35302(^T) = CCUG 29674(^T)</td>
<td></td>
<td>AJ781433</td>
<td>Z49737</td>
</tr>
<tr>
<td>L. tucsonensis ATCC 49180(^T) = CCUG 31119(^T)</td>
<td></td>
<td>AJ781432</td>
<td>Z32644</td>
</tr>
<tr>
<td>L. wadsworthi ATCC 33877(^T) = CCUG 16413(^T)</td>
<td></td>
<td>AJ781431</td>
<td>Z49738</td>
</tr>
<tr>
<td>L. worsleienis ATCC 49508(^T) = CCUG 44924(^T)</td>
<td></td>
<td>AJ781430</td>
<td>Z49739</td>
</tr>
</tbody>
</table>
might be important for structural stabilization and/or be involved in interactions with other factors in the cell. However, this remains to be investigated.

In *Bacillus subtilis*, RNase P system chemical cleavage footprinting (Loria et al., 1998) and nucleotide analogue modification interference studies (NAIM; Rox et al., 2002) suggest that P12 and the C5 protein are in close proximity. Given that the RNase P holoenzyme consists of one RNA subunit and one protein subunit, neither of the two three-dimensional models are consistent with an interaction of C5 with both P12 and P3 (Chen et al., 1998; Tsai et al., 2003 and references therein). However, Fang et al. (2001) observed that the *B. subtilis* RNase P holoenzyme formed tetramers in solution that might rationalize the footprinting and NAIM data. For *E. coli* RNase P holoenzyme, no tetramers in solution have so far been observed (Fang et al., 2001). Nevertheless, cross-linking between the RNA and the protein subunits in *E. coli* RNase P holoenzyme has been observed (Sharkady & Nolan, 2001). The residues at positions 144 and 145 in *L. pneumophila* sg1 correspond to the nucleotides in *E. coli* RNase P RNA that cross-linked to the C5 protein. In this study, we observed major structural variation in P12 in RNase P RNA derived from a large number of *Legionella* species with respect to size (i.e. number of base pairs), loop size and the number of bulges and their position in P12 (Fig. 1). Taken together, if P12 constitutes a binding site for the RNase P protein, it might be reoriented as proposed elsewhere (Sharkady & Nolan, 2001) or perhaps these data might reflect flexibility in the structure of RNase P RNA. Moreover, there is a possibility of identifying C5 amino acid residues that bind to the RNA in P12 by characterizing the *rnpA* gene from the corresponding *Legionella* species and looking for co-variation.

Another interesting structural feature is the suggested stem–loop structure between P16 and P17. This structural element has previously been described in RNase P RNA derived from *Planctomycetes* and α-Proteobacteria (Brown, 1999), but it has not been observed in RNase P RNA from γ-Proteobacteria. Although the function of this stem–loop is not known, it is located very close to the domain of RNase P.
RNA that interacts with the 3’ RCCA motif of the precursor substrates (interacting residues at position 284:G and 285:G in L. pneumophila ATCC 33152 T sg 1 in Fig. 1) (Kirsebom & Svard, 1994). In the structural model of the RNase P holoenzyme, the RNase P protein is positioned close to the P15/16 region and therefore raises the possibility that the stem–loop structure between P16 and P17 functions as an anchoring site for the RNase P protein. However, its position in the three-dimensional structure cannot be such that it interferes with the binding of the substrate.

**Phylogenetic analysis**

Site-to-site rate variation was modelled separately for each of the genes rnpB, 16S rRNA, rpoB and mip. For the rnpB and rpoB genes, the most adequate model was TrNef + I + Γ, i.e. a Tamura Nei model (Tamura & Nei, 1993), with equal base frequencies and invariant sites plus a gamma rate distribution. For the mip gene, the GTR + I + Γ model, i.e. General Time Reversible (Lanave et al., 1984) with invariant sites and a gamma rate distribution (Yang, 1994), was shown to be the most suitable. For the 16S rRNA gene, the most appropriate model was found to be HKY + I + Γ (Hasegawa et al., 1985). In the Bayesian combined analysis, each gene was assigned its corresponding optimal model with independent parameters and GTR was used in place of TrNef (due to limitation in MrBayes).

The MP, NJ and Bayesian analyses of each gene separately, and for the four genes combined, resulted in similar topologies, although differences were observed. Tree topologies showed that the number of conflicts, above a Bayesian posterior probability of 95, was highest for the comparison between rpoB and mip (eight conflicts), followed by the 16S rRNA gene and rpoB (five conflicts), rnpB and mip (four conflicts) and rpoB and rnpB (three conflicts). There were two conflicts between the 16S rRNA gene sequence and mip and between 16S rRNA and rnpB.

![Fig. 3. Majority rule consensus tree from the Bayesian analysis of the combined data-set comprising rnpB, mip, 16S rRNA and rpoB gene sequences (posterior probabilities >0.9). Numbers below branches indicate bootstrap values >50% obtained by NJ and by parsimony analysis (in parentheses).](image-url)
Node support (posterior probabilities for Bayesian analyses and bootstrap proportions for MP and NJ) was used to evaluate the phylogenetic utility of the information embedded in the gene sequences. The majority rule consensus trees from the Bayesian analyses of *rnpB* and 16S rRNA gene sequences are shown in Fig. 2. Branches with Bayesian posterior probabilities of <0.90 are collapsed and the numbers below the branches indicate the score obtained by bootstrap analysis using NJ and parsimony algorithms.

Bayesian analysis of 16S rRNA gene sequences from 150 different species representing approximately 55 genera showed that the *Legionella* species examined form a well-supported clade and thus constitute a monophyletic group (data not presented).

The phylogenetic analysis of combined datasets (one comprising *rnpB* together with *mip* and the other with all four genes) included 39 *Legionella* species, of which 37 had sequences available for all four genes. The phylogenetic analysis of the four genes together (Fig. 3) and *rnpB* together with *mip* (not presented) revealed very similar branching with 18 and 17 well-supported nodes, respectively, i.e. with posterior probabilities of ≥0.95. The numbers of well-supported nodes in the analyses of each gene sequence alone were 13 for *mip*, 12 for 16S rRNA, 11 for *rnpB* and five for *rpoB*. The well-supported nodes together comprised 29 species in the analysis of all four genes, 28 species for *rnpB* together with *mip*, 25 species for *rnpB*, 22 species for *mip*, 19 species for 16S rRNA and 11 species for *rpoB*. Interestingly, the well-supported nodes derived from *rnpB* and *mip* gene sequences comprised more species than the well-supported nodes from 16S rRNA and *rpoB* gene sequences and therefore contributed most of the phylogenetic signal in the dataset including all four genes.

In our analysis of *rpoB*, we found five fewer nodes of substantial support compared with a previous study (Ko et al., 2002; data not shown). In their presented NJ trees, there were several nodes that were in conflict compared with an analysis of *mip* gene sequences. In contrast, such conflicts were almost absent in our comparison between *rpoB* and the other gene sequences. Thus *rpoB* did not improve the resolution of the phylogenetic tree regarding species, but showed six nodes with significant support within the *L. pneumophila* species. This is more than was found for the *rnpB* and *mip* phylogeny, where only four nodes were observed. However, the *rpoB* tree is...
The results from the Shannon–Wiener analysis are presented in Fig. 4(a) as the SW index for each nucleotide position along the genes. The protein coding genes mip and rpoB have a regular high variation, mainly in nucleotides corresponding to the third position of codons, while the rnpB and 16S rRNA genes have conserved domains interspersed with variable and hypervariable regions. The proportions of nucleotide positions with certain information content are shown in Fig. 4(b) and the histograms show that the 16S rRNA gene has a lower fraction of nucleotide positions with high information content compared with rnpB.

Differentiation of Legionella species

In rnpB, the nucleotide positions with high information content are located in certain loops surrounded by highly conserved regions, a trait that makes rnpB a promising candidate for the differentiation of bacterial species. A minor evaluation of identification based on rnpB sequence was performed on 15 non-L. pneumophila clinical and environmental isolates (Fig. 5). In 12 cases, rnpB-based identification was in agreement with previous phenotypic determinations in a reference laboratory. Two isolates originally identified as L. micdadei had rnpB genes identical to those from L. micdadei ATCC 33218 and L. maceachernii ATCC 35300 and could not be further typed by using rnpB. For the three isolates with discrepant results, two showed rnpB sequences that were identical to type strains other than those indicated by phenotypic identification. The third isolate, 1829/00, was identified as Legionella gormanii at the reference laboratory from which it was obtained. However, here we show that the rnpB of this isolate had ten discrepant nucleotide positions compared with the rnpB of L. gormanii ATCC 33297. The 16S rRNA gene sequence was 98.8% similar to that of L. gormanii ATCC 33297 and sequencing of mip generated a sequence identical to that of the type strain, indicating that the correct identity of the isolate was L. gormanii.

These data suggest that the rnpB gene may be useful for the identification of most Legionella species. However, further investigation on intraspecies sequence heterogeneity is required. The 16S rRNA gene sequence is commonly used for the differentiation of closely related species. It is well-characterized for a wide variety of organisms and displays enough variable and conserved regions in order to be useful. However, 16S rRNA genes can occur as multiple heterogeneous copies in the genome (Clayton et al., 1995; Nubel et al., 1996; Ueda et al., 1999; Wang & Wang, 1997) and this can lead to misidentification, due to chimerical sequences, if this heterogeneity is not considered in the design of the assay for genotypic identification. Furthermore, recombination in 16S rRNA genes has been reported as a potential cause of erroneous species identification (Schouls et al., 2003). The 16S rRNA gene is quite large and consists of approximately 1500 bp and may therefore be laborious to sequence when full-length sequencing is necessary. The mip and rpoB genes are protein-encoding and thus the nucleotide at the third position of each codon is often variable (Fig. 4b). This results in the absence of highly conserved regions and therefore hampers the design of primers in assays for genotypic identification.

For the Legionella genus, sequencing of rnpB in combination with other genes can contribute to improved species identification and could be used in MLST. Characterization of rnpB from type strains of all known members of the genus Legionella will further determine the potential of rnpB for species identification. When compared with other genes commonly used for the discrimination of bacteria, the advantages of rnpB lie in the combination of conserved and highly variable sequence regions. Furthermore, rnpB is a
single copy gene which has been shown to be useful in the
differentiation of closely related bacteria (Herrmann et al.,
2000; Svard et al., 1994; Tapp et al., 2003).

To conclude, the sequence variation of rnpB separated 37 of
39 Legionella species included in this study. The variation in
putative secondary structures provides the possibility of
understanding the interaction between RNase P RNA and
the C5 protein. Phylogenetic analysis showed that rnpB
clustered more species in nodes with strong branch support
than did any of the other three genes, and when combined
with mip the phylogenetic utility was further enhanced.
Thus, the combination of strictly conserved and hypervari-
able elements suggests rnpB as useful for species identifica-
tion and for the phylogenetic analysis of closely related
species.

REFERENCES

Comparative assessment of virulence traits in Legionella spp.
Microbiology 149, 631–641.


Brenner, D. J., Steigerwald, A. G., Eppele, P. & 7 other authors
(1988). Legionella pneumophila serogroup Lansing 3 isolated from a
patient with fatal pneumonia, and descriptions of L. pneumophila
subsp. pneumophila subsp. nov., L. pneumophila subsp. frearsi subsp.
nov., and L. pneumophila subsp. pascualei subsp. nov. J Clin Microbiol
26, 1695–1703.

27, 314.

Comparative photocross-linking analysis of the tertiary structures of
Escherichia coli and Bacillus subtilis RNase P RNAs. EMBO J 17,
1515–1525.

(1995). Intraspecific variation in small-subunit rRNA sequences in
prokaryotic taxa. International Journal of Systematic and Evolutionary
Microbiology 45, 595–599.

De Gheldere, Y., Maes, N., Presti, F. L., Etienne, J. & Struelens, M.
(2001). Rapid identification of clinically relevant Legionella spp. by
analysis of transfer DNA intergenic spacer length polymorphism.

Fang, X. W., Yang, X. J., Littrell, K., Niranjanakumari, S.,
The Bacillus subtilis RNase P holoenzyme contains two RNase P

Immunol 189, 7–11.

Fields, B. S., Benson, R. F. & Besser, R. E. (2002). Legionella and
Legionnaires’ disease: 25 years of investigation. Clin Microbiol Rev
15, 506–526.

genetically and chemically distinct group of bacteria. Proposal to
transfer Legionella maceachernii (Brenner et al.) to the genus
Tatlockia, as Tatlockia maceachernii comb. nov. Syst Appl Microbiol
14, 52–56.

use of 16S ribosomal RNA analyses to investigate the phylogeny of
the family Legionellaceae. J Gen Microbiol 137, 1215–1222.

Gaia, V., Fry, N. K., Harrison, T. G. & Peduzzi, R. (2003). Sequence-
typing of Legionella pneumophila serogroup 1 offers the potential
for true portability in legionellosis outbreak investigation. J Clin Microbiol
41, 2932–2939.


Hasegawa, M., Kishino, H. & Yano, T. (1985). Dating of the human-
ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol

Heilig, J. H., Uldum, S. A., Bernander, S., Luck, P. C., Wewalka, G.,
Abraham, B., Gaia, V. & Harrison, T. G. (2003). Clinical utility of
urinary antigen detection for diagnosis of community-acquired,
travel-associated, and nosocomial legionnaires’ disease. J Clin Microbiol
41, 838–840.

Herrmann, B., Pettersson, B., Everett, K. D., Mikkelsen, N. E. &
RNase P RNA in the order Chlamydiales. Int J Syst Evol Microbiol 50,
149–158.

and hypothesis testing using maximum likelihood. Annu Rev Ecol
Sys 28, 437–466.


Johansson, K. E., Pettersson, B., Uhlen, M., Gunnarsson, A.,
agent of granulocytic ehrlichiosis in Swedish dogs and horses
by direct solid phase sequencing of PCR products from the 16S rRNA

Jonas, D., Meyer, H. G., Matthes, P., Hartung, D., Jahn, B., Daschner,
different genotyping methods for investigation of nosocomial
outbreaks of Legionnaires’ disease in hospitals. J Clin Microbiol 38,
2284–2291.

Escherichia coli RNase P RNA and its substrate. EMBO J 13,
4870–4876.

beta-subunit gene (rpoB) sequences for the molecular differentiation

method for calculating evolutionary substitution rates. J Mol Evol 20,
86–93.

Recognition of a pre-trNA substrate by the Bacillus subtilis RNase P

Massire, C., Jaeger, L. & Westhof, E. (1998). Derivation of the three-
dimensional architecture of bacterial ribonuclease P RNAs from

Medical Microbiology. St Louis, MI: Mosby.

Nebel, U., Engelen, B., Felske, A., Snaidr, J., Wieshuber, A.,
heterogeneities of genes encoding 16S rRNAs in Paenibacillus
polymyxa detected by temperature gradient gel electrophoresis.

macrophage-like cells by Legionella species that have not been


