DNA sequence variability at the rplX locus of *Bacillus subtilis*

PAUL M. SHARP,* NIAMH C. NOLAN, NIAMH NI CHOLMAIN and KEVIN M. DEVINE

Department of Genetics, Trinity College, Dublin 2, Ireland

(Received 4 July 1991; revised 12 September 1991; accepted 20 September 1991)

The pattern and extent of DNA sequence variability at the rplX locus (encoding ribosomal protein L24) has been investigated in nine strains of *Bacillus subtilis*. Overall, there is a very low level of nucleotide diversity, even at silent sites, which is probably due to selection among synonymous codons. By analogy with *Escherichia coli*, there may also be some effect of the relative proximity of rplX to the chromosomal origin of replication. The small number of nucleotide substitutions are non-randomly distributed: all of the synonymous changes are in valine codons. From the sequence differences the strains can be divided into two groups, which are not coincident with their previous classification; this observation is consistent with recombination among strains.

Introduction

An understanding of the population structure and diversity of bacterial species is of obvious medical, economic and environmental relevance, but so far prokaryotes (and microbes in general) have received much less attention than animal and plant species. Multilocus enzyme electrophoresis (MLEE) has been applied to a number of bacteria, with the results generally suggesting that prokaryotes have clonal population structures (Selander *et al.*, 1985, 1987; Piffaretti *et al.*, 1989). Studies of the extent and pattern of DNA sequence diversity have been fewer in number, but have suggested that genetic exchange is more prevalent than the MLEE results might indicate (DuBose *et al.*, 1988; Dowson *et al.*, 1989). Thus, the most fundamental question concerns the extent of recombination among prokaryotes (Maynard Smith & Levin, 1980), which raises the question as to whether it is meaningful to discuss bacterial 'species'. If 'species' do exist, how much DNA sequence variability is there among different isolates?

Not surprisingly, the most extensively studied species has been *Escherichia coli*. MLEE results have been interpreted as indicating a largely clonal pattern of divergence (Selander & Levin, 1980; Ochman & Selander, 1984; Selander *et al.*, 1987). However, DNA sequence analyses of the phoA locus (DuBose *et al.*, 1988) and the trp region (Milkman & Stoltzfus, 1988; Milkman & Bridges, 1990) have revealed recombination of short DNA segments among closely related strains; thus it has been suggested that *E. coli* does comply with the traditional species concept. The extent of DNA sequence diversity among strains of *E. coli* varies among the small number of different loci examined, and it is not clear to what extent this is determined by natural selection. Of course, the prokaryotic kingdom is so diverse that there is no reason to believe that this (as yet incomplete) picture of *E. coli* will be applicable to other bacteria.

*Bacillus subtilis* is potentially an interesting species in which to investigate population genetics because other aspects of its biology are well understood, and because it is so distantly related to *E. coli* (Woese, 1987). Like *E. coli*, there is a well defined physical and genetic map of the *B. subtilis* chromosome (Piggot *et al.*, 1990), and more than 4% of the genome has been sequenced (Sharp *et al.*, 1990). *B. subtilis* is very different from *E. coli* with respect to its life history and ecology. Of particular importance to the question of genetic divergence within and among 'species' is the development of competence in *B. subtilis*, a process which does not occur naturally in *E. coli*. Competence is the ability of cells to bind, process and take up exogenous DNA into the cell. This process is subject to a complex series of nutritional and temporal control mechanisms (Dubnau, 1989) and (under laboratory conditions) only occurs in approximately 20% of cells as they enter stationary phase. It must be stressed that the extent to which this process occurs in the natural habitat is not known. Nevertheless, this potential for genetic exchange means that it is interesting to investigate DNA sequence divergence and the extent to which

*Author for correspondence. e-mail: PSHARP@VAX1.TCD.IE.*

*Abbreviations*: CAI, codon adaptation index; $F_{opt}$, frequency of optimal codons; $K$, number of synonymous substitutions per site.

The nucleotide sequence data reported in this paper have been submitted to GenBank/EMBL and have been assigned the accession numbers M81748 and M81749.
genetic recombination has occurred among wild isolates of \textit{B. subtilis}. In this study, we examine DNA sequence divergence at the rplX locus (encoding ribosomal protein L24) in nine \textit{B. subtilis} strains of diverse origins.

**Methods**

\textit{Bacterial strains.} \textit{E. coli} TG1 [K12Δ(lac-pro) supE thi hsdR F\textsuperscript{-} traD-36 proAB lacI lacZ ΔM15] was obtained from Amersham. The nine isolates of \textit{Bacillus subtilis} investigated in this study are outlined in Table 1, and were collected as follows. Five previously characterized strains (S032, S316, S317, S322, S340) were obtained from F. G. Priest (Heriot-Watt University, Edinburgh, UK), and the identification of these strains in different culture collections is indicated in Table 1; references to these strains can be found in Priest \textit{et al.} (1988). Two strains from the Chinese National Culture Collection (BSG33-1, BSG67), originally isolated from the Gobi Desert, were obtained from T. Trautner (Max-Planck-Institut für Molekular Genetik, Berlin, Germany). These two strains were typed at the National Collections of Industrial and Marine Bacteria (Aberdeen), where one (BSG67) was identified as belonging to the recently defined species \textit{B. atrophaeus} (Nakamura, 1989). JH642, a derivative of 168, was obtained from J. A. Hoch (Scripps Clinic and Research Foundation, La Jolla, Calif., USA). W23 was obtained from the Bacillus Genetic Stock Center (Ohio State University, Columbus, Ohio, USA).

\textit{DNA manipulations.} Chromosomal DNA was isolated from each strain according to Rodriguez \& Tait (1983). Using the published sequence of the \textit{B. subtilis} 168 spo ribosomal protein operon (Henkin \textit{et al.}, 1989), two oligonucleotides, 5'-CTAGCTCCGAAAAGTGATC-3' and 5'-CTTTTCTTTAAGGCGGTTCAT-3', were selected from the 3'-end of \textit{rplN} and the complementary strand of the 5'-end of \textit{rplE}, the genes preceding and following \textit{rplX}, respectively. These oligonucleotides were synthesized on an Applied Biosystems 391 PCRMate DNA synthesizer and were used as primers for a polymerase chain reaction (PCR). Each PCR contained: 100 ng chromosomal DNA, a 10 \(\mu\)M concentration of each primer, 2-7 mM-MgCl\(_2\), 100 mM-Tris/HCl, pH 9, 500 mM-KCl, 1% (v/v) Triton X-100, 0-1% (v/v) gelatin, 200 \(\mu\)M-dNTP mix and 2.5 U of Taq polymerase. The ends of the amplified DNA fragments were made blunt using the Klenow fragment of DNA polymerase. The DNA fragments were then ligated to pGEM5Zf (Promega Corp.), which had previously been linearized with EcoRV, and the ligation mix was transformed into \textit{E. coli} TG1. Insert-containing plasmid DNA was prepared by the method of Birnboim \& Doly (1979). The DNA sequence of the insert was determined with the Promega T7 Sequencing Kit, using universal forward and reverse, SP6, T7 and appropriate synthetic oligonucleotides as sequencing primers. Double-stranded template plasmid DNA was sequenced as outlined in the Promega Protocols and Applications Guide. Except where otherwise specified, all molecular biological methods were carried out according to Maniatis \textit{et al.} (1982).

**Results and Discussion**

\textit{Nucleotide sequence diversity}

The \textit{rplX} region of the chromosome has been cloned and sequenced from nine strains of \textit{B. subtilis} (Fig. 1). The sequence of strain JH642 is identical to the 168 sequence already published (Henkin \textit{et al.}, 1989), except at one nucleotide in the putative ribosome-binding site 5' to \textit{rplX}. The RBS for 168 was reported as GGTGG, whereas the same region is GGAGG in JH642 and all other strains examined here. Bearing in mind that JH642 was derived from 168, it seems probable that the substitution of A by T in 168 is an artefact. Among the other eight strains investigated there are only five different sequences: strains JH642, S316, S322 and BSG33-1 are identical, as are S340 and BSG67. For one of the strains showing most divergence from JH642 (S317, with six differences) two separate PCR-generated clones were sequenced. These two sequences were identical, suggesting that errors produced by \textit{Taq} polymerase do not contribute significantly to the variation observed.

Six of the 103 codons of the \textit{rplX} gene vary among the strains. In the non-coding region between \textit{rplN} and \textit{rplX} (5' to \textit{rplX}) one of 37 nucleotides is variable, and there is one insertion/deletion event; the latter is probably an insertion in the ancestor of the 168, S316, S322, BSG33-1 clad (see below). There is no variation in the non-coding region 3' to \textit{rplX} (5' to \textit{rplE}).

Two of the differences among the \textit{rplX} sequences occur at second codon positions, and thus cause amino acid replacements (Fig. 1). This protein is quite highly conserved, as judged by the 47% amino acid identity between \textit{B. subtilis} and \textit{E. coli}. The Ile→Thr replacement at residue 12 is less conservative than the Ile→Leu replacement seen between the Gram-positive and Gram-negative species. The Lys→Arg replacement at residue 103 would be considered a conservative change, but this residue is conserved as Lys in \textit{E. coli}. Thus both of these replacements may be at least slightly deleterious; we note that each change is found in one strain only.

The number of nucleotide differences, and the nucleotide diversity (nucleotide differences per site) for \textit{rplX}, and for the entire region sequenced, are

**Table 1. \textit{B. subtilis} strains used**

<table>
<thead>
<tr>
<th>Strain name</th>
<th>Designation</th>
<th>Identification method</th>
</tr>
</thead>
<tbody>
<tr>
<td>JH642</td>
<td>(derived from 168)</td>
<td></td>
</tr>
<tr>
<td>W23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S032</td>
<td>NCIB 8055</td>
<td>ATCC 6461 = NRS-275</td>
</tr>
<tr>
<td>S316</td>
<td>NCIB 3610</td>
<td>ATCC 6051 = NRS-744</td>
</tr>
<tr>
<td>S317</td>
<td>NCIB 8054</td>
<td>ATCC 6633 = B-765</td>
</tr>
<tr>
<td>S322</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S340</td>
<td>NCIB 8802</td>
<td></td>
</tr>
<tr>
<td>BSG33-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BSG67*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Typed as \textit{Bacillus atrophaeus}
### rplX variability in *Bacillus subtilis*

**Fig. 1.** Nucleotide sequences of the *rplX* gene from nine strains of *B. subtilis*. Only differences from the JH642 sequence are shown (- indicates no difference), with non-synonymous changes in bold. The insertion/deletion difference between JH642 and the other strains is indicated by +. The putative ribosome-binding site for *rplX* (Henkin et al., 1989) is underlined. The sequences of strains S316, S322 and BSG33-1 are identical to JH642; the reported sequence of 168 (Henkin et al., 1989) differs at one nucleotide within the ribosome-binding site. The sequence of BSG67 is identical to S340.
presented in Table 2; the insertion/deletion event is weighted equally with a nucleotide substitution. The greatest divergence is a little under 2%, and is found between strain W23 and other strains (except S317). The gene (values for the whole sequence are similar). The present in Table 2; the insertion/deletion event is (or among all nine strains) is 0.010 (or 0.006) in the weighted equally with a nucleotide substitution. The extent of divergence obviously depends on the strains chosen, and on the loci examined (see below), and a realistic estimate of nucleotide diversity within the species B. subtilis awaits further investigation.

The majority of the nucleotide changes occur at silent sites, as almost universally seen in both intra- and interspecific comparisons, and as expected if the majority of sequence polymorphisms and evolutionary changes occur at the less constrained sites, a view now widely held in the light of the neutral mutation theory (Kimura, 1983). Unexpectedly, all four silent nucleotide changes occur in Val codons. There are 15 Val residues in ribosomal protein L24, constituting 14.6% of the protein, which is more than twice the average among a wide range of proteins (Collins & Coulson, 1990), but this fact alone is not sufficient to explain the distribution of variability. In the rplX gene there are 52 4-fold degenerate sites (i.e. positions where any substitution is silent; see Table 3), and the presence of all four differences among the 15 codons for one amino acid is unusual. The probability that all four changes occur at (different) Val codons is approximately 15/52 × 14/51 × 13/50 × 12/49 = 0.0050. However, we would have been interested if all changes had involved any amino acid (not just Val), and so similar probabilities should be calculated for each, and summed: the probability becomes 0.0064. This test is conservative, because it ignores the lack of variability at 48 2-fold degenerate sites in rplX. To include the 2-fold degenerate sites, they could each be considered as 1/3 of a fully (4-fold) degenerate site, since one of the three possible nucleotide substitutions is synonymous (Li et al., 1985; for simplic-
possible that the Val codons form some sort of mutational hotspot. The four Val codon changes are of three different types (A : T ↔ T : A, A : T ↔ G : C and G : C ↔ T : A; bearing in mind that it is not known whether the primary mutational event involved the sense or complementary DNA strand). Mutation patterns can be influenced by nearest neighbour nucleotides, and there is some evidence that this affects codon usage in lowly expressed genes in B. subtilis (Shields & Sharp, 1987). In DNA, the third position of a Val codon has a T to the 5' side: none of the nine other synonymous positions with a 5' T (in Phe, Leu and Ile codons) in rplX are variable. The four variable sites are interesting in that they are among seven Val codons followed by a 3' A; none of the Phe, Leu or Ile codons are followed by an A. Thus, it is silent sites occurring in the context T-X-A which are variable at rplX. However, since a similar pattern of variability is not seen at other loci (see below), it is difficult at this stage to know how to interpret this observation.

Relative degree of divergence at rplX

It is possible to make some assessment of the divergence of the rplX gene relative to other loci sequenced in different strains of B. subtilis. A region at the chromosomal replication terminus has been sequenced in both 168 (Carrigan et al., 1987: Ahn & Wake, 1991) and W23 (Lewis & Wake, 1989; Ahn & Wake, 1991). Two genes located at terC can be compared in their entirety between these two strains (Ahn & Wake, 1991): rtp, encoding a DNA-binding protein, and orf257, encoding a protein with strong similarity to the pyrroline 5-carboxylate reductases of E. coli and Pseudomonas aeruginosa. A short 3' fragment of another gene, designated orf238, and encoding a protein with some similarity to the product of the Rhizobium meliloti nodG gene, can also be compared. The three genes at terC have similar levels of nucleotide divergence between 168 and W23 (Table 4), and are each 3-4 times more divergent than rplX. Of course, nucleotide divergence can occur at non-synonymous and synonymous (silent) sites; the increased divergence in orf257 is partly due to less constraint on the protein sequence of its product, but in rtp and orf238 all of the changes are synonymous. To quantify the divergence at silent sites, we have calculated $K_s$, the number of synonymous substitutions per site, by the method of Li et al. (1985), which includes a statistical correction for multiple hits. The values of $K_s$ for the terC genes are again at least 3 times that for rplX (Table 4). Unlike rplX, there is no evidence of an excess of nucleotide differences at Val codons in the terC genes.

By analogy with E. coli, two factors may contribute to the lower divergence at rplX than at terC. First, codon usage bias in these genes, assessed by either the CAI or $F_{op}$, is much lower than in rplX (Table 4). Thus, the difference in degree of divergence may be attributable to differential constraint on synonymous codon usage, as seen in comparisons between E. coli and Salmonella typhimurium (Sharp & Li, 1987b). Second, rplX is located at 12° on the B. subtilis chromosome (Piggot et al., 1990), quite close to the origin of replication; the extent of silent site divergence between E. coli and S. typhimurium is greatly reduced near oriC (Sharp et al., 1989).

The bglC gene encoding endo-$\beta$-1,4-glucanase has been investigated in four different strains of B. subtilis [see Cantwell et al. (1988) and references therein]. These sequences are highly divergent: average nucleotide diversity among the four strains is $0.100$. One of the strains examined was designated as ATCC 6633, which has been classified as being the same as S317 (Table 1). The ATCC 6633 sequence is not complete, but is clearly the most divergent among the four strains, with a nucleotide divergence of around $0.145$ from each of the other three strains. The average nucleotide diversity among the other three strains investigated (DLG, IFO3034 and PAP115) is $0.056$. Thus, while the relationships of these three strains to each other (and to the strains examined here) are unknown, nevertheless the bglC gene appears to be far more divergent than rplX. The map location of bglC is unknown, but codon usage in bglC (CAI = 0.43, $F_{op}$ = 0.30) is much less biased than in rplX, and this alone could account for the greater variability at bglC. There is no sign of hypervariability at Val codons in bglC.

This discussion of relative divergence of different genes may be inappropriate, if segments of the chromosome in each strain do not all have the same history. That is, if extensive recombination occurs among strains of B. subtilis then different extents of divergence may simply reflect differences in the length of time since chromoso-

<table>
<thead>
<tr>
<th>Gene . . .</th>
<th>rplX</th>
<th>orf238</th>
<th>rtp</th>
<th>orf257</th>
</tr>
</thead>
<tbody>
<tr>
<td>Map position</td>
<td>12</td>
<td>180</td>
<td>180</td>
<td>180</td>
</tr>
<tr>
<td>Codon bias CA1</td>
<td>0.72</td>
<td>0.52</td>
<td>0.44</td>
<td>0.39</td>
</tr>
<tr>
<td>$F_{op}$</td>
<td>0.69</td>
<td>0.41</td>
<td>0.36</td>
<td>0.25</td>
</tr>
<tr>
<td>No. of codons</td>
<td>104</td>
<td>28</td>
<td>123</td>
<td>258</td>
</tr>
<tr>
<td>Nucleotide differences</td>
<td>5</td>
<td>5</td>
<td>22</td>
<td>40</td>
</tr>
<tr>
<td>Divergence</td>
<td>0.016</td>
<td>0.057</td>
<td>0.060</td>
<td>0.052</td>
</tr>
<tr>
<td>$K_s$</td>
<td>0.06</td>
<td>0.33</td>
<td>0.35</td>
<td>0.20</td>
</tr>
</tbody>
</table>
strains as *B. atrophaeus*. All of the *B. atrophaeus* strains, and only those strains, exuded a brownish-black pigment, suggesting that the latter is a diagnostic character for this species. Among the strains examined here, BSG67 exudes a pigment, and was typed as *B. atrophaeus* (see Methods). However, the *rplX* gene of BSG67 falls within the 168 cluster (Fig. 2). In contrast, W23 and S317 (neither of which exude pigment) are distinct from 168 at *rplX*. Additionally, strains S316 and S317, by their similarity to NRS-744 and NRS-765, respectively (see Table 1), both fall within Nakamura’s group 3 of *B. subtilis*, and are expected to be distinct from S032 (similar to NRS-275, and thus in group 2). Yet S316 and S032 lie within the 168 group, while S317 is with W23. Thus the pigment production trait seems to be completely concordant with overall DNA relatedness (Nakamura, 1989), but the *rplX* gene is not, indicating that different regions of the chromosome have undergone recombination. This might be anticipated, since *B. subtilis* is naturally competent. Duncan *et al.* (1989) have reported exchange between wild isolates of *B. subtilis* and *B. licheniformis* in soil cultures, although the stability and fitness of the recombinants was questionable. Indeed, Duncan *et al.* (1989) point out that the observed dichotomy between these two sympatic species would be hard to explain if extensive successful recombination occurred.

In conclusion, the *rplX* genes show low divergence among strains of *B. subtilis* isolated from around the world. The degree of divergence is lower than observed for other genes, which is consistent with selection among synonymous codons constraining divergence in highly expressed genes. Even though the number of variable sites is small, they are sufficient to suggest that the relationships among these *rplX* sequences are different from those inferred from other characteristics (pigment production and DNA hybridization), suggesting that recombination has occurred among strains of *B. subtilis*.

We would like to thank Fergus Priest and Tom Trautner for generously providing bacterial strains, and Peter Young and Michael Bulmer for statistical advice. This study was supported by EOLAS grant SC/151/89 to P. M. S. and K. M. D.

**References**


