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 *et al.*, 1991), 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*a*). *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_{op}, frequency of optimal codons; K_s 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.

0001-7024 © 1992 SGM

genetic recombination has occurred among wild isolates of *B. subtilis*. In this study, we examine DNA sequence divergence at the *rplX* locus (encoding ribosomal protein L24) in nine *B. subtilis* strains of diverse origins.

Methods

Bacterial strains. E. coli TG1 [K12Δ *(lac-pro) supE thi hsdR* F′ *traD-36 proAB lacI lacZ* ΔM15] was obtained from Amersham. The nine isolates of *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 each of these strains in different culture collections is indicated in Table 1; references to these strains can be found in Priest *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 *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).

DNA manipulations. Chromosomal DNA was isolated from each strain according to Rodriguez & Tait (1983). Using the published sequence of the *B. subtilis* 168 *spc* ribosomal protein operon (Henkin *et al.*, 1989), two oligonucleotides, 5′-CTAGCTCCAGAAGTTATC-TAA-3′ and 5′-CTTTTCTTTAAGGCGGTTCAT-3′, were selected from the 3′-end of *rplN* and the complementary strand of the 5′-end of *rplE*, the genes preceding and following *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 μM 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% (w/v) gelatin, 200 μM-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 *Eco*RV, and the ligation mix was transformed into *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 *et al.* (1982).

Results and Discussion

Nucleotide sequence diversity

The *rplX* region of the chromosome has been cloned and sequenced from nine strains of *B. subtilis* (Fig. 1). The sequence of strain JH642 is identical to the 168 sequence already published (Henkin *et al.*, 1989), except at one nucleotide in the putative ribosome-binding site 5′ to *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 *Taq* polymerase do not contribute significantly to the variation observed.

Table 1. *B. subtilis strains used*

The first name is the designation used here. Synonyms were identified from Priest *et al.* (1988) and Nakamura (1989).

Strain nomenclature
JH642 (derived from 168)
W23
S032 = NCIB 8055 = ATCC 6461 = NRS-275
S316 = NCIB 3610 = ATCC 6051 = NRS-744
S317 = NCIB 8054 = ATCC 6633 = B-765
S322
S340 = NCIB 8802
BSG33-1
BSG67*

* Typed as *Bacillus atrophaeus*

Six of the 103 codons of the *rplX* gene vary among the strains. In the non-coding region between *rplN* and *rplX* (5′ to *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 clade (see below). There is no variation in the non-coding region 3′ to *rplX* (5′ to *rplE*).

Two of the differences among the *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 *B. subtilis* and *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 *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 the *rplX* gene, and for the entire region sequenced, are

```
                                           rplN
                                           ter
                                   JH642   TAA TTGAAATAAATGCCTTACTCAAGGAGGTGCGATCAGG
                                   S340    --- - -----------------------------------
                                   S032    --- - -----------------------------------
                                   W23     --- - ---------------G-------------------
                                   S317    --- - ---------------G-------------------
                                                +
rplX
Met His Val Lys Lys Gly Asp Lys Val Met Val Ile Ser Gly Lys Asp Lys Gly Lys Gln
ATG CAT GTA AAA AAA GGC GAT AAA GTT ATG GTT ATC TCT GGT AAA GAT AAA GGC AAA CAA
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
--- --- --T --- --- --- --- --- --A --- --- -C- --- --- --- --- --- --- --- ---
--- --- --T --- --- --- --- --- --A --- --- --- --- --- --- --- --- --- --- ---

Gly Thr Ile Leu Ala Ala Phe Pro Lys Lys Asp Arg Val Leu Val Glu Gly Val Asn Met
GGA ACA ATC CTT GCT GCT TTC CCT AAA AAG GAC CGC GTT TTA GTT GAA GGT GTT AAC ATG
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --G --- ---
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --G --- ---

Val Lys Lys His Ser Lys Pro Thr Gln Ala Asn Pro Gln Gly Gly Ile Ser Asn Gln Glu
GTA AAG AAA CAC TCT AAA CCA ACT CAA GCT AAC CCT CAA GGC GGT ATT TCT AAT CAA GAG
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
--G --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
--G --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---

Ala Pro Ile His Val Ser Asn Val Met Pro Leu Asp Pro Lys Thr Gly Glu Val Thr Arg
GCG CCA ATT CAT GTA TCA AAC GTT ATG CCG CTC GAT CCT AAA ACA GGT GAA GTG ACT CGC
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---

Val Gly Tyr Lys Val Glu Asp Gly Lys Lys Val Arg Val Ala Lys Lys Ser Gly Gln Val
GTA GGA TAC AAA GTG GAA GAT GGC AAA AAA GTT CGT GTA GCA AAA AAA TCT GGG CAA GTT
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---

                                           rplE
Leu Asp Lys ter                            Met
CTA GAT AAA TAG TATTAAAAGGAAGGGAGGTCTATCTC ATG   JH642
--- --- --- --- -------------------------- ---   S340
--- --- -G- --- -------------------------- ---   S032
--- --- --- --- -------------------------- ---   W23
--- --- --- --- -------------------------- ---   S317
```

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.

Table 2. *Nucleotide divergence of rplX among strains*

Values are the number of base substitutions (and divergence per site in brackets); values (below diagonal) for *rplX* coding sequence only, and (above diagonal) for the entire region sequenced (see Fig. 1). Strains S316, S322 and BSG33-1 are identical to JH642; BSG67 is identical to S340.

	JH642	S340	S032	S317	W23
JH642		1 (0·003)	2 (0·005)	6 (0·016)	7 (0·019)
S340	0 (0·000)		1 (0·003)	5 (0·013)	6 (0·016)
S032	1 (0·003)	1 (0·003)		6 (0·016)	7 (0·019)
S317	4 (0·013)	4 (0·013)	5 (0·016)		1 (0·003)
W23	5 (0·016)	5 (0·016)	6 (0·019)	1 (0·003)	

Table 3. *Codon usage in B. subtilis 168 rplX*

'Optimal' codons are underlined, 'non-optimal' codons are indicated by an asterisk (*).

Phe	UUU	0	Ser	UCU	4	Tyr	UAU	0	Cys	UGU	0
	UUC	1		UCC*	0		UAC	1		UGC	0
Leu	UUA	1		UCA	1	ter	UAA	0	ter	UGA	0
	UUG	0		UCG*	0		UAG	1	Trp	UGG	0
Leu	CUU	1	Pro	CCU	3	His	CAU	2	Arg	CGU	1
	CUC	1		CCC*	0		CAC	1		CGC	2
	CUA	1		CCA	2	Gln	CAA	5		CGA*	0
	CUG	0		CCG	1		CAG	0		CGG*	0
Ile	AUU	2	Thr	ACU	2	Asn	AAU	1	Ser	AGU	0
	AUC	2		ACC*	0		AAC	3		AGC	0
	AUA*	0		ACA	2	Lys	AAA	16	Arg	AGA	0
Met	AUG	4		ACG	0		AAG	2		AGG*	0
Val	GUU	8	Ala	GCU	3	Asp	GAU	5	Gly	GGU	4
	GUC	0		GCC*	0		GAC	1		GGC	4
	GUA	5		GCA	1	Glu	GAA	3		GGA	2
	GUG	2		GCG	1		GAG	1		GGG*	1

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 average nucleotide diversity among the five sequences (or among all nine strains) is 0·010 (or 0·006) in the *rplX* gene (values for the whole sequence are similar). 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 simplicity, the third positions of Ile codons are classified as 2-fold degenerate); then the *rplX* gene contains 68 4-fold degenerate site 'equivalents'. Thus, the clustering of changes in Val codons seems highly significant. However, it is difficult to gauge the real significance of this observation, because the particular hypothesis being tested has been suggested by the data.

A potential explanation of non-random changes at silent sites is that they are constrained by natural selection. Selection among synonymous codons has been inferred to occur in *B. subtilis* (Shields & Sharp, 1987), and indeed other ribosomal protein genes have highly biased codon usage. The Codon Adaptation Index (CAI) was developed as a measure of the codon bias in a gene (Sharp & Li, 1987*a*), in terms of the relative 'fitness' of each of its constituent codons; this relative fitness is assessed from the frequency of each codon in a reference set of genes expressed at very high levels. The CAI for *rplX* is 0·72, placing it among the most highly biased genes in *B. subtilis* (Sharp *et al*., 1990*a*). To consider the effect of codon selection on *rplX* sequences it is simpler to consider the occurrence of 'optimal' codons (defined in Sharp *et al*., 1990*b*). Codon usage in the 168 *rplX* gene is shown in Table 3. The frequency of optimal codons, F_{op}, is 0·69, confirming that this is one of the most highly biased genes known in this species; only five out of 186 *B. subtilis* genes considered by Sharp *et al*. (1990*b*) have F_{op} values higher than 0·70.

However, codon bias seems insufficient to explain the clustering of changes at Val codons. First, Val residues in *rplX* are encoded by three different codons (Table 3), suggesting that substitutions among these codons may be relatively unconstrained. Second, usage of (for example) Gly codons is distributed among all four synonyms (Table 3), yet none of the 11 Gly codons are variable. It is 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.

Table 4. *Divergence between B. subtilis strains 168 and W23*

Codon usage bias is estimated by two indices: CAI, Codon Adaptation Index, and F_{op}, Frequency of optimal codons – in each case values are the average for the two strains. Divergence is presented as the nucleotide divergence, and as K_S, the estimated number of synonymous substitutions per site.

Gene . . .	*rplX*	*orf238*	*rtp*	*orf257*
Map position	12	180	180	180
Codon bias CAI	0·72	0·52	0·44	0·39
F_{op}	0·69	0·41	0·36	0·25
No. of codons	104	28	123	258
Nucleotide differences	5	5	22	40
Divergence	0·016	0·057	0·060	0·052
K_S	0·06	0·33	0·35	0·20

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, 1987*b*). 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-β-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-

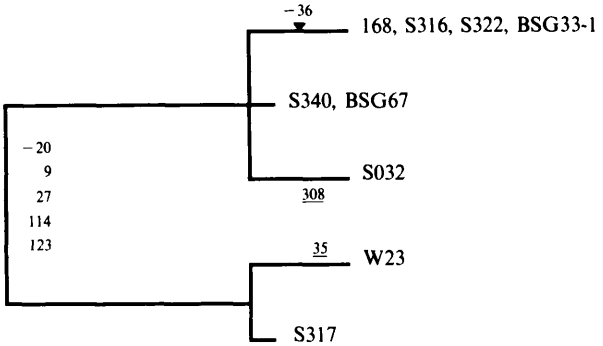

Fig. 2. Relationship among the *rplX* genes of the nine *B. subtilis* strains examined. Strain designations appear at the right end of branches. Horizontal branch lengths are approximately proportional to the number of differences (the insertion is weighted equally with a substitution). Numbers on branches indicate the nucleotide sites at which nucleotide changes have occurred, with those involving an amino acid replacement underlined, and the single nucleotide insertion indicated as a triangle (the first base of the *rplX* start codon is taken as position 1). Changes at the five sites near the root of the tree may have occurred on either branch leading from the common ancestor.

mal segments shared a common ancestor. In fact, there is some indication that recombination has occurred among *B. subtilis* strains (see below). However, the 168–W23 comparison is the most divergent among the *rplX* sequences examined here, and yet it is lower than all divergence values for *terC* or *bglC*; this suggests that there has indeed been stronger constraint on divergence at *rplX*.

Relationships among strains of B. subtilis

The nucleotide differences at *rplX* can be used to investigate the evolutionary relationships among these strains – at least for the *rplX* locus. This proviso arises because it is not clear how prevalent recombination is in different species of bacteria (Maynard Smith *et al.*, 1991). Using the maximum parsimony principle (Fitch, 1977), the strains of *B. subtilis* examined here fall into two groups (Fig. 2), one including 168 (and six other strains), the other including W23 (and S317). We may consider just four (representative) strains (168, S032, W23 and S317) and ask whether the phylogenetically informative sequence variations (i.e. excluding differences which are found in only one of the sequences examined) significantly support the grouping in Fig. 2. There are three possible (unrooted) relationships among four strains, yet all five informative sites support the grouping of 168 with S032, and W23 with S317; following Li & Gouy (1990), the probability that this grouping is incorrect is approximately $P = 0{\cdot}01$.

The relationships inferred for the *rplX* sequences (Fig. 2) are different from those which might be expected. First, diversity does not appear to be related to geographical origin, since strains BSG33-1 and BSG67, isolated from the Gobi Desert in Central Asia, are identical to 168 at *rplX*. Second, Nakamura (1989) has examined 52 strains of *B. subtilis* by DNA hybridization, reflecting overall DNA relatedness. He found a major subdivision among the strains, with one group being quite distinct from the others; he reclassified these 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 B-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 sympatric 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

Ahn, K. S. & Wake, R. G. (1991). Variations and coding features of the sequence spanning the replication terminus of *Bacillus subtilis* 168 and W23 chromosomes. *Gene* **98**, 107–112.

Birnboim, H. C. & Doly, J. (1979). A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Research* **7**, 1513–1523.

Cantwell, B. A., Sharp, P. M., Gormley, E. & McConnell, D. J. (1988). Molecular cloning of *Bacillus* beta-glucanases. In *Biochemistry and Genetics of Cellulose Degradation*, pp. 181–201. Edited by J.-P. Aubert, P. Beguin & J. Millet. London: Academic Press.

Carrigan, C. M., Haarsma, J. A., Smith, M. T. & Wake, R. G. (1987). Sequence features of the replication terminus of the *Bacillus subtilis* chromosome. *Nucleic Acids Research* **15**, 8501–8509.

Collins, J. F. & Coulson, A. F. W. (1990). Significance of protein sequence similarities. *Methods in Enzymology* **183**, 474–487.

Dowson, C. G., Hutchison, A., Brannigan, J. A., George, R. C., Hansman, D., Linares, J., Tomasz, A., Maynard Smith, J. & Spratt, B. G. (1989). Horizontal transfer of penicillin-binding protein genes in penicillin-resistant clinical isolates of *Streptococcus pneumoniae*. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 8842–8846.

Dubnau, D. (1989). The competence regulon of *Bacillus subtilis*. In *Regulation of Prokaryotic Development*, pp. 147–166. Edited by I. Smith, R. A. Slepecky & P. Setlow. Washington, DC: American Society for Microbiology.

DuBose, R. F., Dykhuizen, D. E. & Hartl, D. L. (1988). Genetic exchange among natural isolates of bacteria: recombination within the *phoA* gene of *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America* **85**, 7036–7040.

Duncan, K. E., Istock, C. A., Graham, J. B. & Ferguson, N. (1989). Genetic exchange between *Bacillus subtilis* and *Bacillus licheniformis*: variable hybrid sterility and the nature of bacterial species. *Evolution* **43**, 1585–1609.

Fitch, W. M. (1977). On the problem of discovering the most parsimonious tree. *American Naturalist* **111**, 223–257.

Henkin, T. M., Moon, S. H., Mattheakis, L. C. & Nomura, M. (1989). Cloning and analysis of the *spc* ribosomal operon of *Bacillus subtilis*: comparison with the *spc* operon of *Escherichia coli*. *Nucleic Acids Research* **17**, 7469–7486.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press.

Lewis, P. J. & Wake, R. G. (1989). DNA and protein sequence conservation at the replication terminus in *Bacillus subtilis* 168 and W23. *Journal of Bacteriology* **171**, 1402–1408.

Li, W.-H. & Gouy, M. (1990). Statistical tests of molecular phylogenies. *Methods in Enzymology* **183**, 645–659.

Li, W.-H., Wu, C.-I. & Luo, C.-C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution* **2**, 150–174.

Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982). *Molecular Cloning: a Laboratory Manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.

Maynard Smith, J., Dowson, C. G. & Spratt, B. G. (1991). Localized sex in bacteria. *Nature, London* **349**, 29–31.

Milkman, R. & Bridges, M. M. (1990). Molecular evolution of the *Escherichia coli* chromosome. III. Clonal frames. *Genetics* **126**, 505–517.

Milkman, R. & Stoltzfus, A. (1988). Molecular evolution of the *Escherichia coli* chromosome. II. Clonal segments. *Genetics* **120**, 359–366.

Nakamura, L. K. (1989). Taxonomic relationships of black-pigmented *Bacillus subtilis* strains and a proposal for *Bacillus atrophaeus* sp. nov. *International Journal of Systematic Bacteriology* **39**, 295–300.

Ochman, H. & Selander, R. K. (1984). Evidence for clonal population structure in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America* **81**, 198–201.

Piffaretti, J.-C., Kressbuch, H., Aeschbacher, M., Bille, J., Bannerman, E., Musser, J. M., Selander, R. K. & Rocourt, J. (1989). Genetic characterization of clones of the bacterium *Listeria monocytogenes* causing epidemic disease. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 3818–3822.

Piggot, P. J., Amjad, M., Wu, J.-J., Sandoval, H. & Castro, J. (1990). Genetic and physical maps of *Bacillus subtilis* 168. In *Molecular Biological Methods for Bacillus*, pp. 557–569. Edited by C. R. Harwood & S. M. Cutting. Chichester: John Wiley.

Priest, F. G., Goodfellow, M. & Todd, C. (1988). A numerical classification of the genus *Bacillus*. *Journal of General Microbiology* **134**, 1847–1882.

Rodriguez, R. L. & Tait, R. C. (1983). *Recombinant DNA Techniques: An Introduction*. Reading, MA: Addison-Wesley.

Selander, R. K. & Levin, B. R. (1980). Genetic diversity and structure in *Escherichia coli* populations. *Science* **210**, 545–547.

Selander, R. K., McKinney, R. M., Whittam, T. S., Bibb, W. F., Brenner, D. J., Nolte, F. S. & Pattison, P. E. (1985). Genetic structure of populations of *Legionella pneumophila*. *Journal of Bacteriology* **163**, 1021–1037.

Selander, R. K., Caugant, D. A. & Whittam, T. S. (1987). Genetic structure and variation in natural populations of *Escherichia coli*. In *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, pp. 1625–1648. Edited by F. C. Neidhardt, J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter & H. E. Umbarger. Washington, DC: American Society for Microbiology.

Sharp, P. M. & Li, W.-H. (1987*a*). The Codon Adaptation Index – a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research* **15**, 1281–1295.

Sharp, P. M. & Li, W.-H. (1987*b*). The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Molecular Biology and Evolution* **4**, 222–230.

Sharp, P. M., Shields, D. C., Wolfe, K. H. & Li, W.-H. (1989). Chromosomal location and evolutionary rate variation in Enterobacterial genes. *Science* **246**, 808–810.

Sharp, P. M., Higgins, D. G., Shields, D. C. & Devine, K. M. (1990*a*). Protein-coding genes: DNA sequence database and codon usage. In *Molecular Biological Methods for Bacillus*, pp. 557–569. Edited by C. R. Harwood & S. M. Cutting. Chichester: John Wiley.

Sharp, P. M., Higgins, D. G., Shields, D. C., Devine, K. M. & Hoch, J. A. (1990*b*). *Bacillus subtilis* gene sequences. In *Genetics and Biotechnology of Bacilli*, vol 3, pp. 89–98. Edited by M. M. Zukowski, A. T. Ganesan & J. A. Hoch. San Diego: Academic Press.

Shields, D. C. & Sharp, P. M. (1987). Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. *Nucleic Acids Research* **15**, 8023–8040.

Woese, C. R. (1987). Bacterial evolution. *Microbiological Reviews* **51**, 221–271.