The *dnaB*–*pheA* (256°–240°) region of the *Bacillus subtilis* chromosome containing genes responsible for stress responses, the utilization of plant cell walls and primary metabolism

Anil Wipat,¹ Noel Carter,² S. Caroline Brignell,^{1,2} B. Jane Guy,^{1,2}
Karen Piper,² June Sanders,² Peter T. Emmerson² and Colin R. Harwood¹

Author for correspondence: Colin Harwood. Tel: +44 191 222 7708. Fax: +44 191 222 7736.
e-mail: colin.harwood@ncl.ac.uk

School of Microbiological,
Immunological and
Virological Sciences¹ and
Department of
Biochemistry & Genetics,²
University of Newcastle
upon Tyne, Framlington
Place, Newcastle upon
Tyne NE2 4HH, UK

Within the framework of the international programme to sequence the genome of *Bacillus subtilis* strain 168, we were allocated the region between *dnaB* (256°) and *pheA* (240°). The sequencing of this region is now complete and we report our primary analysis of the 114 kb region containing 114 ORFs. In addition to previously characterized genes, we have identified genes involved in the utilization of plant cell wall polysaccharides, stress responses and the metabolism of amino acids, cell walls, DNA and fatty acids. We also discuss various structural and physical features, including the orientation of genes with respect to replication, putative start and stop codons, ribosome binding sites and *p*-independent transcription terminators.

Keywords: *Bacillus subtilis*, genome sequencing, Long Range PCR, stress genes, polysaccharide utilization

INTRODUCTION

The international programme to sequence the genome of *Bacillus subtilis* is a joint project involving European, Japanese and US research groups. To date a total of ca 3.3 Mb of the genome has been sequenced, with the remaining 0.9 Mb expected to be completed in 1997. The amenability of *B. subtilis* to genetic manipulation, combined with the availability of extensive expertise on its biochemistry and physiology, make this bacterium a valuable organism in which to investigate the properties of genes for which functions cannot be readily ascribed by the usual methods.

We were assigned the region from *dnaB* (256°) to *pheA* (240°), estimated on the basis of the genetic linkage map (Anagnostopoulos *et al.*, 1993) to be approximately 180 kb in length. Sequencing of the region is now complete and, at 114 kb, is considerably shorter than originally estimated. The region encodes genes for a wide variety of cellular processes including sporulation, germination, protein synthesis, the utilization of plant cell wall polysaccharides, the TCA cycle, stress responses and the

metabolism of amino acids, cell walls, DNA and fatty acids.

Here we report our primary analysis of the entire region in which we identify the main features of the ORFs not already represented in the databases. In addition, we present data on a region of limited coding capacity, part of which is highly homologous at the DNA level to a previously reported non-coding region from *Escherichia coli* (Raha *et al.*, 1993). We also describe a sequencing strategy that was used to overcome problems associated with the cloning of AT-rich sequences in *E. coli*. (Glaser *et al.*, 1993; Ogasawara *et al.*, 1994). The latter has involved random sequencing of a combination of λ phage inserts and Long Range (LR) PCR products, and directed sequencing from both using specific oligonucleotide primers. Initial selection of λ phage clones was greatly facilitated by the use of a collection of yeast artificial chromosomes containing 95% of the *B. subtilis* chromosome (Azevedo *et al.*, 1993).

METHODS

Bacterial strains, bacteriophages and plasmids. Chromosomal DNA for sequencing was prepared (Harwood & Archibald, 1990) from *B. subtilis* strain 168 (Anagnostopoulos & Spizizen, 1961), kindly provided by Dr C. Anagnostopoulos (Jouy-en-Josas, France). Clones of bacteriophage λ Gem11 were

Abbreviation: LR PCR, Long Range PCR.

The EMBL accession number for the nucleotide sequence reported in this paper is Z75208.

propagated and titred in *E. coli* KW251 (Promega). DNA fragments were cloned into the general purpose vector pUC18 (Vieira & Messing, 1987) and amplified in *E. coli* XL1-Blue (Stratagene).

Media and growth conditions. LB broth and LB agar were used for the growth of *B. subtilis* and *E. coli*. LB agar containing 0.6% agarose was used as an overlay agar for plating and titration of λ phage.

Cloning, generation of libraries and DNA preparation. A λ phage library containing chromosomal DNA from *B. subtilis* strain 168 was prepared in λ Gem11. Chromosomal DNA was partially digested with *Sau*3A to give fragments predominately in the 9–23 kb range and size-fractionated by centrifugation through a 10–40% (w/v) sucrose gradient. Fractions containing fragments of between 9 and 23 kb were ligated to *Bam*HI-digested λ Gem11 arms (Promega) and packaged into phage heads (Promega).

A subset of λ phage clones containing DNA from within the *pheA* (240°) to *dnaB* (256°) region were isolated by plaque hybridization (Sambrook *et al.*, 1989) using as a probe 32 P-labelled DNA from pYAC10-8, a yeast artificial chromosome vector containing *B. subtilis* chromosomal DNA within this region (Azevedo *et al.*, 1993). This subset of λ phage clones was then re-probed with previously cloned and sequenced genes to map the insert fragments.

Additional sources of specific chromosomal DNA were obtained by PCR amplification. Chromosomal fragments up to 25 kb in length were amplified by LR PCR using the GeneAmp XL PCR Kit (Perkin Elmer), as directed by the manufacturer. PCR was performed as 10 separate 100 μ l reactions from oligonucleotide primer pairs of at least 20 nucleotides. The reaction consisted of an initial denaturing step of 1 min at 94 °C, followed by 16 cycles of 30 s at 94 °C and 10 min at 66 °C, and 12 cycles of 30 s at 94 °C and 10 min at 66 °C; the latter was extended for 15 s for each cycle. Finally, the sample was incubated at 72 °C for 10 min to complete the reaction. The separate reaction products were pooled prior to fragmentation.

B. subtilis chromosomal DNA, either cloned in λ or as a LR PCR product, was randomly fragmented and cloned into pUC18. Random overlapping fragments of DNA were generated with DNaseI, essentially using the methods of Sambrook *et al.* (1989). Briefly, template DNA was partially digested with DNaseI in the presence of manganese chloride, repaired using T4 DNA polymerase and filled-in using the Klenow fragment of *E. coli* DNA polymerase I. The resulting fragments were size-fractionated on a 10–40% sucrose gradient and fragments between 0.5 kb and 2 kb were ligated into *Sma*I-digested and dephosphorylated pUC18 (Ready-to-Go; Pharmacia). Ligated DNA was transformed into *E. coli* XL1-Blue.

λ DNA was isolated using the method of Sambrook *et al.* (1989). Plasmid DNA for sequencing was prepared either by a modification of the alkaline lysis method (Birnboim & Doly, 1979) or by use of a 96-well format kit (AGTC). PCR products for sequencing were purified by agarose gel electrophoresis and recovered using QIAquick columns (Qiagen).

DNA hybridization. DNA hybridization was performed following the procedures described in Sambrook *et al.* (1989).

DNA sequencing. pUC18 clones containing *B. subtilis* DNA were sequenced using universal forward and reverse primers. Sequences were obtained using the *Taq* Dye Terminator and Dye Primer Cycle Sequencing Kits and Model 373A Sequencer

from Applied Biosystems. DNA sequence was determined on both strands with a mean redundancy of six, except for limited re-sequencing of regions already in the databases.

Two strategies were used to complete the gaps between the resulting sequence islands. PCR products spanning the gaps were either sequenced directly using custom primers or were used as probes to isolate specific clones from a pUC18 plasmid library of randomly generated clones containing smaller (100–500 bp) chromosomal fragments.

Data handling and computer analysis. DNA sequence assembly was performed using SEQUENCHER (Gene Codes Corp.), a Macintosh-based DNA sequence assembly and analysis program. The DNA sequence was analysed for ORFs using the MACMOLLY TETRA (Soft Gene GmbH) software package. DNA and amino acid homology searches were performed using the FASTA (Pearson & Lipman, 1988) and BLAST (Altschul *et al.*, 1990) packages, respectively.

RESULTS AND DISCUSSION

Sequencing and analysis of the region from *dnaB* to *pheA*

The region between *dnaB* and *pheA* is 114 kb in length and its physical organization is shown in Fig. 1. Prior to our sequencing studies, four significant contigs of 6.9 kb (*dnaB*–*thrS*; Putzer *et al.*, 1990; Ogasawara *et al.*, 1986), 8.9 kb (*trx*–*gerE*; Chen *et al.*, 1989; Magnusson *et al.*, 1986; Phillips *et al.*, 1987; Cutting & Mandelstam, 1986), 4.2 kb (*ilvBNC*; Grandoni *et al.*, 1992) and 31 kb (*lon*–*pheA*; Schmidt *et al.*, 1994) had been sequenced. In addition, the 6 kb *pheST* region had been sequenced but not mapped (Brakhage *et al.*, 1990).

Initial probing of the λ phage library with *gerM* produced λ phage SC32, containing a 15 kb insert extending from the 3' end of *gerE* to the middle of *leuA*. The gap between *leuA* and *lonA* was spanned with a LR PCR product from *leuA* to *hemX*. A clone containing *pheST*, obtained previously from random cloning of pYAC10-8 DNA, was used to probe the λ phage library for clones within the region between *thrS* and *trx*, resulting in the isolation of λ phage SC9. The sequence obtained from λ SC9 was used to obtain LR PCR products linking *yseC* to *thrS* on one side and *pheT* to *trx* on the other.

The region encodes 113 ORFs for putative polypeptides of at least 65 amino acids and with appropriately located putative ribosome binding sites and initiation codons. An additional polypeptide (LysCb), the β -subunit of aspartokinase II, is identical to the C-terminal domain of the α -subunit (LysCa) of this enzyme (Zhang *et al.*, 1990). These ORFs account for 92% of the potential coding capacity of the DNA. The ORFs range in size from 67 residues (RpmL) to 881 residues (ValS), with a mean length of 307 amino acid residues. The largest non-coding region, between the divergent genes *ysnD* and *ilvB*, is 816 bp in length. Ninety-eight (86%) of the ORFs are orientated in the same direction as the movement of the replication complex, which is in agreement with the observations of others (e.g. Glaser *et al.*, 1993; Ogasawara *et al.*, 1994). Table 1 shows the sizes, the putative ribosome binding

sites and initiation codons for the newly identified ORFs. Thirty-nine putative *p*-independent transcription terminators were identified within the region, and their locations (Fig. 1) and sequences (Table 2) determined. Where transcription patterns are convergent, both orientations of the putative terminators (e.g. *thrS* and *ysaA*) are given in Table 2. All three putative start codons were used (ATG, 76%; TTG, 17%; GTG, 6%), together with the highly unusual ATT start codon used for *infC*, encoding translation initiation factor 3 (IF3). There was a strong bias towards TAA as a stop codon (65%), with TGA (25%) and TAG (10%) being used much less frequently.

A number of genes originally mapped to this region (Anagnostopoulos *et al.*, 1993) were either not found or not recognized. These include: *relA*, encoding the stringent factor; *gerH*, encoding a germination defect; *ten*, enhancement of transfection by phage SP82; *phoT*, encoding a constitutive alkaline phosphatase phenotype; *spoVIA*, encoding a stage 6 sporulation defect; and *pfk*, encoding a phosphofructokinase. In addition, the locations and/or order of a number of genes have been revised.

Homology search and properties of newly identified ORFs

Protein databases were searched for homologies to the newly identified ORFs by use of the BLAST algorithm (Altschul *et al.*, 1990). The revealed homologies are shown in Table 3 and any special features indicated below. The putative products of the following ORFs showed no significant homology to any proteins in the databases up to 1 April 1996; *yscA*, *yscB*, *ysdA*, *ysdB*, *yseB*, *ysfE*, *ysfA*, *ysgB*, *ysbA* and *ysbA*.

ysaA. The putative product of *ysaA* shows homology with a hypothetical protein of *E. coli* encoded by *yjgG*, located in the *rimI*-*prfC* intergenic region (Burland *et al.*, 1995).

lytS and lytT. The *lytS* and *lytT* gene products showed significant homology with the products of the *lytS* and *lytR* genes of *Staphylococcus aureus* (Brunskill & Bayles, 1996). Since *lytR* has been used previously in *B. subtilis* to designate a gene encoding a regulator of the major vegetative phase autolysin, the amidase LytC (Lazarevic *et al.*, 1992), we have labelled the *S. aureus* *lytR* homologue as *lytT*. Genes *lytS* and *lytT* are likely to encode the sensor-kinase and response-regulator components, respectively, of a two-component regulatory pathway involved in controlling the rate of autolysis.

ysbA and ysbB. The putative products of the *ysbA* and *ysbB* genes show homology to those of *ywbH* and *ywbG* of *B. subtilis* (Glaser *et al.*, 1993), *yohJ* and *yohK* of *E. coli*, and *yohJ* and *yohK* of *Haemophilus influenzae* (Fleischmann *et al.*, 1995). Although the functions of these proteins are not known, they are all likely to be transmembrane proteins.

infC, rpmI and rplT. The products of the *infC*, *rpmI* and *rplT* genes show extensive homology to the products of the *infC*, *rpmI*, *rplT* operon of *B. stearothermophilus*. This

operon encodes translation initiation factor IF3 and the structural genes for ribosomal proteins L35 and L20 (Pon *et al.*, 1989). The order of genes in *B. subtilis* matches that of *B. stearothermophilus* and both are located in the region between the *thrS* and *pheS* genes. Interestingly, the unusual ATT start codon, reported for the *infC* gene in *B. stearothermophilus* (Pon *et al.*, 1989) and *E. coli* (Gold *et al.*, 1984), is present in *B. subtilis*. As in *B. stearothermophilus*, these genes may form a single transcriptional unit in *B. subtilis* since *rplT* is followed by a putative *p*-independent transcription terminator.

ysdC. The putative product of the *ysdC* gene shows homology to the product of the *celM* gene of *Clostridium thermocellum* (Kobayashi *et al.*, 1993), a putative protein encoded from the *frv* operon of *E. coli* (Reizer *et al.*, 1994) and the *pepA* gene of *Lactococcus lactis* (l'Anson *et al.*, 1995). The *celM* and *pepA* genes encode an endoglucanase and glutamyl-aminopeptidase, respectively. Finally, residues 2–31 of YsdC match 27 out of 30 residues of a 30-residue N-terminal fragment of a well-characterized thermophilic aminopeptidase of *B. stearothermophilus* (Stoll *et al.*, 1973).

abnA. The *abnA* gene lies at the 5' end of a series of genes involved in the utilization of arabinan and arabinose. Its putative protein product shows homology with an arabinan-endo 1,5- α -L-arabinase from *Aspergillus niger* (Flippin *et al.*, 1993), particularly throughout the C-terminal half of the proteins. The N terminus of AbnA has the structure of a typical *Bacillus* signal peptide of 32 residues in length (Simonen & Palva, 1993).

araA, araB and araD. The putative protein products of *araA*, *araB* and *araD* show very strong homology to L-arabinose isomerase, L-ribulokinase and L-ribulose-5-phosphate 4-epimerase, respectively, from *E. coli* (Lee *et al.*, 1986) and a number of other bacteria. These genes therefore form an operon for L-arabinose catabolism. The operon from *B. subtilis* has been cloned and mapped previously (Sa-Nogueira *et al.*, 1988; Sa-Nogueira & de Lencastre, 1989). The homology data confirm that the gene order in *B. subtilis*, (*araABD*; Sa-Nogueira & de Lencastre, 1989), is different from that of *E. coli*, (*araBAD*; Lee *et al.*, 1986).

yseA. YseA shows homology to a hypothetical protein product of *yrdG*, the thiaminase I region of *B. subtilis*, *pnpP* of *Saccharomyces cerevisiae* encoding 4-nitrophenyl-phosphatase (Kaneko *et al.*, 1989) and the *nagD* gene product of *E. coli* (Plumbridge, 1989).

yseC, yseD, yseE and abfA. The putative products of *yseC*, *yseD* and *yseE* show homology to the products of three genes in the *lacEFGZK* operon of *Agrobacterium radiobacter* (Williams *et al.*, 1992), encoding components of an ABC-type (Higgins *et al.*, 1986) binding-protein-dependent lactose transporter. YseC shows homology to the periplasmic sugar binding protein, LacE, and to the substrate binding proteins of other transporters. In common with the substrate binding proteins of other Gram-positive ABC transporters, YseC appears to be a lipoprotein and has the structure of a typical lipoprotein

Fig. 1. For legend see facing page.

Fig. 1. Organization of the *B. subtilis* chromosome from *dnaB* (256°) to *pheA* (240°). The dashed lines represent the source of DNA template used to generate the sequence, either cloned λ phage insert or LR PCR product. The diagram is orientated in the direction of the replication complex (left to right). *NotI* sites are indicated on the diagram; no *SfiI* sites were found (Itaya, 1993). Thick black arrows represent the putative ORFs, together with their direction of transcription, and are named according to the rules described by Demerec *et al.* (1968). Shaded arrows indicate genes which have been cloned and sequenced previously. The loop structures show putative ρ -independent transcription terminators.

signal peptide, with an Ala-Gly processing site followed by Cys-Ser residues (Simonen & Palva, 1993). YseD and YseE show, respectively, homology to two integral membrane proteins, namely LacF and LacG. However, no ORF was found with homology to LacK, a membrane-peripheral ATP binding protein. The product of the *abfA* gene shows very high homology with the α -L-arabinofuranosidase gene from *Streptomyces lividans* (Manin *et al.*, 1994). We have cloned and over-expressed *abfA* in *E. coli* and shown the gene product to exhibit α -L-arabinofuranosidase activity (unpublished). Interestingly, *abfA* does not appear to encode a signal peptide, suggesting a requirement for an uptake system for its substrate, a possible function of the proteins encoded by *yseC*, *yseD* and *yseE*.

***cstA*.** The putative product of the *cstA* gene shows homology to the product of a well-studied carbon-starvation-induced (csi) *cstA* gene (Schlutz & Martin, 1991) and a putative csi gene *yjiY* (Burland *et al.*, 1995) from *E. coli*. Both are thought to be transmembrane proteins.

***ysfD* and *ysfC*.** The putative products of the *ysfD* and *ysfC* genes show homology with the products of the *glcF* and *glcD* genes, respectively, of *E. coli* (Molina *et al.*, 1994). Genes *glcF* and *gluD* encode subunits of glycolate oxidase, with the former being an Fe-S protein. *E. coli* glycolate oxidase has three subunits; however, no gene encoding a product with homology to the third subunit (*glcE*) was found in this region.

***ysfB*.** The putative product of the *ysfB* gene shows homology to hypothetical products of the *yaeG* genes of *E. coli* (Fujita *et al.*, 1994) and *Haemophilus influenzae* (Fleischmann *et al.*, 1995). The function of these gene products is unknown.

***ysgA*.** The putative product of the *ysgA* gene shows some homology to SpoU-type rRNA methyltransferases (Koonin & Rudd, 1993), including that from *Mycoplasma genitalium* (Fraser *et al.*, 1995), suggesting a role for this protein in RNA metabolism. Since this ORF lies immediately upstream of the *pheST* operon, encoding phenylalanyl-tRNA-synthetase, it may be involved in the methylation of tRNA.

Table 1. Main features of newly identified ORFs in the region of the *B. subtilis* genome between *dnaB* and *pheA*

ORF	Size (aa)	Endpoints (nt)	Translation start*	Description/comments
<i>ysaA</i>	261	7528–6746	<u>AGGAGaagatc</u> ATG	<i>E. coli</i> <i>yijG</i> homologue
<i>lytS</i>	594	7695–9476	<u>AGGtGacatga</u> ATG	<i>Staphylococcus aureus</i> <i>lytS</i> homologue
<i>lytR</i>	242	9454–10179	<u>AgGGAGaagaac</u> ATG	<i>Staphylococcus aureus</i> <i>lytT</i> homologue
<i>ysbA</i>	147	10313–10753	<u>AAaGAGGTgaagaaa</u> ATG	<i>B. subtilis</i> <i>ywbH</i> homologue
<i>ysbB</i>	232	10775–11470	<u>AGGAGGTtcttcac</u> ATG	<i>B. subtilis</i> <i>ywbGr</i> homologue
<i>yscB</i>	211	12134–11502	<u>AAGGgaGcctgacac</u> ATG	No significant homologues
<i>yscA</i>	75	12408–12184	<u>AGGtGagaagcgg</u> GTG	No significant homologues
<i>infC</i>	174	12580–13101	<u>AtGGAGGTggctc</u> ATT	<i>B. stearothermophilus</i> <i>infC</i> homologue
<i>rpmI</i>	67	13114–13314	<u>AAGGAGGaaatccc</u> ATG	<i>B. stearothermophilus</i> <i>rpmI</i> homologue
<i>rplT</i>	120	13346–13705	<u>AGGAGGgaaattca</u> ATG	<i>B. stearothermophilus</i> <i>rplT</i> homologue
<i>ysdA</i>	90	13762–14031	<u>AAGGAtGaggcaac</u> ATG	No significant homologues
<i>ysdB</i>	131	14439–14047	<u>AgGGAGGTgccttcacaca</u> ATG	No significant homologues
<i>ysdC</i>	362	14623–15708	<u>AAGGAGGgtttggaac</u> ATG	<i>B. stearothermophilus</i> aminopeptidase homologue
<i>abnA</i>	314	15905–16846	<u>AAGGAGGTtgaacaaa</u> TTG	<i>Aspergillus niger</i> arabinase homologue
<i>araA</i>	499	17054–18550	<u>AAaGgGGcagtttac</u> ATG	L-Arabinose isomerase
<i>araB</i>	561	18564–20246	<u>AGGgGGatgtcac</u> ATG	L-Ribulokinase
<i>araD</i>	230	20263–20952	<u>AAGGgtGatggagc</u> ATG	L-Ribulose-5-phosphate 4-epimerase
<i>yseA</i>	273	20930–21748	<u>AcGGAGcaa</u> ATG	<i>B. subtilis</i> <i>ygdG</i> homologue
<i>yseB</i>	395	21745–22929	<u>AGGAGGggattctgat</u> ATG	No significant homologues
<i>yseC</i>	434	22960–24261	<u>AaGgGGTaaacag</u> ATG	Possible substrate binding lipoprotein
<i>yseD</i>	314	24297–25238	<u>GGAGGTtctc</u> ATG	Possible integral membrane component of ABC transporter
<i>yseE</i>	282	25239–26084	<u>AAGGgGGaggataa</u> ATG	Possible integral membrane component of ABC transporter
<i>abfA</i>	501	26103–27605	<u>AAGGAGGaaacgtgtg</u> ATG	<i>Streptomyces lividans</i> arabinofuranosidase homologue
<i>cstA</i>	599	27757–29553	<u>AAGGaGaTgggagaaa</u> ATG	Carbon-starvation-induced transmembrane protein
<i>yseE</i>	81	29726–29968	<u>AcagGGTctacgata</u> TTG	No significant homologues
<i>yseD</i>	445	31341–30007	<u>AcagGGTctacgata</u> TTG	<i>E. coli</i> glycolate oxidase Fe-S protein subunit homologue
<i>yseC</i>	471	32750–31338	<u>GGAGGcggaag</u> ATG	<i>E. coli</i> glycolate oxidase protein subunit homologue
<i>yseB</i>	369	33959–32853	<u>AgGGgGaTgggtgaa</u> ATG	<i>E. coli</i> <i>yaeG</i> homologue
<i>yseA</i>	72	34243–34028	<u>AGGgGGaacgcc</u> ATG	No significant homologues
<i>ysgA</i>	249	34362–35108	<u>AAGGAGcgccacatagaca</u> GTG	Possible methyltransferase
<i>ysgB</i>	314	39904–38963	<u>AAGGAGaTtttac</u> GTG	No significant homologues
<i>ysbA</i>	86	40038–40295	<u>AtGGAGGgaaacg</u> TTG	No significant homologues
<i>ysbB</i>	178	40302–40835	<u>AAGGAttgaacaact</u> ATG	Weak <i>Haemophilus influenzae</i> <i>cvpA</i> homologue
<i>ysbC</i>	571	40909–42621	<u>GGGgGGTatgat</u> ATG	DNA polymerase β homologue
<i>ysbD</i>	786	42642–44999	<u>AAGGAGGctcacacaatc</u> GTG	<i>mutS/hexA</i> homologue
<i>ysbE</i>	135	45014–45418	<u>AAGGgaGTatggcc</u> ATG	No significant homologues
<i>lefA</i>	561	45607–47289	<u>GGAGGTttt</u> ATG	<i>E. coli</i> <i>lefA</i> homologue
<i>ysiA</i>	585	47394–47978	<u>AgGaAGGaacatgaaca</u> TTG	Possible regulatory protein
<i>ysiB</i>	259	47993–48769	<u>AAGGAGtTctgtctc</u> ATG	<i>Clostridium acetobutylicum</i> crotonase homologue
<i>etfB</i>	258	48784–49557	<u>AAGGgGaTatgatc</u> ATG	β subunit of an electron transfer flavoprotein
<i>etfA</i>	326	49593–50570	<u>AGGgGGatgttaga</u> ATG	α subunit of an electron transfer flavoprotein
<i>asd</i>	496	50787–52274	<u>AAGGAatgatcagaggc</u> ATG	<i>Bacteroides ovatus</i> arabinosidase homologue
<i>ysmB</i>	147	61453–61893	<u>AgtGAGGgatggacaa</u> ATG	Possible MarR-type transcriptional regulator
<i>racE</i>	273	61901–62719	<u>AAtGAGGcgatt</u> TTG	Glutamate racemase
<i>ysnA</i>	199	64796–65392	<u>AAGacaGtaagc</u> TTG	Homology to unknown <i>E. coli</i> and <i>Haemophilus influenzae</i> proteins
<i>ysnB</i>	172	65402–65917	<u>AAGGAGGaagc</u> TTG	No significant homologues
<i>ysnF</i>	141	66570–66148	<u>AAGaAGaaaaaacg</u> ATG	No significant homologues
<i>ysnG</i>	89	66179–66445	<u>GGttGTcggtct</u> ATG	No significant homologues
<i>ysnE</i>	152	67646–67191	<u>AGGAGGaaataaaa</u> TTG	<i>Azospirillum brasilense</i> acetyltransferase homologue
<i>ysnD</i>	112	68150–67815	<u>AAGGtGgTacc</u> ATG	No significant homologues
<i>leuA</i>	519	72243–73799	<u>GGtGGTctccg</u> TTG	2-Isopropylmalate synthase
<i>leuC</i>	473	74968–76386	<u>AgGGAGGaaataaag</u> ATG	3-Isopropylmalate dehydratase large subunit
<i>leuD</i>	200	76399–76998	<u>AAGGAGtgcccgag</u> ATG	3-Isopropylmalate dehydratase small subunit
<i>ysaA</i>	311	77187–78119	<u>GtGGaaaaagg</u> ATG	No significant homologues
<i>tig</i>	425	78347–79621	<u>GGAGGGaaacac</u> ATG	Trigger factor homologue
<i>clpX</i>	421	79894–81156	<u>AAGgGGTgaaaga</u> ATG	ClpX stress-induced protease
<i>ysaC</i>	615	81206–80592	<u>GGaAGagttgc</u> ATG	No significant homologues
<i>lonB</i>	553	81308–82966	<u>AGGAGGgaccacc</u> ATG	ATP-dependent Lon protease homologue

* Underlined, ribosome binding site 16S rRNA consensus; bold, start codon.

Table 2. Putative ρ -independent transcription terminators within the *dnaB* (256°) to *pheA* (240°) region of the *B. subtilis* chromosome

Gene	Putative ρ -independent transcription terminator*
<i>ytxC</i>	tattcAACCTCGtCCCTtTcAtAGGGgCCgGGTTtttatagc
<i>thrS</i>	aaaatAAAatAAAAAAGCATGATCTCaTtgAaGAGATCATGCTTTTTTatTTTctcta
<i>ysaA</i>	tagagAAAatAAAAAAGCATGATCTGtTcaAtGAGATCATGCTTTTTTatTTTatttt
<i>lytR</i>	gaaggAATtgAAAAAGCtGCCCAtatttgATGGGCGGCTTTTTgcATTt
<i>ysbB</i>	ggaggAtAAGCCAAGGcTgaatGCCTTGGCTTtTttatttta
<i>yscB</i>	ataaaAaAGCCAAGGcattcaGCCTTGGCTTtTcctccgac
<i>rplT</i>	tatagAGCCGCTCTcCaGcAGAGCGGCTttttctatataaa
<i>ysdA</i>	cgcaaAAAAAGCAGAAgAGCGTTaAACGCTtTTCTGCTTTTTTataaatcaaa
<i>ysdB</i>	tttatAAAAAGCAGAAaAGCGTTtAACGCTcTTCTGCTTTTTttgagagttt
<i>abnA</i>	attgaAtAAAAAGcCgGGctctgcccCCgGcCTTTTTtTaaagaa
<i>abfA</i>	tagcaAAGCCGGAGAttctTCTCCGGCTTgtctttcaac
<i>ysfE</i>	gcacaAAAGCCCAAAACAatgtcTGTTTTGGGCTTTtctcatg
<i>ysfD</i>	atgagAAAGCCCAAAACAgacatTGTTTTGGGCTTTgtgtcgcg
<i>ysgA</i>	gccacAAcTCGTCCTTGTACAAGGGACGgGTTTTttttatttt
<i>pheT</i>	ctgatAAAAAAGCTTGCAGAgaaaTCTGCAAGCTTTTTTctatg
<i>ysgB</i>	catagAAAAAAGCTTGCAGAttctTCTGCAAGCTTTTTTatcag
<i>ysbB</i>	ggcatAAAAAACTTCTCttGtCcGAGAAAGTTTTTTTcaagt
<i>ysbE</i>	ttatgAGAAAAAGAAAGATtAcCTTTCTTTTTCTttttttgcg
<i>etfA</i>	aaaggAAAtCCCGGACTTTaAAAGTCCGGGtTTTttcatattataaa
<i>asd</i>	cgcgCAGAGCGCCGgagctcatgCCGGCGCTCTTttcaggttttaa
<i>lysCab</i>	atgacAATCAAAAAGGCGGgactatgcagtcacCGCCTTTTTTGATTtacct
<i>yslB</i>	agtgtAATCAAAAAGGCGGgtgactgcagatgcCCGCCTTTTTTGATTgtcat
<i>yslA</i>	tacatAAATtAtAaTTTTTcgaattgattgattGAAAAaTtTtATTTtatca
<i>sdhB</i>	tataaTGTaagAAAAAACCTCTtCCgcatGGgAGAGGTTTTTTtaaACAataag
<i>gerM</i>	atactAggAAGAAAGAGGtttagCGttcCGgctgCCTCTTTCTTatTcaatt
<i>ysnC</i>	gaaaaTgTCATGAGGAAGAAAAATTTCTTCTCATGAaAaaaattct
<i>ysnF</i>	accgaCAAccaAAAAACCTCAAATCCtaaatGGATTtTGAGGTTTTTcttTTGgtac
<i>ysnE</i>	gctttAtAAgGATAGATAAGtgaggggaaggtctgtCTTATCTATCtTTtTtata
<i>leuD</i>	cttcaAGcCtgAAAAAGCgGCCGGtatttgcCGGCgGCTTTTTtGcCTggtga
<i>tig</i>	agtacTAatAAAACAgGGCgCGAATcATTcGtGCCtTGTTTTcgTacaatt
<i>clpX</i>	taagcAcAAACCTCCTGAGTGgTAaCACTCAGGAGGTTTTtTtgca
<i>lonB</i>	cgtttAAACCCTATTtTgAtAATAGGGTTTttttcatga
<i>ysxC</i>	atgatAAACCGGTAGAGGATCCTCTACCGGTTTtatttatcgt
<i>ysxD</i>	aaataAAACCGGTAGAGGATCCTCTACCGGTTTatcattttttgat
<i>hemL</i>	gagtgAAAACCGGtatCAAGGAcTCCTTGtg_CCGGTTTTttcgt
<i>ysxE</i>	ccgcgAAAGAGCTtTTCGTCCtTtAcAgGGAAtGAAGAGCTCTTTtttcgtct
<i>valS</i>	aatgaAAGcCggGGCGGTTTCTagAGAAACCGCttGtCTTataaaa
<i>minD</i>	caaagAgAAgAAtCTGACAAAGCATATGCTgTGTCAGgTTtTtTttgtttttgc
<i>rpmA</i>	gattcAAAAAACTCCgGTCgatgatGACtGGAGTTTTTTttgca
<i>obg</i>	ggctgAAaAGAGAGCCTGGtgCCAGGCTCTCTaTTttaaa
<i>pheA</i>	aattaTAAAAAAGCCCaCTAGaGGGCTTTTTTtAgctt

* Inverted repeats are indicated by upper case letters.

ysbB. The putative product of the *ysbB* gene shows weak homology to colicin V production proteins, the products of the *cvpA* genes in *Haemophilus influenzae* (Fleischmann *et al.*, 1995) and *E. coli* (Fath *et al.*, 1989). The *ysbB* product also shows homology to FliR (Carpenter *et al.*, 1993), required for the biosynthesis of flagella. All of these proteins appear to be integral membrane proteins.

ysnC, ysdD and yshE. Genes *ysnC*–*ysnE* appear to form an operon of 4.6 kb with putative ρ -independent terminators

at each end. The putative product of the *ysnC* gene shows homology to DNA polymerase β from a number of organisms, including *Crithidia sciculata* (Torri & Englund, 1995), humans and rats. The putative product of the *ysnD* gene shows homology with the product of the *mutS* gene in *Thermus aquaticus* (Biswas & Hsieh, 1996) and the *hexA* gene of *Streptococcus pneumoniae* (Priebe *et al.*, 1988). *mutS* and *hexA* gene products are, respectively Gram-negative and Gram-positive homologues of enzymes responsible for heteroduplex DNA base mismatch repair that operates

Table 3. Homologues of newly identified/sequenced ORFs in the region of the *B. subtilis* genome between *dnaB* and *pheA*

ORF	BLAST score	Genes encoding homologous products	Function/activity of homologue	Reference/database number
<i>ysaA</i>	3.2e ⁻¹⁴	<i>yjiG</i> of <i>E. coli</i>	Hypothetical, unknown	Burland <i>et al.</i> (1995)
<i>lysS</i>	2.1e ⁻¹⁸⁵	<i>lysS</i> of <i>Staphylococcus aureus</i>	Sensor kinase, autolysins	Brunskill & Bayles (1996)
<i>lysT</i>	4.1e ⁻⁶⁷	<i>lysR</i> of <i>Staphylococcus aureus</i>	Response regulator, autolysin	Brunskill & Bayles (1996)
<i>ysbA</i>	1.3e ⁻²¹	<i>ywbH</i> of <i>B. subtilis</i>	Transmembrane protein	Glaser <i>et al.</i> (1993)
	9.2e ⁻¹⁰	<i>yohJ</i> of <i>E. coli</i>	Transmembrane protein	PIR: P33372
	1.6e ⁻⁸	<i>yohJ</i> of <i>Haemophilus influenzae</i>	Transmembrane protein	Fleischmann <i>et al.</i> (1995)
<i>ysbB</i>	7.9e ⁻⁴³	<i>ywbG</i> of <i>B. subtilis</i>	Transmembrane protein	Glaser <i>et al.</i> (1993)
	3.3e ⁻³⁵	<i>yohK</i> of <i>E. coli</i>	Transmembrane protein	PIR: P33373
	1.2e ⁻²⁵	<i>yohK</i> of <i>Haemophilus influenzae</i>	Transmembrane protein	Fleischmann <i>et al.</i> (1995)
<i>infC</i>	4.1e ⁻⁷⁶	<i>rpmI</i> of <i>B. stearothermophilus</i>	Translation initiation factor IF3	Pon <i>et al.</i> (1989)
<i>rpmI</i>	3.3e ⁻³⁴	<i>rpmI</i> of <i>B. stearothermophilus</i>	Ribosomal proteins L35	Pon <i>et al.</i> (1989)
<i>rplT</i>	1.6e ⁻⁴²	<i>rplT</i> of <i>B. stearothermophilus</i>	Ribosomal proteins L20	Pon <i>et al.</i> (1989)
<i>ysdC</i>	2.7e ⁻⁹²	<i>celM</i> of <i>Clostridium thermocellum</i>	Endoglucanase	Kobayashi <i>et al.</i> (1993)
	3.1e ⁻⁸²	<i>fruX</i> of <i>E. coli</i>	Hypothetical, unknown.	Reizer <i>et al.</i> (1994)
	2.3e ⁻⁷⁴	<i>pepA</i> of <i>Lactococcus lactis</i>	Glutamyl-aminopeptidase	I'Anson <i>et al.</i> (1995)
<i>abnA</i>	1.1e ⁻²⁹	<i>abnA</i> of <i>Aspergillus niger</i>	Arabinan-endo 1,5- α -L-arabinase	Flipphi <i>et al.</i> (1993)
<i>araA</i>	6.0e ⁻¹⁹³	<i>araA</i> of <i>E. coli</i>	L-Arabinose isomerase	Lee <i>et al.</i> (1986)
<i>araB</i>	1.3e ⁻⁹⁰	<i>araB</i> of <i>E. coli</i>	L-Ribulokinase	Lee <i>et al.</i> (1986)
<i>araD</i>	3.4e ⁻⁹⁶	<i>araD</i> of <i>E. coli</i>	L-Ribulose-5-phosphate 4-epimerase	Lee <i>et al.</i> (1986)
<i>yseA</i>	3.2e ⁻²¹	<i>yzdG</i> of <i>B. subtilis</i>	Hypothetical, unknown.	PIR: P46351
	2.2e ⁻¹⁹	<i>pnpP</i> of <i>Saccharomyces cerevisiae</i>	4-Nitrophenyl-phosphatase	Kaneko <i>et al.</i> (1989)
	1.4e ⁻¹⁵	<i>nagD</i> of <i>E. coli</i>	N-Acetyl glucosamine metabolism	Plumbridge (1989)
<i>yseC</i>	2.1e ⁻¹⁰	<i>lacE</i> of <i>Agrobacterium radiobacter</i>	Binding-protein-dependent lactose transporter	Williams <i>et al.</i> (1992)
<i>yseD</i>	1.3e ⁻³⁸	<i>lacF</i> of <i>Agrobacterium radiobacter</i>	Integral membrane protein lactose transporter	Williams <i>et al.</i> (1992)
<i>yseE</i>	1.1e ⁻⁵⁹	<i>lacG</i> of <i>Agrobacterium radiobacter</i>	Integral membrane protein lactose transporter	Williams <i>et al.</i> (1992)
<i>abfA</i>	3.3e ⁻¹⁸⁸	<i>abfA</i> of <i>Streptomyces lividans</i>	α -L-Arabinofuranosidase	Manin <i>et al.</i> (1994)
<i>cstA</i>	1.5e ⁻¹⁶⁸	<i>cstA</i> of <i>E. coli</i>	Transmembrane carbon-starvation-induced protein	Schlutz & Matin (1991)
	1.4e ⁻¹⁷⁴	<i>yjiY</i> of <i>E. coli</i>	Transmembrane carbon-starvation-induced protein	Burland <i>et al.</i> (1995)
<i>ysfD</i>	1.7e ⁻⁵³	<i>glcF</i> of <i>E. coli</i>	Fe-S protein subunits of glycolate oxidase	Molina <i>et al.</i> (1994)
<i>ysfC</i>	5.8e ⁻¹¹⁰	<i>glcD</i> of <i>E. coli</i>	Subunits of glycolate oxidase	Molina <i>et al.</i> (1994)
<i>ysfB</i>	3.2e ⁻³⁷	<i>yaeG</i> of <i>E. coli</i>	Hypothetical, unknown.	Fujita <i>et al.</i> (1994)
	6.9e ⁻³⁵	<i>yaeG</i> of <i>Haemophilus influenzae</i>	Hypothetical, unknown.	Fleischmann <i>et al.</i> (1995)
<i>ysgA</i>	1.3e ⁻²⁰	<i>yacO</i> of <i>B. subtilis</i>	SpoU-type rRNA methyltransferase	Ogasawara <i>et al.</i> (1994)
	8.6e ⁻¹⁵	<i>yjiH</i> of <i>E. coli</i>	SpoU-type rRNA methyltransferase	Burland <i>et al.</i> (1995)
<i>ysbB</i>	3.1e ⁻⁷	<i>cvpA</i> of <i>Haemophilus influenzae</i>	Integral membrane colicin V production protein	Fleischmann <i>et al.</i> (1995)
	0.0051	<i>cvpA</i> of <i>E. coli</i>	Integral membrane colicin V production protein	Fath <i>et al.</i> (1989)
	0.005	<i>fliR</i> of <i>E. coli</i>	Integral membrane protein required for the biosynthesis of flagella	Carpenter <i>et al.</i> (1993)
<i>ysbC</i>	1.9e ⁻¹⁴	gi66274 protein of <i>Critidia fasciculata</i>	DNA polymerase β	Torri & Englund (1995)
<i>ysbD</i>	1.8e ⁻²³	<i>mutS</i> of <i>Thermus aquaticus</i>	Responsible for heteroduplex DNA base mismatch repair	Biswas & Hsieh (1996)
	2.4e ⁻²⁶	<i>hexA</i> of <i>Streptococcus pneumoniae</i>	Responsible for heteroduplex DNA base mismatch repair	Priebe <i>et al.</i> (1988)
<i>lcfA</i>	4.4e ⁻¹³⁷	<i>lcfC</i> of <i>E. coli</i>	Long-chain acyl-coenzyme synthetase	Black <i>et al.</i> (1992)

Table 3 (cont.)

ORF	BLAST score	Genes encoding homologous products	Function/activity of homologue	Reference/database number
<i>ysiA</i>	4.0e ⁻¹²	<i>mtrR</i> of <i>Neisseria gonorrhoeae</i>	Regulatory protein of the HTH-TetR family	Pan & Spratt (1994)
<i>ysiB</i>	6.9e ⁻⁶²	<i>crt</i> of <i>Clostridium acetobutylicum</i>	Crotonase (butyryl-CoA hydratase)	Boynton <i>et al.</i> (1996)
<i>etfB</i>	1.4e ⁻⁶⁷	<i>etfB</i> of <i>Paracoccus denitrificans</i>	α subunit of electron transfer flavoprotein	Bedzyk <i>et al.</i> (1993)
<i>etfA</i>	4.8e ⁻⁴³	<i>etfA</i> of <i>Paracoccus denitrificans</i>	β subunit of electron transfer flavoprotein	Bedzyk <i>et al.</i> (1993)
<i>asd</i>	5.0e ⁻²⁰	<i>asd</i> of <i>Bacteroides ovatus</i>	Arabinan degradation	PIR: U15179
<i>ysmB</i>	3.5e ⁻⁶	<i>papX</i> of <i>E. coli</i>	MarR family of transcriptional regulators	Ariza <i>et al.</i> (1994)
<i>racE</i>	2.6e ⁻⁸³	<i>racE</i> of <i>B. sphaericus</i>	Glutamate racemase	PIR: U26733
<i>ysnA</i>	2.3e ⁻⁵⁶	ORF o197 protein of <i>E. coli</i>	Hypothetical protein	PIR: U28377
<i>ysnE</i>	4.2e ⁻¹⁷	<i>iaaT</i> of <i>Azospirillum brasilense</i>	Putative acetyltransferase	Zimmer <i>et al.</i> (1991)
<i>leuA</i>	4.0e ⁻¹⁴¹	<i>leuA</i> of <i>E. coli</i>	2-Isopropylmalate synthase	Yura <i>et al.</i> (1992)
<i>leuC</i>	2.3e ⁻¹⁰⁹	<i>leuC</i> of <i>E. coli</i>	3-Isopropylmalate dehydratase large subunit	Yura <i>et al.</i> (1992)
<i>leuD</i>	2.4e ⁻²⁰³	<i>leuD</i> of <i>E. coli</i>	3-Isopropylmalate dehydratase small subunit	Yura <i>et al.</i> (1992)
<i>tig</i>	1.1e ⁻⁷²	<i>tig</i> of <i>Haemophilus influenzae</i>	Trigger factor	Fleischmann <i>et al.</i> (1995)
<i>clpX</i>	1.4e ⁻¹⁷⁴	<i>clpX</i> of <i>E. coli</i>	Stress proteases	Gottesman <i>et al.</i> (1993)
<i>lonB</i>	6.4e ⁻¹⁶	Homology to adjacent <i>lonA</i> of <i>B. subtilis</i>	ATP-dependent Lon protease gene	PIR: U18229

to correct the nascent and donor DNA strands after replication and, in the latter case, transformation. The product of the final gene in the putative operon, YshE, shows no significant homology to any proteins in the databases.

lcfA, ysiA, ysiB, etfB and etfA. This group of genes appears to encode proteins involved in fatty acid metabolism. The putative product of the *lcfA* gene shows homology with the long-chain acyl-coenzyme synthetase from *E. coli* (Black *et al.*, 1992). The putative product of the *ysiA* gene shows homology with a number of regulatory proteins of the HTH-TetR family (Waters *et al.*, 1983). The putative *ysiB* gene product shows a high degree of homology with crotonase (butyryl-CoA hydratase) from *Clostridium acetobutylicum* (Boynton *et al.*, 1996), involved in the pathway for converting acetyl-CoA to butyryl-CoA. The protein also shows good homology to enoyl-CoA hydratases from a variety of organisms. The *etfA* and *etfB* gene products show very high homology with α and β subunits of electron transfer flavoprotein proteins from a number of bacterial species (e.g. *Paracoccus denitrificans*; Bedzyk *et al.*, 1993). Interestingly, the gene encoding crotonase from *C. acetobutylicum* is also in an operon which includes the α and β subunits of an electron transfer flavoprotein (Boynton *et al.*, 1996).

asd. The protein encoded by *asd* shows striking homology with an arabinosidase gene from *Bacteroides ovatus*, a gene responsible for arabinan degradation.

ysmB and racE. The product of the *ysmB* gene shows homology with PapX, PrsX and other members of the

MarR family of transcriptional regulators (Ariza *et al.*, 1994). *racE* shows very high homology with glutamate racemase from a variety of bacterial species, including *B. sphaericus*. The *B. subtilis* *racE* gene product has been studied in this laboratory and its glutamate racemase activity confirmed (unpublished). We have also shown that *ysmB* and *racE* are transcribed as a single transcriptional unit and are currently investigating the possibility that YsmB has a role in the expression of this operon.

gerM. The *gerM* gene has been described previously (Slynn *et al.*, 1994). However, we have extended the C terminus of this protein from 210 to 367 aa.

ysnA and ysnB. The putative product of the *ysnA* gene shows good homology with putative proteins of unknown function from *E. coli* (ORF o197) and *Haemophilus influenzae* (HI0260; Fleischmann *et al.*, 1995). The product of the *ysnB* gene shows no significant homology with any entries in the database.

ysnD, ysnE, ysnF and ysnG. The *ysnDEF* genes are transcribed in the opposite direction to that of chromosomal replication and lie in a region of limited coding capacity, part of which shows a striking homology to DNA within a large non-coding region of *E. coli*. The putative products of *ysnD*, *ysnF* and *ysnG* show no significant homologies with any entries in the databases; however, the product of *ysnE* shows significant homology with a putative acetyltransferase required for indolacetic acid biosynthesis and located within the *trpGDC* gene cluster of *Azospirillum brasilense* (Zimmer *et al.*, 1991).

Interestingly, a similar ORF has been identified in the gap between two closely related flagellar gene clusters in *E. coli* and which also includes a gene (*amyA*) encoding an α -amylase related to the liquefying amylase of *Bacillus* spp. (Raha *et al.*, 1992, 1993). Gene *ysnE* and its *E. coli* homologue show > 60% DNA sequence homology throughout the extent of their ORFs, but the degree of homology falls rapidly beyond the ends of the ORFs.

leuA, leuB, leuC and leuD. The *leuABCD* gene cluster, responsible for leucine biosynthesis, has been previously mapped and cloned in *B. subtilis* (Nagahari & Sekaguchi, 1978). The sequence of *leuB* (formerly *leuC*) has been described previously (Imai *et al.*, 1987). *leuA*, *leuC* and *leuD* were designated on the basis of the very high homology of their products with 2-isopropylmalate synthase, 3-isopropylmalate dehydratase large subunit and 3-isopropylmalate dehydratase small subunit, respectively, of a large number of other bacterial species.

tig. *tig* encodes a product which shows high homology to trigger factor from a variety of bacterial species, including *Haemophilus influenzae* (Fleischmann *et al.*, 1995), *E. coli* (Aldea *et al.*, 1989) and *Campylobacter jejuni* (Griffiths *et al.*, 1995). Although believed to be involved in cell cycle events (Guthrie & Wickner, 1990), the precise function of *Tig* remains to be determined.

clpX and ysoC. The product of the *clpX* gene shows extensive homology to the ClpX stress proteases from a variety of bacterial species including *E. coli* (Gottesman *et al.*, 1993) and *Haemophilus influenzae* (Fleischmann *et al.*, 1995). We have examined the expression of *clpX* gene under a variety of stress conditions and mapped its σ^A -type promoter (Gerth *et al.*, 1996). *ysoC* runs on the opposite strand to *clpX* starting just downstream of its 3' terminus and terminating approximately half way through. It is very rich in Ser and Thr residues and shows no significant homology to proteins in the databases.

lonB. The putative product of the *lonB* gene shows extensive homology to the product of the adjacent ATP-dependent Lon protease gene, now designated *lonA*. The genes for these putative isozymes are transcribed in the same direction, but separated by a structure resembling a *p*-independent transcription terminator. The 3' end of the *lonB* gene has been located and described previously (Riethdorf *et al.*, 1994) and the sequence submitted by Ye & Wong (PIR accession number U18229).

ACKNOWLEDGEMENTS

This work was funded by the European Commission under the Biotechnology Programme, contract number BIO2-CT93-0272. We are grateful to the following for clones and/or sequence information: Drs C. Anagnostopoulos, B. M. Corfe, S. D. Ehrlich, M. Hecker, L. Hederstedt, R. Losick, Anne Moir, H. Paulus, H. Putzer, R. Schmidt and P. Serror. We also wish to acknowledge the excellent technical support of R. M. Liddell of the Molecular Biology Facility, University of Newcastle upon Tyne.

REFERENCES

- Aldea, M., Garrido, T., Hernandez-Chico, C., Vicente, M. & Kushner, S. R. (1989). Induction of a growth-phase-dependent promoter triggers transcription of *bolA*, an *Escherichia coli* morphogene. *EMBO J* **8**, 3923–3931.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic Local Alignment Search Tool. *J Mol Biol* **215**, 403–410.
- Anagnostopoulos, C. & Spizizen, J. (1961). Requirements for transformation of *Bacillus subtilis*. *J Bacteriol* **81**, 741–746.
- Anagnostopoulos, C., Piggot, P. J. & Hoch, J. A. (1993). The genetic map of *Bacillus subtilis*. In *Bacillus subtilis and Other Gram-positive Bacteria: Biochemistry, Physiology and Molecular Genetics*, pp. 425–461. Edited by A. L. Sonenshein, J. A. Hoch & R. Losick. Washington, DC: American Society for Microbiology.
- Ariza, R. R., Cohen, S. P., Bachhawat, N., Levy, S. B. & Demple, B. (1994). Repressor mutations in the *marRAB* operon that activate oxidative stress genes and multiple antibiotic resistance in *Escherichia coli*. *J Bacteriol* **176**, 143–148.
- Azevedo, V., Alvarez, E., Zumstein, E., Damiani, G., Sgarrella, V., Ehrlich, S. D. & Serror, P. (1993). An ordered collection of *Bacillus subtilis* DNA segments cloned in yeast artificial chromosomes. *Proc Natl Acad Sci USA* **90**, 6047–6051.
- Bedzyk, L. A., Escudero, K. W., Gill, R. E., Griffin, K. J. & Freman, F. E. (1993). Cloning, sequencing, and expression of the genes encoding subunits of *Paracoccus denitrificans* electron transfer flavoprotein. *J Biol Chem* **268**, 20211–20217.
- Birnboim, H. C. & Doly, J. (1979). A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Res* **7**, 1513–1523.
- Biswas, I. & Hsieh, P. (1996). Identification and characterization of a thermostable *mutS* homolog from *Thermus aquaticus*. *J Biol Chem* **271**, 5040–5048.
- Black, P. N., Dirusso, C. C., Metzger, A. K. & Heimert, T. L. (1992). Cloning, sequencing, and expression of the *fadD* gene of *Escherichia coli* encoding acyl coenzyme A synthetase. *J Biol Chem* **267**, 25513–25520.
- Boynton, Z. L., Bennett, G. N. & Rudolph, F. B. (1996). Cloning, sequencing and expression of clustered genes encoding beta-hydroxybutyryl-coenzyme A (CoA) dehydrogenase, crotonase and butyryl-CoA dehydrogenase from *Clostridium acetobutylicum* ATCC 824. *J Bacteriol* **178**, 3015–3024.
- Brakhage, A. A., Wozny, M. & Putzer H. (1990). Structure and nucleotide sequence of the *Bacillus subtilis* phenylalanyl-tRNA synthetase genes. *Biochimie* **72**, 725–734.
- Brunskill, E. W. & Bayles, K. W. (1996). Identification and molecular characterization of a putative regulatory locus that affects autolysis in *Staphylococcus aureus*. *J Bacteriol* **178**, 611–618.
- Burland, V. D., Plunkett, G., Sofia, H. J., Daniels, D. L. & Blattner F. R. (1995). Analysis of the *Escherichia coli* genome VI: DNA sequence of the region from 92.8 through 100 minutes. *Nucleic Acids Res* **23**, 2105–2119.
- Carpenter, P. B., Zuberi, A. R. & Ordal, G. W. (1993). *Bacillus subtilis* flagellar proteins FliP, FliQ, FliR and FliH are related to *Shigella flexneri* virulence factors. *Gene* **137**, 243–245.
- Chen, N.-Y., Zhang, J.-J. & Paulus, H. (1989). Chromosomal location of the *Bacillus subtilis* aspartokinase II gene and nucleotide sequence of the adjacent genes homologous to *uvrC* and *trx* of *Escherichia coli*. *J Gen Microbiol* **135**, 2931–2940.
- Cutting, S. & Mandelstam, J. (1986). The nucleotide sequence and

the transcription during sporulation of the *gerE* gene of *Bacillus subtilis*. *J Gen Microbiol* **132**, 3013–3024.

Demerec, M., Abelberg, E. A., Clark, A. J. & Hartman, P. E. (1968). Proposal for a uniform nomenclature in bacterial genetics. *J Gen Microbiol* **50**, 1–14.

Fath, M. J., Mahanty, H. K. & Kolter, R. (1989). Characterization of a *purF* operon mutation which affects colicin V production. *J Bacteriol* **171**, 3158–3161.

Fleischmann, R. D. and others (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512.

Flippin, M. J., Panneman, H., van der Veen, P., Visser, J. & de Graaff, H. (1993). Molecular cloning, expression and structure of the endo-1,5- α -L-arabinase gene of *Aspergillus niger*. *Appl Microbiol Biotechnol* **40**, 318–326.

Fraser, C. M. and others (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403.

Fujita, N., Mori, H., Yura T. & Ishihama, A. (1994). Systematic sequencing of the *Escherichia coli* genome: analysis of the 2.4–4.1 min (110,917–193,643 bp) region. *Nucleic Acids Res* **22**, 1637–1639.

Gerth, U., Wipat, A., Harwood, C. R., Carter, N., Emmerson, P. T. & Hecker, M. (1996). Sequence and transcription analysis of *clpX*—a class III heat shock gene of *Bacillus subtilis*. *Gene* (in press).

Glaser, P., Kunst, F., Arnaud, M., Coudart, M.-P., Gonzales, W., Hullo, M.-F., Ionescu, M., Lubochinsky, B., Marcelino, L., Moszer, I., Presecan, E., Santana, M., Schneider, E., Schweizer, J., Vertes, A., Rapoport, G. & Danchin, A. (1993). *Bacillus subtilis* genome project: cloning and sequencing of the 97 kb region from 325° to 333°. *Mol Microbiol* **10**, 371–384.

Gold, L., Stormo, G. & Saunders, R. (1984). *Escherichia coli* translational initiation factor IF3: a unique case of translational regulation. *Proc Natl Acad Sci USA* **81**, 7061–7065.

Gottesman, S., Clark, W. P., de Crecy-Lagard, V. & Maurizi, M. R. (1993). ClpX, an alternative subunit for the ATP-dependent Clp protease of *Escherichia coli*. Sequence and *in vivo* activities. *J Biol Chem* **268**, 22618–22626.

Grandoni, J. A., Zahler, S. A. & Calvo, J. M. (1992). Transcriptional regulation of the *ilv-leu* operon of *Bacillus subtilis*. *J Bacteriol* **174**, 3212–3219 (erratum *J Bacteriol* **174**, 4863).

Griffiths, P. L., Park, R. W. & Connerton, I. F. (1995). The gene for *Campylobacter* trigger factor: evidence for multiple transcription start sites and protein products. *Microbiology* **141**, 1359–1367.

Guthrie, B. & Wickner, W. (1990). Trigger factor depletion or overproduction causes defective cell division but does not block protein export. *J Bacteriol* **172**, 5555–5562.

Harwood, C. R. & Archibald, A. R. (1990). Growth, maintenance and general techniques. In *Molecular Biological Methods for Bacillus*, pp. 1–21. Edited by C. R. Harwood & S. M. Cutting. Chichester: John Wiley & Sons.

Higgins, C. F., Hiles, I. D., Salmond, G. P. C., Gill, D. R., Downie, J. A., Evans, I. J., Holland, I. B., Gray, L., Buckel, S. D., Bell, A. W. & Hermondson, M. A. (1986). A family of related ATP-binding subunits coupled to many distinct biological processes in bacteria. *Nature* **323**, 448–450.

I'Anson, K. J. A., Movahedi, S., Griffin, H. G., Gasson, M. J. & Mulholland, F. (1995). A non-essential glutamyl aminopeptidase is required for optimal growth of *Lactococcus lactis* MG1363 in milk. *Microbiology* **141**, 2873–2881.

Imai, R., Sekiguchi, T., Nosoh, Y. & Tsuda, K. (1987). The nucleotide sequence of the 3-isopropylmalate dehydrogenase gene from *Bacillus subtilis*. *Nucleic Acids Res* **15**, 4988.

Itaya, M. (1993). Physical map of the *Bacillus subtilis* 168 chromosome. In *Bacillus subtilis and Other Gram-positive Bacteria: Biochemistry, Physiology and Molecular Genetics*, pp. 463–472. Edited by A. L. Sonenshein, J. A. Hoch & R. Losick. Washington, DC: American Society for Microbiology.

Kaneko, Y., Toh-e, A., Banno, I. & Oshima, Y. (1989). Molecular characterization of a specific *p*-nitrophenylphosphatase gene, PHO13, and its mapping by chromosome fragmentation in *Saccharomyces cerevisiae*. *Mol Gen Genet* **220**, 133–139.

Kobayashi, T., Romaniec, M. P. M., Barker, P. J., Gerngross, U. T. & Demain, A. L. (1993). Nucleotide sequence of gene *celM* encoding a new endoglucanase (CelM) of *Clostridium thermocellum* and purification of the enzyme. *J Ferment Bioeng* **76**, 251–256.

Koonin, E. V. & Rudd, K. E. (1993). SpoU protein of *Escherichia coli* belongs to a new family of putative rRNA methylases. *Nucleic Acids Res* **21**, 5519.

Lazarevic, V., Margot, P., Suldo, B. & Karamata, D. (1992). Sequencing and analysis of the *Bacillus subtilis* *lytRABC* divergon: a regulatory unit encompassing the structural genes of the *N*-acetylmuramoyl-L-alanine amidase and its modifier. *J Gen Microbiol* **138**, 1949–1961.

Lee, N., Gielow, W., Martin, R., Hamilton, E. & Fowler, A. (1986). The organization of the *araBAD* operon of *Escherichia coli*. *Gene* **47**, 231–244.

Magnusson, K., Phillips, M. K., Guest, J. R. & Rutberg, L. (1986). Nucleotide sequence of the gene for cytochrome *b₅₅₈* of the *Bacillus subtilis* succinate dehydrogenase. *J Bacteriol* **166**, 1067–1071.

Manin, C., Shareek, F., Morosoli, R. & Kluepfel, D. (1994). Purification and characterization of an α -L-arabinofuranosidase from *Streptomyces lividans* 66 and DNA sequence of the gene (*abfA*). *Biochem J* **302**, 443–449.

Molina, I., Pellicer, M. T., Badia, J., Aguilar, J. & Baldoma, L. (1994). Molecular characterisation of *Escherichia coli* malate synthase G differentiation with the malate synthase isoenzyme. *Eur J Biochem* **224**, 541–548.

Nagahari, K. & Sekaguchi, K. (1978). Cloning of the *Bacillus subtilis* leucine A, B and C genes with *Escherichia coli* plasmids and the expression of the *leuC* gene in *Escherichia coli*. *Mol Gen Genet* **158**, 263–270.

Ogasawara, N., Nakai, S. & Yoshikawa, H. (1994). Systematic sequencing of the 180 kilobase region of the *Bacillus subtilis* chromosome containing the replication origin. *DNA Res* **1**, 1–14.

Ogasawara, N., Moriya, S., Mazza, P. G. & Yoshikawa, N. (1986). Nucleotide sequence and organisation of the *dnaB* and neighbouring genes on the *Bacillus subtilis* chromosome. *Nucleic Acids Res* **14**, 9989–9999.

Pan, W. & Spratt, B. G. (1994). Regulation of the permeability of the gonococcal cell envelope by the mtr system. *Mol Microbiol* **11**, 769–775.

Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* **85**, 2444–2448.

Phillips, M. K., Hederstedt, L., Hasnain, S., Rutberg, L. & Guest, J. R. (1987). Nucleotide sequence encoding the flavoprotein and iron-sulfur protein subunits of the *Bacillus subtilis* PY79 dehydrogenase complex. *J Bacteriol* **169**, 864–873.

Plumbridge, J. A. (1989). Sequence of the *nagBACD* operon in *Escherichia coli* K12 and pattern of transcription within the *nag* regulon. *Mol Microbiol* **3**, 505–515.

Pon, C. L., Brombach, M., Thamm, S. & Gualerzi, C. O. (1989). Cloning and characterization of a gene cluster from *Bacillus*

- stearothermophilus* comprising *infC*, *rpmI* and *rplT*. *Mol Gen Genet* **218**, 355–357.
- Priebe, S. D., Hadi, S. M., Greenberg, B. & Lacks, S. A. (1988). Nucleotide sequence of the *hexA* gene for DNA mismatch repair in *Streptococcus pneumoniae* and homology of *hexA* to *mutS* of *Escherichia coli* and *Salmonella typhimurium*. *J Bacteriol* **170**, 190–196.
- Putzer, H., Brakhage, A. & Grunberg-Manago, M. (1990). Independent genes for two threonyl-tRNA synthetases in *Bacillus subtilis*. *J Bacteriol* **172**, 4593–4602.
- Raha, M., Kawagishi, I., Müller, V., Kihara, M. & MacNab, R. M. (1992). *Escherichia coli* produces a cytoplasmic α -amylase, AmyA. *J Bacteriol* **174**, 6644–6652.
- Raha, M., Kihara, M., Kawagishi, I. & MacNab, R. M. (1993). Organization of the *Escherichia coli* and *Salmonella typhimurium* chromosomes between flagella regions IIIa and IIIb, including a large non-coding region. *J Gen Microbiol* **139**, 1401–1407.
- Reizer, J., Michotey, V., Reizer, A. & Saier, M. H. (1994). Novel phosphotransferase system genes revealed by bacterial genome analysis: unique putative fructose- and glucoside-specific systems. *Protein Sci* **3**, 440–450.
- Riethdorf, S., Völker, U., Gerth, U., Winkler, A., Engelmann, S. & Hecker, M. (1994). Cloning, nucleotide sequence, and expression of the *Bacillus subtilis* *lon* gene. *J Bacteriol* **176**, 6518–6527.
- Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning: a Laboratory Manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.
- Sa-Nogueira, I. & de Lencastre, H. (1989). Cloning and characterization of *araA*, *araB* and *araD*, the structural genes for L-arabinose utilization in *Bacillus subtilis*. *J Bacteriol* **171**, 4088–4091.
- Sa-Nogueira, I., Paveia, H. & de Lencastre, H. (1988). Isolation of constitutive mutants for L-arabinose utilization in *Bacillus subtilis*. *J Bacteriol* **170**, 2855–2857.
- Schmidt, R., Decatur, A. L., Rather, P. N., Moran, C. P. & Losick, R. (1994). *Bacillus subtilis* Lon protease prevents inappropriate transcription of genes under the control of the sporulation transcription factor sigma(G). *J Bacteriol* **176**, 6528–6537.
- Schultz, E. & Matin, A. (1991). Molecular and functional characterization of a carbon starvation gene of *Escherichia coli*. *J Mol Biol* **218**, 129–140.
- Simonen, M. & Palva, I. (1993). Protein secretion in *Bacillus* species. *Microbiol Rev* **57**, 109–137.
- Slynn, G. M., Sammons, R. L., Smith, D. A., Moir, A. & Corfe, B. M. (1994). Molecular genetical and phenotypical analysis of the *gerM* spore germination gene of *Bacillus subtilis* 168. *FEMS Microbiol Lett* **121**, 315–320.
- Stoll, E., Ericsson, L. H. & Zuber, H. (1973). The function of the two subunits of thermophilic aminopeptidase I. *Proc Natl Acad Sci USA* **70**, 3781–3784.
- Torri, A. F. & Englund, P. T. (1995). A DNA polymerase beta in the mitochondrion of the trypanosomatid *Crithidia fasciculata*. *J Biol Chem* **270**, 3495–3497.
- Vieira, J. & Messing, J. (1987). Production of single-stranded plasmid DNA. *Methods Enzymol* **153**, 3–15.
- Waters, S. H., Rogowsky, P., Grinsted, J., Altenbuchner, J. & Schmitt, R. (1983). The tetracycline resistance determinant of RPI and Tn1721: nucleotide sequence analysis. *Nucleic Acids Res* **11**, 6089–6105.
- Williams, S. G., Greenwood, J. A. & Jones, C. W. (1992). Molecular analysis of the *lac* operon encoding the binding-protein-dependent lactose transport system and β -galactosidase in *Agrobacterium radiobacter*. *Mol Microbiol* **6**, 1755–1768.
- Yura, T., Mori, H., Nagai, H., Nagata, T., Ishihama, A., Fujita, N., Isono, K., Mizobuchi, K. & Nakata, A. (1992). Systematic sequencing of the *Escherichia coli* genome: analysis of the 0–2.4 min region. *Nucleic Acids Res* **20**, 3305–3308.
- Zhang, J. J., Hu, F. M., Chen, N. Y. & Paulus, H. (1990). Comparison of the three aspartokinase isozymes in *Bacillus subtilis* Marburg and 168. *J Bacteriol* **172**, 701–708.
- Zimmer, W., Aparicio, C. & Elmerich, C. (1991). Relationship between tryptophan biosynthesis and indole-3-acetic acid production in *Azospirillum*: identification and sequencing of a *trpGDC* cluster. *Mol Gen Genet* **229**, 41–51.

Received 8 May 1996; revised 5 July 1996; accepted 12 July 1996.