The *Bacillus subtilis* 168 chromosome from *sspE* to *katA*

N. J. Cummings and I. F. Connerton

Author for correspondence: I. F. Connerton. Tel: +44 118 935 7000. Fax: +44 118 926 7917.
e-mail: ian.connerton@bbsrc.ac.uk

Department of Food
Macromolecular Science,
Institute of Food Research,
Earley Gate, Whiteknights
Road, Reading RG6 6BZ,
UK

We have cloned and sequenced a 24.5 kb region of the *Bacillus subtilis* 168 chromosome spanning the *sspE* and *katA* genes. The region contains a ribosomal RNA operon, *rrnD*, a tRNA gene set, *trnD* and 17 ORFs, 16 with putative ribosome-binding sites. Four of the ORFs (ORF2, ORF14, ORF16 and ORF17) match to known *B. subtilis* genes (*sspE*, *thiA*, *senS* and *katA*). Eight of the remaining ORF products show similarities with proteins present in the databases, including an ATP-binding transport protein, a glutamate-1-semialdehyde aminotransferase, a thiol-specific antioxidant protein, a mitomycin radical oxidase and a ferric uptake regulation protein.

Keywords: *Bacillus subtilis*, genome sequencing, glutamate-1-semialdehyde aminotransferase, *rrnD* operon, *trnD* cluster

We have sequenced 24.5 kb of the *Bacillus subtilis* 168 chromosome from a region adjacent to a *NotI* linking clone (pNEXT47; Itaya & Tanaka, 1991). pNEXT47 is located at 947 kb on the physical map corresponding to the *sspE* marker on the genetic map. The sequence spans the region *sspE*–*katA* following the direction of replication.

B. subtilis 168 was used as a source of chromosomal DNA which was prepared by methods described by Maniatis *et al.* (1982). A partial *Sau3A* genomic λ library in λ GEM11 (Promega) was provided by C. Harwood (University of Newcastle upon Tyne, UK) and used as a source of genomic clones. Genomic and λ clone DNAs served as templates for PCR and Long Range PCR amplifications using the Expand Long Template PCR System (Boehringer Mannheim). Such amplified DNA fragments were used as gene probes, as sources of DNase I-generated shot-gun clones and for direct sequencing.

To clone the entire 24.5 kb region several cloning strategies were used. Firstly, genomic libraries were screened by DNA/DNA hybridization of plaque lifts with [³²P]-labelled PCR fragments of the genes *sspE*, *phoA* and *senS* amplified from *B. subtilis* 168 genomic DNA. The library screen, using the *senS* probe, produced nine independent positive plaques. DNA was isolated from four of the nine plaques, all produced a diagnostic *senS* PCR fragment with the *senS*-specific

primers. Using primers based on the T7 and SP6 promoters the genomic DNA inserts of these clones were amplified using the Expand Long Template PCR system. None of the clones produced PCR fragments using the *sspE* or *phoA* primer pairs, two genes believed to be located in this region. In the absence of any additional information, the largest insert fragment was selected for DNA sequencing (clone 2c.2). The 2c.2 insert was amplified in 10 independent reactions which were pooled prior to shot-gun cloning. The size of the λ clone was verified by amplifying various sections directly from genomic DNA by PCR using specific primers. Sequencing was performed by a strategy which used a combination of random DNase I fragments, direct sequencing from the 2c.2 PCR product and direct sequencing of internal PCR fragments produced from template *B. subtilis* 168 genomic DNA. This strategy generated a contig of 14.962 kb which spanned the 16S RNA–*katA* genomic region.

To extend the sequence to *sspE*, we adopted a Long Range PCR strategy, using genomic DNA as the template, one primer based on sequence adjacent to the 16S RNA and the second based on the *sspE* sequence. These primers produced a fragment, 16S-S, which was approximately 9.5 kb. Ten independent PCR amplifications of this fragment were pooled and shot-gun cloned. Both direct sequencing of the PCR product and data obtained from the random clones were used to produce the completed sequence which overlapped the previous region by 800 bp.

The region upstream of *sspE* was obtained by

The EMBL accession number for the sequence reported in this paper is Z82044.

Table 1. Putative ORFs in the 24.5 kb sequence from the *sspE* to *katA* region

ORF	Gene	Position and direction of transcription	Start/stop	Bases	Codons	Shine–Dalgarno consensus sequence (upper case) and initiation codon (bold)
1	<i>ygaA</i>	615 > 1367	AUG/UA	753	251	GGAGaTGAggaaaa atg
2	<i>sspE</i>	1436 > 1690	AUG/UA	255	85	GGAGGTGATaaca atg
3	<i>ygaB</i>	1779 > 2123	AUG/UA	345	115	GaAGGTctaCaga atg
4	<i>ygaC</i>	2267 > 2584	AUG/UA	318	106	AGAAAGtAGGgGAgaa atg
5	<i>ygaD</i>	2857 > 4626	AUG/UA	1770	590	AcAtGgGcTcAcacacat atg
6	<i>ygaE</i>	4703 < 5764	AUG/UA	1062	354	GAAtGcAcGcacaac atg
7	<i>gsaB</i>	5984 < 7273	UUG/UA	1290	430	TAtaGGAGGttttttctt gtt
8	<i>ygaF</i>	7426 > 7668	AUG/UA	243	81	GGAGGTaAcgta atg
9	<i>ygaG</i>	8023 > 8460	AUG/UA	438	146	AGAAGaGAGGTGcatgacg atg
10	<i>ygaH</i>	8701 > 9033	AUG/UA	333	111	GctGaGAcCga atg
11	<i>ygxA</i>	9057 > 9941	AUG/UA	885	295	AttGGAGGaatagtg atg
12	<i>ygaI</i>	16909 < 17685	AUG/UA	777	259	AaAAAGGgGaaGATCtaac atg
13	<i>ygaJ</i>	18427 > 19182	AUG/UA	756	252	GcttGGcGGgtgTga atg
14	<i>thiA</i>	19433 > 21205	AUG/UA	1773	591	AGAAAGGgGaaGgaaaAGta atg
15	<i>ygaK</i>	21243 < 22685	UUG/UA	1443	481	GAGGgtgTtgacactgtatcaga ttg
16	<i>senS</i>	22849 > 23046	AUG/UA	198	66	AAGGGgGGattattg atg
17	<i>katA</i>	23073 < 24524	AUG/UA	1452	484	Data not available

Fig. 1. For legend see facing page.

Table 2. Similarity of predicted ORF products to proteins in the database

ORF	Homologous protein	Identity/positives observed	BLAST score
1	7- α -Hydroxysteroid dehydrogenase (HsdH) (<i>Escherichia coli</i>)	42 % / 58 % in 82 residues	141
		43 % / 70 % in 30 residues	76
		24 % / 49 % in 57 residues	69
2	Partial spore protein (SspE) (<i>B. subtilis</i>)		
5	Hypothetical ATP transport in AcdA N-terminal region (<i>B. subtilis</i>)	55 % / 72 % in 225 residues	622
		24 % / 49 % in 279 residues	309
	Leukotoxin secretion ATP-binding protein (LktP) (<i>Pasteurella haemolytica</i>)	51 % / 70 % in 229 residues	614
	Cyclolysin secretion ATP-binding protein (CyaB) (<i>Bordetella pertussis</i>)	50 % / 67 % in 241 residues	572
7	Glutamate-1-semialdehyde aminotransferase (Gsa) (<i>Synechococcus</i> 6301)	56 % / 56 % in 1120 residues	1181
	Glutamate-1-semialdehyde aminotransferase (HemL) (<i>Salmonella typhimurium</i>)	56 % / 56 % in 922 residues	992
		60 % / 60 % in 105 residues	
	Glutamate-1-semialdehyde 2,1-aminomutase (HemL) (<i>B. subtilis</i>)	58 % / 58 % in 800 residues	1012
8	Bacterioferritin co-migratory protein (<i>Escherichia coli</i>)	40 % / 62 % in 61 residues	146
	Natural killer cell enhancing factor b (<i>Homo sapiens</i>)	58 % / 69 % in 43 residues	135
	Thiol-specific antioxidant protein (<i>Rattus norvegicus</i>)	53 % / 67 % in 43 residues	129
9	Ferric uptake regulation protein (<i>Yersinia pestis</i>)	31 % / 62 % in 54 residues	104
		28 % / 48 % in 56 residues	
		40 % / 55 % in 27 residues	
11	Partial hypothetical protein (YgxA) (<i>B. subtilis</i>)	100 % over terminal 157 bases	824
13	Peptidase E (PepE) (<i>Salmonella typhimurium</i>)	38 % / 55 % in 51 residues	81
14	ThiA		
15	Mitomycin radical oxidase (<i>Streptomyces lavendulae</i>)	36 % / 54 % in 77 residues	125
		33 % / 50 % in 60 residues	80
		23 % / 50 % in 53 residues	73
		23 % / 48 % in 56 residues	66
		56 % / 68 % in 16 residues	53
		24 % / 46 % in 41 residues	43
16	SenS		
17	KatA

Fig. 2. Changes in the relative positions of the genetic markers located downstream of *sspE*. The upper section is based on genetic data (Zeigler, 1993); the lower section is based on the sequence data covering the 24 kb region.

sequencing the *NotI* linking clone pNEXT47 (Itaya & Tanaka, 1991). The insert was sub-cloned into pBlue-script and sequenced. The position of the pNEXT47 insert was confirmed by a PCR reaction using *B. subtilis* 168 genomic DNA as template and primers specific to the pNEXT47 and 16S-S sequence data. The linking clone added just over 1 kb to the overall contig and overlapped the 16S-S sequence by 600 bp. This resulted in a final sequence of 24 524 kb determined in both strands with a mean base redundancy of 5.

All sequences were determined in both strands using DyeDeoxy Terminator Cycle Sequencing Kits with 373A and 373A STRETCH automated sequencers (Applied

Fig. 1. Restriction map and organization of the ORFs in the *sspE*-*katA* region of the *B. subtilis* 168 chromosome. The upper parts of the map show the location of the ORFs and their direction of transcription, together with the location of the rRNA and tRNA genes, and relevant *NotI* and *I-CeuI* restriction sites. The lower sections show the position of the *NotI* linking clone pNEXT47 (solid bar) with the overlapping diagnostic PCR product used to confirm its location (horizontal stripes) and the positions of the PCR products amplified from λ clones (vertical stripes) and genomic DNA templates (diagonal stripes) are also indicated.

Biosystems). These reactions were primed with either long versions of the M13 universal primers or custom-synthesized oligonucleotides.

DNA sequences obtained were compiled using the fragment assembly programs in the GCG package (Devereux *et al.*, 1984). The compiled sequences were analysed for the location of possible ORFs. Protein sequences of each ORF were used to search the relevant databases using the MoBiCS version of BLAST (Altschul *et al.*, 1990).

Analysis of the 24.5 kb sequence revealed the tRNA gene set *trnD*, containing 16 tRNA genes with anticodons corresponding to 15 aa, Leu being represented twice (Wawrousek *et al.*, 1984), and the rRNA operon *rrnD* containing 16S, 23S and a 5S minor rRNA species (Raué *et al.*, 1977). The *rrnD* operon features an I-CeuI restriction site which has been used to identify and map the relative positions of the rRNA operons (Toda & Itaya, 1995). A search of the remaining sequence in all six reading frames revealed 17 ORFs. All the ORFs, with the exception of *kata* (in the absence of upstream sequence data), were preceded by a Shine–Dalgarno sequence complementary to the 3' end of the 16S rRNA of *B. subtilis* (3' UCUUCCUCCACUAG 5'; McLaughlin *et al.*, 1981). Information on the position, size, direction of transcription and putative ribosome-binding sites of each of the 17 ORFs is given in Table 1.

Of the 17 ORFs, four had been identified previously: ORF2 as *sspE* (Hackett & Setlow, 1987; EMBL M16184); ORF 14 as *thiA* (EMBL U26178); ORF16 as *senS* (Wang & Doi, 1990; EMBL M34826) and ORF17 as *kata* (EMBL M80796). Only minor differences between these data and this sequence were noted. However, upon translation of the *thiA* gene reported here, residues 170–174 are changed from RLFLP to AIIPS due to frame shifts, residue 333 is changed from L to F and residue 547 from G to C. Five of the ORFs present show no strong homology to any protein sequence in the databases. The remaining eight do show significant similarity and these data are presented in Table 2.

Probably the most interesting predicted protein product is that of ORF7, which starts with the alternative initiation codon TTG. Its deduced product shows 50% sequence identity at the amino acid level with the glutamate-1-semialdehyde aminotransferase from *Synechococcus* (Grimm *et al.*, 1991). Expression of glutamate-1-semialdehyde aminotransferase is of interest as it represents a second isozymic activity in *B. subtilis*. The product of the *hemL* gene of *B. subtilis* is also a glutamate-1-semialdehyde aminotransferase (Hansson *et al.*, 1991), with which ORF7 shares 48% sequence identity. The *hemAXCDBL* operon of *B. subtilis* encodes the enzymes responsible for the biosynthesis of the tetrapyrrole uroporphyrinogen III from glutamate. Internal deletion of the *hemL* gene, however, did not result in a growth requirement for haem or its precursor 5-aminolevulinic acid (Hansson *et al.*, 1991). It would seem likely that the second glutamate-1-

semialdehyde aminotransferase gene reported here can fulfill this function *in vivo*.

Not many genetic markers have been located in this region (Zeigler, 1993). However, the relative map positions of those that are present are incorrect. The changes in the relative positions are shown in Fig. 2. The genetic markers *comO* (formally *com-18*; Biauudet *et al.*, 1996) and *recH*, which also map in this region, could not be easily ascribed to any of the ORFs located within the sequence.

ACKNOWLEDGEMENTS

This work was supported by the Commission of the European Communities Biotechnology Programme (contract BI02-CT94-2011)

REFERENCES

- 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.
- Biauudet, V., Samson, F., Anagnostopoulos, C., Ehrlich, S. D. & Bessires, P. (1996). Computerized genetic map of *Bacillus subtilis*. *Microbiology* **142**, 2669–2729.
- Devereux, J., Haeberli, P. & Smithies, O. (1984). A comprehensive set of sequence analysis programmes for the VAX. *Nucleic Acids Res* **12**, 387–395.
- Grimm, B., Bull, A. & Breu, V. (1991). Structural genes of glutamate-1-semialdehyde aminotransferase for porphyrin synthesis in a cyanobacterium and *E. coli*. *Mol Gen Genet* **225**, 1–10.
- Hackett, R. H. & Setlow, P. (1987). Cloning, nucleotide sequencing, and genetic mapping of the gene for small, acid-soluble spore protein γ of *B. subtilis*. *J Bacteriol* **169**, 1985–1992.
- Hansson, M., Rutberg, L., Schrder, I. & Hederstedt, L. (1991). The *Bacillus subtilis* *hemAXCDBL* gene cluster, which encodes enzymes of the biosynthetic pathway from glutamate to uroporphyrinogen III. *J Bacteriol* **173**, 2590–2599.
- Itaya, M. & Tanaka, T. (1991). Complete physical map of the *Bacillus subtilis* 168 chromosome constructed by a gene-directed mutagenesis method. *J Mol Biol* **169**, 1911–1916.
- McLaughlin, J. R. Murray, C. L. & Rabinowitz, J. C. (1981). Unique features in the ribosome-binding site sequence of the Gram-positive *Staphylococcus aureus* β -lactamase gene. *J Biol Chem* **250**, 11283–11291.
- Maniatis, T., Sambrook, J. & Fritsch, E. F. (1982). *Molecular Cloning: a Laboratory Manual*. Cold spring Harbor, NY: Cold Spring Harbor Laboratory.
- Rau, H. A., Rosner, A. & Planta, R. J. (1977). Heterogeneity of the genes coding for 5S RNA in three related strains of the genus *Bacillus*. *Mol Gen Genet* **156**, 185–193.
- Toda, T. & Itaya, M. (1995). I-CeuI recognition sites in the *rrn* operons of the *B. subtilis* 168 chromosome: inherent landmarks for genome analysis. *Microbiology* **141**, 1937–1945.
- Wang, L.-F. & Doi, R. H. (1990). Complex character of *senS*, a novel gene regulating expression of extracellular-protein genes of *B. subtilis*. *J Bacteriol* **172**, 1939–1947.

Wawrousek, E., Narasimhan, N. & Hansen, N. (1984). Two large clusters with thirty-seven transfer RNA genes adjacent to ribosomal RNA gene sets in *Bacillus subtilis*. *J Biol Chem* **259**, 3694–3702.

Zeigler, D. R. (1993). Genetic maps: locus maps of complex genomes. In *Genetic Map of Bacillus subtilis*, Book 2, 6th edn,

2.34–2.59. Edited by S. J. O'Brian. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.

Received 14 November 1996; revised 2 February 1997; accepted 17 February 1997.