The *Bacillus subtilis* genome from *gerBC* (311°) to *licR* (334°)

E. Presecan,¹ I. Moszer,¹ L. Boursier,¹ H. Cruz Ramos,¹ V. de la Fuente,² M.-F. Hullo,^{1,2} C. Lelong,¹ S. Schleich,¹ A. Sekowska,¹ B. H. Song,^{1†} G. Villani,¹ F. Kunst,² A. Danchin¹ and P. Glaser¹

Author for correspondence: Philippe Glaser. Tel: +33 1 45 68 84 41. Fax: +33 1 45 68 89 48.
e-mail: pglaser@pasteur.fr

^{1,2} Unité de Régulation de l'Expression Génétique¹, and Unité de Biochimie Microbienne², Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France

As part of the international project to sequence the *Bacillus subtilis* genome, the DNA region located between *gerBC* (311°) and *licR* (334°) was assigned to the Institut Pasteur. In this paper, the cloning and sequencing of 176 kb of DNA and the analysis of the sequence of the entire 271 kb region (6.5% of the *B. subtilis* chromosome) is described; 273 putative coding sequences were identified. Although the complete genome sequences of seven other organisms (five bacteria, one archaeon and the yeast *Saccharomyces cerevisiae*) are available in public databases, 65 genes from this region of the *B. subtilis* chromosome encode proteins without significant similarities to other known protein sequences. Among the 208 other genes, 115 have paralogues in the currently known *B. subtilis* DNA sequences and the products of 178 genes were found to display similarities to protein sequences from public databases for which a function is known. Classification of these genes shows a high proportion of them to be involved in the adaptation to various growth conditions (non-essential cell wall constituents, catabolic and bioenergetic pathways); a small number of the genes are essential or encode anabolic enzymes.

Keywords: genome sequencing, *Bacillus subtilis*, intergenic repeated element, orthologues/paralogues, spore coat

INTRODUCTION

The systematic sequencing of the chromosome of the Gram-positive bacterium *Bacillus subtilis* is being conducted by an international consortium of European and Japanese laboratories (Kunst *et al.*, 1995; Ogasawara & Yoshikawa, 1996). By April 1997, 95% of the sequence was determined, and the analysis of several long regions has been published (e.g. Ogasawara *et al.*, 1994; Albertini *et al.*, 1995; Mizuno *et al.*, 1996; Soldo *et al.*, 1996; Sorokin *et al.*, 1996; Wipat *et al.*, 1996; Yoshida *et al.*, 1996). The completion of this genome sequence will be the first complete genetic determination of a sporulating *Bacillus* sp. with a high level of competence and recombination potential. As for the other completely

known genomes, the Gram-negative bacteria *Haemophilus influenzae* (Fleischmann *et al.*, 1995) and *Escherichia coli* (O'Brien, 1997), the mycoplasmas *Mycoplasma genitalium* (Fraser *et al.*, 1995) and *Mycoplasma pneumoniae* (Himmelreich *et al.*, 1996), the cyanobacterium *Synechocystis* PCC 6803 (Kaneko *et al.*, 1996), the archaeon *Methanococcus jannaschii* (Bult *et al.*, 1996) and the yeast *Saccharomyces cerevisiae* (Goffeau *et al.*, 1996), about one-third of the genes in the *B. subtilis* chromosome are 'unknown' genes (Moszer *et al.*, 1996), i.e. they encode proteins without similarities to proteins with a known function. To elucidate their role, an original international research project of systematic functional analysis was set up, using the approach of reverse genetics for which *B. subtilis* is particularly suited (Harwood & Wipat, 1996). Accessibility of the whole genetic information of an organism associated with the knowledge of other genome sequences opens new possibilities for sequence analysis, for example classification of genes according to their codon usage (Moszer *et al.*, 1995). It becomes possible to study the distribution of nucleotide motifs along the chromosome

† Present address: Department of Biology, Teachers College, Kyung-Pook University, TAEGU, 702-701, South Korea.

Abbreviation: PTS, phosphotransferase system.

The EMBL/GenBank/DDBJ accession numbers for the sequence reported in this paper are Z49782, Z49792, Z49884, Z80355, Z82987, Z83337, Z92952, Z92953, Z92954 and Z93767.

and thereby get insights into regulatory processes (Hénaut *et al.*, 1996), and to compare chromosomal gene clusters conserved during evolution in different organisms. These analyses may reveal functions for unknown genes and new functions for already known genes (Danchin, 1996). For this kind of analysis appropriate computer tools will be required.

Several lines of investigation have been opened or speeded up as a direct consequence of the recent availability of *B. subtilis* sequence data. An interesting area of research in *B. subtilis* is the sporulation pathway (for a recent review see Stragier & Losick, 1996). The knowledge of the whole genome sequence will contribute to the discovery of new genes involved in this bacterial cell differentiation process. In particular, several important sporulation loci were discovered in the described region (Perego *et al.*, 1994, 1996; Londoño-Vallejo & Stragier, 1995; Karow *et al.*, 1995).

This paper describes the features of the *B. subtilis* chromosomal region initially assigned to our group (Kunst *et al.*, 1995). This 271 kb region between *gerBC* (311°) and *licR* (334°) covers 6.5% of the chromosome. It includes a region of 97 kb which has been sequenced and published (Glaser *et al.*, 1993).

METHODS

Bacterial strains. The same isolate of *B. subtilis* strain 168 provided by C. Anagnostopoulos is used by all the participant groups in the European *B. subtilis* genome sequencing project. This strain was the source for the chromosomal DNA used to construct a lambda library (Kunst & Devine, 1991) and for all the other cloning methods. Direct cloning by plasmid rescue was performed in the *E. coli* *pcnB* strain TP611 (Glaser *et al.*,

1993). Strain P2392 was used for propagation of the lambda library (Silhavy *et al.*, 1984). *E. coli* strains used for DNA sequencing were XL-1 Blue (Bullock *et al.*, 1987) for dsDNA preparations and TG1 for ssDNA preparations (Gibson, 1984).

Media. Luria-Bertani (LB) medium was used for standard cultures of *B. subtilis* and *E. coli* (Miller, 1972). 2YT medium was used for M13 recombinant phage preparation. Lambda stocks were prepared from LB plates (15 g agarose l⁻¹) using LB overlays containing 7 g agarose l⁻¹. Antibiotics were added to the following final concentrations, when necessary: for *E. coli*, 100 mg ampicillin l⁻¹; for *B. subtilis*, 5 mg chloramphenicol l⁻¹, 10 mg erythromycin l⁻¹.

Phages and plasmids. Gene libraries of *B. subtilis* DNA were constructed in Lambda FIX II (Stratagene) (Kunst & Devine, 1991). DNA sequences were determined from subclones in the phage M13mp8 (Messing & Vieira, 1982) or in pUC18 (Yanisch-Perron *et al.*, 1985). The plasmid rescue method was performed after an initial cloning step in pDIA5304 (Glaser *et al.*, 1993) or pDIA5305. pDIA5305 was obtained by insertion of the pC194 *cat* gene (Horinouchi & Weisblum, 1982) into the unique *NaeI* restriction site of pMTL22 (Chambers *et al.*, 1988). The high-copy-number *B. subtilis* replicon pMTL500E (Swinfield *et al.*, 1990) was used for marker rescue in *B. subtilis*.

DNA manipulations. *E. coli* transformation was generally performed as described by Chung & Miller (1988). Shotgun libraries in M13mp8 or in pUC18 were used to transform *E. coli* XL-1 Blue as described by Hanahan (1983). Recombinant plasmids were transferred to *E. coli* TP611 by the calcium chloride transformation method (Sambrook *et al.*, 1989). *B. subtilis* cells were transformed as described by Kunst & Rapoport (1995). Southern blots, and plaque and colony transfers were performed as described by Sambrook *et al.* (1989). Membranes were further hybridized with non-radioactively labelled probes (Boehringer DIG-dUTP labelling).

Fig. 1. Cloning of the *licR-gerBC* region. The thin upper line represents the genetic map with divisions representing chromosomal degrees. Relevant genetic markers are indicated. Underneath this line, the physical map is represented by a thicker line. Restriction site positions are deduced from the sequence and are in good agreement with the physical mapping of Itaya & Tanaka (1991). The thick dotted line represents the 97 kb region previously sequenced (Glaser *et al.*, 1993). The short dotted line represents the single *B. subtilis* plasmid derivative. Numbers refer to the *E. coli* pDIA plasmids used for sequencing the newly sequenced region.

Table 1. Boundaries of the 17 cloned fragments

Plasmid/phage	Beginning*	End*
pDIA 5363	1 (<i>Xba</i> I)	7851 (<i>Eco</i> RV)
97 kb ^a	6543	103557
pDIA 5364	101448 (<i>Eco</i> RV)	108016 (<i>Hind</i> III)
pDIA 5365	107788 (<i>Not</i> I)	128777 (<i>Eco</i> RV)
pDIA 5348	126849 (<i>Xba</i> I)	142330 (<i>Ssp</i> I)
pDIA 5353	142069 (<i>Kpn</i> I)	152949 (<i>Eco</i> RI)
pPP41 ^b	151210 (<i>Bgl</i> II)	157518 (<i>Bgl</i> II)
pDIA 5354	156207 (<i>Pst</i> I)	166327 (<i>Sal</i> I)
pDIA 5355	162280 (<i>Pst</i> I)	168828 (<i>Eco</i> RV)
pDIA 5329	167912 (<i>Not</i> I)	178399 (<i>Hind</i> III)
<i>λnarA</i>	177261 (<i>Sau</i> 3A)	195141 (<i>Sau</i> 3A)
pDIA 5366	194576 (<i>Hinc</i> II)	205685 (<i>Sac</i> I)
pDIA 5367	204509 (<i>Pst</i> I)	216491 (<i>Kpn</i> I)
pDIA 5368	214619 (<i>Swa</i> I)	228235 (<i>Xba</i> I)
pDIA 5369	227320 (<i>Sac</i> II)	241670 (<i>Apa</i> I)
pDIA 5370	240641 (<i>Bgl</i> II)	255853 (<i>Sac</i> I)
pDIA 5371	253248 (<i>Kpn</i> I)	263953 (<i>Eco</i> RI)
pDIA 5372	263330 (<i>Sac</i> II)	271513 (<i>Pst</i> I)

* First base of the restriction site.

References: ^aGlaser *et al.* (1993); ^bTrach *et al.* (1988).

Chromosome walking experiments involved either a plasmid rescue method, using integrative plasmids (Niaudet *et al.*, 1982), or a marker rescue method after chromosomal integration of a selective marker (*Cm*^R) (Glaser *et al.*, 1993). Fifteen walking steps (Fig. 1 and Table 1) were performed as follows. A single walking step was performed from the *ywaA* end of the 97 kb contig. The resulting plasmid (pDIA5363) contains the whole *lic* operon (Tobisch *et al.*, 1997). From plasmid pPP41 containing the *spoOF* gene (Trach *et al.*, 1988), four successive walking steps were performed in the direction of the *pta* gene, leading to plasmids pDIA5353, pDIA5348, pDIA5365 and pDIA5364. pDIA5364 has a 2.1 kb overlap with the 97 kb contig. In the other direction a single walking step was possible (pDIA5354). From *λnarA*, only a single step could be performed in the direction of the pPP41 insert (pDIA5329). The region (only 1.6 kb long) between the pDIA5354 and pDIA5329 inserts could not be cloned into *E. coli* by this method. It was therefore cloned in *B. subtilis* using the pMTL500E plasmid (pDIA5355). From the other end of *λnarA*, five successive walking steps were performed, giving plasmids pDIA5366–5370. Finally, two walking steps were performed from the cloned *gerBC* gene, giving plasmids pDIA5372 and pDIA5371. Plasmids pDIA5370 and pDIA5371 overlap by 2.6 kb.

Shotgun cloning and DNA sequencing. Fifteen shotgun sequencing experiments were performed independently. The corresponding strategy has already been described (Moszer *et al.*, 1991; Glaser *et al.*, 1993).

Data handling and computer analysis. DNA sequences were compiled using the XBAP program of R. Staden (Dear & Staden, 1991). Sequences were analysed using the DNA Strider 1.1 software (Marck, 1988). The BLAST program (Altschul *et al.*, 1990) was used to search for similar sequences in the non-redundant database (5 April 1997) at the Institut Pasteur computer centre. The CLUSTAL W program was used for the multiple alignment of the repeated DNA element (Higgins

et al., 1992). *B. subtilis* sequences were handled using the *SubtiList* database (Moszer *et al.*, 1995).

RESULTS AND DISCUSSION

Cloning and sequencing of the *gerBC*–*licR* region

We have previously described the cloning of a 97 kb region from *ywaA* to *ywfO* as four lambda clones, three *E. coli* plasmids and two *B. subtilis* plasmids (Glaser *et al.*, 1993). For the remaining part of the assigned region, one fragment was cloned as a lambda insert isolated using the cloned *narA* gene as a probe (*λnarA*) (Santana *et al.*, 1994). To obtain the rest of the region, a plasmid rescue walking strategy was used (Glaser *et al.*, 1993). Successive walking steps were performed, starting from ends of the cloned fragments of the region as described in Methods. The new regions were obtained as 15 *E. coli* plasmids, one lambda clone and one *B. subtilis* plasmid (Fig. 1 and Table 1). The sequences of 15 fragments were determined after shotgun subcloning. Only the 1.6 kb region of plasmid pDIA5355 was sequenced by directed subcloning and primer walking. The extensively studied *spoOF* region (Trach *et al.*, 1988) was only partially resequenced. The chromosomal region from *licR* to *gerBC* is 271518 bp long and overlaps the regions sequenced in Dr Y. Fujita's laboratory in Hiroshima (Japan) (Yoshida *et al.*, 1996) and in Dr D. Karamata's laboratory in Lausanne (Switzerland) (Soldo *et al.*, 1996).

Identification of coding sequences and general analysis

The putative coding sequences in the six reading frames were identified according to previously described rules (Glaser *et al.*, 1993), and 273 coding sequences were predicted. They were 39 codons (*phrF*) to 1228 codons (*narG*) long with a mean length of 875 bp (Fig. 2). Compared to other systematic sequencing projects, a relatively large number of small genes was predicted: 19 genes were shorter than 100 codons. Of these, 10 encode proteins with known function (*dltC*, *rpmE*, *atpE*, *spoIIID*, *sbo*) or similar to known proteins (*ywbE*, *ywhB*, *phrF*, *ywmG*, *ywqI*), and for nine (*ywcE*, *ywdA*, *ywhR*, *ywjC*, *ywkF*, *ywmE*, *ywqO*, *ywsA*, *ywtC*) no counterpart could be identified in the databases. A majority of the genes are transcribed in the direction of the replication fork movement (189 of 273, 69%), but their distribution is not uniform. Indeed, three regions with a majority of genes in the opposite direction could be defined (from *ywaA* to *ywbH*, from *ywhB* to *sbo*, and from *ywqM* to *rsbR*; see Fig. 2). Because of the limited size of the region, the significance of this observation could not be assessed and analysis awaits the complete genome sequence.

Operon organization and non-coding regions

The 273 coding sequences were grouped into 147 putative transcription units by analysis of the intergenic regions, mainly by the identification of the putative

Fig. 2. Coding sequence organization in the *B. subtilis* *licR-gerBC* chromosomal region. A number was assigned to each gene corresponding to its rank in the sequence. Genes are represented by arrows with different fill-in motifs, to distinguish them according to their classification (see Table 2): ■, cell envelope and cellular processes; ▨, intermediary metabolism; ▩, information pathways; ▧, other functions; ▦, miscellaneous; ▤, unknown genes similar to other genes of unknown function; □, unknown genes without similarities. The direction of the replication complex is from left to right. Putative transcription terminators are indicated by ϕ.

transcription terminators and by taking into account their length. Eighty-two of these transcription units were predicted to consist of a single gene. Fifty-seven may be involved in translational coupling as assessed from the relative positions of the ribosome-binding site and the upstream gene stop codon. The longest operon is the *sps* operon, which is comprised of 11 genes. In the region, several operons of known function, including the *atp* (Santana *et al.*, 1994), the *narGHJI* (Cruz Ramos *et al.*, 1995) and the *qox* (Santana *et al.*, 1992) operons, have a conserved organization in a broad range of bacteria. However, in the case of the urease operon, only the structural genes *ureABC* were found, and genes encoding nickel accessory proteins, required for nickel incorporation into the enzyme, are missing (Cruz Ramos *et al.*, 1997). In the thermophilic *Bacillus* sp. strain TB-90, these genes (*ureEFGDH*) map downstream from the *ureC* gene in the *ure* operon (Maeda *et al.*, 1994). In some cases, such as *ywbHG* in *B. subtilis* and *yohJK* in *E. coli*, the transcriptional organization of unknown genes is also conserved. Three genes with unknown function, *ywqH*, *ywqI* and *ywqJ*, are similar to three other *B. subtilis* genes, *ysiB*, *ysiC* and *ysiD*, respectively, and in both cases the three genes seem to be organized in an operon.

As for other long segments of the *B. subtilis* chromosome, the overall coding capacity of the region is high (88%). Only 10 intergenic regions are longer than 500 bp. In seven cases they are located between divergent coding sequences: *dltA* and *ywaB* (613 bp), *ywcB* and *ywcC* (957 bp), *vpr* and *ywcl* (530 bp), *ywfN* and *ywfO* (601 bp), *ywmA* and *ywmB* (529 bp), *ywqB* and *ywqC* (581 bp), *rbsR* and *ywsC* (784 bp), and in three cases between genes in the same orientation: *galT* and *qoxA* (835 bp), *mmr* and *thrZ* (595 bp), *ywmG* and *ureA* (1000 bp). Most of these regions may be regulatory. Alternatively, they may contain short genes devoid of typical ribosome-binding site sequences, or may be transcribed as non-translated RNA.

Systematic comparison with known nucleotide sequences using the intergenic regions as query sequences

revealed that two intergenic regions (*ywmE*–*narQ* and *ywbN*–*ywbO*) contained a repeated element. This sequence was found six times in the publicly available *B. subtilis* sequences (Fig. 3). This element, named *Bs-rep* was described in 1994 as being repeated three times in the sequences available at that time (Popham & Setlow, 1994). The six repeats are located in less than one-third of the chromosome (from position 3230 kb to position 513 kb) and are all in the same orientation with respect to the replication fork movement. Comparisons with sequence databases show that this element is extremely well conserved in two sequences from the phylogenetically closely related bacterium: *Bacillus licheniformis* (Fig. 3). Furthermore, two classes could be distinguished on the basis of sequence similarities: *ydbT*–*ydcA* and *ywmE*–*narQ* in one group and the remaining six sequences including the two *B. licheniformis* sequences in the other. The nature (IS element or scar of a mobile element) and the possible role of this repeated sequence are unknown.

Gene function analysis

At the beginning of the sequencing project, only eight genes were known in this region of the *B. subtilis* chromosome. Since then, independent of our sequencing project, 25 other genes have been characterized. Most are involved in two aspects of bacterial life: carbon and nitrogen source utilization and sporulation. *sacXY* (Zukowski *et al.*, 1990) and *sacTPA* (Fouet *et al.*, 1986, 1987) encode proteins involved in saccharose metabolism. The *rbs* operon is involved in ribose utilization (Woodson & Devine, 1994). Two exocellular proteases are encoded by the *epr* (Bruckner *et al.*, 1990) and *vpr* (Sloma *et al.*, 1991) genes. The *als* operon (*alsRSD*; Renna *et al.*, 1993) encodes proteins required for the synthesis of acetoin as a by-product, and the *nrgAB* operon (Wray *et al.*, 1994) encodes an ammonium permease (NrgA). Seven sporulation genes have been analysed: *spoOF* (Yoshikawa *et al.*, 1986), *spoIIQ*

Fig. 3. Alignment of *Bs-rep* repeats. Gene names indicate the genes bracketing each of the repeated elements. *Bl.*, *B. licheniformis*. *B. subtilis* sequences were extracted from the SubtilList database (Moszer *et al.*, 1995) and those of *B. licheniformis* from EMBL (accession nos Z80222 and D31955). Strictly conserved nucleotides found in at least four sequences are represented as black boxes.

Fig. 4. Sequence comparisons between the *cotB* and the *cotG* genes. (a) Kyte-Doolittle plots of the amino acid hydropathy (Kyte & Doolittle, 1982). (b) Distribution of the acidic (A) and basic (B) amino acids. (c) Alignments of the 13 or 14 amino acid repeated sequences of each protein and the corresponding consensus sequences.

(Londoño-Vallejo *et al.*, 1997), *spoIID* (Lopez-Diaz *et al.*, 1986), *spoIIID* (Kunkel *et al.*, 1989), *cotG* (Sacco *et al.*, 1995), *cotH* (Naclerio *et al.*, 1996) and *cotB* (Donovan *et al.*, 1987).

Only the first 72 codons of *cotB*, which encodes a protein located in the outer coat of the spore, were known. The analysis of its complete sequence (Fig. 4) revealed that this protein may have two domains. The 150 residues C-terminal domain is hydrophilic and has a low complexity organization (two types of repeated unit of 13 and 14 amino acids, respectively), similar to that of CotG (Sacco *et al.*, 1995; Fig. 4). Although the sequence identity is low, both types of repeated unit are serine-rich and highly charged. The similar organization may reflect a conserved structural domain in both coat proteins. It has been reported that the CotB protein is absent from a *cotG* mutant strain (Sacco *et al.*, 1995). It is therefore possible that both proteins are part of a complex or molecular assembly involving these repeated structures.

Eight other genes with various activities from the *licR-gerBC* region have been described by other groups. They determine or encode methylenomycin A resistance (*mmr*; Putzer *et al.*, 1992), one of the two fructose-bisphosphate aldolases (*fbaA*; Trach *et al.*, 1988; Mitchell *et al.*, 1992), the minor threonine-tRNA synthetase (*thrZ*; Putzer *et al.*, 1990), the non-essential RNA polymerase δ subunit (*rpoE*; Lampe *et al.*, 1988),

the transcription termination factor Rho (*rho*; Quirk *et al.*, 1993), the CTP synthase (*ctrA*; Trach *et al.*, 1988), the thymidine kinase (*tdk*; Trach *et al.*, 1988) and a morphogene gene (*mbi*; Abhayawardhane & Stewart, 1995).

Fifty-four genes discovered during the systematic sequencing of this part of the chromosome were analysed either in our laboratory, or in collaboration with other groups. Several loci are involved in energy conversion: the *atp* operon (encoding the nine subunits of the F_0F_1 ATPase; Santana *et al.*, 1994), the *gox* operon (encoding the four subunits of the major aa_3 -type terminal oxidase; Santana *et al.*, 1992), the *narGHJI*, *narK* and *narX* genes (involved in the utilization of nitrate as an alternative terminal electron acceptor; Cruz Ramos *et al.*, 1995; Glaser *et al.*, 1995), and *fnr* (a regulatory gene for anaerobic respiration; Cruz Ramos *et al.*, 1995). Five loci are involved in carbon and nitrogen source utilization: lichenan degradation (the *lic* operon; Tobisch *et al.*, 1997), fatty acid utilization (acyl-CoA dehydrogenase, *acdA*; V. de la Fuente and others, unpublished), acetate production as a by-product (phosphotransacetylase, *pta*; E. Presecan and others, unpublished), urea utilization (the *ure* operon; Cruz Ramos *et al.*, 1997), and arginine, ornithine and proline degradation (the *roc* operon; Calogero *et al.*, 1994). The *dlt* and the *sps* operons are involved in the synthesis of cell wall components: D-alanylation of cell wall lipoteichoic and

Table 2. Classification of the genes of the *B. subtilis* *licR*–*gerBC* region

Gene name	Product size (aa)	Total no. paralogues	BLAST score	Genes encoding similar protein/similar gene products*	Function/activity of the gene product or similar protein	Gene no.†
1. Cell envelope and cellular processes						
1.1. Cell wall components						
<i>dltE</i>	252	9	2.4e–21	<i>ydfG</i> , <i>Escherichia coli</i> (sw:P39831)	D-Alanine esterification of lipoteichoic acid	7
			7.5e–18	<i>trn2</i> , <i>Hyoscyamus niger</i> (sw:P50164)		
<i>dltD</i>	392	0	1.6e–42	<i>dltD</i> , <i>Staphylococcus aureus</i> (embl:D86240)	D-Alanine esterification of lipoteichoic acid	8
<i>dltC</i>	78	0	1.6e–31	<i>dltC</i> , <i>Staphylococcus aureus</i> (embl:D86240)	D-Alanine esterification of lipoteichoic acid (carrier)	9
<i>dltB</i>	395	0	3.5e–159	<i>dltB</i> , <i>Staphylococcus aureus</i> (embl:D86240)	D-Alanine esterification of lipoteichoic acid	10
<i>dltA</i>	503	10	2.3e–153	<i>dltA</i> , <i>Lactobacillus casei</i> (sw:P35854)	D-Alanine esterification of lipoteichoic acid	11
<i>ywhK</i>	451	1	4.4e–20	<i>orf492</i> , <i>Methanosarcina mazei</i> (embl:X84710)	Surface antigen	114
			1.3e–137	<i>ywhL</i> , <i>Bacillus subtilis</i> (embl:Z80360)		
<i>ywhL</i>	458	1	6.9e–20	<i>orf375</i> , <i>Methanosarcina mazei</i> (embl:X84710)	Surface antigen	115
			7.4e–138	<i>ywhK</i> , <i>Bacillus subtilis</i> (embl:Z80360)		
<i>murA</i>	436	1	5.4e–139	<i>murZ</i> , <i>Synechocystis</i> sp. PCC 6803 (embl:D6400)	UDP-N-acetylglucosamine 1-carboxyvinyl transferase	181
<i>mbi</i>	313	2	9.1e–182	<i>mbi</i> , <i>Bacillus cereus</i> (sw:P32444)	Rod-shape-determining protein	215
<i>ywqC</i>	248	1	1.0e–56	<i>capA</i> , <i>Staphylococcus aureus</i> (sw:P39850)	Required for the biosynthesis of type 1 capsular polysaccharide	230
<i>ywqD</i>	237	1	2.0e–57	<i>capB</i> , <i>Staphylococcus aureus</i> (sw:P39851)	Required for the biosynthesis of type 1 capsular polysaccharide	231
<i>ywqE</i>	254	0	2.6e–62	<i>capC</i> , <i>Staphylococcus aureus</i> (sw:P39852)	Required for the biosynthesis of type 1 capsular polysaccharide	232
<i>ywsC</i>	247	0	2.4e–184	<i>capB</i> , <i>Bacillus anthracis</i> (sw:P19580)	Synthesis of the polyglutamate capsule	266
<i>ywtA</i>	149	0	4.5e–84	<i>capC</i> , <i>Bacillus anthracis</i> (sw:P19581)	Synthesis of the polyglutamate capsule	267
<i>ywtB</i>	380	0	1.6e–124	<i>capA</i> , <i>Bacillus anthracis</i> (sw:P19579)	Synthesis of the polyglutamate capsule	268
<i>ywtD</i>	390	5	2.9e–21	<i>spr</i> , <i>Escherichia coli</i> (embl:D86610)	Lipoprotein precursor	270
			9.8e–20	<i>nlpC</i> , <i>Haemophilus influenzae</i> (sw:P45296)		
1.2. Transport/binding proteins and lipoproteins						
<i>licB</i>	102	1	2.8e–37	<i>celA</i> , <i>Bacillus stearothermophilus</i> (pir:B49898)	β-Glucoside permease IIB component (PTS)	2
<i>licC</i>	452	2	7.8e–253	<i>ywbA</i> , <i>Bacillus subtilis</i> (sw:P39584)	β-Glucoside permease IIC component (PTS)	3
			6.0e–221	<i>celB</i> , <i>Bacillus stearothermophilus</i> (pir:C49898)		
<i>licA</i>	110	1	4.9e–38	<i>celD</i> , <i>Bacillus stearothermophilus</i> (pir:E49898)	β-Glucoside permease IIA component (PTS)	4
<i>ywbF</i>	399	0	2.5e–108	<i>malA</i> , <i>Bacillus stearothermophilus</i> (pir:S43915)	Maltose permease	27
<i>ywcA</i>	516	2	7.5e–174	<i>dhlC</i> , <i>Xanthobacter autotrophicus</i> (pir:S52839)	Na ⁺ -dependent symport	37
<i>ywcJ</i>	256	1	4.5e–42	<i>nirC</i> , <i>Escherichia coli</i> (sw:P11097)	Nitrite transporter	54
<i>sacP</i>	460	4	1.8e–171	<i>ptsB</i> , <i>Klebsiella pneumoniae</i> (sw:P27219)	Sucrose permease IIBC component (PTS)	55
<i>rocC</i>	470	9	3.9e–215	<i>rocE</i> , <i>Bacillus subtilis</i> (sw:P39137)	Arginine utilization; permease	84
<i>ywfA</i>	412	13	1.2e–26	PID:g1742740, <i>Escherichia coli</i> (embl:D90810)	Arabinose polymer transport	85
<i>ywhE</i>	647	3	2.5e–143	<i>mrcA</i> , <i>Synechocystis</i> sp. PCC 6803 (embl:D90913)	Penicillin-binding protein	108
<i>ywhQ</i>	239	23	6.8e–33	<i>ecsA</i> , <i>Bacillus subtilis</i> (sw:P55339)	ABC-type transporter	120
<i>narK</i>	395	1	7.4e–43	<i>nasA</i> , <i>Bacillus subtilis</i> (sw:P42432)	Nitrite extrusion protein	126
			1.2e–66	MTCY04C12.22c, <i>Mycobacterium tuberculosis</i> (embl:Z81360)		
<i>ywjA</i>	575	26	3.7e–136	<i>ygaD</i> , <i>Bacillus subtilis</i> (embl:Z82044)	ABC-type transporter	135
<i>ywkB</i>	319	0	3.9e–37	<i>mdcE</i> , <i>Klebsiella pneumoniae</i> (embl:U56096)	Unknown transporter	154
<i>ywoA</i>	193	0	1.4e–25	<i>bcrC</i> , <i>Bacillus licheniformis</i> (sw:P42334)	Bacitracin transport permease protein	204
<i>nrgA</i>	404	0	1.1e–128	<i>amt</i> , <i>Mycobacterium tuberculosis</i> (sw:Q10968)	Ammonium transporter	206
<i>ywoD</i>	452	7	2.4e–28	<i>qacA</i> , <i>Staphylococcus aureus</i> (sw:P23215)	Drug-resistance protein	209
			1.7e–36	MTCY3G12.01, <i>Mycobacterium tuberculosis</i> (embl:Z79702)		
<i>ywoE</i>	490	1	1.4e–79	<i>fur4</i> , <i>Schizosaccharomyces pombe</i> (embl:X98696)	Uracil permease	210
			9.3e–135	PID:g1773191, <i>Escherichia coli</i> (embl:U82664)		
<i>ywoG</i>	396	9	2.7e–40	<i>norA</i> , <i>Synechocystis</i> sp. PCC 6803 (embl:D90909)	Quinolone-resistance protein	212
<i>ywrA</i>	178	1	6.5e–16	<i>chrA</i> , <i>Pseudomonas aeruginosa</i> (sw:P14285)	Essential for chromate resistance chromate transport	243
<i>ywrB</i>	197	1	7.0e–16	<i>chrA</i> , <i>Pseudomonas aeruginosa</i> (sw:P14285)	Essential for chromate resistance chromate transport	244
<i>ywrK</i>	442	1	2.8e–79	<i>arsB</i> , <i>Staphylococcus xylosus</i> (sw:Q01255)	Arsenical pump membrane protein	253
<i>rbsB</i>	305	0	2.1e–96	<i>rbsB</i> , <i>Escherichia coli</i> (sw:P02925)	Ribose transport operon; D-ribose-binding protein	260
<i>rbsC</i>	322	0	2.8e–117	<i>rbsC</i> , <i>Haemophilus influenzae</i> (sw:P44736)	Ribose transport operon; D-ribose-binding protein	261
<i>rbsA</i>	493	9	1.0e–149	<i>xylG</i> , <i>Escherichia coli</i> (sw:P37388)	Ribose transport operon; ATP-binding protein	262
<i>rbsD</i>	131	0	5.2e–39	<i>rbsD</i> , <i>Haemophilus influenzae</i> (sw:P44734)	Ribose transport operon; D-ribose-binding protein	263
<i>ywtG</i>	457		5.3e–143	<i>yxbC</i> , <i>Bacillus subtilis</i> (sw:P46333)	Metabolite transporter	273

Table 2 (cont.)

Gene name	Product size (aa)	Total no. paralogues	BLAST score	Genes encoding similar protein/similar gene products*	Function/activity of the gene product or similar protein	Gene no.†
1.4. Membrane bioenergetics; electron-transport chain; ATP synthase						
<i>qoxA</i>	322	1	2.2e-40	<i>qox-2</i> , <i>Acetobacter acetii</i> (sw:P50653)	Quinol oxidase polypeptide II precursor	43
<i>qoxB</i>	649	1	7.5e-247	<i>cyoB</i> , <i>Escherichia coli</i> (sw:P18401)	Quinol oxidase polypeptide I	44
<i>qoxC</i>	204	1	5.1e-70	<i>cox-3</i> , <i>Thermophilic bacterium PS-3</i> (sw:Q03439)	Quinol oxidase polypeptide III	45
<i>qoxD</i>	124	1	3.0e-09	<i>cyoD</i> , <i>Escherichia coli</i> (sw:P18403)	Quinol oxidase polypeptide IV	46
<i>ywcG</i>	249	1	1.4e-67	<i>frp</i> , <i>Vibrio harveyi</i> (embl:U08996)	NADPH:flavin oxidoreductase	49
<i>narG</i>	1228	0	0.0	<i>narG</i> , <i>Escherichia coli</i> (sw:P09152)	Nitrate reductase α subunit	130
<i>narH</i>	487	0	2.9e-229	<i>narH</i> , <i>Escherichia coli</i> (sw:P11349)	Nitrate reductase β subunit	131
<i>narJ</i>	184	0	3.3e-15	<i>narJ</i> , <i>Thiosphaera pantotropha</i> (pir:S61308)	Nitrate reductase protein J	132
			4.3e-11	<i>narJ</i> , <i>Escherichia coli</i> (sw:P11351)		
<i>narI</i>	223	0	1.8e-31	<i>narV</i> , <i>Escherichia coli</i> (sw:P19316)	Nitrate reductase subunit γ	133
			1.3e-62	MTCY04C12.21c, <i>Mycobacterium tuberculosis</i> (embl:Z81360)		
<i>atpI</i>	126	0	2.7e-38	<i>atpI</i> , <i>Bacillus megaterium</i> (sw:P20598)	ATP synthase I subunit	170
<i>atpB</i>	244	0	1.2e-128	<i>atpB</i> , <i>Bacillus megaterium</i> (sw:P20600)	ATP synthase a subunit	171
<i>atpE</i>	70	0	2.0e-30	<i>atpE</i> , <i>Bacillus megaterium</i> (sw:P20603)	ATP synthase c subunit	172
<i>atpF</i>	170	0	3.4e-67	<i>atpF</i> , <i>Bacillus megaterium</i> (sw:P20601)	ATP synthase b subunit	173
<i>atpH</i>	181	0	1.5e-59	<i>atpH</i> , <i>Bacillus stearothermophilus</i> (sw:P42008)	ATP synthase δ subunit	174
<i>atpA</i>	502	0	1.3e-300	<i>atpA</i> , <i>Bacillus megaterium</i> (sw:P17674)	ATP synthase α subunit	175
<i>atpG</i>	287	0	1.4e-143	<i>atpG</i> , <i>Bacillus megaterium</i> (sw:P20602)	ATP synthase γ subunit	176
<i>atpD</i>	473	0	1.5e-288	<i>atpD</i> , <i>Bacillus caldotenax</i> (sw:P41009)	ATP synthase β subunit	177
<i>atpC</i>	132	0	3.7e-60	<i>atpC</i> , <i>Bacillus megaterium</i> (sw:P12699)	ATP synthase ϵ subunit	178
<i>narQ</i>	262	0	1.2e-36	<i>fdhD</i> , <i>Wolinella succinogenes</i> (sw:P28181)	Unknown	186
<i>yurO</i>	175	2	6.3e-71	<i>yurK</i> , <i>Bacillus subtilis</i> (sw:P54439)	Hypothetical NAD(P)H oxidoreductase	257
1.5. Mobility and chemotaxis						
<i>flhO</i>	270	3	7.8e-20	<i>flgG</i> , <i>Caulobacter crescentus</i> (sw:Q06172)	Unknown; similar to flagellar hook-basal body	216
<i>flhP</i>	268	2	8.4e-45	<i>flgG</i> , <i>Bacillus subtilis</i> (sw:P23446)	Unknown; similar to flagellar hook-basal body	217
1.7. Cell division						
<i>ywcF</i>	393	2	4.5e-39	<i>rodA</i> , <i>Haemophilus influenzae</i> (sw:P44468)	Rod-shape-determining protein	48
1.8. Sporulation						
<i>spsA</i>	256	1	8.1e-69	<i>cgeD</i> , <i>Bacillus subtilis</i> (sw:P42092)	Spore-coat polysaccharide	69
<i>spsB</i>	472	0		No significant similarities	Spore-coat polysaccharide	70
<i>spsC</i>	389	2	3.4e-155	MJ1066, <i>Methanococcus jannaschii</i> (pir:A64433)	Spore-coat polysaccharide	71
<i>spsD</i>	289	0		No significant similarities	Spore-coat polysaccharide	72
<i>spsE</i>	373	0	3.1e-82	MJ1065, <i>Methanococcus jannaschii</i> (pir:H64432)	Spore-coat polysaccharide	73
<i>spsF</i>	239	0	1.2e-46	MJ1063, <i>Methanococcus jannaschii</i> (pir:F64432)	Spore-coat polysaccharide	74
<i>spsG</i>	222	0	7.9e-14	MJ1062, <i>Methanococcus jannaschii</i> (pir:E64432)	Spore-coat polysaccharide	75
<i>spsH</i>	117	0	3.5e-05	<i>flaR</i> , <i>Caulobacter crescentus</i> (embl:U27302)	Spore-coat polysaccharide	76
<i>spsI</i>	246	3	4.8e-65	<i>rhsA</i> , <i>Sphingomonas S88</i> (embl:U51197)	Spore-coat polysaccharide (glucose-1-phosphate thymidyltransferase)	77
<i>spsJ</i>	315	2	7.5e-115	<i>rfbB</i> , <i>Leptospira interrogans</i> (embl:U61226)	Spore-coat polysaccharide (UDP-N-acetylglucosamine epimerase)	78
<i>spsK</i>	432	0	2.5e-66	<i>rfbD2</i> , <i>Synechocystis</i> sp. PCC 6803 (embl:D90899)	Spore-coat polysaccharide (dTDP-6-deoxy-L-mannose-dehydrogenase)	79
<i>phrF</i>	39	1	2.3e-08	<i>phr</i> , <i>Bacillus subtilis</i> (embl:D50453)	Phosphatase regulator peptide	112
<i>rapF</i>	381	8	6.1e-118	<i>rapB</i> , <i>Bacillus subtilis</i> (embl:Z81356)	Response regulator aspartate phosphatase	113
<i>spoOF</i>	124	8	6.1e-62	<i>spoOF</i> , <i>Bacillus thuringiensis</i> (sw:P52942)	Sporulation	145
<i>spoIIR</i>	224	0		No significant similarities	Sporulation	161
<i>spoIID</i>	343	1	5.3e-180	SP2D, <i>Bacillus amyloliquefaciens</i> (sw:P13251)	Sporulation	182
<i>rapB</i>	377	8	1.4e-137	<i>rapA</i> , <i>Bacillus subtilis</i> (sw:Q00828)	Response regulator aspartate phosphatase	188
<i>spoIIQ</i>	283	0		No significant similarities	Sporulation	202
<i>spoIIID</i>	93	0		No significant similarities	Sporulation	214
<i>rapD</i>	354	8	1.3e-37	<i>rapC</i> , <i>Bacillus subtilis</i> (embl:D50453)	Response regulator aspartate phosphatase	218
<i>cotG</i>	195	0		No significant similarities	Spore-coat protein G	249
<i>cotH</i>	362	0		No significant similarities	Spore-coat protein H	250
<i>cotB</i>	380	0		No significant similarities	Spore-coat protein B	251
2. Intermediary metabolism						
2.1. Metabolism of carbohydrates, sugar and amino sugars						
<i>licH</i>	442	2	1.4e-147	<i>celF</i> , <i>Escherichia coli</i> (sw:P17411)	6-Phospho- β -glucosidase	5
<i>gspA</i>	286	0	3.0e-20	<i>lgtA</i> , <i>Rhizobium leguminosarum</i> (embl:X94963)	Galactosyltransferase	18
<i>ywbA</i>	444	2	8.9e-254	<i>licB</i> , <i>Bacillus subtilis</i> (sw:P46317)	Cellobiose permease IIC component (PTS)	22
<i>ywbC</i>	126	1	2.8e-10	<i>lguL</i> , <i>Lycopersicon esculentum</i> (tomato) (sw:Q42891)	Lactoylglutathione lyase	24

Table 2 (cont.)

Gene name	Product size (aa)	Total no. paralogues	BLAST score	Genes encoding similar protein/similar gene products*	Function/activity of the gene product or similar protein	Gene no.†
<i>galK</i>	390	0	6.0e–122	<i>gal-1</i> , <i>Lactobacillus helveticus</i> (sw: Q00052)	Galactokinase	41
<i>galT</i>	513	0	2.0e–122	<i>gal-7</i> , <i>Butyrivibrio fibrilvolvens</i> (sw: P45981)	Galactose-1-phosphate uridylyltransferase	42
<i>sacA</i>	480	2	6.5e–177	<i>surA</i> , <i>Bacillus stearothermophilus</i> (embl: U34872)	Sucrose-6-phosphate hydrolase	56
			1.6e–123	<i>scrB</i> , <i>Salmonella typhimurium</i> (sw: P37075)		
<i>ywfD</i>	255	19	3.7e–69	<i>dhg-2</i> , <i>Bacillus megaterium</i> (sw: P39483)	Glucose dehydrogenase	88
<i>ywfH</i>	259	0	7.3e–36	MTCY09F9.36, <i>Mycobacterium tuberculosis</i> (embl: Z84498)	Short-chain alcohol dehydrogenase family signature	92
			1.1e–28	<i>phbB</i> , <i>Alcaligenes eutrophus</i> (sw: P14697)	Acetoacetyl-CoA reductase	
<i>pta</i>	323	0	2.7e–97	<i>pta</i> , <i>Clostridium acetobutylicum</i> (embl: U38234)	Phosphate acetyltransferase	94
<i>ywbB</i>	62	0	3.6e–12	<i>dmpI</i> , <i>Pseudomonas putida</i> (pir: S24422)	4-Oxalocrotonate tautomerase	105
<i>ywjF</i>	705	0	5.1e–30	MJ0092, <i>Methanococcus jannaschii</i> (pir: D64311)	Fumarate reductase	140
<i>fbaA</i>	285	1	2.7e–111	<i>fbaB</i> , <i>Bacillus subtilis</i> (sw: P42420)	Fructose-1,6-diphosphate aldolase	146
<i>ywjH</i>	186	0	9.8e–72	MJ0960, <i>Methanococcus jannaschii</i> (embl: H64419)	Transaldolase	147
<i>murZ</i>	429	1	3.9e–128	<i>murZ</i> , <i>Synechocystis</i> sp. PCC 6803 (embl: D6400)	UDP-N-acetylglucosamine 1-carboxyvinyltransferase	148
<i>ywjI</i>	321	0	7.3e–112	<i>glpX</i> , <i>Escherichia coli</i> (sw: P28860)	Glycerol metabolism	149
<i>ywkA</i>	582	1	5.8e–183	<i>maoX</i> , <i>Saccharomyces cerevisiae</i> (sw: P36013)	Malate oxidoreductase	153
			2.0e–192	MTCY3G12.02c, <i>Mycobacterium tuberculosis</i> (embl: Z79702)		
<i>ywlF</i>	149	0	7.4e–44	<i>rpiB</i> , <i>Escherichia coli</i> (sw: P37351)	Ribose 5-phosphate isomerase B	166
<i>ywoC</i>	189	0	1.0e–12	<i>dhbB</i> , <i>Bacillus subtilis</i> (sw: P45747)	Isochorismatase	208
			9.3e–52	<i>yecD</i> , <i>Escherichia coli</i> (sw: P37347)		
<i>ywgF</i>	440	0	4.9e–137	<i>slr1299</i> , <i>Synechocystis</i> sp. PCC 6803 (embl: D90916)	UDPglucose dehydrogenase	233
<i>alsS</i>	571	3	6.0e–203	<i>als</i> , <i>Lactococcus lactis</i> (embl: L16975)	Acetolactate synthase (pH 6)	255
<i>alsD</i>	255	0	2.0e–66	<i>aldC</i> , <i>Enterobacter aerogenes</i> (sw: P05361)	Acetolactate decarboxylase	256
<i>rbsK</i>	293	0	3.7e–69	<i>rbsK</i> , <i>Haemophilus influenzae</i> (sw: P44331)	Ribose transport operon; ribokinase	264
2.2. Metabolism of amino acids and related molecules						
<i>ywaA</i>	363	0	1.5e–111	<i>ilvE</i> , <i>Mycobacterium tuberculosis</i> (sw: P10399)	Branched-chain amino acid aminotransferase	6
<i>ywaD</i>	455	0	5.3e–15	<i>ape3</i> , <i>Saccharomyces cerevisiae</i> (sw: P37302)	Proteinase precursor	14
<i>epi</i>	645	3	1.7e–91	<i>aprM</i> , <i>Bacillus</i> sp. strain B18' (pir: A48373)	Minor extracellular serine protease	21
<i>vpr</i>	806	6	2.6e–57	<i>subT</i> , <i>Bacillus amyloliquefaciens</i> (sw: P00782)	Minor extracellular serine protease	51
<i>yweB</i>	425	1	2.5e–148	<i>dhe-2</i> , <i>Clostridium difficile</i> (sw: P27346)	Glutamate dehydrogenase (NAD)	81
			7.3e–231	<i>ypcA</i> , <i>Bacillus subtilis</i> (sw: P50735)		
<i>rocA</i>	515	8	1.9e–166	<i>putA</i> , <i>Synechocystis</i> sp. PCC 6803 (embl: D90913)	1-Pyrroline-5-carboxylate dehydrogenase	82
			5.2e–259	<i>ycgN</i> , <i>Bacillus subtilis</i> (embl: D50453)		
<i>rocB</i>	566	1	6.1e–66	<i>yqjN</i> , <i>Bacillus subtilis</i> (sw: P54551)	Unknown	83
<i>ywfG</i>	399	4	3.9e–97	<i>aspC</i> , <i>Synechocystis</i> sp. PCC 6803 (embl: D90913)	Aspartate transaminase	91
<i>ywbF</i>	276	0	3.0e–96	MJ0313, <i>Methanococcus jannaschii</i> (pir: B64339)	Spermidine synthase	109
<i>ywbG</i>	290	2	2.4e–78	MJ0309, <i>Methanococcus jannaschii</i> (pir: F64338)	Agmatinase	110
<i>glyA</i>	415	0	1.5e–220	<i>glyA</i> , <i>Hyphomicrobium methylovorum</i> (pir: S30382)	Serine hydroxymethyltransferase	168
<i>ureA</i>	105	0	8.0e–46	<i>ureA</i> , <i>Bacillus pasteurii</i> (sw: P41022)	Urease γ subunit	191
<i>ureB</i>	124	0	2.4e–33	<i>ureB</i> , <i>Bacillus pasteurii</i> (sw: P41021)	Urease β subunit	192
<i>ureC</i>	569	0	8.0e–272	<i>ureC</i> , <i>Bacillus pasteurii</i> (sw: P41020)	Urease α subunit	193
<i>nrgB</i>	116	0	3.8e–28	<i>glnB</i> , <i>Klebsiella pneumoniae</i> (sw: P11671)	Nitrogen-regulated PII protein	205
<i>ywrD</i>	525	1	1.2e–87	<i>ggt</i> , <i>Escherichia coli</i> (sw: P18956)	γ -Glutamyltranspeptidase precursor	246
			7.5e–114	MTCY369.18, <i>Mycobacterium tuberculosis</i> (embl: Z80226)		
2.3. Metabolism of nucleotides and nucleic acids						
<i>ctrA</i>	535	0	4.1e–243	MJ1174, <i>Methanococcus jannaschii</i> (pir: E64446)	CTP synthase	143
<i>tdk</i>	195	0	1.3e–54	<i>kitH</i> , <i>Mycoplasma genitalium</i> (sw: P47280)	Thymidine kinase	152
<i>upp</i>	209	0	4.5e–125	<i>upp</i> , <i>Bacillus caldolyticus</i> (embl: X99545)	Uracil phosphoribosyltransferase	169
2.4. Metabolism of lipids						
<i>ywdH</i>	457	8	1.4e–144	<i>slr0091</i> , <i>Synechocystis</i> sp. PCC 6803 (embl: D64004)	Aldehyde dehydrogenase	64
<i>ywiE</i>	500	2	4.8e–120	<i>cls</i> , <i>Escherichia coli</i> (sw: P31071)	Cardiolipin synthase	134
<i>ywjE</i>	398	2	2.6e–55	<i>ywnE</i> , <i>Bacillus subtilis</i> (embl: Y08559)	Cardiolipin synthase	139
			2.5e–78	<i>ywnE</i> , <i>Bacillus subtilis</i> (embl: Y08559)		
			8.3e–45	<i>cls</i> , <i>Escherichia coli</i> (sw: P31071)		
<i>acdA</i>	379	3	2.3e–157	<i>acdS</i> , <i>Clostridium acetobutylicum</i> (sw: P52042)	Acyl-CoA dehydrogenase	141
<i>ywnE</i>	482	2	1.2e–61	<i>cls</i> , <i>Escherichia coli</i> (sw: P31071)	Cardiolipin synthetase	198
			2.9e–120	<i>ywiE</i> , <i>Bacillus subtilis</i> (sw: P45860)		

Table 2 (cont.)

Gene name	Product size (aa)	Total no. paralogues	BLAST score	Genes encoding similar protein/similar gene products*	Function/activity of the gene product or similar protein	Gene no.†
<i>ywpB</i>	132	0	1.3e-43	<i>fabZ</i> , <i>Neisseria meningitidis</i> (embl:U79481)	(3R)-Hydroxymyristoyl-(acyl carrier protein) dehydratase saturated fatty acid biosynthesis	219
2.5. Metabolism of coenzymes and prosthetic groups						
<i>ywaB</i>	311	0	1.1e-21	<i>menA</i> , <i>Haemophilus influenzae</i> (sw:P44739)	Menaquinone biosynthesis 1,4-dihydroxy-2-naphthoate octaprenyltransferase	12
<i>ywbJ</i>	272	0	2.0e-27	<i>thi6</i> , <i>Saccharomyces cerevisiae</i> (sw:P41835)	Thiamin biosynthetic bifunctional enzyme	31
<i>ywbK</i>	222	1	4.1e-74	HI0417, <i>Haemophilus influenzae</i> (pir:A64151)	Thiamin-phosphate pyrophosphorylase	32
			1.2e-31	<i>thi4</i> , <i>Schizosaccharomyces pombe</i> (sw:P40386)		
<i>ywjB</i>	174	1	1.1e-53	HI0415, <i>Haemophilus influenzae</i> (pir:A64152)	Dihydrofolate reductase	136
			5.0e-09	<i>dyrA</i> , <i>Thermotoga maritima</i> (pir:JC4116)		
<i>ywkE</i>	288	0	1.2e-26	<i>yypA</i> , <i>Bacillus subtilis</i> (sw:P37508)	Protoporphyrinogen oxidase	158
<i>narA</i>	341	0	4.6e-67	<i>hemK</i> , <i>Escherichia coli</i> (pir:I83570)	Nitrate assimilation and anaerobic respiration	187
3. Information pathways						
3.1. DNA replication						
<i>ywpH</i>	113	1	3.8e-45	<i>ssb</i> , <i>Bacillus subtilis</i> (sw:P37455)	ssDNA-binding protein	225
3.2. DNA restriction/modification and repair						
<i>ung</i>	225	0	2.4e-84	<i>ung</i> , <i>Escherichia coli</i> (sw:P12295)	Uracil-DNA glycosylase	63
<i>ywjD</i>	320	0	8.4e-35	<i>uvr1</i> , <i>Neurospora crassa</i> (pir:S55262)	UV-endonuclease	138
3.5. RNA synthesis; initiation; regulation; elongation; termination						
<i>licR</i>	641	2	2.3e-48	<i>mtlR</i> , <i>Bacillus stearothermophilus</i> (embl:U18943)	Transcriptional regulator of the <i>lic</i> operon	1
<i>ywaE</i>	171	13	1.1e-05	<i>slyA</i> , <i>Salmonella typhimurium</i> (sw:P40676)	Transcriptional regulator MarR family	16
<i>sacY</i>	280	2	9.2e-107	<i>surT</i> , <i>Bacillus stearothermophilus</i> (embl:U34873)	Transcriptional regulator BglG family	19
			1.0e-98	<i>sacT</i> , <i>Bacillus subtilis</i> (sw:P26212)	Transcriptional regulator of <i>sacY</i>	20
<i>sacX</i>	459	4	4.2e-172	<i>sacP</i> , <i>Bacillus subtilis</i> (sw:P05306)		
<i>ywbI</i>	301	15	1.2e-47	<i>cynR</i> , <i>Escherichia coli</i> (sw:P27111)	Transcriptional activator LysR family	30
<i>ywcC</i>	223	0	3.3e-07	<i>arpA</i> , <i>Streptomyces griseus</i> (embl:D49782)	Transcriptional regulator (A factor receptor)	39
<i>sacT</i>	276	2	6.5e-120	<i>surT</i> , <i>Bacillus stearothermophilus</i> (embl:U34873)	Transcriptional regulator BglG family	53
			2.2e-97	<i>sacY</i> , <i>Bacillus subtilis</i> (sw:P15401)	Transcriptional regulator LysR family	95
<i>ywfK</i>	299	15	6.2e-44	<i>rbcR</i> , <i>Synechocystis</i> sp. PCC 6803 (embl:D64004)		
<i>ywbA</i>	139	11	1.8e-09	<i>ysmB</i> , <i>Bacillus subtilis</i> (embl:Z75208)	Transcriptional regulator MarR family	104
<i>ywbH</i>	157	1	3.9e-07	<i>orfX</i> , <i>Pseudomonas aeruginosa</i> (embl:U49666)	Transcriptional regulator	111
<i>fnr</i>	238	0	2.2e-16	<i>ntcA</i> , <i>Synechococcus</i> sp. PCC 7942 (sw:P29283)	Anaerobic regulatory protein	127
<i>rpoE</i>	173	0		No significant similarities	RNA polymerase δ subunit	142
<i>rho</i>	427	0	1.4e-170	<i>rho</i> , <i>Escherichia coli</i> (sw:P03002)	Transcription termination factor	150
<i>ywnD</i>	257	8	1.7e-60	<i>tipA</i> , <i>Streptomyces lividans</i> (sw:P32184)	Transcriptional activator MerR family	197
<i>ywoH</i>	137	9	6.5e-07	<i>slyA</i> , <i>Escherichia coli</i> (sw:P55740)	Transcriptional regulator MarR family	213
<i>ywpl</i>	258	3	1.2e-42	<i>agaR</i> , <i>Escherichia coli</i> (sw:P42902)	Transcriptional regulator DeoR family	226
<i>ywqA</i>	922	1	1.6e-66	<i>mot1</i> , <i>Saccharomyces cerevisiae</i> (sw:P32333)	Probable helicase of the Snf2/Rad54 family	228
			1.4e-156	<i>sl1366</i> , <i>Synechocystis</i> sp. PCC 6803 (embl:D90916)	Transcriptional regulator LysR family	240
<i>ywqM</i>	293	15	3.8e-38	<i>gltR</i> , <i>Bacillus subtilis</i> (embl:U79494)		
<i>ywrC</i>	158	4	8.3e-23	MJ0151, <i>Methanococcus jannaschii</i> (pir:H64318)	Transcriptional regulator AsnC family	245
<i>alsR</i>	302	15	1.3e-56	<i>xapR</i> , <i>Escherichia coli</i> (sw:P23841)	Transcriptional regulator LysR family	254
<i>rbsR</i>	326	11	3.1e-58	<i>scrR</i> , <i>Pediococcus pentosaceus</i> (sw:P43472)	Transcription regulator LacI family	265
<i>ywtF</i>	322	2	6.1e-50	<i>lytR</i> , <i>Bacillus subtilis</i> (sw:Q02115)	Attenuator for itself and <i>lytABC</i> operon	272
3.7. Protein synthesis; ribosomal proteins; aminoacyl-tRNA synthetases; initiation; elongation; termination						
<i>tyrZ</i>	413	1	3.9e-135	<i>tyrS</i> , <i>Haemophilus influenzae</i> (sw:P43836)	Minor tyrosyl-tRNA synthetase	15
<i>thrZ</i>	638	1	1.8e-241	<i>thrS</i> , <i>Bacillus subtilis</i> (sw:P18255)	Minor threonyl-tRNA synthetase	103
<i>argS</i>	556	0	6.7e-161	<i>argS</i> , <i>Brevibacterium lactofermentum</i> (sw:P41253)	Arginyl-tRNA synthetase	125
<i>rpmE</i>	66	1	5.4e-26	<i>rpmE</i> , <i>Mycobacterium leprae</i> (sw:P45834)	Ribosomal protein L31	151
<i>prfA</i>	356	1	3.7e-130	<i>prf</i> , <i>Mycoplasma capricolum</i> (embl:50100)	Peptide-chain-releasing factor	157
4. Other functions						
4.1. Adaptation to atypical conditions						
<i>ywpC</i>	130	0	1.4e-48	<i>mscL</i> , <i>Escherichia coli</i> (sw:P23867)	Large conductance mechanosensitivity channel	220
4.3. Antibiotic production						
<i>sbo</i>	43	0		No significant similarities	Antibiotic peptide, subtilisin A	123
4.4. Drug/analogue sensitivity						
<i>mmr</i>	466	3	2.8e-151	<i>mmr</i> , <i>Streptomyces coelicolor</i> (sw:P11545)	Methylenomycin A resistance protein	102

Table 2 (cont.)

Gene name	Product size (aa)	Total no. paralogues	BLAST score	Genes encoding similar protein/similar gene products*	Function/activity of the gene product or similar protein	Gene no.†
<i>ywnH</i>	163	0	1.5e–15 9.4e–56	<i>bar</i> , <i>Streptomyces coelicolor</i> (pir: JH0246) sll1647, <i>Synechocystis</i> sp. PCC 6803 (embl: D90903)	Phosphinotricine N-acetyltransferase	201
5. Miscellaneous						
<i>ywcH</i>	333	3	1.4e–09	<i>lxa-1</i> , <i>Photobacterium leiognati</i> (sw: P09140)	Alkanal monooxygenase α -chain, bacterial luciferase	50
<i>ywdL</i>	181	0	6.3e–61 1.2e–05	<i>yhbW</i> , <i>Escherichia coli</i> (sw: P45529) <i>bnc1</i> , <i>Brassica napus</i> (sw: P33523)	Seed storage protein	68
<i>ywlE</i>	150	0	1.2e–12	<i>walE</i> , <i>Escherichia coli</i> (embl: U38473)	Acid phosphatase	165
Genes without known function						
<i>ywcC</i>	210	1	9.7e–15	Gene in <i>mleR</i> region of <i>Lactococcus lactis</i> (embl: M26949)	Unknown	13
<i>ywbD</i>	396	0	1.8e–35	PID: g1787201, <i>Escherichia coli</i> (embl: AE000198)	Unknown	25
<i>ywbG</i>	241	3	7.7e–51	<i>yohK</i> , <i>Escherichia coli</i> (sw: P33373)	Unknown	28
<i>ywbH</i>	128	3	2.4e–16	<i>yohJ</i> , <i>Escherichia coli</i> (sw: P33372)	Unknown	29
<i>ywbL</i>	481	0	4.6e–09	slr0964, <i>Synechocystis</i> sp. PCC 6803	Unknown	33
<i>ywbM</i>	385	0	3.9e–67	PID: g1787254, <i>Escherichia coli</i> (embl: AE000203)	Unknown	34
<i>ywbN</i>	416	0	5.0e–40	PID: g1787255, <i>Escherichia coli</i> (embl: AE000203)	Unknown	35
<i>ywcD</i>	127	0	6.5e–06	<i>rflB</i> , <i>Escherichia coli</i> (embl: D90865)	Unknown	40
<i>ywdB</i>	271	1	1.0e–70	Y416, <i>Haemophilus influenzae</i> (sw: P44697)	Unknown	58
<i>ywdF</i>	268	0	1.0e–14	H11696, <i>Haemophilus influenzae</i> (pir: D64175)	Unknown	62
<i>ywdJ</i>	460	0	7.5e–41	PID: g1786723 <i>Escherichia coli</i> (embl: AE000157)	Unknown	66
<i>ywdK</i>	113	0	3.1e–19	<i>ygdD</i> , <i>Haemophilus influenzae</i> (sw: P45019)	Unknown	67
<i>yweA</i>	154	0	7.9e–17	<i>orf1</i> , upstream <i>gbsAB</i> operon <i>Bacillus subtilis</i> (embl: U47861)	Unknown	80
<i>ywfM</i>	296	2	1.4e–14	<i>yicL</i> , <i>Escherichia coli</i> (sw: P31437)	Unknown	97
<i>ywfO</i>	433	0	1.7e–39	MG461, <i>Mycoplasma genitalium</i> (sw: P39651)	Unknown	99
<i>ywgB</i>	156	0	2.0e–43	<i>orf156</i> , <i>clpA</i> region <i>Lactococcus lactis</i> (embl: L36907)	Unknown	101
<i>ywhC</i>	219	0	2.4e–09	MJ0611, <i>Methanococcus jannaschii</i> (pir: C64376)	Unknown	106
<i>ywhN</i>	426	0	7.5e–12	<i>cpn-60</i> , <i>Synechocystis</i> sp. PCC 6803 (embl: D64001)	Unknown	117
<i>ywiA</i>	448	0	8.8e–14	MTCY210.10 <i>Mycobacterium tuberculosis</i> (embl: Z84395)	Unknown	122
<i>ywiC</i>	239	0	2.3e–47	H11626, <i>Haemophilus influenzae</i> (sw: P44278)	Unknown	128
<i>ywkD</i>	128	2	3.3e–53	<i>yaeR</i> , <i>Escherichia coli</i> (sw: P52096)	Unknown	156
<i>ywlC</i>	346	0	1.4e–88	<i>sua5</i> , <i>Saccharomyces cerevisiae</i> (sw: P32579)	Unknown	163
<i>ywlD</i>	185	0	6.8e–22	<i>ywlD</i> , <i>Escherichia coli</i> (embl: D90826)	Unknown	164
<i>ywmC</i>	227	1	1.6e–64	<i>ywmD</i> , <i>Bacillus subtilis</i> (embl: Z81356)	Unknown	183
<i>ywmD</i>	224	1	1.4e–64	<i>ywmC</i> , <i>Bacillus subtilis</i> (embl: Z81356)	Unknown	184
<i>ywmH</i>	62	0	4.1e–07	<i>yjbJ</i> , <i>Escherichia coli</i> (sw: P32691)	Unknown	190
<i>ywnC</i>	127	1	2.3e–30	<i>ywnG</i> , <i>Bacillus subtilis</i> (embl: Y08559)	Unknown	196
<i>ywnG</i>	172	1	4.3e–29	<i>ywnC</i> , <i>Bacillus subtilis</i> (embl: Y08559)	Unknown	200
<i>ywoF</i>	468	0	5.3e–06	MJ1396, <i>Methanococcus jannaschii</i> (pir: C64474)	Unknown	211
<i>ywpD</i>	278	0	7.5e–29	<i>yehU</i> , <i>Escherichia coli</i> (embl: D90868)	Unknown	221
<i>ywpJ</i>	285	5	9.8e–15	<i>lytS</i> , <i>Staphylococcus aureus</i> (embl: L42945)	Unknown	227
			3.7e–34	<i>ycsE</i> , <i>Bacillus subtilis</i> (sw: P42962)		
			3.6e–30	<i>yxeH</i> , <i>Bacillus subtilis</i> (sw: P54947)		
			3.3e–26	<i>yidA</i> , <i>Escherichia coli</i> (sw: P09997)		
<i>ywqH</i>	140	1	2.1e–22	<i>ybhA</i> , <i>Escherichia coli</i> (sw: P21829)	Unknown	235
			0.0019	<i>yxiB</i> , <i>Bacillus subtilis</i> (sw: P42294)		
			9.4e–10	<i>yxiC</i> , <i>Bacillus subtilis</i> (sw: P42295)		
			6.2e–143	<i>yqcG</i> , <i>Bacillus subtilis</i> (sw: P45942)		
<i>ywqI</i>	86	3	5.8e–09	<i>yxiD</i> , <i>Bacillus subtilis</i> (sw: P42296)	Unknown	237
			8.4e–19	<i>yjaF</i> , <i>Escherichia coli</i> (sw: P32679)		
			6.1e–10	MJ1083, <i>Methanococcus jannaschii</i> (pir: B64435)		
			2.2e–66	<i>ywpJ</i> , <i>Bacillus subtilis</i> (embl: Z83337)		
<i>ywqJ</i>	602	7	7.3e–31	<i>ycsE</i> , <i>Bacillus subtilis</i> (sw: P42962)	Unknown	239
			9.7e–33	<i>yxeH</i> , <i>Bacillus subtilis</i> (sw: P54947)		
			2.6e–27	<i>yidA</i> , <i>Escherichia coli</i> (sw: P09997)		
			9.8e–22	<i>ybhA</i> , <i>Escherichia coli</i> (sw: P21829)		
<i>ywqL</i>	238	0	8.4e–19	<i>yjaF</i> , <i>Escherichia coli</i> (sw: P32679)	Unknown	239
<i>ywqN</i>	181	0	6.1e–10	MJ1083, <i>Methanococcus jannaschii</i> (pir: B64435)	Unknown	241
<i>ywtE</i>	286	7	2.2e–66	<i>ywpJ</i> , <i>Bacillus subtilis</i> (embl: Z83337)	Unknown	271

* Accession numbers are for SWISS-PROT (sw), Protein Information Resource (pir) or EMBL (embl).

†Numbers refer to Fig. 2.

teichoic acids (*dlt* operon; Perego *et al.*, 1995) and synthesis of a spore-coat polysaccharide (*sps* operon; M. F. Hullo and others, unpublished). Two sporulation loci have also been characterized: *spoII*R (Karow *et al.*, 1995; Londoño-Vallejo & Stragier, 1995) and *rapB* (Perego *et al.*, 1996). Finally, the uracil phosphoribosyl-transferase involved in the pyrimidine salvage pathway (*upp*; Martinussen *et al.*, 1995) and a minor alternative tyrosine-tRNA synthetase (*tyrZ*; Glaser *et al.*, 1991) of this *B. subtilis* chromosomal region have been studied.

In the latest version of the *B. subtilis* genetic map, there are 96 markers located between *gerBC* and *licR* (*celR*) (Biaudet *et al.*, 1996). The relative positions of the markers determined by genetic methods are in good agreement with the position of the corresponding genes on the chromosome. Only eight genetic markers could not be unambiguously associated to a gene. *furC* and *furE* (resistance to 5-fluorouracil) may both correspond to *upp*. Alternatively, one of these loci and *azpB* (resistance to azopyrimidines) may lie in *ywoE*, which encodes a protein similar to uracil permeases. For the other five loci, *flaC*, *outF*, *divII*, *bac* and *gneA*, no obvious gene could be allotted.

The predicted gene products of the *licR-gerBC* region were compared with all known protein sequences available in public sequence libraries (5 April 1997). The 273 genes of the region were thereby classified into five groups: (i) genes with known function (non- γ gene) encoding proteins similar to known proteins (94 genes); (ii) orphan genes (for which no similar sequence has been found) with known function (9 genes); (iii) genes with unknown function but encoding proteins similar to proteins with known function (75 genes); (iv) unknown genes encoding proteins similar to proteins with unknown function (39 genes) and (v) orphan genes with unknown function (56 genes). Although this classification is subjective, the proportion of genes in classes (iv) and (v) is similar to that found for other eubacterial genomes (35%) using the BLAST program. Although the complete genomes of five bacteria, one archaeon and the yeast *S. cerevisiae* are known, no phylogenetically related genes could be detected by this method for more than 20% of the genes in this region of the *B. subtilis* chromosome. In 10 cases (*spsA*, *rocB*, *phrF*, *ywmC*, *ywmD*, *ywnC*, *ywnG*, *ywqH*, *ywqI*, *ywqJ*), counterparts were detected only in *B. subtilis* sequences (paralogue orphan genes). These genes, together with the class (ii) and (v) genes, may constitute a kind of signature for *B. subtilis* and related species. It is however not possible to know whether the absence of detectable similarities reflects a recent origin of these genes or whether the relatedness to other proteins has been obscured by evolution.

The 217 genes in classes (i), (ii), (iii) and (iv) were further classified according to Moszer *et al.* (1996) and are presented in Table 2 (see also the arrow fill-in motifs in Fig. 2). No member was found for 10 categories [numbers in parentheses represent classes and subclasses according to Moszer *et al.* (1996)]: sensors (1.3),

Table 3. Orthologues of the *B. subtilis* genes from the *licR-gerBC* region

Genome	Total number of genes	Number of orthologues*
<i>Escherichia coli</i>	4285	158
<i>Synechocystis</i> PCC 6803	3168	94
<i>Haemophilus influenzae</i>	1743	92
<i>Methanococcus jannaschii</i>	1682	59
<i>Saccharomyces cerevisiae</i>	6200	54
<i>Mycoplasma genitalium</i>	470	27

* Number of genes in the *licR-gerBC* region having an orthologue in each genome.

secretion (1.6), germination (1.9), competence (1.10), metabolism of phosphate (2.6) and of sulphur (2.7), DNA recombination (3.3), RNA modification (3.6), protein modification (3.8) and folding (3.9). The distribution of genes into the remaining classes is consistent with what has been described for the whole genome (Moszer *et al.*, 1996). However, there are several striking features. About half of the genes in the *licR-gerBC* region belong to the 'cell envelope and cellular processes' class of genes, with a large proportion in the 'cell wall components' subclass (the *sps* operon in the sporulation subclass is also involved in the synthesis of wall components). The region contains very few genes encoding anabolic enzymes (in the 'intermediary metabolism' class): *glyA*, *narA* and the *thiC* operon. The 'information pathways' class contains mainly regulatory genes and very few essential genes. Three aminoacyl-tRNA synthetase genes are found in this region, but only *argS* is essential (A. Sekowska and others, unpublished), *thrZ* and *tyrZ* encode minor synthetases of unknown physiological function (Glaser *et al.*, 1991; Putzer *et al.*, 1990). Other than *argS*, only *rpmE* and *rho*, encoding the ribosomal protein L31 and a transcription terminator factor, respectively, are important genes of this class.

Orthologues and paralogues in the *gerBC-licR* region

We have searched for *B. subtilis* sequences similar to sequences present in six known genomes (orthologues). The number of genes of the *B. subtilis* *licR-gerBC* region having an orthologue in each of these genomes is presented in Table 3. The largest number of orthologues (58%) was with *E. coli* sequences. Note however that the *E. coli* genome contains more than twice as many genes as the *H. influenzae* genome. A striking result of this analysis is the small number (27) of genes in the *licR-gerBC* region having an orthologue in *Mycoplasma genitalium*. This may be due to the information content of this part of the *B. subtilis* chromosome (see above). Indeed mycoplasmas have minimal genomes which appear to have evolved from Gram-positive bacteria by elimination of non-essential genes (Maniloff, 1983). The

Fig. 5. Chromosomal distribution of the paralogs of the genes in the *licR-gerBC* region. Bars represent the frequency of paralogs (number of paralogs/corresponding number of genes) found in 100 kb segments of the chromosome. The darker bars are in the *licR-gerBC* region.

number of orthologues with the archaeon (*Methanococcus jannaschii*) and with the eukaryote (*S. cerevisiae*) genomes are similar (49 and 54) but correspond to different genes (data not shown).

We systematically compared the 273 putative gene products with protein sequences deduced from known *B. subtilis* DNA sequences. Of these, 115 genes (42%) are similar to at least one other *B. subtilis* gene (paralogs). Forty-five genes have only one counterpart. In contrast, 17 genes have more than eight paralogs (Table 2). These genes encode proteins which belong to large families of regulatory or transporter proteins. For example, three genes in this region of the chromosome encode regulatory proteins of the LysR family and there are 13 other members of this class in the remaining known sequences. We performed a preliminary analysis of the distribution of the paralogs along the genome map, considering only genes which are not part of large families of proteins (Fig. 5). The distribution is not uniform, presenting a higher density of paralogs in or near our 271 kb region (between 3600 and 4100 kb). In several cases paralogs are in tandem repeat, in the same transcriptional unit (*ywbH* and *ywbG* or *ywbK* and *ywbL*), or in two transcriptional units (*ywmC* and *ywmD*). Some protein families are particularly well represented in this part of the chromosome. The three known genes putatively encoding cardiolipin synthase are found in a 70 kb region: *ywiE* and *ywnE* are transcribed in the same orientation and *ywjE* is transcribed in the opposite direction. Three of the eight known members of the newly discovered regulatory aspartyl phosphatase family (Perego *et al.*, 1996) are encoded in this part of the chromosome (*rapA*, *rapD*, *rapF*).

Conclusion

The systematic sequencing approach has considerably accelerated the study of the genetic content and the gene function analysis of the *licR-gerBC* region of the *B. subtilis* chromosome. This region appears to contain genes mainly involved in adaptation to growth conditions. In particular, it carries a high density of loci responsible for the utilization of various carbon and nitrogen sources and alternative terminal electron acceptors. The synthesis of non-essential cell wall components and their modification may reflect adaptation to physical and chemical variations in *B. subtilis*' ecological niche: the soil and the surface of plant leaves (Priest, 1993). The small number of genes having an orthologue in the *Mycoplasma genitalium* genome, considered as a minimal genome with very little adaptive capacity, together with the large number of paralogs in this part of the chromosome are consistent with this observation. Our analysis also supports the hypothesis of a functional organization of the *B. subtilis* chromosome. The systematic analysis of the unknown genes will confirm or invalidate this view.

ACKNOWLEDGEMENTS

We thank Dr J. Hoch for providing plasmid pPP41. This work was supported by the Institut Pasteur, the European Commission (Biotechnology contract B102-CT-930272), the Ministère de l'Éducation Nationale, de l'Enseignement Supérieur et de la Recherche, the Groupement de Recherche et d'Étude sur les Génomes (GREG) (grant 1993-no. 20) and the Centre National de la Recherche Scientifique (GDR 1029 and URA 1300). H.C.R. was supported by a fellowship from the Universidad Nacional Autónoma de México.

REFERENCES

- Abhayawardhane, Y. & Stewart, G. C. (1995). *Bacillus subtilis* possesses a second determinant with extensive sequence similarity to the *Escherichia coli mreB* morphogene. *J Bacteriol* 177, 765–773.
- Albertini, A. M., Caramori, T., Scoffone, F., Scotti, C. & Galizzi, A. (1995). Sequence around the 159° region of the *Bacillus subtilis* genome: the *pksX* locus spans 33.6 kb. *Microbiology* 141, 299–309.
- 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.
- Biaudet, V., Samson, F., Anagnostopoulos, C., Ehrlich, S. D. & Bessières, P. (1996). Computerized genetic map of *Bacillus subtilis*. *Microbiology* 142, 2669–2729.
- Bruckner, R., Shoseyov, O. & Doi, R. H. (1990). Multiple active forms of a novel serine protease from *Bacillus subtilis*. *Mol Gen Genet* 221, 486–490.
- Bullock, W. O., Fernandez, J. M. & Short, J. M. (1987). A high efficiency plasmid transforming *recA* *Escherichia coli* strain with β -galactosidase selection. *BioTechniques* 5, 376.
- Bult, C. J., White, O., Olsen, G. J. & 37 other authors (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* 273, 1058–1073.
- Calogero, S., Gardan, R., Glaser, P., Schweizer, J., Rapoport, G. &

- Débarbouillé, M. (1994).** RocR, a novel regulatory protein controlling arginine utilization in *Bacillus subtilis*, belongs to the NtrC/NifA family of transcriptional activators. *J Bacteriol* **176**, 1234–1241.
- Chambers, S. P., Prior, S. E., Barstow, D. A. & Minton, N. P. (1988).** The pMTL *nic*⁻ cloning vectors. 1. Improved pUC polylinker regions to facilitate the use of sonicated DNA for nucleotide sequencing. *Gene* **68**, 139–149.
- Chung, C. T. & Miller, R. H. (1988).** A rapid and convenient method for the preparation and storage of competent bacterial cells. *Nucleic Acids Res* **16**, 3580.
- Cruz Ramos, H., Boursier, L., Moszer, I., Kunst, F., Danchin, A. & Glaser, P. (1995).** Anaerobic transcription activation in *Bacillus subtilis*: identification of distinct FNR-dependent and -independent regulatory mechanisms. *EMBO J* **14**, 5984–5994.
- Cruz Ramos, H., Glaser, P., Wray, L. V. & Fisher, S. H. (1997).** The *Bacillus subtilis* *ureABC* operon. *J Bacteriol* **179**, 3371–3373.
- Danchin, A. (1996).** mRNA turnover and DNA synthesis: a lesson from bacterial genome comparisons. *Mol Microbiol* **20**, 895–897.
- Dear, S. & Staden, R. (1991).** A sequence assembly and editing program for efficient management of large projects. *Nucleic Acids Res* **19**, 3907–3911.
- Donovan, W., Zheng, L. B., Sandman, K. & Losick, R. (1987).** Genes encoding spore coat polypeptides from *Bacillus subtilis*. *J Mol Biol* **196**, 1–10.
- Fleischmann, R. D., Adams, M. D., White, O. & 37 other authors (1995).** Whole-genome random sequencing of *Haemophilus influenzae* Rd. *Science* **269**, 496–512.
- Fouet, A., Klier, A. & Rapoport, G. (1986).** Nucleotide sequence of the sucrase gene of *Bacillus subtilis*. *Gene* **45**, 221–225.
- Fouet, A., Arnaud, M., Klier, A. & Rapoport, G. (1987).** *Bacillus subtilis* sucrose-specific enzyme II of the phosphotransferase system: expression in *Escherichia coli* and homology to enzymes II from enteric bacteria. *Proc Natl Acad Sci USA* **84**, 8773–8777.
- Fraser, C. M., Gocayne, J. D., White, O. & 26 other authors (1995).** The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403.
- Gibson, T. J. (1984).** *Studies on the Epstein-Barr virus genome*. PhD thesis. University of Cambridge.
- Glaser, P., Kunst, F., Débarbouillé, M., Vertès, A., Danchin, A. & Dedonder, R. (1991).** A gene encoding a tyrosine-tRNA synthetase is located near *sacS* in *Bacillus subtilis*. *DNA Seq* **1**, 251–261.
- Glaser, P., Kunst, F., Arnaud, M. & 14 other authors (1993).** *Bacillus subtilis* genome project: cloning and sequencing of the 97 kb region from 325° to 333°. *Mol Microbiol* **10**, 371–384.
- Glaser, P., Danchin, A., Kunst, F., Zuber, P. & Nakano, M. M. (1995).** Identification and isolation of a gene required for nitrate assimilation and anaerobic growth of *Bacillus subtilis*. *J Bacteriol* **177**, 1112–1115.
- Goffeau, A., Barrell, B. G., Bussey, H. & 13 other authors (1996).** Life with 6000 genes. *Science* **274**, 563–567.
- Hanahan, D. (1983).** Studies on transformation of *Escherichia coli* with plasmids. *J Mol Biol* **16**, 557–580.
- Harwood, C. R. & Wipat, A. (1996).** Sequencing and functional analysis of the genome of *Bacillus subtilis* strain 168. *FEBS Lett* **389**, 84–87.
- Hénaut, A., Rouxel, T., Gleizes, A., Moszer, I. & Danchin, A. (1996).** Uneven distribution of GATC motifs in the *Escherichia coli* chromosome, its plasmids and its phages. *J Mol Biol* **257**, 574–585.
- Higgins, D. G., Bleasby, A. J. & Fuchs, R. (1992).** CLUSTAL V – Improved software for multiple sequence alignment. *Comput Appl Biosci* **8**, 189–191.
- Himmelreich, R., Hilbert, H., Plagens, H., Pirkel, E., Li, B. & Herrmann, R. (1996).** Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res* **24**, 4420–4449.
- Horinouchi, S. & Weisblum, B. (1982).** Nucleotide sequence and functional map of pC194, a plasmid that specifies inducible chloramphenicol resistance. *J Bacteriol* **150**, 815–825.
- Itaya, M. & Tanaka, T. (1991).** Complete physical map of the *Bacillus subtilis* 168 chromosome constructed by a gene-directed mutagenesis method. *J Mol Biol* **220**, 631–648.
- Kaneko, T., Sato, S., Kotani, H. & 21 other authors (1996).** Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC 6803. 2. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res* **3**, 109–136.
- Karow, M. L., Glaser, P. & Piggot, P. J. (1995).** Identification of a gene, *spoIIIR*, that links the activation of σ^E to the transcriptional activity of σ^F during sporulation in *Bacillus subtilis*. *Proc Natl Acad Sci USA* **92**, 2012–2016.
- Kunkel, B., Kroos, L., Poth, H., Youngman, P. & Losick, R. (1989).** Temporal and spatial control of the mother-cell regulatory gene *spoIIID* of *Bacillus subtilis*. *Genes Dev* **3**, 1735–1744.
- Kunst, F. & Devine, K. (1991).** The project of sequencing the entire *Bacillus subtilis* genome. *Res Microbiol* **142**, 905–912.
- Kunst, F. & Rapoport, G. (1995).** Salt stress is an environmental signal affecting degradative enzyme synthesis in *Bacillus subtilis*. *J Bacteriol* **177**, 2403–2407.
- Kunst, F., Vassarotti, A. & Danchin, A. (1995).** Organization of the European *Bacillus subtilis* genome sequencing project. *Microbiology* **141**, 249–255.
- Kyte, J. & Doolittle, R. F. (1982).** A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**, 105–132.
- Lampe, M., Binnie, C., Schmidt, R. & Losick, R. (1988).** Cloned gene encoding the δ subunit of *Bacillus subtilis* RNA polymerase. *Gene* **67**, 13–19.
- Londoño-Vallejo, J.-A. & Stragier, P. (1995).** Cell-cell signaling pathway activating a developmental transcription factor in *Bacillus subtilis*. *Genes Dev* **9**, 503–508.
- Londoño-Vallejo, J.-A., Fréhel, C. & Stragier, P. (1997).** *spoIIQ*, a forespore-expressed gene required for engulfment in *Bacillus subtilis*. *Mol Microbiol* **24**, 29–39.
- Lopez-Diaz, I., Clarke, S. & Mandelstam, J. (1986).** *spoIID* operon of *Bacillus subtilis*: cloning and sequence. *J Gen Microbiol* **132**, 341–354.
- Maeda, M., Hidaka, M., Nakamura, A., Masaki, H. & Uozumi, T. (1994).** Cloning, sequencing and expression of thermophilic *Bacillus* sp. strain TP-90 urease gene complex in *Escherichia coli*. *J Bacteriol* **176**, 432–442.
- Maniloff, J. (1983).** Evolution of wall-less prokaryotes. *Annu Rev Microbiol* **37**, 477–499.
- Marck, C. (1988).** ‘DNA Strider’: a ‘C’ program for the fast analysis of DNA and protein sequences on the Apple Macintosh family of computers. *Nucleic Acids Res* **16**, 1829–1836.
- Martinussen, J., Glaser, P., Andersen, P. S. & Saxild, H. H. (1995).** Two genes encoding uracil phosphoribosyltransferase are present in *Bacillus subtilis*. *J Bacteriol* **177**, 271–274.
- Messing, J. & Vieira, J. (1982).** A new pair of M13 vectors for selecting either DNA strand of double-digest restriction fragments. *Gene* **19**, 269–276.

- Miller, J. H. (1972). *Experiments in Molecular Genetics*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.
- Mitchell, C., Morris, P. W., Lum, L., Spiegelman, G. & Vary, J. C. (1992). The amino acid sequence of a *Bacillus subtilis* phosphoprotein that matches an *orfY*–*tsr* coding sequence. *Mol Microbiol* 6, 1345–1349.
- Mizuno, M., Masuda, S., Takemaru, K., Hosono, S., Sato, T., Takeuchi, M. & Kobayashi, Y. (1996). Systematic sequencing of the 283 kb 210°–232° region of the *Bacillus subtilis* genome containing the *skin* element and many sporulation genes. *Microbiology* 142, 3103–3111.
- Moszer, I., Glaser, P. & Danchin, A. (1991). Multiple IS insertion sequences near the replication terminus in *Escherichia coli* K-12. *Biochimie* 73, 1361–1374.
- Moszer, I., Glaser, P. & Danchin, A. (1995). *SubtiList*: a relational database for the *Bacillus subtilis* genome. *Microbiology* 141, 261–268.
- Moszer, I., Kunst, F. & Danchin, A. (1996). The European *Bacillus subtilis* genome sequencing project: current status and accessibility of the data from a new World Wide Web site. *Microbiology* 142, 2987–2991.
- Naclerio, G., Baccigalupi, L., Zilhao, R., De Felice, M. & Ricca, E. (1996). *Bacillus subtilis* spore coat assembly requires *cotH* gene expression. *J Bacteriol* 178, 4375–4380.
- Niaudet, B., Goze, A. & Ehrlich, S. D. (1982). Insertional mutagenesis in *Bacillus subtilis*: mechanism and use in gene cloning. *Gene* 19, 277–284.
- O'Brien, C. (1997). Entire *E. coli* genome sequenced – at last. *Nature* 385, 472.
- Ogasawara, N. & Yoshikawa, H. (1996). The systematic sequencing of the *Bacillus subtilis* genome in Japan. *Microbiology* 142, 2993–2994.
- 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.
- Perego, M., Hanstein, C., Welsh, K. M., Djavakhishvili, T., Glaser, P. & Hoch, J. A. (1994). Multiple protein-aspartate phosphatases provide a mechanism for the integration of diverse signals in the control of development in *Bacillus subtilis*. *Cell* 79, 1047–1055.
- Perego, M., Glaser, P., Minutello, A., Strauch, M. A., Leopold, K. & Fischer, W. (1995). Incorporation of D-alanine into lipoteichoic acid and wall teichoic acid in *Bacillus subtilis*: identification of genes and regulation. *J Biol Chem* 270, 15598–15606.
- Perego, M., Glaser, P. & Hoch, J. A. (1996). Aspartyl-phosphate phosphatases deactivate the response regulator components of the sporulation signal transduction system in *Bacillus subtilis*. *Mol Microbiol* 19, 1151–1157.
- Popham, D. L. & Setlow, P. (1994). Cloning, nucleotide sequence, mutagenesis, and mapping of the *Bacillus subtilis* *pbpD* gene, which codes for penicillin-binding protein 4. *J Bacteriol* 176, 7197–7205.
- Priest, F. G. (1993). Systematics and ecology of *Bacillus*. In *Bacillus subtilis and Other Gram-positive Bacteria: Biochemistry, Physiology, and Molecular Genetics*. Edited by A. L. Sonenshein, J. A. Hoch & R. Losick. Washington, DC: American Society for Microbiology.
- Putzer, H., Brakhage, A. & Grunberg-Manago, M. (1990). Independent genes for two threonyl-tRNA synthetases in *Bacillus subtilis*. *J Bacteriol* 172, 4593–4602.
- Putzer, H., Gendron, N. & Grunberg-Manago, M. (1992). Co-ordinate expression of the two threonyl-tRNA synthetase genes in *Bacillus subtilis*: control by transcriptional antitermination involving a conserved regulatory sequence. *EMBO J* 11, 3117–3127.
- Quirk, P., Dunkley, E., Lee, P. & Krulwich, T. (1993). Identification of a putative *Bacillus subtilis* *rho* gene. *J Bacteriol* 175, 647–654.
- Renna, M. C., Najimudin, N., Winik, L. R. & Zahler, S. A. (1993). Regulation of the *Bacillus subtilis* *alsS*, *alsD*, and *alsR* genes involved in post-exponential-phase production of acetoin. *J Bacteriol* 175, 3863–3875.
- Sacco, M., Ricca, E., Losick, R. & Cutting, S. (1995). An additional GerE-controlled gene encoding an abundant spore coat protein from *Bacillus subtilis*. *J Bacteriol* 177, 372–377.
- Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning: a Laboratory Manual*, 2nd edn. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.
- Santana, M., Kunst, F., Hullo, M. F., Rapoport, G., Danchin, A. & Glaser, P. (1992). Molecular cloning, sequencing, and physiological characterization of the *qox* operon from *Bacillus subtilis* encoding the *aa*₃-600 quinol oxidase. *J Biol Chem* 267, 10225–10231.
- Santana, M., Ionescu, M. S., Vertes, A., Longin, R., Kunst, F., Danchin, A. & Glaser, P. (1994). *Bacillus subtilis* F₀F₁ ATPase: DNA sequence of the *atp* operon and characterization of *atp* mutants. *J Bacteriol* 176, 6802–6811.
- Silhavy, T. J., Berman, M. L. & Enquist, L. W. (1984). *Experiments with Gene Fusions*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.
- Sloma, A., Rufo, G. A., Jr, Theriault, K. A., Dwyer, M., Wilson, S. H. & Pero, J. (1991). Cloning and characterization of the gene for an additional extracellular serine protease of *Bacillus subtilis*. *J Bacteriol* 173, 6889–6895.
- Soldo, B., Lazarevic, V., Mauël, C. & Karamata, D. (1996). Sequence of the 305°–307° region of the *Bacillus subtilis* chromosome. *Microbiology* 142, 3079–3088.
- Sorokin, A., Azevedo, V., Zumstein, E., Galleron, N., Ehrlich, S. D. & Serron, P. (1996). Sequence analysis of the *Bacillus subtilis* chromosome region between the *serA* and *kdg* loci cloned in a yeast artificial chromosome. *Microbiology* 142, 2005–2016.
- Stragier, P. & Losick, R. (1996). Molecular genetics of sporulation in *Bacillus subtilis*. *Annu Rev Genetics* 30, 297–341.
- Swinfield, T. J., Oultram, J. D., Thompson, D. E., Brehm, J. K. & Minton, N. P. (1990). Physical characterisation of the replication region of the *Streptococcus faecalis* plasmid pAMB1. *Gene* 87, 79–90.
- Tobisch, S., Glaser, P., Krüger, S. & Hecker, M. (1997). Identification and characterization of a new β -glucoside utilization system in *Bacillus subtilis*. *J Bacteriol* 179, 496–506.
- Trach, K., Chapman, J., Piggot, P., Lecoq, D. & Hoch, J. (1988). Complete sequence and transcriptional analysis of the *spoOF* region of the *Bacillus subtilis* chromosome. *J Bacteriol* 170, 4194–4208.
- Wipat, A., Carter, N., Brignell, S. C., Guy, B. J., Piper, K., Sanders, J., Emmerson, P. T. & Harwood, C. R. (1996). 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. *Microbiology* 142, 3067–3078.
- Woodson, K. & Devine, K. M. (1994). Analysis of a ribose transport operon from *Bacillus subtilis*. *Microbiology* 140, 1829–1838.
- Wray, L. V., Atkinson, M. R. & Fisher, S. H. (1994). The nitrogen-regulated *Bacillus subtilis* *nrgAB* operon encodes a membrane protein and a protein highly similar to the *Escherichia coli* *glnB*-encoded PII protein. *J Bacteriol* 176, 108–114.

Yanisch-Perron, C., Vieira, J. & Messing, J. (1985). Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. *Gene* **33**, 103–119.

Yoshida, K., Shindo, K., Sano, H., Seki, S., Fujimura, M., Yanai, N., Miwa, Y. & Fujita, Y. (1996). Sequencing of a 65 kb region of the *Bacillus subtilis* genome containing the *lic* and *cel* loci, and creation of a 177 kb contig covering the *gnt-sacXY* region. *Microbiology* **142**, 3113–3123.

Yoshikawa, H., Kazami, J., Yamashita, S., Chibazakura, T., Sone, H., Kawamura, F., Oda, M., Isaka, M., Kobayashi, Y. & Saito, H.

(1986). Revised assignment for the *Bacillus subtilis* *spo0F* gene and its homology with *spo0A* and with two *Escherichia coli* genes. *Nucleic Acids Res* **14**, 1063–1072.

Zukowski, M., Miller, L., Cogswell, P., Chen, K., Aymerich, S. & Steinmetz, M. (1990). Nucleotide sequence of the *sacS* locus of *Bacillus subtilis* reveals the presence of two regulatory genes. *Gene* **90**, 153–155.

.....
Received 16 April 1997; revised 2 June 1997; accepted 6 June 1997.