Published online September 21, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 16 ***4937–4944***
doi:10.1093/nar/gkh830

Genome sequence of *Symbiobacterium thermophilum*, an uncultivable bacterium that depends on microbial commensalism

Kenji Ueda*, Atsushi Yamashita[1], Jun Ishikawa[2], Masafumi Shimada, Tomo-o Watsuji, Kohji Morimura, Haruo Ikeda[1], Masahira Hattori[1,3] and Teruhiko Beppu

Life Science Research Center, College of Bioresource Sciences, Nihon University, 1866 Kameino, Fujisawa, Kanagawa 252-8510, Japan, [1]Kitasato Institute for Life Science, Kitasato University, 1-15-1 Kitasato, Sagamihara, Kanagawa 228-8555, Japan, [2]Department of Bioactive Molecules, National Institute of Infectious Diseases, 1-23-1 Toyama, Shinjyuku-ku, Tokyo 162-8640, Japan and [3]RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan

Received June 7, 2004; Revised July 22, 2004; Accepted August 27, 2004 DDBJ/EMBL/GenBank accession no. AP006840

ABSTRACT

Symbiobacterium thermophilum **is an uncultivable bacterium isolated from compost that depends on microbial commensalism. The 16S ribosomal DNA-based phylogeny suggests that this bacterium belongs to an unknown taxon in the Gram-positive bacterial cluster. Here, we describe the 3.57 Mb genome sequence of *S.thermophilum*. The genome consists of 3338 protein-coding sequences, out of which 2082 have functional assignments. Despite the high G + C content (68.7%), the genome is closest to that of Firmicutes, a phylum consisting of low G + C Gram-positive bacteria. This provides evidence for the presence of an undefined category in the Gram-positive bacterial group. The presence of both *spo* and related genes and microscopic observation indicate that *S.thermophilum* is the first high G + C organism that forms endospores. The *S.thermophilum* genome is also characterized by the widespread insertion of class C group II introns, which are oriented in the same direction as chromosomal replication. The genome has many membrane transporters, a number of which are involved in the uptake of peptides and amino acids. The genes involved in primary metabolism are largely identified, except those that code several biosynthetic enzymes and carbonic anhydrase. The organism also has a variety of respiratory systems including Nap nitrate reductase, which has been found only in Gram-negative bacteria. Overall, these features suggest that *S.thermophilum* is adaptable to and thus lives in various environments, such that its growth requirement could be a substance or a physiological condition that is generally available in the natural environment rather than a highly specific substance that is present only in a limited niche. The genomic information from *S.thermophilum* offers new insights into microbial diversity and evolutionary sciences, and provides a framework for characterizing the molecular basis underlying microbial commensalism.**

INTRODUCTION

Molecular ecological studies have suggested that a majority of environmental microbes are still uncultured (1). Uncultured microbes include not only those in dormant states but also organisms whose appropriate culture conditions are unknown (2). The latter includes microbes whose growth depends on commensalism with animals, plants and microbes. Elucidating the genetic information of uncultivated microorganisms should provide significant insight not only into biotechnology but also into microbial physiology and evolutionary sciences. In accordance with these views, the current mass-sequencing studies have embarked on screening for unknown microbial genomes among environmental DNA libraries (3,4).

Symbiobacterium thermophilum is a thermophilic bacterium found in a commensal submerged culture that was derived from compost (5). This bacterium is characterized by a marked growth dependence on microbial commensalism; it does not grow by itself under standard culture conditions; however, when cocultured with *Bacillus* sp. strain S, it propagates up to 5×10^8 cells/ml (6). Molecular phylogeny using the 16S ribosomal DNA (rDNA) sequence has indicated that *S.thermophilum* belongs to an unknown taxonomic group in the Gram-positive bacterial cluster (7). The current 16S rDNA database content suggests that the presence of this bacterium and related organisms is still poorly recognized, probably due to the technical problems involved in its isolation. Meanwhile, our ecological study has revealed the potential phylogenetic diversity and the wide distribution of *Symbiobacterium* and related bacteria in the natural environment (8).

Here, we sequenced the whole genome of *S.thermophilum*. We find that despite the high G+C content (68.7%), the genome

*To whom correspondence should be addressed. Tel: +81 466 84 3937; Fax: +81 466 84 3935; Email: ueda@brs.nihon-u.ac.jp

is closest to that of Firmicutes, a group of low G + C Gram-positive bacteria. These results provide evidence for the potential genetic diversity in unknown microorganisms.

METHODS

Genome sequencing

The genomic DNA of *S.thermophilum* IAM14863 was isolated, using a standard technique, from pure cells cultured as described previously (6). Shotgun libraries were prepared using *Escherichia coli* DH12S, a host suited for the stable cloning of methylated DNA (Invitrogen). The entire genome sequence was obtained from a combination of 69 767 end sequences (providing 8.2-fold coverage) from a pUC118 genomic shotgun library (2–5 kb), using dye terminator chemistry on automated DNA sequencers (MegaBACE1000, Amersham Biosciences and ABI3700, Applied Bio systems). Sequence assembly was accomplished using the PHRED/PHRAP software on Consed (9). Gap closure was performed by PCR direct sequencing, using primers designed to anneal to each end of the neighboring contigs. Regions containing rDNA and group II introns were independently cloned and sequenced, and the flanking sequences were assembled with the contig sequences to obtain the finished sequence.

Informatics and comparative genomics

Transfer RNA (tRNA)-encoding regions were predicted by tRNAscan-SE (10). Potential protein-coding sequences (CDSs) were predicated using the Glimmer (11) and Genaris (T. Nishi, manuscript in preparation) programs. The predicated protein sequences were searched against a non-redundant protein database using BLASTP (12). The protein sorting signal of each CDS was analyzed by a PSORT program (13). The results were used for the manual annotation of the CDSs, which were finally assigned using FramePlot program (14) that is optimized to handle genomic sequences. The fully annotated genome sequence is available on our website (http://hp.brs.nihon-u.ac.jp/~projects/genome/STH/) and the DDBJ/EMBL/GenBank databases under the accession no. AP006840. The genomic information of various organisms was collected from the website of the National Center of Biological Information (NCBI; http://www.ncbi.nlm.nih.gov). Orthologs were identified as reciprocal best-hit pairs by using the BLASTP program. The origin and orientation of replication was identified by GC skew (15).

Condition for spore formation

Pure cells of *S.thermophilum*, prepared in a manner similar to that described above, were inoculated into 5 ml T2Y1 medium (containing 2% Bacto Tryptone, 1% Bacto Yeast Extract and 0.5% NaCl) at $\sim 1 \times 10^4$ cells/ml and cultured stationary in a silicone-plugged test tube (diameter 18 mm) for 7 days at 60°C. This condition yielded sporulating cells at 0.1–0.5%. A much higher yield (~20%) was obtained by dialysis culture using a hollow fiber module (pore size, 0.05 μm; fiber diameter, 0.5 mm; type, polysulfone) (MicroKros Module, Toyobo, Japan). This was performed by dialyzing 100 ml culture broth inoculated in a manner similar to that described above, against 1 l fresh T2Y1 medium throughout the 7 days culture period at 60°C and a flow rate of 1 ml/min.

RESULTS

General features

The random sequencing strategy indicated that *S.thermophilum* has a circular chromosome consisting of 3 566 135 bp with 68.7% G + C (Figure 1) and no extrachromosomal element. The GC skew (15) clearly indicated the direction of replication and the position of the replication origin (*oriC*). The *oriC* contains AT-rich repeated sequences, which probably serve as the binding sites for DnaA encoded immediately downstream from *oriC*.

There are six rRNA operons (16S–23S–5S) in *S.thermophilum* and all of these are oriented in the same direction as the chromosomal replication. Although *S.thermophilum* should be classified into the phylum Actinobacteria, based on the results of the 16S rDNA-based phylogeny and high G + C content (7), the structure of the 23S rRNA of *S.thermophilum* lacks the specific insertion for that group of bacteria (16). Phylogenetic analysis of the 5S rRNA gene demonstrated that the *S.thermophilum* branch is located within the low G + C bacterial group of the Gram-positive cluster.

The *S.thermophilum* genome contains 3338 CDSs (Figure 1; Table 1), which comprises 89% of the chromosome. Predicted protein sequences were searched against a non-redundant protein database, and biological roles were assigned to 2082 (62%) CDSs. A noteworthy finding was that the *S.thermophilum* proteins demonstrated a greater similarity to the proteins from Firmicutes, including Bacillae and Clostridia (low G + C Gram positives), than to those from Actinobacteria. Forty-seven percent of *S.thermophilum* proteins showed top match similarity to proteins from Firmicutes. Among these, *Thermoanaerobacter tengcongensis*, an anaerobic thermophile with 36.7% G+C (17), was the best-hit organism. A comparative analysis by a CDS similarity matrix search indicated that the *S.thermophilum* genome does not display even a partial structural similarity to other prokaryotic genomes sequenced thus far.

Movable genetic elements and horizontal gene transfer

A marked structural feature of the *S.thermophilum* genome was the presence of numerous copy numbers of group II introns. There are 22 copies carrying intact CDSs for maturase and four truncated variants. The nucleotide sequences of the copies are almost identical. This is the first instance of a prokaryotic genome inserted with such a high copy number of group II introns carrying maturase CDSs. A phylogenetic analysis indicated that the maturase belongs to the bacterial class C, as defined by Zimmerly *et al*. (18), and is very closely related to that identified in *Bacillus halodurans* (accession no. AP001507). As with other class C group II introns, all the copies of *S.thermophilum* are located in intergenic regions. Each of these is preceded by an inverted repeat structure, which probably functions as a Rho-independent transcriptional terminator. Interestingly, with the exception of STH5616, all the others exist in the same direction as the chromosomal replication (Figure 1).

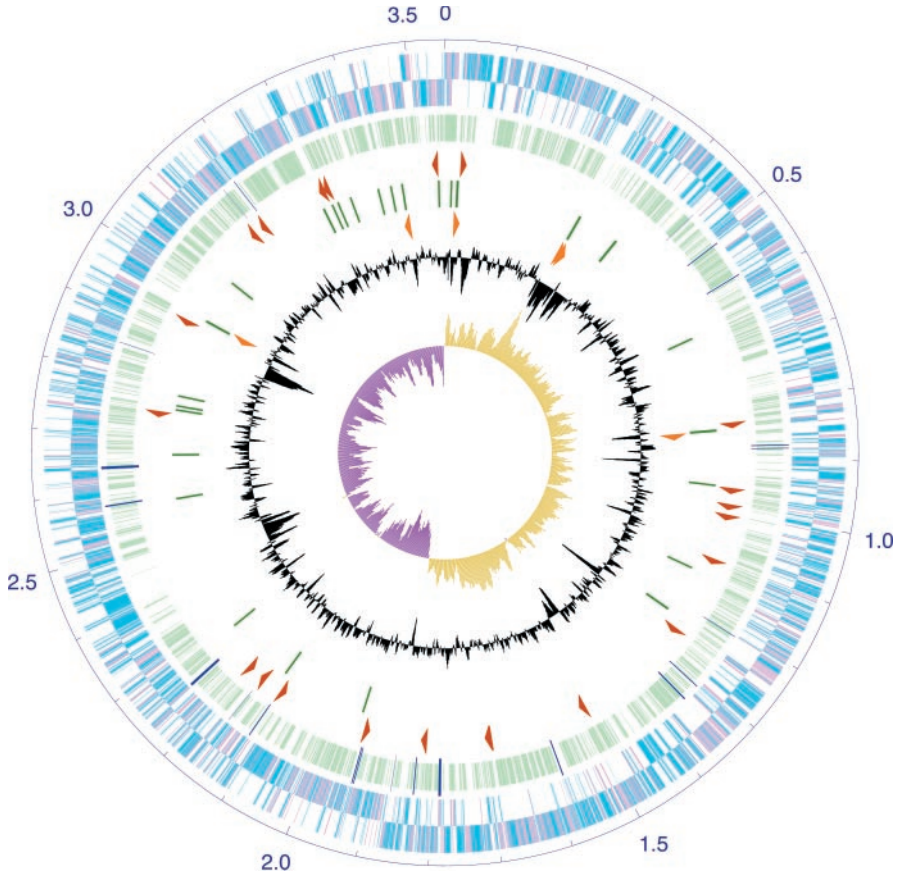

Figure 1. Circular representation of the *S.thermophilum* genome. The circles (numbered 1–8, from outside to inside) indicate: predicted protein-coding regions on the plus (circle 1) and minus (circle 2) strands; proteins, orthologs of which are present in *B.subtilis* (circle 3); group II introns (circle 4); tRNA genes (circle 5); rRNA genes (circle 6); percentage G + C (circle 7); GC skew (circle 8). Proteins showing top matches to proteins from Firmicutes are colored with pink in circles 1 and 2. CDSs related to endospore formation are shown by the wide blue bars in circle 3. The orientation of group II introns and rRNA genes are indicated by arrowheads.

Table 1. General feature of the *S.thermophilum* genome

Complete genome size	3 566 135 bp
G + C content	68.7
Total number of CDS	3338
Average CDS size	929
Percentage coding	89.3
Number of rRNA operons (16S–23S–5S)	6
Number of tRNA genes	98
Group II introns	27
Proteins similar to proteins of known function	2082
Conserved hypothetical proteins	663
Conserved domain proteins	113
Hypothetical proteins	480

S.thermophilum also carries several transposons and insertion sequences, some of which appear to be related to the acquisition of foreign genes. For example, the low G + C content region containing the lipopolysaccharide (LPS) biosynthetic gene cluster (STH2707-2718) is flanked by transposon-like sequences. A similar structure is also present in the region encoding type II restriction and modification enzymes, XmaI (STH333) and XmaI methylase (STH334). The chromosomal DNA of *S.thermophilum* cannot be digested with XmaI and isozymes in spite of the presence of 5743 XmaI recognition sites.

Cell structure

The traditional Gram-stain result indicates that *S.thermophilum* is Gram negative (5). The enterobacteria-like ability of *S.thermophilum* that leads to tryptophanase (19) and tyrosinase (20) production also suggests that this bacterium is Gram negative. However, *S.thermophilum* lacks the major Gram-negative membrane biosynthesis proteins, such as LPS:glycosyltransferase and polysaccharide transporters. This is consistent with the results of the 16S rDNA phylogeny, which indicate that *S.thermophilum* belongs to the Gram-positive bacterial group. *S.thermophilum* contains proteins associated with the S-layer (STH61, 969, 1321, 2197, 2492 and 3168) and this corresponds to our previous electron microscopy observations (7). Gram variability was observed in several S-layer-coated bacteria, such as *Bacillus* and *Clostridium* (21). As mentioned above, *S.thermophilum* has a putative, horizontally transferred LPS biosynthetic gene cluster.

The basic components for cell division common to Eubacteria are found in the *S.thermophilum* genome. The *ftsA/E/H/W/X/Z* genes were identified along with *minCDE*, thereby, specifying the position and formation of the constricting ring through their concerted function. The *mreBCD* (STH372-4) gene, which is located adjacent to the *min* locus, may determine the shape of the *S.thermophilum*

Table 2. *spo* and related genes of *S.thermophilum*

Vegetative	Stage 0	Stage II	Stage III	Stage IV	Stage V	Maturation	Germination
abrB (3253)	*spo0A* (0225, 1832)	*spoIIAA AB AC*[a] (1814-2)	*spoIIIAA AB AC AD AE AF AG AH* (1862-55)	*spoIVA* (1684)	*spoVAC AD AE* (0202-4)	*spmA B* (1786-7)	*gerKA KB KC* (1642-40, 2106-4, 2452-4)
	spo0H[a] (3113)	*spoIID* (0101, 1164)	*spoIIID* (0103)	*spoIVB* (1833)	*spoVB* (2392, 2694, 3232, 3233)	*cotH* (0783)	*cwlC* (0329)
		spoIIE (3200)	*spoIIIE* (1561, 1972)	*sigK*[a] (1958)	*spoVC* (3238)	*cotN* (1182)	*yaaH* (0478, 0537, 1982)
		spoIIGA[a]*GB* (1220-1)	*spoIIIG*[a] (1222)		*spoVD* (1205, 1991)	*ytaA* (0778)	*ydhD* (1790)
		spoIIM (1823)	*spoIIIJ jag* (3337-8)		*spoVE* (1209, 2914)	*yraD* (2020) (1181)	*sleB* (1089, 2958, 3266)
		spoIIP (0481)			*spoVG* (3241)	*sspA* (1431)	
		spoIIR (1223)			*spoVK* (1745)	*sspC* (1430, 1432, 2932)	
					spoVR (0724, 1170)	*sspF* (0419, 1720, 2659)	
					spoVS (1454, 1776)		
					spoVT (3234)		

[a]CDSs for sporulation-specific sigma factors.

cells. *S.thermophilum* has a complete flagella biosynthesis gene cluster; however, it lacks the parts for the outer-membrane-spanning components, such as FlgH (L-ring) and FlgI (P-ring). This is consistent with the Gram-positive membrane structure of *S.thermophilum*.

A remarkable finding was that a set of genes involved in endospore formation is present in the *S.thermophilum* genome (Table 2). Endospore-forming bacteria have so far been found only in two classes of Firmicutes, Bacilli and Clostridia. We previously described *S.thermophilum* as non-spore forming after observing the pure cells cultivated under the optimal conditions for their proliferation (7). We re-examined the morphological features of *S.thermophilum* and found that under specific culture conditions, the organism forms endospore-like cellular structures (Figure 2). *S.thermophilum* possesses all the sporulation-specific sigma factors, σ^H (Spo0H), σ^E (SpoIIGA), σ^F (SpoIIAC), σ^G (SpoIIIG) and σ^K (SigK), together with their specific regulators (22); however, it lacks most of the *spo0* and related genes involved in the zero-stage regulation (23). The average G + C content of sporulation-related CDSs is 89.5%.

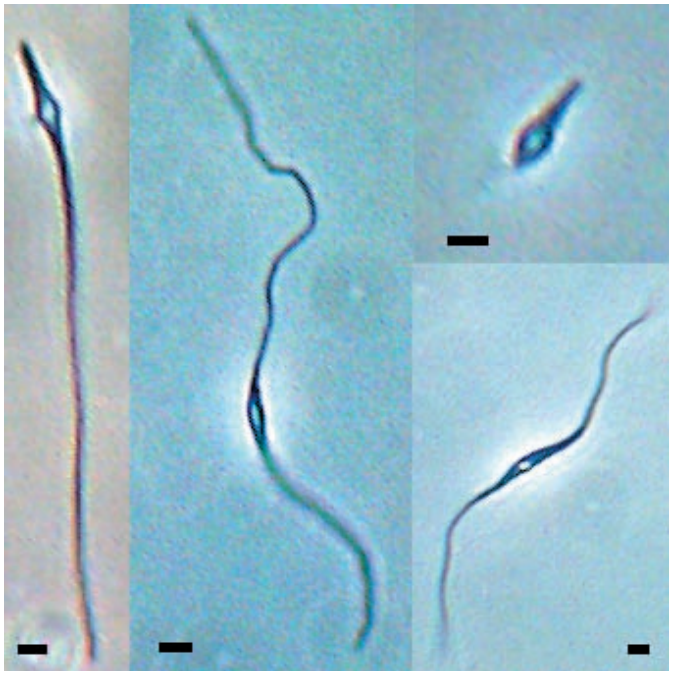

Figure 2. Optical micrograph of *S.thermophilum* cells forming a putative spore structure. Typical cells containing refractile bodies are photographed. Bar, 1 μm.

Solute uptake

There are several membrane transporters in *S.thermophilum* (Figure 3), among which the predominant transport mechanism is ATP-coupled solute efflux. The phylogeny of the substrate-binding domains is roughly parallel to that of the families defined in other Eubacteria (24) and displays a marked expansion of domains specific for oligopeptides and amino acids. A similar predominance of ABC transporters and oligopeptide permeases was reported previously with *Thermotoga maritima* (25). The other transporters include those uptaking lactate (STH96), uracil (STH226), gluconate (STH2301) and formate (STH3297). Carriers importing ions, such as Zn^{2+} (STH1625), Mg^{2+} (STH539), NH^{4+} (STH229), $PO4^{2-}$ (STH235-8, 751-46, 764-1) and $SO4^{2-}$ (STH873-5), are also present. All transporters are of the bacterial type, except for TnaT, which is the Na^+-dependent neurotransmitter transporter involved in tryptophan uptake (26). Efflux transporters that confer resistance against arsenate (STH680-1), chromate (STH1010) and zinc and/or cadmium (STH1625) and nine transporters for drug-resistance are also found in *S.thermophilum*.

Metabolism

The genes for primary metabolism were mostly identified in the *S.thermophilum* genome. The main glucose degradation

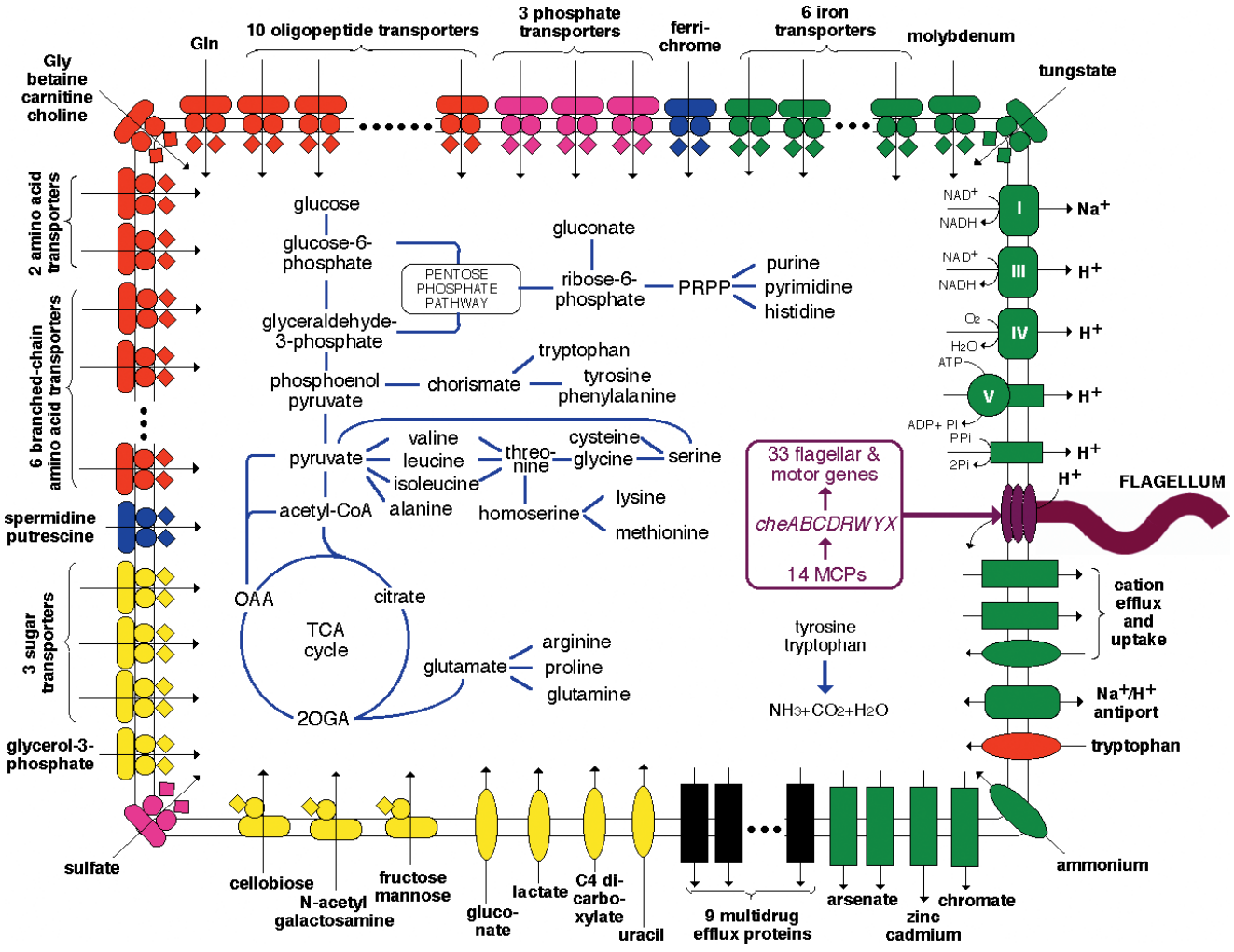

Figure 3. Overview of metabolism and transport in *S.thermophilum*. Pathways for energy production and the metabolism of organic compounds, acids and aldehydes are shown. Transporters are grouped by substrate specificity according to their role category: amino acids/peptides/amines/purines (red), carbohydrates (yellow), cations (green), anions (pink), drugs (black) and others (blue). Arrows indicate the direction of transport. Energy coupling mechanisms are also shown. PRPP, phosphoribosyl–pyrophosphate; ATP, adenosine triphosphate; ADP, adenosine diphosphate; MCP, methyl-accepting chemotaxis protein; OAA, oxalacetic acid; 2OGA, 2-oxoglutaric acid.

pathway of *S.thermophilum* is the non-oxidative branch of the pentose-phosphate glycolytic pathway (Figure 3). This organism lacks the Entner–Doudoroff pathway; however, it carries the genes for metabolizing glycerol, gluconate, cellobiose and *N*-acetylgalactosamine, as well as amino acids, such as tyrosine and tryptophan. It does not possess cellulose- and amylose-degrading enzymes. The constituents of the TCA and urea cycles, along with the genes for Co-SH-dependent ferredoxin oxidoreductases specific for aldehydes (STH2051, 2598, 2886, 3283), pyruvate (STH3262 and 3264) and 2-oxoacid (STH1316-7), are present in *S.thermophilum*.

S.thermophilum possesses the enzymes required for the biosynthesis of all essential amino acids, with the exception of a few enzymes required in methionine and lysine biosynthesis, which are also unknown in other bacteria. Similarly, biosynthetic enzymes for folate, nicotinamide and adenosylcobalamin, are completely identified, while those for biotin, thiamine, heme, pantothenate and pyridoxal are partially unknown. Large parts of the biosynthetic pathways for riboflavin and menaquinone are also unknown. None of the *S.thermophilum* proteins show similarity to carbonic anhydrase, which catalyses the conversion between carbon dioxide and bicarbonate.

Respiration

S.thermophilum proliferates under both aerobic and anaerobic conditions (6). The *S.thermophilum* genome has a variety of respiratory enzymes. There are operons that encode subunits of cytochrome *c* oxidase (STH2096-8) and sodium-dependent NADH:quinone oxidoreductase (STH896-901). This microorganism contains both aerobic (STH427) and anaerobic (STH1984-6) glycerol-3-phosphate dehydrogenase and two operons encoding the NADH dehydrogenase I complex (STH1586-98 and STH2777-67). Formate serves as an electron donor during the anaerobic respiration of *S.thermophilum* via the functioning of two formate dehydrogenase complexes (STH2602-597 and STH3098-103). For anaerobic electron acceptor reactions, there are two types of succinate dehydrogenase [cyanobacteria-type (STH2637-40) and *Bacillus*-type

(STH3176-4)] and two dimethyl sulfoxide reductases (STH713-1 and STH2332-0). It should be noted that *S. thermophilum* is the first Gram-positive bacterium that retains the Nap nitrate reductase gene cluster (STH918-5). The Nar nitrate reductase (STH2056) is also present in *S.thermophilum*. These systems probably enable *S.thermophilum* to grow by nitrate respiration, as reported for *Symbiobacterium toebii*, a close relative of *S.thermophilum* (27).

Replication, recombination and DNA repair

S.thermophilum has three paralogs for *dnaE* that encode the α-subunit of DNA polymerase III holocomplex (STH659, 1512 and 1885). The presence of multiple *dnaE* genes is known in several infective and symbiotic bacteria, such as *Mycobacterium* and *Mesorhizobium*. These genes are related to error-prone replication and facilitate DNA sequence heterogeneity (28). The *dinP* homolog that encodes DNA polymerase IV is not found in *S.thermophilum*. Meanwhile, *S.thermophilum* has two paralogs for DNA polymerase I (STH848, 2679). The *rec* family genes present in *S.thermophilum* are *recA*, *recF*, *recN*, *recO* and *recR*, and *S.thermophilum* lacking *recBCD* probably utilizes resolvasome encoded by *ruvABC* (STH1159-61) to process the Holliday junction. Although *S.thermophilum* has the LexA protein, which implies the presence of the SOS regulon, the genome does not contain the *uvr* and *umu* homologs that are involved in SOS DNA repair. An analogous situation to such genetic impairment of DNA repair systems was reported in *Buchnera* sp., an endocellular symbiont of aphids (29). *S.thermophilum* possesses *mfd*, which encodes a transcription-repair coupling factor (STH3235) and a uracil mismatch repair protein (STH940).

Transcription and translation

The five subunits (α, β, β′, δ and ω) of the core RNA polymerase in *S.thermophilum* are encoded by *rpoA/B/C/E/Z*, respectively. The *S.thermophilum* genome codes for 27 sigma factors, including 18 ECF (extra-cytoplasmic function), five sporulation-specific (σ^H, σ^E, σ^F, σ^G and σ^K), one flagella biosynthesis-specific (σ^{FliA}), one major (σ^{RpoD}), one minor (STH1162) and one enhancer-dependent (σ^{54}) sigma factor. *S.thermophilum* does not possess the stress-response family of sigma factors as represented by σ^B of *Bacillus subtilis*, which are involved in general stress-response.

Transcription termination in *S.thermophilum* is intriguing since this organism does not have the Rho and Nus family proteins with the exception of the NusB homolog (STH1849). A sequence motif search by the TransTerm program (30) predicted 437 Rho-independent transcriptional terminators. It is evident that the number is far lesser than the number of transcriptional units, implying the presence of an alternative mechanism for transcriptional termination in this organism. The motif of the T-box antitermination system identified previously in the mRNA leader region for the leucyl-tRNA synthase gene (STH444) (31), was also found in the corresponding position of tRNA synthase genes for glutamine (STH16), valine (STH366), lysine (STH525), phenylalanine (STH1105), tyrosine (STH1122), isoleucine (STH1231), proline(STH1502), alanine (STH2000) and methionine (STH3252).

S.thermophilum possesses a complete set of ribosomal proteins and each component has a single copy number with the exception of the S4 subunit, which has two copies. The *S. thermophilum* genome is characterized by the presence of many tRNAs (98 copies) specifying 43 codons, including UGA for selenocysteine. The tRNA synthase for all amino acids except glutamine is also present. The transamidation of Glu-tRNA^{Gln} to Gln-tRNA^{Gln} by Glu-tRNA^{Gln} aminotransferase (STH2820-2) could be an alternative synthesis mechanism, as observed in other bacteria.

Response to environmental stimuli

There are 23 pairs of sensor histidine kinases and response regulators in *S.thermophilum*, which are probably involved in responding to environmental stimuli. Among these, eight pairs are located close to the ABC transporter genes and this suggests that they play a role in cellular processes involving membrane transport. For example, the pair encoded by STH750 and 749 are located in a phosphate ABC transporter operon, thereby, suggesting their role in phosphate response. A two-component signaling pathway assembled from *che* gene products (CheA/B/C/D/R/W/Y/X) and 13 methyl-accepting chemotactic transducer proteins (MCPs) that probably regulate chemotaxis are also identified. The majority of the MCPs of *S.thermophilum* show a distinct similarity to those identified in *Desulfitobacterium hafniense* and *B.halodurans*.

Protein secretion

S.thermophilum possesses four type I and three type II signal peptidases and 729 proteins contain signal sequences. In addition to the SecA-dependent mechanism, *S.thermophilum* has a type III secretion system assembled from Fli and FlhA/B proteins associated with the flagellum assembly; however, it lacks several of the competence genes, which are widely distributed in *Bacillus*.

DISCUSSION

This study revealed the unique features of the *S.thermophilum* genome. A comparative analysis indicated that the genome does not exhibit structural similarity to that of other prokaryotes known thus far. The taxonomic features of *S.thermophilum* were of interest because of both its unique physiological properties that depend on microbial commensalism and the fact that it is left unrecognized by microbiologists. Although current bacterial systematics should classify the organism into Actinobacteria based on its high GC content (7), the *S.thermophilum* proteins showed marked similarity to those from Bacilli and Clostridia (low GC Gram-positive bacteria). This suggests that *S.thermophilum* shared a close common ancestor with Bacilli/Clostridia and evolved from it under selective pressure, which has driven a noticeable genetic drift in the third letter GC content of the CDSs. Although genomic GC content is an important marker in the taxonomy of Gram-positive bacteria, it is inappropriate for the evaluation of the taxonomic and evolutionary characteristics of *S.thermophilum*. Future study on other bacterial genomes may reveal additional instances of 'Actinobacteria' belonging to 'Firmicutes' and vice versa.

A remarkable finding related to the above-mentioned results was that the *S.thermophilum* genome contains a set of genes

involved in endospore formation, and that the organism is actually capable of forming endospores. Till date, the general understanding has been that the ability to form endospores is distributed only in the Bacilli and Clostridia, and hence the sporulation-related genes known thus far have a low GC content. On the other hand, the GC content of the *S.thermophilum* counterparts was as high as that of the entire genome, which provides the first genetic evidence of the presence of high GC endospore-forming bacterium. The cluster structures of functionally related *spo* genes characterized in *B.subtilis* are highly conserved in *S.thermophilum*, which suggests that the proteins encoded by each cluster also have a concerted function in *S.thermophilum*. *S.thermophilum* retains the genes related to the so-called sigma cascade, which regulates endospore development (22). This suggests that each step in spore formation is controlled in a manner similar to that of *B.subtilis*. Meanwhile, *S.thermophilum* lacks genes for stage zero regulation (23), which suggests that the decision mechanism for the onset of spore formation is different from that of *B.subtilis*. This might reflect the differences in the environmental niches in which these organisms live and/or the environmental and physiological signals to which they need to respond to, in order to initiate endospore formation. Currently, the germination conditions of the putative spores are unknown, and therefore, we do not possess the data pertaining to the physiological properties of the spore-like cellular structure, such as tolerance against excess temperature and dryness.

Another distinctive feature of the *S.thermophilum* genome was the presence of a large number of group II introns. Group II introns are unique genetic elements that were initially identified in the organellar genomes of plants and eukaryotes (32). These genetic elements are also widespread among bacterial genomes, and this has resulted in the current understanding that they originate in bacteria and, subsequently, spread to the organelles. The group II introns identified in the bacterial genome are classified into five categories according to their sites of insertion (18). Among these, the class C group II introns are found in intergenic regions preceded by Rho-independent transcriptional terminators (32). All copies in the *S.thermophilum* genome were also found downstream of the inverted repeat structures, which strongly reinforces the notion that the insertion mechanism of the type of group II intron is related to the structure or function of the transcriptional terminator. In addition, we found that almost all the copies of *S.thermophilum* are oriented in the same direction as the chromosomal replication deduced from the GC skew analysis. This implies that the insertion of the type of group II intron is under a mechanism coupled to DNA replication, which directs the maturase CDS in the same orientation as the chromosomal replication. The high copy number of almost identical DNA sequences may have mediated the genetic rearrangement during the evolution of the genome. The widespread insertion of group II introns was previously reported in the genome of *Thermosynechococcus elongatus* (33).

The unveiling of the *S.thermophilum* genome indicated that the organism does not possess any large-scale genetic lesion for primary metabolism like that observed with the genome of symbiotic microorganisms, such as *Buchnera* sp. (29) and *Lactobacillus johnsonii* (34). Furthermore, the presence of a wide variety of respiratory enzymes suggests that the energy metabolism in this organism is highly adaptable to different environments. Meanwhile, we found that the genome appears to lack the genes for carbonic anhydrase. Our present study revealed that the impaired yet distinct self-growth of *S.thermophilum* occurs when CO_2 gas was introduced into the culture broth (K. Ueda and T. Beppu, manuscript in preparation). It is known that carbonic anhydrase deficiency confers an essential requirement for high CO_2 concentrations in *Ralstonia eutropha* (35) and *E.coli* (36,37). Therefore, we speculate that the requirement for CO_2 by *S.thermophilum* is due to the deficiency of carbonic anhydrase. The CO_2 concentration sufficient for *S.thermophilum* growth in culture media may be achieved by the growth of *Bacillus* sp., such that *S.thermophilum* proliferates in the commensal culture with *Bacillus*. CO_2 supply can be one of the important factors that support the commensal growth of *S.thermophilum*.

The *S.thermophilum* genome is characterized by the presence of numerous genes that encode solute uptake systems. Several ABC transporters that show a marked diversity in the solute binding domains, including those specifying amino acids and peptides, are also present. Our recent study has shown that miscellaneous peptidic fractions, generated by an extracellular protease of *Bacillus* sp., stimulate unbalanced yet remarkable growth of *S.thermophilum* (K. Ueda and T. Beppu, unpublished data). *S.thermophilum* may favor the utilization of amino acids and/or peptides exported from the outside of the cell rather than their synthesis using its own biosynthetic mechanisms. Alternatively, a specific peptide may act as a growth-stimulating agent for the bacterium. A secreted protease is not present in *S.thermophilum*, indicating that the organism could depend on the proteolytic activity of the cognate *Bacillus* sp.

In addition to ABC transporters, elucidation of the *S.thermophilum* genome indicated the presence of a wide variety of proteins related to adaptive response, including high copy numbers of ECF sigma factors, two-component regulatory systems and MCPs. Overall, these features suggest that the organism lives in environments that are subjected to relatively frequent physiological changes, rather than in a confined niche under specific and stable conditions. This is consistent with our ecological study, which has shown that *S.thermophilum* and related organisms are widespread in compost, soil, animal feces and feeds (8) and marine environments (K. Ueda and T. Beppu, unpublished data). We speculate that the requirements of the free-living bacterium could be the complex physiological conditions rather than simple nutrition or cofactors. An adequate condition, which includes an increased CO_2 concentration and supply of peptides, could be successfully created by the simultaneous growth of cognate *Bacillus* sp. in laboratory culture. This genomic information will assist future comprehensive biochemical characterizations of the housekeeping gene products of *S.thermophilum*, which may reveal not only the molecular basis underlying the commensalophilic property of *S.thermophilum*, but may also help to develop new culture techniques for unknown microorganisms.

ACKNOWLEDGEMENTS

We thank N. Ogasawara, F. Kawamura, T. Kudo, H. Nishida and Y. Shiga for helpful discussions. We also thank K. Furuya, C. Yoshino, Y. Yamashita and A. Nakazawa for technical assistance. This study was supported by the JSPS Research

for the Future Program, 21st century COE Program and the Grant-in-aid for scientific research (accession no. 15013255) by MEXT, Japan.

REFERENCES

1. Amann,R.I., Ludwig,W. and Schleifer,K.H. (1995) Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiol. Rev.*, **59**, 143–169.
2. Barer,M.R. and Harwood,C.R. (1999) Bacterial viability and culturability. *Adv. Microb. Physiol.*, **41**, 93–137.
3. Tyson,G.W., Chapman,J., Hugenholtz,P., Allen,E.E., Ram,R.J., Richardson,P.M., Solovyev,V.V., Rubin,E.M., Rokhsar,D.S. and Banfield,J.F. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428**, 37–43.
4. Venter,J., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,J.A., Wu,D., Paulsen,I., Nelson,K.E., Nelson,W.*et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.
5. Suzuki,S., Horinouchi,S. and Beppu,T. (1988) Growth of a tryptophanase-producing thermophile, *Symbiobacterium thermophilum* gen. nov., sp. nov., is dependent on co-culture with a *Bacillus* sp. *J. Gen. Microbiol.*, **134**, 2353–2362.
6. Ohno,M., Okano,I., Watsuji,T., Kakinuma,T., Ueda,K. and Beppu,T. (1999) Establishing the independent culture of a strictly symbiotic bacterium *Symbiobacterium thermophilum* from its supporting *Bacillus* strain. *Biosci. Biotechnol. Biochem.*, **63**, 1083–1090.
7. Ohno,M., Shiratori,H., Park,M.J., Saitoh,Y., Kumon,Y., Yamashita,N., Hirata,A., Nishida,H., Ueda,K. and Beppu,T. (2000) *Symbiobacterium thermophilum* gen. nov., sp. nov., a symbiotic thermophile that depends on co-culture with a *Bacillus* strain for growth. *Int. J. Syst. Evol. Microbiol.*, **50** (Pt 5), 1829–1832.
8. Ueda,K., Ohno,M., Yamamoto,K., Nara,H., Mori,Y., Shimada,M., Hayashi,M., Oida,H., Terashima,Y., Nagata,M.*et al.* (2001) Distribution and diversity of symbiotic thermophiles, *Symbiobacterium thermophilum* and related bacteria, in natural environments. *Appl. Environ. Microbiol.*, **67**, 3779–3784.
9. Gordon,D., Desmarais,C. and Green,P. (2001) Automated finishing with autofinish. *Genome Res.*, **11**, 614–625.
10. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
11. Delcher,A.L., Harmon,D., Kasif,S., White,O. and Salzberg,S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, **27**, 4636–4641.
12. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
13. Horton,P. and Nakai,K. (1996) A probabilistic classification system for predicting the cellular localization sites of proteins. *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, St. Louis, CA, AAAI Press, pp. 109–115.
14. Ishikawa,J. and Hotta,K. (1999) FramePlot: a new implementation of the frame analysis for predicting protein-coding regions in bacterial DNA with a high G + C content. *FEMS Microbiol. Lett.*, **174**, 251–253.
15. Grigoriev,A. (1998) Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res.*, **26**, 2286–2290.
16. Roller,C., Ludwig,W. and Schleifer,K.H. (1992) Gram-positive bacteria with a high DNA G + C content are characterized by a common insertion within their 23S rRNA genes. *J. Gen. Microbiol.*, **138** (Pt 6), 1167–1175.
17. Bao,Q., Tian,Y., Li,W., Xu,Z., Xuan,Z., Hu,S., Dong,W., Yang,J., Chen,Y., Xue,Y.*et al.* (2002) A complete sequence of the *T. tengcongensis* genome. *Genome Res.*, **12**, 689–700.
18. Zimmerly,S., Hausner,G. and Wu,X. (2001) Phylogenetic relationships among group II intron ORFs. *Nucleic Acids Res.*, **29**, 1238–1250.
19. Hirahara,T., Suzuki,S., Horinouchi,S. and Beppu,T. (1992) Cloning, nucleotide sequences, and overexpression in *Escherichia coli* of tandem copies of a tryptophanase gene in an obligately symbiotic thermophile, *Symbiobacterium thermophilum*. *Appl. Environ. Microbiol.*, **58**, 2633–2642.
20. Hirahara,T., Horinouchi,S. and Beppu,T. (1993) Cloning, nucleotide sequence, and overexpression in *Escherichia coli* of the beta-tyrosinase gene from an obligately symbiotic thermophile, *Symbiobacterium thermophilum*. *Appl. Microbiol. Biotechnol.*, **39**, 341–346.
21. Beveridge,T.J. (1990) Mechanism of Gram variability in select bacteria. *J. Bacteriol.*, **172**, 1609–1620.
22. Stragier,P. and Losick,R. (1996) Molecular genetics of sporulation in *Bacillus subtilis*. *Annu. Rev. Genet.*, **30**, 297–241.
23. Hoch,J.A. and Varughese,K.I. (2001) Keeping signals straight in phosphorelay signal transduction. *J. Bacteriol.*, **183**, 4941–4949.
24. Saier,M.H.,Jr (1994) Computer-aided analyses of transport protein sequences: gleaning evidence concerning function, structure, biogenesis, and evolution. *Microbiol. Rev.*, **58**, 71–93.
25. Nelson,K.E., Clayton,R.A., Gill,S.R., Gwinn,M.L., Dodson,R.J., Haft,D.H., Hickey,E.K., Peterson,J.D., Nelson,W.C., Ketchum,K.A.*et al.* (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature*, **399**, 323–329.
26. Androutsellis-Theotokis,A., Goldberg,N.R., Ueda,K., Beppu,T., Beckman,M.L., Das,S., Javitch,J.A. and Rudnick,G. (2003) Characterization of a functional bacterial homologue of sodium-dependent neurotransmitter transporters. *J. Biol. Chem.*, **278**, 12703–12709.
27. Rhee,S.K., Jeon,C.O., Bae,J.W., Kim,K., Song,J.J., Kim,J.J., Lee,S.G., Kim,H.I., Hong,S.P., Choi,Y.H.*et al.* (2002) Characterization of *Symbiobacterium toebii*, an obligate commensal thermophile isolated from compost. *Extremophiles*, **6**, 57–64.
28. Boshoff,H.I., Reed,M.B., Barry,C.E.,III and Mizrahi,V. (2003) DnaE2 polymerase contributes to *in vivo* survival and the emergence of drug resistance in *Mycobacterium tuberculosis*. *Cell*, **113**, 183–193.
29. Shigenobu,S., Watanabe,H., Hattori,M., Sakaki,Y. and Ishikawa,H. (2000) Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature*, **407**, 81–86.
30. Ermolaeva,M.D., Khalak,H.G., White,O., Smith,H.O. and Salzberg,S.L. (2000) Prediction of transcription terminators in bacterial genomes. *J. Mol. Biol.*, **301**, 27–33.
31. Grundy,F.J. and Henkin,T.M. (1999) A regulatory system hitherto found only in Gram-positive bacteria in a Gram-negative bacterium that grows only in co-culture with a *Bacillus* strain. *Mol. Microbiol.*, **33**, 667–668.
32. Dai,L. and Zimmerly,S. (2002) Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. *Nucleic Acids Res.*, **30**, 1091–1102.
33. Nakamura,Y., Kaneko,T., Sato,S., Ikeuchi,M., Katoh,H., Sasamoto,S., Watanabe,A., Iriguchi,M., Kawashima,K., Kimura,T.*et al.* (2002) Complete genome structure of the thermophilic cyanobacterium *Thermosynechococcus elongatus* BP-1. *DNA Res.*, **9**, 123–130.
34. Pridmore,R.D., Berger,B., Desiere,F., Vilanova,D., Barretto,C., Pittet,A.C., Zwahlen,M.C., Rouvet,M., Altermann,E., Barrangou,R.*et al.* (2004) The genome sequence of the probiotic intestinal bacterium *Lactobacillus johnsonii* NCC 533. *Proc. Natl Acad. Sci. USA*, **101**, 2512–2517.
35. Kusian,B., Sultemeyer,D. and Bowien,B. (2002) Carbonic anhydrase is essential for growth of *Ralstonia eutropha* at ambient CO_2 concentrations. *J. Bacteriol.*, **184**, 5018–5026.
36. Merlin,C., Masters,M., McAteer,S. and Coulson,A. (2003) Why is carbonic anhydrase essential to *Escherichia coli*? *J. Bacteriol.*, **185**, 6415–6424.
37. Hashimoto,M. and Kato,J. (2003) Indispensability of the *Escherichia coli* carbonic anhydrases YadF and CynT in cell proliferation at a low CO_2 partial pressure. *Biosci. Biotechnol. Biochem.*, **67**, 919–922.