Extremophiles (2008) 12:415–429
DOI 10.1007/s00792-008-0145-y

ORIGINAL PAPER

Complete nucleotide sequence of pGS18, a 62.8-kb plasmid from *Geobacillus stearothermophilus* strain 18

Milda Stuknyte · Simone Guglielmetti · Diego Mora · Nomeda Kuisiene · Carlo Parini · Donaldas Citavicius

Received: 7 November 2007 / Accepted: 22 January 2008 / Published online: 28 February 2008

Abstract The complete nucleotide sequence (62.8 kb) of pGS18, the largest sequenced plasmid to date from the species *Geobacillus stearothermophilus*, was determined. Computational analysis of sequence data revealed 65 putative open reading frames (ORFs); 38 were carried on one strand and 27 were carried on the other. These ORFs comprised 84.1% of the pGS18 sequence. Twenty-five ORFs (38.4%) were assigned to putative functions; four ORFs (6.2%) were annotated as pseudogenes. The amino acid sequences obtained from 29 ORFs (44.6%) had the highest similarity to hypothetical proteins of the other microorganisms, and seven (10.8%) had no significant similarity to any genes present in the current open databases. Plasmid replication region, strongly resembling that of the theta-type replicon, and genes encoding three different plasmid maintenance systems were identified, and a putative discontinuous transfer region was localized. In addition, we also found several mobile genetic elements and genes, responsible for DNA repair, distributed along the whole sequence of pGS18. The alignment of pGS18 with two other large indigenous plasmids of the genus *Geobacillus* highlighted the presence of well-conserved segments and has provided a framework that can be exploited to formulate hypotheses concerning the molecular evolution of these three plasmids.

Keywords *Geobacillus stearothermophilus* · Plasmid · Complete nucleotide sequence · Comparison

Communicated by F. Robb.

M. Stuknyte (✉) · N. Kuisiene · D. Citavicius
Department of Plant Physiology and Microbiology,
Faculty of Natural Sciences, Vilnius University,
Ciurlionio 21/27, 03101 Vilnius, Lithuania
e-mail: Milda.Stuknyte@gf.vu.lt

S. Guglielmetti · D. Mora · C. Parini
Department of Food Science and Microbiology,
Industrial Microbiology Section, Faculty of Agriculture,
University of Milan, Via Celoria 2, 20133 Milan, Italy

Introduction

Plasmids are important in the overall physiology and contribute directly in the adaptation of the bacteria to their natural environments (Ochman et al. 2000). Members of the gram-positive thermophilic genus *Geobacillus* produce a variety of industrially and biotechnologically important thermostable enzymes and antibacterial substances (McMullan et al. 2004). Several of them are plasmid-encoded (Mielenz 1983; Stahl 1991). Plasmids were isolated only from a limited number of *Geobacillus* spp. strains (Bingham et al. 1979; Feng et al. 2007; Hoshino et al. 1985; Imanaka et al. 1981, 1982, 1984; Khalil et al. 2003; Liao et al. 1986; Mielenz 1983; Nakayama et al. 1993; Stahl 1991; Takami et al. 2004) and their sizes vary from 1.9 to 108.0 kb. Native *Geobacillus* spp. plasmids are of particular interest, because they can be used in developing genetic engineering tools, which would allow manipulating the genome of this bacterium. The plasmid biology of this genus is still poorly understood. Until now, there are only three completely sequenced *Geobacillus* spp. plasmids published: cryptic pSTK1 (1,883 bp) from *G. stearothermophilus*, pHTA426 (47,890 bp) from *G. kaustophilus* and pLW1071 (57,693 bp) from *G. thermodenitrificans* (Feng et al. 2007; Nakayama et al. 1993; Takami et al. 2004).

Up to date, only a few shuttle vectors between *G. stearothermophilus*, *B. subtilis* and *E. coli* have been generated: pSTE33, pRP9 and pNW33N (De Rossi et al.

1994; Mee and Welker unpublished results; Narumi et al. 1993). All of them are based on rolling-circle-replicating replicons, and none is based on theta. Analysis of new large plasmids may provide novel theta replicon types for genetic engineering of *G. stearothermophilus*.

In the current report, we present the complete nucleotide sequence and an initial analysis of pGS18, a large (62.8 kb) indigenous plasmid from *G. stearothermophilus* strain 18 that was isolated from the oil-fields near the Baltic Sea in Lithuania (Kuisiene unpublished results), and its in-silico comparison with the other two large *Geobacillus* spp. plasmid sequences available to date.

Materials and methods

Bacterial strains and culture conditions

Geobacillus stearothermophilus strain 18 was grown on NA (Nutrient agar, Biokar Diagnostics, France) plates at 60°C. *Escherichia coli* XL1-Blue strain used in transformation experiments was grown in SOB medium supplemented with 100 μg/ml of ampicillin at 37°C (Sambrook and Russell 2001).

Plasmid DNA isolation

Cells of *Geobacillus stearothermophilus* strain 18 were washed twice with modified STE buffer [200 mM NaCl, 10 mM Tris–HCl (pH 8.0), 1 mM EDTA (pH 8.0)], and native plasmid DNA was isolated using NucleoBond® BAC 100 extraction kit (Macherey-Nagel, Germany) following the subsequent purification by CsCl–ethidium bromide gradient centrifugation (Sambrook and Russell 2001).

Plasmid DNA library construction and sequencing strategies

The sequencing strategy used to assemble the complete plasmid sequence was based on the primer walking procedure. Two restriction enzyme—*Eco*RI and *Hin*dIII—fragment libraries in pUC19 were constructed from the pGS18 plasmid. Recombinant plasmids were introduced into Inoue "ultra-competent" *E. coli* XL1-Blue, as described by Sambrook and Russell (2001). Transformants were selected by blue/white screening on X-Gal/IPTG (40 and 400 μg/ml, respectively) indicator plates containing ampicillin (100 μg/ml). Templates for DNA sequencing were prepared using NucleoSpin® Plasmid extraction kit (Macherey-Nagel, Germany). DNA sequencing was performed with the deoxy chain terminator principle (Sanger et al. 1977) in the DNA Sequencing Centre of the Institute of Biotechnology (Lithuania) by use of the Big-Dye® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, USA) and 3130*xl* Genetic Analyzer (Applied Biosystems, USA). Plasmid DNA from the libraries was sequenced in both directions by primer walking with a series of custom primers and universal primers from MBI Fermentas (Lithuania) and metabion international AG (Germany). Primers for primer walking were designed using *Vector NTI Advance*™ *9.0* (InforMax, USA) software and *PrimerSelect* module from *Lasergene 6* (DNASTAR, USA). Direct sequencing from pGS18 was used for closing sequence gaps. Hundred and twenty-one oligonucleotides were designed for primer walking of pGS18 to close gaps and increase overall coverage to an average of 3×.

DNA sequence analysis

Sequence assembly was done with *Vector NTI Advance*™ *9.0* software (InforMax, USA). The predicted protein-coding regions were initially defined by searching for ORFs longer than 100 codons. The potential coding regions were confirmed with prokaryotic gene finder *GeneMark.hmm 2.4 for Prokaryotes* (Lukashin and Borodovsky 1998) at http://opal.biology.gatech.edu/GeneMark/gmhmm2_prok.cgi using a hidden Markov model (HMM) pretrained on the *G. kaustophilus* HTA426 genome. Additionally, the whole pGS18 sequence was translated in all six reading frames with *Vector NTI Advance*™ *9.0* software (InforMax, USA). The resulting amino acid sequences (longer than 33 amino acids) were manually analyzed taking attention to the sequence stretches that were not covered by *GeneMark.hmm*. Promoter prediction was done using *Promoter Prediction by Neural Network* (http://www.fruitfly.org/seq_tools/promoter.html). Promoter −10 and −35 position determination was accomplished using *BPROM* provided by SoftBerry (http://www.softberry.com/berry.phtml). To include possible ORFs, the presence of putative promoter (scores ≥0.75 and the presence of significant −10 and −35 promoter positions) and/or putative Shine-Dalgarno sites had to be determined. Potential ORFs were subsequently manually analyzed by database searches using the *BLAST* suite of programs [including blastx, clusters of orthologous groups (COG) and conserved domain database (CDD)] (Altschul et al. 1990), which incorporate domains imported from the simple modular architecture research tool (SMART) (Letunic et al. 2006) and the protein family database (Pfam) (Finn et al. 2006) using standard default *BLAST* parameters. *E* values for the most relevant homologs are presented in Table 1. The CD-search service was applied

Table 1 ORFs of pGS18: the most relevant homologies and comparison with pHTA426 and pLW1071

pGS18 ORF name[a]	Position in sequence		Protein length (aa)	Most relevant homolog	Organism (plasmid)	Number of identities/ number examined (%)	*E* value	GenBank accession number	pHTA426		pLW1071	
	s[b]	bp							ORF[a]	id %[c]	ORF[a]	id %[c]
#1	+	105–902	265	Chromosome partitioning protein	*Geobacillus kaustophilus* HTA426 (pHTA426)	264/265 (99)	5e−151	YP_145847	#36	99	–	–
#2	+	886–1,173	95	Hypothetical protein GKP35	*Geobacillus kaustophilus* HTA426 (pHTA426)	94/95 (98)	1e−46	YP_145846	#35	98	–	–
#3	+	1,464–2,744	426	Replication protein	*Geobacillus kaustophilus* HTA426 (pHTA426)	422/426 (99)	0.0	YP_145815	#04	99	–	–
#4	–	3,851–3,630	73	Hypothetical protein	*Geobacillus* phage GBSV1	45/71 (63)	7e−19	YP_764495	–	–	–	–
#5	+	3,938–4,783	281	Late competence protein	*Geobacillus kaustophilus* HTA426 (pHTA426)	280/281 (99)	4e−155	YP_145816	#05	99	–	–
#6	+	4,794–5,009	71	Hypothetical protein GKP06	*Geobacillus kaustophilus* HTA426 (pHTA426)	69/71 (97)	2e−33	YP_145817	#06	97	–	–
#7	–	5,597–5,214	127	Hypothetical protein GKP07	*Geobacillus kaustophilus* HTA426 (pHTA426)	126/127 (99)	1e−68	YP_145818	#07	99	–	–
#8	–	6,241–6,032	69	Integrase/recombinase Xer (N-terminal truncation)	*Geobacillus kaustophilus* HTA426 (pHTA426)	67/69 (97)	3e−32	YP_145824	#13	97	#3492	97
#9	–	6,447–6,268	59	DNA-damage repair protein, ImpB/MucB/ SamB family GTNG_2069 (C-terminal truncation)	*Geobacillus thermodenitrificans* NG80-2	32/34 (94)	2e−11	YP_001126166	–	–	–	–
#10	+	6,941–8,839	632	Type IIs modification methyltransferase	*Geobacillus kaustophilus* HTA426 (pHTA426)	627/632 (99)	0.0	YP_145819	#08	99	–	–
#11	+	8,861–10,555	564	Type IIs restriction endonuclease	*Geobacillus kaustophilus* HTA426 (pHTA426)	562/564 (99)	0.0	YP_145820	#09	99	–	–
#12	–	12,410–10,644	588	Hypothetical protein GKP10	*Geobacillus kaustophilus* HTA426 (pHTA426)	586/588 (99)	0.0	YP_145821	#10	99	–	–
#13	–	12,966–12,430	178	Hypothetical protein GKP11	*Geobacillus kaustophilus* HTA426 (pHTA426)	177/178 (99)	2e−101	YP_145822	#11	99	–	–
#14	–	13,404–13,150	84	No significant similarity					–	–	–	–
#15	–	14,357–13,443	304	Putative integrase/ recombinase XerD	*Geobacillus thermodenitrificans* NG80-2	291/304 (95)	3e−166	YP_001126164	#13	93	#3492	94
#16	+	14,627–16,075	482	Transposase of ISBst12-like element	*Geobacillus kaustophilus* HTA426	479/482 (99)	0.0	YP_146148	#12	58	–	–

Table 1 continued

pGS18 ORF name[a]	Position in sequence		Protein length (aa)	Most relevant homolog	Organism (plasmid)	Number of identities/number examined (%)	*E* value	GenBank accession number	pHTA426		pLW1071	
	s[b]	bp							ORF[a]	id %[c]	ORF[a]	id %[c]
#17	+	16,805–17,962	385	Protein of unknown function DUF455	*Dechloromonas aromatica* RCB	60/214 (289)	8e−15	YP_284345	–	–	–	–
#18	+	18,084–18,413	109	Transposase IS3/IS911	*Bacillus weihenstephanensis* KBAB4	49/98 (50)	1e−19	ZP_01183073	–	–	–	–
#19[d]	+	18,413–19,277							–	–	–	–
#20	–	20,369–19,647	240	Hypothetical protein OB0303	*Oceanobacillus iheyensis* HTE831	107/237 (45)	2e−56	NP_691224	–	–	–	–
#21	+	20,522–21,187	221	No significant similarity					–	–	–	–
#22	+	21,168–21,479	103	No significant similarity					–	–	–	–
#23	+	21,476–23,248	590	Methyltransferase type 11	*Alkaliphilus metalliredigenes* QYMF	63/234 (26)	2e−09	YP_001319591	–	–	–	–
#24	+	23,271–23,414	47	No significant similarity					–	–	–	–
#25	+	23,509–24,729	406	Hypothetical protein MJ1157	*Methanocaldococcus jannaschii* DSM 2661	688/293 (23)	1e−08	NP_248152	–	–	–	–
#26	+	24,756–25,499	247	Purine and other phosphorylases, family 1	*Azotobacter vinelandii* AvOP	66/216 (30)	7e−19	ZP_00419195	–	–	–	–
#27	+	25,505–26,764	419	Predicted transporter protein	*Clostridium kluyveri* DSM 555	110/403 (27)	5e−31	EDK34268	–	–	–	–
#28	+	26,766–28,040	424	UvrB/UvrC protein:AAA ATPase, central region: Clp, N terminal	*Chlorobium limicola* DSM 245	168/442 (38)	1e−75	ZP_00511878	–	–	–	–
#29[d]	–	28,976–28,378							–	–	–	–
#30	+	29,349–29,579	76	No significant similarity					–	–	–	–
#31	–	30,439–29,822	205	Hypothetical protein fc35	Uncultured bacterium	76/184 (41)	7e−31	CAI78847	–	–	–	–
#32	–	31,182–30,451	243	Hypothetical protein PY05513	*Plasmodium yoelii yoelii* str. 17XNL	39/158 (24)	0.036	XP_725982	–	–	–	–
#33	+	31,615–32,055	146	DNA repair protein RadC	*Listeria monocytogenes* str. 4b H785	102/146 (69)	7e−54	ZP_00231288	–	–	#3449	84
#34	+	32,247–32,738	163	Hypothetical protein GTNG_3450	*Geobacillus thermodenitrificans* NG80-2 (pLW1071)	162/163 (99)	5e−89	YP_001127528	–	–	#3450	99
#35	–	33,461–33,030	143	Hypothetical protein GTNG_3451	*Geobacillus thermodenitrificans* NG80-2 (pLW1071)	106/145 (73)	2e−55	YP_001127529	–	–	#3451	73

Table 1 continued

pGS18 ORF name[a]	Position in sequence		Protein length (aa)	Most relevant homolog	Organism (plasmid)	Number of identities/ number examined (%)	*E* value	GenBank accession number	pHTA426		pLW1071	
	s[b]	bp							ORF[a]	id %[c]	ORF[a]	id %[c]
#36	+	33,673–34,176	167	Hypothetical protein GTNG_3452	*Geobacillus thermodenitrificans* NG80-2 (pLW1071)	166/167 (99)	5e−95	YP_001127530	–	–	#3452	99
#37[d]	–	35,705–34,198							–	–	#3453	87
#38	–	36,390–35,722	222	Hypothetical protein GTNG_3454	*Geobacillus thermodenitrificans* NG80-2 (pLW1071)	165/225 (73)	4e−87	YP_001127532	–	–	#3454	73
#39	–	37,449–36,406	347	TraL	*Geobacillus thermodenitrificans* NG80-2 (pLW1071)	317/347 (91)	3e−177	YP_001127533	–	–	#3455	91
#40	–	37,708–37,442	88	No significant similarity					–	–	–	–
#41	–	38,277–37,729	182	Hypothetical protein GTNG_3456	*Geobacillus thermodenitrificans* NG80-2 (pLW1071)	157/182 (86)	1e−87	YP_001127534	–	–	#3456	86
#42	–	39,302–38,274	342	TraG	*Geobacillus thermodenitrificans* G80-2 (pLW1071)	316/344 (91)	0.0	YP_001127535	–	–	#3457	91
#43	–	41,152–39,299	617	TrsE	*Geobacillus thermodenitrificans* G80-2 (pLW1071)	606/617 (98)	0.0	YP_001127536	–	–	#3458	98
#44	–	42,215–41,169	348	TrsD	*Geobacillus thermodenitrificans* G80-2 (pLW1071)	338/348 (97)	0.0	YP_001127537	–	–	#3459	97
#45	–	42,452–42,228	74	No significant similarity								
#46[d]	–	43,200–42,575							–	–	#3460	90
#47	+	43,575–44,183	202	Hypothetical protein pBM19_p03	*Bacillus methanolicus* MGA3 (pBM19)	103/169 (60)	3e−53	NP_957650	–	–	#3462	37
#48	+	44,489–45,301	270	Hypothetical cytosolic protein	*Bacillus thuringiensis* ser. *israelensis* ATCC 35646	77/156 (49)	8e−28	ZP_00739834	–	–	–	–
#49	+	45,524–46,252	242	Hypothetical protein GTNG_3466	*Geobacillus thermodenitrificans* NG80-2 (pLW1071)	228/242 (94)	6e−129	YP_001127544	–	–	#3466	94
#50	+	46,287–46,766	159	Hypothetical protein GTNG_3467	*Geobacillus thermodenitrificans* NG80-2 (pLW1071)	156/158 (98)	2e−86	YP_001127545	#37	66	#3467	98

Table 1 continued

pGS18 ORF name[a]	Position in sequence		Protein length (aa)	Most relevant homolog	Organism (plasmid)	Number of identities/ number examined (%)	*E* value	GenBank accession number	pHTA426		pLW1071	
	s[b]	bp							ORF[a]	id %[c]	ORF[a]	id %[c]
#51	–	47,195–46,812	127	Hypothetical protein GTNG_3468	*Geobacillus thermodenitrificans* NG80-2 (pLW1071)	124/127 (97)	8e−66	YP_001127546	–	–	#3468	97
#52	–	48,362–47,211	383	Hypothetical protein GTNG_3469	*Geobacillus thermodenitrificans* NG80-2 (pLW1071)	380/383 (99)	0.0	YP_001127547	–	–	#3469	99
#53	–	49,468–48,821	215	Hypothetical protein GTNG_3475	*Geobacillus thermodenitrificans* NG80-2 (pLW1071)	212/215 (98)	7e−112	YP_001127553	–	–	#3475	98
#54	–	49,802–49,485	105	Hypothetical protein GTNG_3476	*Geobacillus thermodenitrificans* NG80-2 (pLW1071)	102/105 (97)	5e−51	YP_001127554	–	–	#3476	97
#55	+	50,193–52,124	643	Nickase TraA	*Geobacillus thermodenitrificans* G80-2 (pLW1071)	628/643 (97)	0.0	YP_001127555	–	–	#3477	97
#56	+	52,143–54,278	711	TrsK	*Geobacillus thermodenitrificans* NG80-2 (pLW1071)	708/711 (99)	0.0	YP_001127556	–	–	#3478	99
#57	+	54,300–56,420	706	TrsI	*Geobacillus thermodenitrificans* NG80-2 (pLW1071)	704/706 (99)	0.0	YP_001127557	–	–	#3479	99
#58	+	56,644–57,342	232	LtrC	*Geobacillus thermodenitrificans* NG80-2 (pLW1071)	208/224 (92)	3e−113	YP_001127558	–	–	#3480	92
#59	+	57,415–57,978	187	Hypothetical protein GK2912	*Geobacillus kaustophilus* HTA426	184/187 (98)	1e−109	YP_148765	–	–	–	–
#60	+	58,120–59,370	416	IS1604-like transposase	*Geobacillus kaustophilus* HTA426	401/416 (96)	0.0	YP_146633	–	–	–	–
#61	+	59,363–60,163	266	Hypothetical protein GK0881	*Geobacillus kaustophilus* HTA426	260/266 (97)	3e−146	YP_146734	–	–	–	–
#62	+	60,358–60,750	130	LtrC (N-terminal truncation)	*Geobacillus thermodenitrificans* NG80-2 (pLW1071)	123/126 (97)	1e−65	YP_001127558	–	–	#3480	97
#63	+	60,767–61,123	118	Hypothetical protein GTNG_3481	*Geobacillus thermodenitrificans* NG80-2 (pLW1071)	116/118 (98)	1e−62	YP_001127559	–	–	#3481	98

Table 1 continued

pGS18 ORF name[a]	Position in sequence		Protein length (aa)	Most relevant homolog	Organism (plasmid)	Number of identities/ number examined (%)	*E* value	GenBank accession number	pHTA426		pLW1071	
	S[b]	bp							ORF[a]	id %[c]	ORF[a]	id %[c]
#64	+	61,134–61,958		Hypothetical protein GTNG_3482	*Geobacillus thermodenitrificans* NG80-2 (pLW1071)	270/274 (98)	1e−154	YP_001127560	–	–	#3482	98
#65	+	62,414–62,830	138	Hypothetical protein GKP37	*Geobacillus kaustophilus* HTA426 (pHTA426)	137/138 (99)	1e−74	YP_145848	#37	99	#3467	66

[a] ORF numbers are abbreviated forms originating from their formal designations: pGS18_ORF##, GKP##, GTNG_####

[b] Strand

[c] Percentages of amino acid identities

[d] Pseudogene

to identify conserved domains in amino acid sequences (Marchler-Bauer and Bryant 2004). Prediction of transmembrane helices in putative proteins was performed by *TMHMM* software (Krogh et al. 2001) and helix-turn-helix motifs by *Helix-Turn-Helix Motif Prediction* program (Dodd and Egan 1990). IS-elements were annotated using the *IS Finder* database (http://www-is.biotoul.fr/is.html). Sequence alignments were performed using *MEGA 3.1* (Kumar et al. 2004), LFASTA (Duret et al. 1996; Pearson and Lipman 1988) and *AlignX* module of *Vector NTI Advance*$^{\text{TM}}$ *9.0* with standard default parameters.

Nucleotide sequence accession numbers

The complete nucleotide sequence of pGS18 has been deposited in the EMBL Nucleotide Sequence Database under accession number AM886060. The previously reported two restriction fragments of pGS18 (AM501412 and AM501413; Stuknyte et al. 2007) correspond to 42,569–35,986 bp and 53,377–48,329 bp coordinates of a complimentary strand in the complete plasmid sequence, respectively. The sequences of *G. kaustophilus* HTA426 plasmid pHTA426 and *G. thermodenitrificans* NG80-2 plasmid pLW1071 were obtained from GenBank under accession numbers NC_006509 and NC_009329, respectively.

Results and discussion

Overall features of pGS18

pGS18 is a circular plasmid of 62,830 bp with a G + C content of 40.02%. A total of 65 putative ORFs were identified. Thirty-eight ORFs were found to be transcribed from (+) strand and 27 ORFs in the opposite orientation (Fig. 1). The entire putative coding sequence might account for as much as 84.1% of the total pGS18 sequence. Of the 65 predicted ORFs, 25 (38.4%) were assigned to putative functions and four (6.2%) were annotated as pseudogenes. The amino acid sequences obtained from 29 ORFs (44.6%) had the highest similarity to hypothetical proteins of the other microorganisms (mainly *G. kaustophilus* HTA426 and *G. thermodenitrificans* NG80-2), and seven ORFs (10.8%) had no significant similarity to any genes present in the current open databases (Table 1).

Plasmid replication region and genes encoding three different plasmid maintenance systems were identified; in addition, a region of a possible transfer was determined. Several transfer genes that were initially described in our previous paper (Stuknyte et al. 2007) encoded proteins homologous to the *G. thermodenitrificans* NG80-2 pLW1071 conjugative system components. We also predicted several mobile genetic elements and genes, responsible for DNA repair, distributed along the whole sequence of pGS18.

Replication protein coding region

ORF3 starts with TTG at position 1,464 and extends 1,281 nucleotides to TGA stop codon at nucleotide 2,744. Putative RNA polymerase consensus regions at nucleotides 1,381–1,386 (−35) and 1,402–1,408 (−10) are located immediately upstream of the coding sequence (Fig. 2). However, ORF3 does not contain a typical ribosomal binding site. A structure reminiscent of a transcription terminator is located downstream of the coding sequence, with a hairpin forming region between nucleotides 2,821 and 2,835. Translation of this open reading frame should result in a polypeptide of 426 amino acids.

Inspection of the primary structure of ORF3-encoded product, deduced from DNA sequence, revealed a helix-turn-helix DNA binding motif (residues 401 to 422), which is frequently observed in DNA binding proteins (Dodd and Egan 1990).

A computer-based comparison of pGS18 ORF3 putatively encoded polypeptide product with primary protein sequences currently available in the GenBank database was performed. A high degree of homology (99% amino acid identities) was observed with the replication protein of the plasmid pHTA426 from the closely related bacterium *G. kaustophilus* (Takami et al. 2004). Rep protein of pHTA426 has additional 13 amino acids at its N-terminal part. These amino acids are absent in the putative Rep of pGS18. Furthermore, the helix-turn-helix motif is occupied by the same conserved region in both Rep proteins. Also, the putative Rep of pGS18 showed homology to a protein (55% identities, 72% positives over 432 amino acids) encoded on the plasmid of the alkaliphilic *Bacillus* strain KSM-KP43 and to replication proteins (39–43% identities, 57–64% positives) of several plasmids of *Clostridium* spp. (with the corresponding accession numbers in the GenBank: ABS42939, Brinkac et al. unpublished results; NP_783730 and NP_783816, Bruggemann et al. 2003; NP_040453, Garnier and Cole 1988; YP_699929 and YP_697960, Myers et al. 2006; YP_209688, Miyamoto et al. unpublished results).

Analysis of the sequence G + C deviation was used to indicate a potential origin of replication as described by Lobry (1996). However, no notable G + C skew was determined on pGS18 (Fig. 1).

Three adjacent repeated sequences of 9 bp (putative iterons) were present upstream of ORF3 (Fig. 2). A putative *dnaA* box (5'-TGTGAATaa-3') differed from that of two other theta replicating plasmids, pAMβ1 and pIP501, only in 2 bp (Bruand et al. 1993). Examination of the 87-bp DNA segment upstream of ORF3 (coordinates 1,043–1,270) that is relatively G + C-rich (74.70% rather than 40.02% for the whole plasmid) revealed an inverted repeat. It is located immediately downstream of AT-rich region, which has several incomplete direct repeats (5'-GAAGA-3') and could be proposed as a strand melting site in the process of plasmid replication initiation. According to its secondary structure and location, this inverted repeat strongly resembles the *ssiA* site of pAMβ1 from *Enterococcus faecalis* (Janniere et al. 1993). Its role is to presumably facilitate lagging-strand synthesis by stimulating formation of a primosome *priA*-dependent complex (del Solar et al. 1998; Janniere et al. 1993).

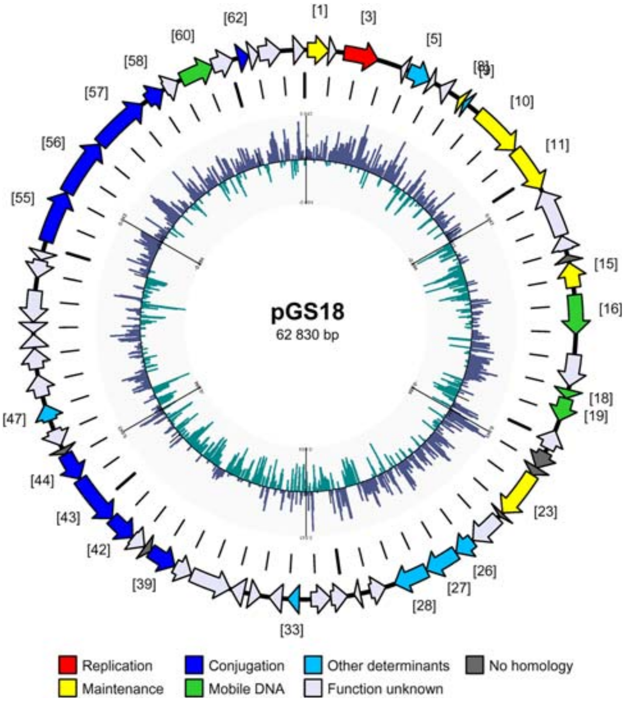

Fig. 1 Circular representation of pGS18. Open reading frames (ORFs) are represented by *block arrows* on the *outer circle*. Predicted functions/homologies are indicated by the color key featured below; several ORF numbers are indicated for reference (see Table 1 for details). The *inner circle* is a circular bar graph of the G + C skew of the plasmid sequence with a window size of 100 and step size of 50. The *second circle* from the center is a graduated size scale with *small tick marks* every 1 kb and *large tick marks* every 10 kb

Two GATC sequences, which are the sites of methylation by the host Dam methylase, upstream of the ORF3 were present. These sequences are usually found at *ori* region on plasmids of gram-negative bacteria (del Solar et al. 1998).

In summary, several features were identified on pGS18, which are typical for theta replicating plasmids: (1) repeated sequences (iterons); (2) *dnaA* box; (3) an AT-rich region containing several imperfect direct repeats; (4) a replication terminator (Meijer et al. 1995). Although these

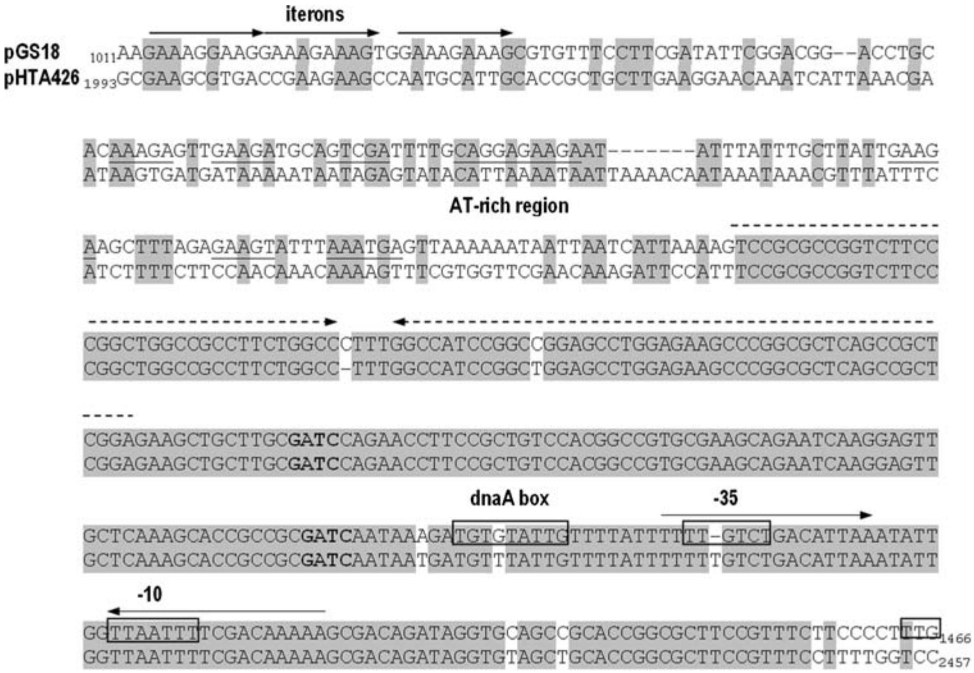
iterons
pGS18 1011AAGAAAGGAAGGAAAGAAAGTGGAAAGAAAGCGTGTTTCCTTCGATATTCGGACGG--ACCTGC
pHTA426 1993GCGAAGCGTGACCGAAGAAGCCAATGCATTGCACCGCTGCTTGAAGGAACAAATCATTAAACGA
ACAAAGAGTTGAAGATGCAGTCGATTTTGCAGGAGAAGAAT-------ATTTATTTGCTTATTGAAG
ATAAGTGATGATAAAAATAATAGAGTATACATTAAAATAATTAAAACAATAAATAAACGTTTATTTC
AT-rich region
AAGCTTTAGAGAAGTATTTAAATGAGTTAAAAAATAATTAATCATTAAAAGTCCGCGCCGGTCTTCC
ATCTTTTCTTCCAACAAACAAAAGTTTCGTGGTTCGAACAAAGATTCCATTTCCGCGCCGGTCTTCC
CGGCTGGCCGCCTTCTGGCCCTTTGGCCATCCGGCCGGAGCCTGGAGAAGCCCGGCGCTCAGCCGCT
CGGCTGGCCGCCTTCTGGCC-TTTGGCCATCCGGCTGGAGCCTGGAGAAGCCCGGCGCTCAGCCGCT
CGGAGAAGCTGCTTGCGATCCAGAACCTTCCGCTGTCCACGGCCGTGCGAAGCAGAATCAAGGAGTT
CGGAGAAGCTGCTTGCGATCCAGAACCTTCCGCTGTCCACGGCCGTGCGAAGCAGAATCAAGGAGTT
dnaA box
-35
GCTCAAAGCACCGCCGCGATCAATAAAGATGTGTATTGTTTTATTTTTT-GTCTGACATTAAATATT
GCTCAAAGCACCGCCGCGATCAATAATGATGTTTATTGTTTTATTTTTTTGTCTGACATTAAATATT
-10
GGTTAATTTTCGACAAAAAGCGACAGATAGGTGCAGCCGCACCGGCGCTTCCGTTTCTTCCCCTTTG1466
GGTTAATTTTCGACAAAAAGCGACAGATAGGTGTAGCTGCACCGGCGCTTCCGTTTCCTTTTGGTCC2457

Fig. 2 Proposed *ori* region of pGS18 and its alignment to the sequence of pHTA426 upstream of the *rep* gene. The *shaded letters* indicate identical nucleotides. Iterons (*solid arrows*); imperfect repeated sequences of AT-rich region (consensus 5'-GAAGA-3'; *underlined*); imperfect inverted repeat (*dashed arrows*) is a putative *ssiA* site of theta replicating plasmids; GATC sequences (in *bold*) are the putative sites of methylation by the host Dam methylase; an inverted repeats (*arrows*); *dnaA* box, pGS18 *rep* start codon and putative −10 and −35 promoter elements (*boxed*)

properties are not absolute proof, they strongly suggest that pGS18 uses the theta mechanism of replication.

Plasmid maintenance system

The stable maintenance of plasmids depends on a number of functions that prevent irreversible plasmid-loss during cell growth and division. Such systems include nonlethal and lethal stability determinants. The former are the active partitioning systems (*par*) and multimer resolution systems (*mrs*), and the later correspond to postsegregational killing systems (PSK) and plasmid-encoded restriction-modification systems (RM) (for a review, see Gerdes et al. 2000).

The predicted products of ORF8 and ORF15 are putative integrases/recombinases Xer with the highest similarity to the corresponding *G. kaustophilus* HTA426 plasmid pHTA426 encoded protein and chromosome encoded protein of *G. thermodenitrificans* NG80-2, respectively (Table 1; Feng et al. 2007; Takami et al. 2004). Both proteins contain a conserved INT_XerDC domain (cd00798). They also belong to the Int family of site-specific recombinases, which have complex recombination sites and require accessory factors. Site-specific recombinases are required for plasmid oligomer resolution in plasmid partitioning process (Gerdes et al. 2000).

Upstream of the ORF3, encoding a possible plasmid replication protein, ORF1 is located. Its predicted product contains a conserved ParA domain (cd02042). Upstream of the ORF1, a significant promoter sequence (*Promoter Prediction by Neural Network* score is 1.00) was found. ParA belongs to a conserved family of bacterial proteins, resolvases, implicated in segregation process. *Par* systems are usually found on low-copy-number plasmids that segregate nonrandomly. They actively distribute the plasmid copies to daughter cells during division and stabilize low copy number replicons to which they are attached, without affecting copy number and without killing plasmid free segregants (Gerdes et al. 2000). The prototype of *par* system was described on ColE1 plasmid form *E. coli* (Leung et al. 1985).

ORF10 and ORF11 encode type IIs modification methyltransferase and type IIs restriction endonuclease, respectively. Also, the C-terminal part of the deduced polypeptide product of ORF23 is similar to methyltransferase of *Alkaliphilus metalliredigens* QYMF (Table 1; Copeland et al. unpublished results). It contains a Methyltransf_11 conserved domain (pfam08241). Members of this family are *S*-adenosyl-L-methionine-dependent methyltransferases. A postsegregational host-killing model for plasmid stabilization that is associated with restriction-modification systems is known (Gerdes et al. 2000;

Kobayashi 2001; Naito et al. 1995). Further genetic analysis is required to probably associate ORF10- and ORF11-encoded products with this function.

Interestingly, pGS18 carries genes of more than one plasmid maintenance system. This may seem superfluous, since one stability system should be sufficient to ensure inheritance of a given plasmid. Although there is no direct evidence to support this, it could be plausibly hypothesized that these plasmids arose after recombination of multiple, initially separate plasmids, each with its own maintenance system (Holcik and Iyer 1997).

The putative transfer (*tra*) genes

pGS18 was found to carry a 6.93- and 8.565-kb-long regions that with respect to gene structure and organization were almost identical (nucleotide sequence identity of 92 and 98%, respectively) to the corresponding regions of pLW1071 that encompass genes associated with conjugative plasmid transfer (Fig. 3). All the ORFs identified on pGS18 as possible conjugative transfer genes were found to have highly similar homologs on pLW1071 (see percentages in Table 1). The partial sequence of pGS18 was previously described, showing the presence of these putative conjugative transfer genes on this plasmid (Stuknyte et al. 2007). Briefly, a nickase TraA, encoded by ORF55, mating pair formation (*mpf*) complex components (TrsG (ORF42), a putative lytic enzyme, TrsD (ORF44) and TrsE (ORF43), putative translocation energy suppliers) and a putative coupling protein TrsK (ORF66) were identified. As well a protein (encoded by ORF39), which is significantly similar to TraL of *S. aureus* subsp. *aureus* USA300, a plasmid pUSA03 horizontal transfer complex protein with an unidentified function, was present. A putative site for the initiation of conjugative plasmid transfer, *oriT*, was located.

After completion of pGS18 sequence, several other putative conjugative transfer elements were also identified. ORF57 was predicted to encode a topoisomerase, which is similar to both the TrsI of *G. thermodenitrificans* plasmid pLW1071 (Feng et al. 2007) as well as the protein encoded by ORF63 of plasmid pAW63 from *Bacillus thuringiensis* (Van der Auwera et al. 2005). The later protein is thought to be a relaxase, function of which is equivalent to that of the VirB/D4 type IV secretion system (T4SS) component VirD2, the key player of the *dtr* system (Grohmann et al. 2003; Van der Auwera et al. 2005).

ORF58 and ORF62 displayed significant homology to LtrC-like (from the *Lactococcus* transfer ORFC) putative conjugative elements from *G. thermodenitrificans* (92 and 97% identity, respectively); *Listeria* sp., *Staphylococcus* sp. and *Lactococcus* sp. (33–59% identity). The function of the LtrC, which was identified in the *tra* region of the *L. lactis* conjugative plasmid pRS01, has not been determined (Mills et al. 1994).

pGS18, as well as pLW1071, do not contain a continuous *tra* region that is always present on conjugative plasmids. Two regions encoding putative *tra* genes are separated with approximately 8 and 14 kb inserts, respectively (Fig. 3). In the case of pGS18, 10 hypothetical proteins (the products of ORF45–ORF54) are encoded in between the regions. FFAS03 program (http://ffas.ljcrf.edu/ffas-cgi/cgi/ffas.pl) (Jaroszewski et al. 2005) was used on these ORFs to improve the sensitivity of detection of distant homologies to T4SS components and to enhance

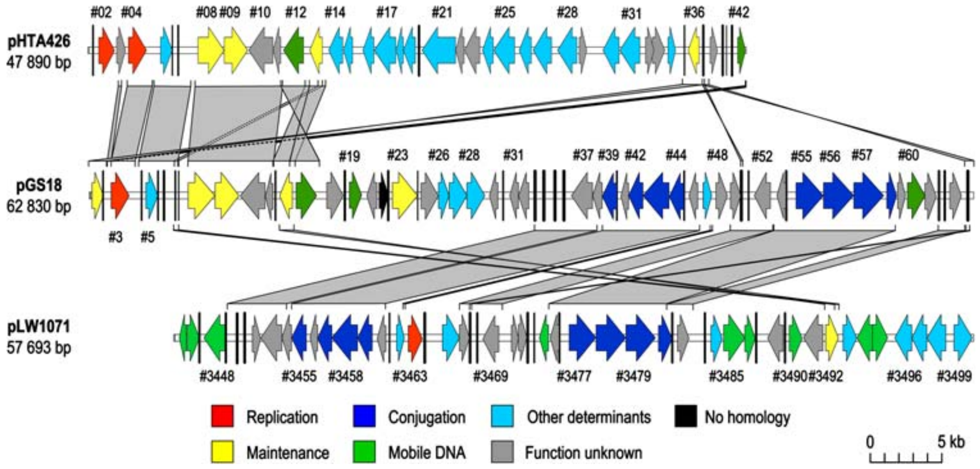

Fig. 3 Linear alignment of pGS18, pHTA426 and pLW1071. ORFs are represented by *block arrows* and *solid lines* (ORFs smaller than 500 nt). Several ORF numbers are indicated for reference on each plasmid just above or below their representation (see Table 1 for details). Predicted functions/homologies are indicated by the color key featured below. Well-conserved segments of the plasmids are paired by *shaded regions*; percentages of polypeptide sequence identities for specific ORF pairs can be found in Table 1. Scale is indicated by the bar in the lower right-hand corner

alignment accuracy (Grynberg et al. 2007). However, no similarities to putative *tra* genes were detected.

Mobile genetic elements

Four ORFs, ORF16, ORF18, ORF19 and ORF60, were identified as constituents of insertion sequences on the basis of their homologies. ORF16 was predicted to encode a transposase of IS*Bst12*-like element (Egelseer et al. 2000). The protein contains a conserved Transposase_25 domain of pfam03050 family. This family includes IS*66* from *Agrobacterium tumefaciens*. However, IS*Bst12* belongs to the group of unclassified insertion sequences (*IS finder* database, see "Materials and methods"). ORF16 is bounded by 11-bp inverted repeats (IRs) and flanked by a directly repeated 8-bp sequence. Thus, it constitutes a complete insertion sequence highly similar to IS*Bst12* (Egelseer et al. 2000).

ORF18 encodes a putative transposase of IS*3* family. It contains helix-turn-helix motif, which is typical for this family proteins. IS*3*-family members generally have two consecutive and partially overlapping reading frames, *orfA* and *orfB*, in relative translational reading phases 0 and −1, respectively. Taking into account this feature, ORF19 could correspond to *orfB*. However, ORF19 does not encode a functional protein due to the nonsense mutation and C-insertion that cause frame-shift mutation in the sequence. The presence of strongly significant rve conserved domain (pfam00665) and Tra5 domain (COG2801) shows that this ORF could have encoded a functional integrase or transposase. ORF19 contains a DDE catalytic domain, which is also a characteristic feature of IS*3* family transposases. Members of IS*3* family are distinguished by lengths between 1,200 and 1,550 bp and inverted terminal repeats in the range of 20–40 bp, which are variable but clearly related (*IS finder* database, see "Materials and methods"). The sum of the lengths of ORF18 and ORF19 is consistent with this observation. Furthermore, imperfect inverted terminal repeats of 21 bp were present. Thus, ORF18 and ORF19 with the flanking sequences could be assigned to the IS*3* family of insertion sequences.

ORF60 encodes a putative transposase with 96% amino acid identities to the IS*1604*-like transposase of *G. kaustophilus* HTA426. Additional search in *IS finder* database (see "Materials and methods") showed it to be highly similar to IS*Bst1* belonging to IS*481* family of insertion sequences. Protein possesses rve (pfam00665) and Mu-transpos_C (pfam09299) conserved domains that are found in various prokaryotic integrases and transposases. The length of the protein-encoded DNA sequence (1,251 bp) consists with that of transposases of IS*481* family (approximately 1 kb) (*IS finder* database, see "Materials and methods"). However, flanking sequences significantly differ. ORF60 is flanked by 7-bp inverted repeats. Nine-base pair direct repeats, surrounding ORF60 and adjacent to IRs, were also present.

DNA damage repair and stress response genes

Several putative proteins (encoded by ORF9, ORF28, ORF33, ORF47), associated with DNA damage repair or stress response, were present on pGS18.

ORF9-predicted product is similar to a DNA damage repair protein of *G. thermodenitrificans* NG80-2 plasmid pLW1071 (Table 1; Feng et al. 2007). It harbors an incomplete conserved Pol_IV_kappa (cd03586) and DinP (COG0389) domains. Pol_IV_kappa is a member of the Y-family of DNA polymerases. Expression of Y-family polymerases is often induced by DNA damage (Goodman 2002). These polymerases are phylogenetically unrelated to classical DNA polymerases. DinP is a nucleotidyltransferase/DNA polymerase involved in DNA repair. It is rarely found that plasmid encodes the DNA polymerase (Qin et al. 2007).

ORF28-deduced polypeptide product has similarities to UvrB/UvrC protein of *Chlorobium limicola* (Table 1; Copeland et al. unpublished results). The UvrABC repair system catalyzes the recognition and processing of DNA lesions (Truglio et al. 2006). ORF28-encoded product harbors a conserved ClpA domain (COG0542). ClpA is an ATP-binding subunit of Clp protease and DnaK/DnaJ chaperones that participate in posttranslational modification, protein turnover and act as chaperones. Clp_N domain (pfam02861) is located at the N-terminal part of ORF28-encoded product. The function of this domain is uncertain, but it may form a protein-binding site. Central part of ORF28 putatively encoded protein harbors AAA domain belonging to pfam0004 family. It is an ATPase family associated with various cellular activities. AAA family proteins often perform chaperone-like functions that assist in the assembly, operation, or disassembly of protein complexes.

The predicted protein product of ORF33 is highly similar to DNA repair protein RadC of *Listeria monocytogenes* (Table 1; Nelson et al. 2004). RadC is required for the repair of DNA strand breaks. RadC functions specifically in recombinational repair that is associated with the replication fork (Saveson and Lovett 1999).

ORF47-deduced product is similar to hypothetical protein p03 of *Bacillus methanolicus* plasmid pBM19 (Table 1; Brautaset et al. 2004). This protein at its N-terminal part contains a conserved helix-turn-helix motif ($_{6}$YSTKDIANIVGIATPTVRKYAQ$_{27}$) characteristic for several transcriptional regulators, YyaN (cd01109), MlrA (cd01104) and MERR (cd00592), which mediate responses to stress in eubacteria.

Other ORFs

The amino acid sequence obtained from the putative ORF5 exhibited 99% identities to the late competence protein of *G. kaustophilus* HTA426 plasmid pHTA426 (Table 1; Takami et al. 2004). It harbors a conserved ComEC domain (COG2333). ComEC has been predicted from its sequence to be a polytopic membrane protein and is required for DNA transport (Provvedi and Dubnau 1999). Members of ComEC family are integral membrane proteins with six trans-membrane helices (Provvedi and Dubnau 1999). Some members of this family (pfam03772) have been shown to be essential for bacterial competence in uptake of extracellular DNA. However, ORF5 predicted product has only one trans-membrane helix.

ORF20 encodes a putative protein of 240 amino acids that showed similarity to hypothetical protein OB0303 from *Oceanobacillus iheyensis* HTE831 (45% identities, 65% positives; Takami et al. 2002). It harbors an N-terminal part of a conserved nucleotidyltransferase domain (pfam01909). Members of this family belong to a large family of nucleotidyltransferases. This family includes kanamycin nucleotidyltransferase, which inactivates antibiotics by catalyzing the addition of a nucleotidyl group onto the drug. Resistance of *G. stearothermophilus* strain 18 to kanamycin was tested. However, this strain was sensitive to the drug (data not shown).

The amino acid sequence obtained from the putative ORF27 exhibited 27% identities and 50% positives to the predicted transporter protein of *Clostridium kluyveri* (Table 1; Seedorf et al. unpublished results). The deduced polypeptide product of ORF27 harbors a conserved DUF894 domain of pfam05977 family. This family consists of several bacterial proteins, many of which are annotated as putative trans-membrane transport proteins. A hydropathy plot of the ORF27 polypeptide sequence revealed 10 potential membrane-spanning sequences, suggesting that the putative ORF27-encoded protein is localized in the cytoplasmic membrane.

Of the 65 predicted ORFs of pGS18, four (6.2%) were annotated as pseudogenes, that is, defunct relatives of known genes that have lost their protein-coding ability resulting from frame-shifts due to nucleotide deletions (Vanin 1985). As it was mentioned previously in this article, ORF19 could have encoded a functional integrase or transposase of insertion sequence IS*3*-family members. ORF29, ORF37 and ORF46 could have encoded hypothetical proteins from *G. kaustophilus* and *G. thermodenitrificans*. Namely, ORF29 is homologous to hypothetical protein GK1107 of *G. kaustophilus* HTA426 encoding gene. Nucleotide sequence alignment of ORF29 and GK1107 showed that ORF29 has a G deletion (between 28,804 and 28,803 nucleotides in the whole pGS18 sequence) compared to GK1107. This determines a frame-shift mutation and probably nonfunctional ORF29. Analogously, ORF37 has a C deletion (between 35,590 and 35,589 nucleotides) compared to GTNG_3453 from *G. thermodenitrificans* NG80-2 plasmid pLW1071, and ORF46 has a T deletion (between 42,822 and 42,821 nucleotides) compared to GTNG_3460 from the same plasmid.

Thirty-six out of the 65 predicted ORFs of the pGS18 (55.4%) encoded hypothetical proteins, which either were similar to proteins with unknown function from other microorganisms (mainly *G. kaustophilus* HTA426 and *G. thermodenitrificans* NG80-2) or had no homology to any proteins present in the current open databases (Table 1). However, this is not a paradox, since for many large low-copy-number and self-transmissible plasmids a significant proportion of their genome is a DNA of unknown function (Thomas 2004).

pGS18 comparison with other sequenced large *Geobacillus* spp. plasmids

We compared the nucleotide sequence of pGS18 with two other large plasmids of *Geobacillus* spp. sequenced to date: pHTA426 of *G. kaustophilus* HTA426 and pLW1071 of *G. thermodenitrificans* NG80-2 (Feng et al. 2007; Takami et al. 2004). Twenty-four (36.91%) of the 65 predicted pGS18 ORFs had sequence similarities to pLW1071 genes; 11 (16.9%) pGS18 ORFs had sequence similarities to pHTA426 genes, and four ORFs (6.2%) were similar to ORFs of both plasmids (Table 1; Fig. 4).

The putative transfer (*tra*) region of pGS18 with the surrounding ORFs was found to be equivalent to that on pLW1071 of *G. thermodenitrificans* NG80-2. Notably, ORF33–ORF46 in this segment of pGS18 corresponded to the coding sequences GTNG_3449–GTNG_3460 on pLW1071 and ORF49–ORF58 to GTNG_3466–GTNG_3469 together with GTNG_3475–GTNG_3480 (Table 1, Fig. 3). As described earlier and by Stuknyte et al. (2007), nine of these ORFs in pGS18 encoded putative conjugation-related proteins. It is interesting to note that neither pGS18 nor pLW1071 contains a continuous *tra* region. Moreover, the intervening sequences in both plasmids are of the different length (7,979 bp in pGS18 and 14,229 bp in pLW1071) and have only partial similarities (Fig. 3).

Sequence analysis revealed similar loci in both *G. stearothermophilus* pGS18 and *G. kaustophilus* pHTA426.

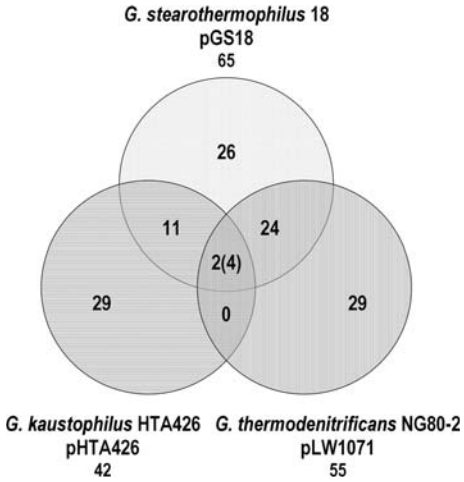

Fig. 4 Venn diagram illustrating the number of putative proteins associated with each organism and the number shared with the intersecting organism. *Dots G. stearothermophilus* 18 pGS18; *horizontal lines G. kaustophilus* HTA426 pHTA426; *vertical lines G. thermodenitrificans* NG80-2 pLW1071. Two orthologs are shared by all three plasmids, and pGS18 has two copies of each of them

They were located in the region encoding plasmid replication and stable maintenance functions (Table 1, Fig. 3). Comparison of the proposed *ori* region of pGS18 to the sequence of pHTA426 upstream of the *rep* gene is shown in Fig. 2. It is obvious that even if replication proteins of both plasmids are almost identical (99% of amino acid identities), the upstream regions differ. The functional analysis has to be performed to elucidate the replication mechanism of both plasmids.

We found two orthologs shared by the three plasmids (Table 1, Fig. 3). The first one is a CDS (ORF50 and ORF65 in pGS18; GKP37 in pHTA426; GTNG_3467 in pLW1071) encoding a hypothetical protein. ORF65- and GKP37-encoded proteins contain a conserved ribbon-helix-helix motif of CopG family (pfam01402). CopG is a transcriptional repressor, a homodimeric protein, that constitutes the smallest natural transcriptional repressor characterized so far and is involved in plasmid replication control (del Solar et al. 2002). The second ortholog shared by all three *Geobacillus* spp. plasmids is a CDS (ORF15 in pGS18; GKP13 in pHTA426; GTNG_3492 in pLW1071) encoding an integrase/recombinase Xer. ORF8 is a putative N-terminal truncation (69 C-terminal amino acids) of this CDS.

The similarity of modular composition of the three *Geobacillus* spp. plasmids provides a framework that can be exploited to formulate hypotheses concerning the molecular evolution of these plasmids. It could be proposed that several recombination events that involved two or more plasmids formed the pGS18. High degree of similarity between certain segments of pGS18 and pLW1071 reveals the similar nature of plasmids present in the two oil-field-residing bacteria, *G. stearothermophilus* 18 and *G. thermodenitrificans* NG80-2 (Feng et al. 2007).

Conclusions

The nucleotide sequence of the 62,830-bp *G. stearothermophilus* plasmid pGS18 reported in this study represents the largest *Geobacillus* spp. plasmid sequence determined so far and the first analyzed putative theta-type replicon from the genus *Geobacillus*. G + C content of pGS18 is 40.02%. It is significantly lower than the G + C content of *G. stearothermophilus* genome (52.2 mol%) and even of the all members of the genus *Geobacillus* (49–58 mol%) (Nazina et al. 2001). A lower G + C composition suggests the possibility of horizontal transfer of DNA from an organism with a lower G + C content.

The pGS18 shows a unique degree of genome coordination encompassing replication, transfer and stable inheritance. However, the presence of those modules on the plasmid genome is not a paradox. Successful plasmids have maximized their segregational stability and minimized the burden they place on their host. They also have acquired an efficient transfer system or occupy the niche where they can exploit an available transfer strategy (Thomas 2004). Most of the evidence points to plasmids being unable to maintain themselves simply by their replication, maintenance and transfer activities. Besides the machinery for their own maintenance and transfer, most plasmids carry genes that confer potentially useful traits on their bacterial hosts. In most cases, these traits are ones that are useful only intermittently or in certain environments (Lilley et al. 2000). The reason why large low-copy-number plasmid survives may be very different from the reasons for the persistence of a small, high-copy-number plasmid. DNA sequencing, particularly with respect to low-copy-number and self-transmissible plasmids, shows that for many a significant proportion of their genome is additional DNA, often of unknown function (Thomas 2004). As it was mentioned earlier, high number (i.e. 36 out of the 65 ORFs) of unknown genes is also noticeable on the pGS18. It is interesting that a large portion of them (13 ORFs) is highly similar to unknown genes from the other oil-field-residing bacterial plasmid, pLW1071 from *G. thermodenitrificans* NG80-2 (Feng et al. 2007). This points to the possible influence of the specific environmental conditions to the possible unknown function conferred by the plasmid.

In conclusion, the pGS18 has been sequenced, showing substantial similarity of modular composition to other big *Geobacillus* spp. plasmids, and it requires further genetic characterization to understand the benefits it provides to the host. Nevertheless, pGS18 could serve as a basis for construction of vectors, based on theta-type replicon.

Acknowledgments We are grateful to Dr. Geraldine Van der Auwera from the University of Louvain (Belgium) for providing scripts of BioPython that were applied to draw Fig. 1. We would like to thank Dr. Juozas Raugalas (Vilnius University, Lithuania) and Dr. Nano So (University of Milan, Italy) for a fruitful collaboration. We also acknowledge Egle Rudokiene in the Sequencing Centre from the Institute of Biotechnology (Vilnius, Lithuania) for automated sequencing. A special thanks to Dr. Rasa Rakauskaite from the University of Maryland (USA) for the critical review of the manuscript. This research was in part supported by grant of the Lithuanian State Science and Studies Foundation (project number T-05112) and by grant 2006-04-19 R-100 from the Science Foundation of Vilnius University.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Bingham AH, Bruton CJ, Atkinson T (1979) Isolation and partial characterization of four plasmids from antibiotic-resistant thermophilic bacilli. J Gen Microbiol 114:401–408

Brautaset T, Jakobsen MO, Flickinger MC, Valla S, Ellingsen TE (2004) Plasmid-dependent methylotrophy in thermotolerant *Bacillus methanolicus*. J Bacteriol 186:1229–1238

Bruand C, Le Chatelier E, Ehrlich SD, Janniere L (1993) A fourth class of theta-replicating plasmids: the pAMβ1 family from gram-positive bacteria. Proc Natl Acad Sci USA 90:11668–11672

Bruggemann H, Baumer S, Fricke WF, Wiezer A, Liesegang H, Decker I, Herzberg C, Martinez-Arias R, Merkl R, Henne A, Gottschalk G (2003) The genome sequence of *Clostridium tetani*, the causative agent of tetanus disease. Proc Natl Acad Sci USA 100:1316–1321

De Rossi E, Brigidi P, Welker NE, Riccardi G, Matteuzzi D (1994) New shuttle vector for cloning in *Bacillus stearothermophilus*. Res Microbiol 145:579–583

del Solar G, Giraldo R, Ruiz-Echevarria MJ, Espinosa M, Diaz-Orejas R (1998) Replication and control of circular bacterial plasmids. Microbiol Mol Biol Rev 62:434–464

del Solar G, Hernandez-Arriaga AM, Gomis-Ruth FX, Coll M, Espinosa M (2002) A genetically economical family of plasmid-encoded transcriptional repressors involved in control of plasmid copy number. J Bacteriol 184:4943–4951

Dodd IB, Egan JB (1990) Improved detection of helix-turn-helix DNA-binding motifs in protein sequences. Nucleic Acids Res 18:5019–5026

Duret L, Gasteiger E, Perriere G (1996) LALNVIEW: a graphical viewer for pairwise sequence alignments. Comput Appl Biosci 12:507–510

Egelseer EM, Idris R, Jarosch M, Danhorn T, Sleytr UB, Sara M (2000) IS*Bst*12, a novel type of insertion-sequence element causing loss of S-layer-gene expression in *Bacillus stearothermophilus* ATCC 12980. Microbiology 146:2175–2183

Feng L, Wang W, Cheng J, Ren Y, Zhao G, Gao C, Tang Y, Liu X, Han W, Peng X, Liu R, Wang L (2007) Genome and proteome of long-chain alkane degrading *Geobacillus thermodenitrificans* NG80-2 isolated from a deep-subsurface oil reservoir. Proc Natl Acad Sci USA 104:5602–5607

Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A (2006) Pfam: clans, web tools and services. Nucleic Acids Res 34:D247–D251

Garnier T, Cole ST (1988) Complete nucleotide sequence and genetic organization of the bacteriocinogenic plasmid, pIP404, from *Clostridium perfringens*. Plasmid 19:134–150

Gerdes K, Ayora S, Canosa I, Ceglowski P, Diaz-Orejas R, Franch T, Gultyaev AP, Bugge Jensen R, Kobayashi I, Macpherson C, Summers D, Thomas CM, Zielenkiewicz U (2000) Plasmid maintenance systems. In: Thomas CM (ed) The horizontal gene pool. Bacterial Plasmids and Gene Spread. Harwood Academic Publishers, Amsterdam, pp 49–86

Goodman MF (2002) Error-prone repair DNA polymerases in prokaryotes and eukaryotes. Annu Rev Biochem 71:17–50

Grynberg M, Li Z, Szczurek E, Godzik A (2007) Putative type IV secretion genes in *Bacillus anthracis*. Trends Microbiol 15:191–195

Grohmann E, Muth G, Espinosa M (2003) Conjugative plasmid transfer in gram-positive bacteria. Microbiol Mol Biol Rev 67:277–301

Holcik M, Iyer VN (1997) Conditionally lethal genes associated with bacterial plasmids. Microbiology 143:3403–3416

Hoshino T, Ikeda T, Narushima H, Tomizuka N (1985) Isolation and characterization of antibiotic-resistance plasmids in thermophilic bacilli. Can J Microbiol 31:339–345

Imanaka T, Ano T, Fujii M, Aiba S (1984) Two replication determinants of an antibiotic-resistance plasmid, pTB19, from a thermophilic bacillus. J Gen Microbiol 130:1399–1408

Imanaka T, Fujii M, Aiba S (1981) Isolation and characterization of antibiotic resistance plasmids from thermophilic bacilli and construction of deletion plasmids. J Bacteriol 146:1091–1097

Imanaka T, Fujii M, Aramori I, Aiba S (1982) Transformation of *Bacillus stearothermophilus* with plasmid DNA and characterization of shuttle vector plasmids between *Bacillus stearothermophilus* and *Bacillus subtilis*. J Bacteriol 149:824–830

Janniere L, Gruss A, Ehrlich SD (1993) Plasmids. In: Sonenshein AL, Hoch JA, Losick R (eds) *Bacillus subtilis* and other gram-positive bacteria: biochemistry, physiology, and molecular genetics. ASM Press, Washington, pp 625–644

Jaroszewski L, Rychlewski L, Li Z, Li W, Godzik A (2005) FFAS03: a server for profile–profile sequence alignments. Nucleic Acids Res 33:284–288

Khalil AB, Anfoka GH, Bdour S (2003) Isolation of plasmids present in thermophilic strains from hot springs in Jordan. World J Microbiol Biotechnol 19:239–241

Kobayashi I (2001) Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. Nucleic Acids Res 29:3742–3756

Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580

Kumar S, Tamura K, Nei M (2004) MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform 5:150–163

Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P (2006) SMART 5: domains in the context of genomes and networks. Nucleic Acids Res 34:D257–D260

Leung DW, Chen E, Cachianes G, Goeddel DV (1985) Nucleotide sequence of the partition function of *Escherichia coli* plasmid ColE1. DNA 4:351–355

Liao H, McKenzie T, Hageman R (1986) Isolation of a thermostable enzyme variant by cloning and selection in a thermophile. Proc Natl Acad Sci USA 83:576–580

Lilley A, Young P, Bailey M (2000) Bacterial population genetics: do plasmids maintain bacterial diversity and adaptation? In: Thomas CM (ed) The horizontal gene pool bacterial plasmids and gene spread. Harwood Academic Publishers, Amsterdam, pp 287–300

Lobry JR (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol 13:660–665

Lukashin AV, Borodovsky M (1998) GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res 26:1107–1115

Marchler-Bauer A, Bryant SH (2004) CD-search: protein domain annotations on the fly. Nucleic Acids Res 32:327–331

McMullan G, Christie JM, Rahman TJ, Banat IM, Ternan NG, Marchant R (2004) Habitat, applications and genomics of the aerobic, thermophilic genus *Geobacillus*. Biochem Soc Trans 32:214–217

Meijer WJ, de Boer AJ, van Tongeren S, Venema G, Bron S (1995) Characterization of the replication region of the *Bacillus subtilis* plasmid pLS20: a novel type of replicon. Nucleic Acids Res 23:3214–3223

Mielenz JR (1983) *Bacillus stearothermophilus* contains a plasmid-borne gene for alpha-amylase. Proc Natl Acad Sci USA 80:5975–5979

Mills DA, Choi CK, Dunny GM, McKay LL (1994) Genetic analysis of regions of the *Lactococcus lactis* subsp. *lactis* plasmid pRS01 involved in conjugative transfer. Appl Environ Microbiol 60:4413–4420

Myers GS, Rasko DA, Cheung JK, Ravel J, Seshadri R, DeBoy RT, Ren Q, Varga J, Awad MM, Brinkac LM, Daugherty SC, Haft DH, Dodson RJ, Madupu R, Nelson WC, Rosovitz MJ, Sullivan SA, Khouri H, Dimitrov GI, Watkins KL, Mulligan S, Benton J, Radune D, Fisher DJ, Atkins HS, Hiscox T, Jost BH, Billington SJ, Songer JG, McClane BA,Titball RW, Rood JI, Melville SB, Paulsen IT (2006) Skewed genomic variability in strains of the toxigenic bacterial pathogen, *Clostridium perfringens*. Genome Res 16:1031–1040

Naito T, Kusano K, Kobayashi I (1995) Selfish behavior of restriction-modification systems. Science 267:897–899

Nakayama N, Narumi I, Nakamoto S, Kihara H (1993) Complete nucleotide-sequence of pSTK1, a cryptic plasmid from *Bacillus stearothermophilus* TK015. Biotechnol Lett 15:1013–1016

Narumi I, Nakayama N, Nakamoto S, Kimura T, Yanagisawa T, Kihara H (1993) Construction of a new shuttle vector pSTE33 and its stabilities in *Bacillus stearothermophilus*, *Bacillus subtilis*, and *Escherichia coli*. Biotechnol Lett 15:815–820

Nazina TN, Tourova TP, Poltaraus AB, Novikova EV, Grigoryan AA, Ivanova AE, Lysenko AM, Petrunyaka VV, Osipov GA, Belyaev SS, Ivanov MV (2001) Taxonomic study of aerobic thermophilic bacilli: descriptions of *Geobacillus subterraneus* gen. nov., sp. nov. and *Geobacillus uzenensis* sp. nov. from petroleum reservoirs and transfer of *Bacillus stearothermophilus*, *Bacillus thermocatenulatus*, *Bacillus thermoleovorans*, *Bacillus kaustophilus*, *Bacillus thermoglucosidasius* and *Bacillus thermodenitrificans* to *Geobacillus* as the new combinations *G. stearothermophilus*, *G. thermocatenulatus*, *G. thermoleovorans*, *G. kaustophilus*, *G. thermoglucosidasius* and *G. thermodenitrificans*. Int J Syst Evol Microbiol 51:433–446

Nelson KE, Fouts DE, Mongodin EF, Ravel J, DeBoy RT, Kolonay JF, Rasko DA, Angiuoli SV, Gill SR, Paulsen IT, Peterson J, White O, Nelson WC, Nierman W, Beanan MJ, Brinkac LM, Daugherty SC, Dodson RJ, Durkin AS, Madupu R, Haft DH, Selengut J, Van Aken S, Khouri H, Fedorova N, Forberger H, Tran B, Kathariou S, Wonderling LD, Uhlich GA, Bayles DO, Luchansky JB, Fraser CM (2004) Whole genome comparisons of serotype 4b and 1/2a strains of the food-borne pathogen *Listeria monocytogenes* reveal new insights into the core genome components of this species. Nucleic Acids Res 32:2386–2395

Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85:2444–2448

Provvedi R, Dubnau D (1999) ComEA is a DNA receptor for transformation of competent *Bacillus subtilis*. Mol Microbiol 31:271–280

Qin T, Hirakawa H, Iida K, Oshima K, Hattori M, Tashiro K, Kuhara S, Yoshida S (2007) Complete nucleotide sequence of pLD-TEX-KL, a 66-kb plasmid of *Legionella dumoffii* TEX-KL strain. Plasmid 58:261–268

Sambrook J, Russell DW (2001) Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467

Saveson CJ, Lovett ST (1999) Tandem repeat recombination induced by replication fork defects in *Escherichia coli* requires a novel factor, RadC. Genetics 152:5–13

Stahl SR (1991) Plasmids in *Bacillus stearothermophilus* coding for bacteriocinogeny and temperature resistance. Plasmid 26:94–107

Stuknyte M, Guglielmetti S, Ricci G, Kuisiene N, Mora D, Parini C, Citavicius D (2007) Identification and *in silico* characterization of putative conjugative transfer genes on *Geobacillus stearothermophilus* plasmids. Ann Microbiol 57:407–414

Takami H, Takaki Y, Uchiyama I (2002) Genome sequence of *Oceanobacillus iheyensis* isolated from the Iheya Ridge and its unexpected adaptive capabilities to extreme environments. Nucleic Acids Res 30:3927–3935

Takami H, Takaki Y, Chee GJ, Nishi S, Shimamura S, Suzuki H, Matsui S, Uchiyama I (2004) Thermoadaptation trait revealed by the genome sequence of thermophilic *Geobacillus kaustophilus*. Nucleic Acids Res 32:6292–6303

Thomas CM (2004) Evolution and population genetics of bacterial plasmids. In: Funnel BE, Phillips GJ (eds) Plasmid biology. ASM Press, Washington, pp 509–528

Truglio JJ, Croteau DL, Van Houten B, Kisker C (2006) Prokaryotic nucleotide excision repair: the UvrABC system. Chem Rev 106:233–252

Van der Auwera GA, Andrup L, Mahillon J (2005) Conjugative plasmid pAW63 brings new insights into the genesis of the *Bacillus anthracis* virulence plasmid pXO2 and of the *Bacillus thuringiensis* plasmid pBT9727. BMC Genomics 6:103

Vanin EF (1985) Processed pseudogenes: characteristics and evolution. Annu Rev Genet 19:253–272