Published online 27 October 2021

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 4 **1**
https://doi.org/10.1093/nargab/lqab096

pLS20 is the archetype of a new family of conjugative plasmids harboured by *Bacillus* species

Jorge Val-Calvo [1], Andrés Miguel-Arribas [1], David Abia [2], Ling Juan Wu [3] and Wilfried J. J. Meijer [1,*]

[1]Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM), C. Nicolás Cabrera 1, Universidad Autónoma de Madrid, Canto Blanco, 28049, Madrid, Spain, [2]Bioinformatics Facility, Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM), C. Nicolás Cabrera 1, Universidad Autónoma de Madrid, Canto Blanco, 28049, Madrid, Spain and [3]Centre for Bacterial Cell Biology, Biosciences Institute, Newcastle University, Newcastle upon Tyne, NE2 4AX, UK

Received March 23, 2021; Revised September 03, 2021; Editorial Decision September 26, 2021; Accepted October 04, 2021

ABSTRACT

Conjugation plays important roles in genome plasticity, adaptation and evolution but is also the major horizontal gene-transfer route responsible for spreading toxin, virulence and antibiotic resistance genes. A better understanding of the conjugation process is required for developing drugs and strategies to impede the conjugation-mediated spread of these genes. So far, only a limited number of conjugative elements have been studied. For most of them, it is not known whether they represent a group of conjugative elements, nor about their distribution patterns. Here we show that pLS20 from the Gram-positive bacterium *Bacillus subtilis* is the prototype conjugative plasmid of a family of at least 35 members that can be divided into four clades, and which are harboured by different *Bacillus* species found in different global locations and environmental niches. Analyses of their phylogenetic relationship and their conjugation operons have expanded our understanding of a family of conjugative plasmids of Gram-positive origin.

INTRODUCTION

Many bacteria possess autonomously replicating extrachromosomal DNA molecules named plasmids, which are composed of essential and—most often—one or more modules that are non-essential to the plasmid (1–3). In this work, we define a module as a DNA region containing all the genes and/or genetic information for performing a function or cellular process. One of the modules that is essential for all plasmids concerns a replication module responsible for regulating (initiation of) DNA replication and thereby controlling the plasmid copy number. Most plasmids replicate via the theta or the rolling-circle mechanism. Large low-copy number theta-replicating plasmids contain a segregation module whose encoded proteins ensure faithful partitioning of the plasmid during cell division. A large subset of plasmids is mobile due to the presence of a non-essential module allowing the plasmid to transfer itself horizontally to other cells via conjugation or mobilization (see below). Genes encoding virulence determinants, toxins or antibiotic resistance are often present on conjugative or mobilizable plasmids, and so can efficiently spread, causing major clinical problems and enormous economic consequences worldwide (4,5). A better understanding of the conjugation and mobilization processes is warranted to help develop drugs and strategies impeding the spread of pernicious genes by these transfer routes.

Conjugation modules comprise all the genes required for performing the following steps in the conjugation process. First, selection of and attachment to the recipient cell. Two, generation of a sophisticated translocation channel, a type IV secretion system (T4SS), connecting the donor and the recipient cells. Three, DNA processing generating the DNA form that is transferred, which in most cases is a single DNA-strand (ssDNA), and subsequent transport of the ssDNA strand through the T4SS. Finally, establishment of the transferred DNA in the recipient cell by converting the ssDNA into a double-stranded DNA molecule. Mobilizable plasmids only encode the proteins involved in the DNA processing reaction and exploit the proteins generated by a conjugative plasmid that are necessary for transfer. Because the conjugation process involves many genes, conjugative plasmids are large and use the theta type of replication during growth. Mobilizable plasmids are much smaller and can use the theta or rolling circle mechanism of replication. Conjugative elements can also be embedded in the bacterial chromosome, which are then named integrated and conjugative elements (ICE). Excellent reviews on the process of conjugation have been published elsewhere (6–13).

*To whom correspondence should be addressed. Tel: +34 91 196 4539; Fax: +34 91 196 4420; Email: wmeijer@cbm.csic.es
Present address: Jorge Val-Calvo, Infection Medicine, University of Edinburgh, Edinburgh, Scotland, UK.

Conjugation occurs in both Gram-positive (G+) and Gram-negative (G-) bacteria. However, conjugative elements from G+ bacteria are less well studied than those from G− and the available information is based on a very limited number of conjugative plasmids that share little similarity between them. Although these studies have provided important insights, results on conjugation systems will have added relevance and impacts if they concern a prototype plasmid representing a family of related plasmids. For instance, comprehensive analysis of multiple related sequences may discriminate essential from non-essential genes and identify conserved regions or residues within a set of orthologous proteins that may be important for function.

The conjugative plasmid pLS20 (≈65 kb) was originally identified from the *Bacillus subtilis* natto strain IFO3335 (14). We chose to study pLS20 because it was the only known conjugative *B. subtilis* plasmid at the time we started, and its replication and partition module had already been identified and characterized (15–17). Contrary to most conjugative systems, pLS20 conjugates efficiently in both solid and liquid media (18–21). Since conjugation in liquid medium is rapid and gives very reproducible results, pLS20 is an attractive system to study conjugation. In the recent years, considerable progress has been made in understanding some aspects of the pLS20 conjugation process, particularly about mating pair formation, regulation of the conjugation genes and the DNA processing reaction (21–29).

Previously, we have shown that the 43 kb plasmid p576 from the *Bacillus pumilus* strain NRS576 isolated from the USA is related to pLS20 that was isolated from a *B. subtilis* natto strain in Japan (30–33). This suggested that pLS20 might be the prototype of a family of related plasmids. Here, we show that this is indeed the case. We have identified 35 plasmids in the public databases that are related to pLS20. The majority of these were not annotated as a plasmid but we reconstructed them from multiple contigs that are present in different next-generation sequencing (NGS) projects. All these 35 plasmids are harboured by different *Bacillus* species, and they could be classified into four clades. The identified plasmids have a conserved organization sharing similar origin, partitioning and conjugation modules, as well as several genes of unknown function. In addition, the majority of these pLS20 family plasmids also contain a DNA repair module. The (dis)similarities between the plasmids are analysed and discussed, providing important information for our understanding of these important and wide spread plasmids, and for future research in conjugation-related studies in general, particularly for those in G+ bacteria.

MATERIALS AND METHODS

Bacterial strains, plasmids and oligonucleotides

Bacterial strains were grown in Lysogeny Broth (LB) medium (34) or plated on LB plates containing 1.5% agar. When appropriate, the following antibiotics were added: ampicillin (100 μg/ml) for *Escherichia coli*. The strains, plasmids and oligonucleotides used are listed in Supplementary Tables S1, S2 and S3, respectively. All oligonucleotides were purchased from Isogen Life Science, The Netherlands.

DNA isolation and large plasmid screening

Bacillus strains were grown overnight at 37°C with shaking (180 RPM) in fresh LB medium in a total volume of 12 ml. The next day, 10 ml of the culture was pelleted in 5 Eppendorf tubes of 2 ml and processed as follows. Bacterial pellets were resuspended in 1 ml of Birnboim solution A (20% sucrose, 10 mM Tris-HCl pH 8.1, 10 mM EDTA and 50 mM NaCl) to which lysozyme was added freshly (2 mg/ml). After freezing the samples at -80°C and subsequent thawing, the tubes were incubated at 37°C with shaking during 15 min. The cells were lysed by adding 33 μl of 1% sarkosyl solution and proteinase K (100 μg/ml) to 1 ml of cell suspension and incubating the mixture at 60°C without shaking during 20 min. Next, the mixture was subjected to a phenol-chloroform extraction step and subsequently precipitated overnight in the presence of 96% ethanol. After washing with 70% ethanol solution, resulting DNA pellets were resuspended in sterile MiliQ water and treated with RNAse A. Finally, DNA samples (1–2 μl) were loaded on a 0.7% agarose gel (1× TBE buffer) without EtBr and run at 30 V for 16 h. Next, gels were stained with a 1xTBE solution containing EtBr for 2 h, followed by a destaining step and finally photographed under ultraviolet light.

Sequencing

Circularization of pBatNRS213. Plasmid pBatNRS213 was circularized as follows. A PCR was performed using total DNA extract of *B. atrophaeus* strain NRS213 as template DNA in combination with the outward oriented oligos oJV216 and oJV217 that hybridize near the ends of the linear contig corresponding to pBatNRS213 plasmid. The PCR product was digested with SalI and ligated into pBluescript II KS (+) vector linearized with the same restriction enzyme. The ligation mixture was used to transform competent *E. coli* XL-blue strain, and transformants were selected on LB agar plates supplemented with ampicillin, IPTG (1 mM) and X-gal (60 μg/ml). Some at randomly chosen colonies that had a white colour were analysed by colony PCR for the presence of an insert using oligos M13Fw-21_ext and M13-R. The positively identified derivative was named pJV40. The insert was sequenced by the Sanger method using oligos M13Fw-21_ext and M13-R.

Next-generation sequencing (NGS) of strain B. amyloliquefaciens *B1895.* A total DNA sample of *B. amyloliquefaciens* B1895 strain was used for sequence determination using the Illumina MiSeq platform. Preparation of samples and sequencing was done in collaboration with Dr Ricardo Ramos Parque Científico de Madrid (Madrid, Spain). DNA libraries were prepared with the NEBNext DNA library prep kit for Illumina (New England Biolabs). Briefly, 1 μg DNA was sonicated, and fragments of □675 or 1075 bp were selected. DNA ends were repaired, A-tailed and then ligated to adapters. Next, fragments were PCR amplified (8 cycles) and quality checked on a DNA 7500 chip on a 2100 Bioanalyzer (Agilent). Library sizes of 800 and 1200 bp were isolated, validated (Bioanalyzer) and titrated with quantitative PCR (qPCR). After denaturation, the libraries were seeded on a flow cell (MiSeq v2, 2×150

bp) at a density of 16 pM. Importantly, sequencing conditions were applied that resulted in in deep sequencing (~4 Gb of sequence data). Adapters and low-quality sequences were removed (35,36). After verifying the quality of the processed data (FastQC), *de novo* assembly was performed using software SPAdes (mode careful active; cov-cutoff auto; *k*-mer sizes 21,33,55,77,99,127; other parameters set as the default parameter) (37,38). PlasmidSPAdes (same parameters as used for the previous assembly) was also used to carry out another complementary assembly to identify the extrachromosomal elements (39). The resulting assemblies were evaluated using Quast to generate summary statistics (no of contigs, total length, GC%, N50, L50) (40). The plasmid was circularized using a dedicated perl script that searches for reads within the original sequencing data that span both ends of a contig (41). For a contig to be considered circular there had to be at least three reads spanning a minimum of 60 bp of each contig end. Final SPAdes assembly was automatically annotated using the PROKKA tool 1.14.5 (42). The annotation of plasmid pBamB1895 was checked manually and names were adapted to the nomenclature used for plasmid pLS20.

Bioinformatic methods

Identification of pLS20-like plasmids in the NCBI database reference genomes. The origin of replication region of pLS20 containing all the information for autonomous DNA replication (i.e. the replication module), spanning bp 64 501 to 65 745 of pLS20cat, accession number AB615352 was used as a query to carried out a BlastN search against the NCBI RefSeq genome database (BlastN algorithm, limited to bacteria (taxid 2)). All sequences with an *e*-value $< 2e\text{-}133$ and a query cover $>50\%$ were retrieved for subsequent analysis. The hits were organized depending on whether they concerned entire plasmid sequences or contigs. The complete sequences of plasmids were aligned using the algorithm directly EMBOSS Stretcher (43) or Nucmer (44). Before aligning, the plasmids were rotated and oriented such that plasmid ended with the first base pair of the *alp7A* gene. The Nucmer results presented as dot plot graphics in Supplementary Figure S1 were generated using mummerplot software. In the case of contigs, all the sequences belonging to the corresponding NGS project were downloaded to identify plasmid-related contigs. For this, a TblastX search was performed against all the sequences of the NGS project using as a query the sequences of the complete plasmids pLS20, pDSYZ, pBglSRCM103574 and p576. Contigs with an *e*-value 0.0 were identified as part of the plasmid. Subsequently, the identified contigs were aligned against the closest related query plasmid using the program Mauve (45). Besides confirming that the contig corresponded to a plasmid sequence, this allowed ordering and orienting the contigs, resulting in partially reconstructed plasmids.

Comparative genetic maps and protein analysis. Comparative studies were performed using a set of representative plasmids whose complete or almost complete sequence was known. The representative plasmids were selected based on the phylogenetic relatedness of the plasmids and corresponded to one plasmid of each (sub)clade (see below). The homologous genes of the selected plasmids were initially detected using the tools 'get_homologues' or 'Blast' (46,47). Initially identified homologies were checked by performing global pairwise alignment to confirm and optimize global alignments between pairs of sequences. The percentages of identity obtained by pairwise alignments were used to determine whether the genes encoding the proteins were considered homologous or not: proteins were considered homologous and not homologous when their identity levels were above 30 and below 15%, respectively. Additional analyses were performed for genes encoding proteins sharing identity levels between 15 and 30% to determine whether these proteins share common features. These analyses included the prediction of (i) transmembrane domains (TMHMM server, http://www.cbs.dtu.dk/services/TMHMM/), (ii) an N-terminal sec-dependent secretion signal (SignalP-5.0 server, http://www.cbs.dtu.dk/services/SignalP/), (iii) coiled coil regions (COILS server, https://embnet.vital-it.ch/software/COILS_form.html) or (iv) any conserved domains or classification to a protein family (InterProScan, www.ebi.ac.uk/Tools/services/web/toolform.ebi?tool=iprscan5&sequence). In addition, proteins were screened for similar secondary structure using the Jpred server (www.compbio.dundee.ac.uk/jpred/) or possible structural relationships using the HHPred server (toolkit.tuebingen.mpg.de/tools/hhpred). Genetic maps were generated using the GenomeDiagram module (Biopython, biopython.org) and edited with Inkscape software to manually annotate and represent the homology relationships on the plasmids maps.

Phylogenetic trees. Phylogenetic trees were constructed using IQ-TREE software (48). In all cases, the IQ-TREE software was allowed to determine the substitution model to be used through the ModelFinder module (49), and the ultrafast bootstrap statistical method (1000 replicates) was applied. The root was placed at the midpoint on the unrooted generated trees. The specific conditions, parameters and statistics applied for each tree were as follows.

pLS20 family phylogenetic tree based on the replication module. This tree was constructed from an alignment generated by the MAFFT algorithm (mafft.cbrc.jp/alignment/server/) using the following settings: E-INS-i method, recommended for sequences with multiple conserved domains and long gaps and discarding the gaps. The likelihood tree was made from 35 sequences each having a length of 842 nucleotides. The model selected by ModelFinder was TPM3u + F + G4. The Log-likelihood of the tree was -4201.04.

pLS20 family phylogenetic tree based on conserved CDS of pLS20-like plasmids. This tree was constructed with the help of the get_homologues and get_phylomarkers programs (46,50). Clusters of orthologous genes (COG) detected by the get_homologues program were filtered on a pipeline in order to identify high-quality markers, whose coding DNA sequences were subsequently concatenated and used to elaborate a likelihood tree using IQ-TREE. The

get_homologues program was applied using the following parameters: COG algorithm, *E*-value < 1e-5 and coverage > 60. The settings of the Get_phylomarkers program were as follows. IQ-TREE algorithm, k 1.0, m 0.6, T high and test all DNA models. The resulting tree was based on an alignment of 34 sequences corresponding to a concatenated 5186 long nucleotide sequence. The CDS from which the concatenated sequence of each plasmid was obtained corresponded to genes *24c* (*alp7A*), *25* (*rap*), *28*, *36*, *38*, *39*, *40*, *58* (*rel*) and *60* (according to the nomenclature of plasmid pLS20). The substitution model was GTR+F+ASC+G4. The Log-likelihood of the tree: -48326.87.

Phylogenetic tree of the Rap proteins. The MAFFT algorithm applying the following settings aligned the sequences of the Rap proteins of the seven pLS20-like plasmids analysed in the comparative study. G-INS-i method, recommended for sequences with global homology, and discarding the gaps. The likelihood tree was made from eight sequences with 386 residues. The model selected by ModelFinder was cpREV+F+I+G4. The Log-likelihood of the tree is -5289.94.

The constructed trees were used to determine (sub)clades. For this purpose, the program 'TreeCluster' was used, selecting the clustering method, Avg Clade, threshold 0.045 (51). This program allows defining clusters from phylogenetic trees applying the same criterion (chosen algorithm) to all the branches of the tree, which facilitates the interpretation and consistency of the clades/clusters.

RESULTS AND DISCUSSION

A family of plasmids containing ori_{pLS20}-like sequences

The replication module of most theta-type replicating plasmids of G+ bacteria contains a gene encoding a replication initiation protein that binds to specific sequences within the origin of replication region, thereby provoking unwinding of a DNA region to allow host-encoded DNA replication proteins to be loaded (52–55). The replication module of pLS20, which we refer to as ori_{pLS20}, is fundamentally different from the other plasmids because it is unusually small (~1.1 kb) and does not contain a typical replication gene (15,56). Very little is known about the mechanism by which pLS20-DNA replication is initiated, but the host-encoded DnaA protein may play a role in replication initiation because the ori_{pLS20} region contains a consensus DnaA binding site, and three additional putative DnaA binding sites containing up to two mismatches with respect to the consensus sequence (15). As a first approach to identify pLS20-related plasmids, we screened public databases using the ori_{pLS20} sequence as query against the RefSeq genome database. This search resulted in 35 significant hits (see Supplementary Table S4). All of them contain DnaA binding sites with a maximum of two mismatches with respect to the consensus sequence that are located at very similar positions in the replication module, supporting the view that DnaA plays a role in DNA replication of this type of origin. Another noteworthy feature is that all the 35 hits corresponded to sequences of *Bacillus* species, indicating that the pLS20-type replication module is limited to the genus *Bacillus*. The absence of a typical DNA replication initiation gene, its small size and the presence of DnaA binding sites suggest that replication of the pLS20-type replication depends on host-encoded proteins, which may explain why this type of replicon has an apparent narrow host range. The availability of the 34 additional sequences of this type of replication module could provide useful information for future studies into the replication initiation mechanisms of pLS20 and related plasmids.

Whereas several sequences were annotated as a plasmid, other hits corresponded to individual contigs of different NGS projects. Analyses and descriptions of the annotated plasmids, and those reconstructed from one or more contigs are presented in the following two subsections.

A: Hits annotated as plasmids

Fifteen of the thirty-five hits sharing similarity with ori_{pLS20} were annotated as a plasmid (see Table 1). Unnamed plasmids were named using the first letter of the genus and the first two letters of the species plus the name of the strain; for simplicity we also use a shortened name (C1-C15) as indicated in Table 1. Plasmids C2 to C10 (81,82,83) were highly similar to pLS20 but three (C1, C2, C8) were anomalously large. These three large plasmids contain duplications that were defined by repeated sequences (see Supplementary Figure S1), which could have resulted from assembly artefacts and so were not included in the comparisons and phylogenetic analyses.

The 10 plasmids that are very similar overall to pLS20 are indicated with a green letter colour in Table 1. One of these is present in the *Bacillus* species strain M4U3P1 (C7), the remaining nine, like pLS20, are present in *B. subtilis* strains. Four of them (C2, C7, C9 and C10) are harboured by strains present in soil samples. The other six (C1, C3–C6 and C8) share another feature with pLS20 being that their host strains are used for the production of human food.

pLS20 and the 10 highly similar plasmids have a GC content of 38%, but the other five plasmids have a higher (39–40%) or lower (37%) GC content, indicating that they have diverged more from pLS20 (see Table 1 and below). Three of these five plasmids are present in strains of *Bacillus velezenzis* (C11, C12 and C15) (84,85), one in *Bacillus glycinifermentans* (C13) and one in *Bacillus pumilus* (C14) (30,31). The *B. pumilus* plasmid corresponds to p576 of *B. pumilus* strain NRS576 that we have previously sequenced in our laboratory (30,31, see also Introduction). p576 shares 42.6% identity at sequence level with pLS20. Plasmid pBglSRCM103574 (C13) harboured by *B. glycinifermentans* SRCM103574 shares 53.5% identity with pLS20. Regarding the three *B. velezensis* plasmids: two of them (C11 and C12) are 99.6% identical to each other, except for a small region of 215 bp, and they share 66% identity with pLS20. The region identified on the third *B. velezensis* plasmid (plasmid p2, C15) covers only 57% of the ori_{pLS20} region. In addition, the plasmid is about three times smaller than that of pLS20, and, importantly, contains a gene whose encoded protein shares 40% similarity with the replication protein of the rolling-circle plasmid pUA140 from a *Streptococcus mutans* strain (57). Based on this, we discarded plasmid p2 (C15) as a member of the pLS20 family of plasmids. In sum-

Table 1. General features of plasmids containing an *ori*$_{pLS20}$-like origin of replication

Hit no *	Plasmid/organism	Size (bp)	CG (%)	ID (%)	Reference
Ref§	**pLS20** / *B. subtilis natto*	64 755	38	100	(20)
C1	**pLDW15** / *B. subtilis* ATCC 21228	85 618 / 64 755**	38	100**	-
C2	**pBsuKH2** / *B. subtilis* KH2	74 165 / 64 755**	38	100**	(81)
C3	**pBsuN1-1** / *B. subtilis* N1-1	64 613	38	99.8	-
C4	**pBsuN2-2** / *B. subtilis* N2-2	64 616	38	99.8	-
C5	**pBsuN3-1** / *B. subtilis* N3-1	64 615	38	99.8	-
C6	**pBsuN4-2** / *B. subtilis* N4-2	64 614	38	99.8	-
C7	**pBspM4U3P1** / *Bacillus sp*. M4U3P1	65 122	38	99.0	-
C8	**pBS333** / *B. subtilis* SRCM100333	82 495 / 65 928**	38	97.8**	-
C9	**pBsu29R7-12**** / *B. subtilis* 29R7-12	64 604	38	99.8	(82)
C10	**pBsuCGMCC2108**** / *B. subtilis natto* CGMCC 2108	65 774	38	98.4	(83)
C11	**pDSYZ** / *B. velezenis* DSYZ	62 485	40	66.6	(84)
C12	**pBveCGMCC11640**** / *B. velezensis* CGMCC 11640	62 700	40	66.5	(85)
C13	**pBglSRCM103574**** / *B. glycinifermentans* SRCM103674	65 273	40	53.5	-
C14	**p576** / *B. pumilus* NRS576	43 328	37	42.6	(30,31)
C15	**p2_DKU_NT_04** / *B. velezensis* DKU_NT_04	18 288	40	22.5	-

Highly similar plasmids are given in the same colour.
Grey letter colour (C15) indicates a possible false positive hit.
*, Arbitrary number to facilitate descriptions in the text (C, complete plasmid sequence known).
**, Plasmid size and % of identity after removal of large sequence duplication.
§, reference plasmid.

mary, pairwise comparisons suggest that 14 of the 15 plasmids identified are related to pLS20.

B: Hits annotated as contigs

Besides the 15 hits annotated as plasmid, 20 other hits corresponded to contigs from different NGS projects. Some of these contigs had a large size indicating that they reflected a large or almost entire region of a plasmid. The sizes of the other contigs, though, were small, suggesting that they could be part of a pLS20-related plasmid whose sequence is fragmented over multiple contigs of the NGS project. The following approach was used to reconstruct, at least partially, a plasmid from these multiple contigs belonging to the same NGS project. First, we generated a small database containing all sequences of the 20 NGS projects from which an *ori*$_{pLS20}$-like sequence was detected. A TblastX search was then launched against this database using as query the complete sequences of plasmids pLS20, pDSYZ, (C11), pBglSRCM103574 (C13) and p576 (C14), which share <80% identity among them. This approach allowed the identification of contigs containing one or several (putative) genes sharing similarity with any of these four plasmids. A disadvantage of this strategy is that plasmid contigs that do not encode an ORF sharing similarity with ORFs encoded by any of the query plasmids will not be recognized. This handicap will particularly affect those NGS projects having a high number of contigs and/or contigs of small sizes. Using this strategy, we succeeded in identifying between 1 and 10 contigs for all of these 20 NGS projects. A summary of the results is given in Table 2 and Supplementary Table S5.

None of the 20 NGS projects, which we identified to contain one or more contigs sharing similarity with the query plasmid(s), were annotated to contain a plasmid. We gave a name to each of these reconstructed plasmids using the nomenclature described above. Because it is likely that these reconstructed plasmids do not cover the complete plasmid sequence (see also below), their names were extended with a dagger (†) or a double dagger (‡) symbol, to distinguish whether the reconstructed plasmid was based on a single or multiple contigs, respectively. In addition, for simplicity we also refer to these (partially) reconstructed plasmids using a shortened name (F1-F20) as indicated in Table 2. For seven of the 20 NGS projects, no additional contigs were identified besides the one containing the *ori*$_{pLS20}$-like sequence, suggesting that these contigs, which have large sizes between 40 and 82 kb, correspond to almost the entire plasmid. This conclusion was supported by completing the sequence of *Bacillus atrophaeus* NRS213 plasmid pBatNRS213† (F9) (see below). In the other 13 NGS projects, up to nine additional contigs were identified that probably correspond to the plasmid. To verify this assumption, the bioinformatics software 'Mauve' was used, which allowed alignment of multiple genomes and to align fragmented genomes against a reference genome (45). The plasmid showing the highest values in the TBlastX alignments was used as reference plasmid. Genome alignments obtained by Mauve revealed that in all cases the identified contigs align with their corresponding reference plasmid along almost their entire lengths, supporting the view that they indeed correspond to plasmid sequences (see Supplementary Figures S2–S6).

Supplemental Table S4 shows that the 35 strains harbouring a pLS20-related plasmid were isolated from different environmental niches at very different geographical locations. Therefore, in summary, these results show that pLS20 can be considered the prototype of a family of related plasmids

Table 2. (Partially) reconstructed plasmids from multiple contigs containing an ori_{pLS20}-like sequence

Plasmid name *	No of contigs	NGS project	Strain	Total size (bp)	Ref **
pBsuVK161‡ (F1)	4	GCF_004323065	*B. subtilis natto* VK161	64 761	pLS20
pBsuSFA-H43† (F2)	1	GCF_003265645	*B. subtilis* SFA-H43	63 941	pLS20
pBsuMiyagi-4‡ (F3)	8	GCF_00747645	*B. subtilis* Miyagi-4	58 469	pLS20
pBsuJRS9‡ (F4)	7	GCA_001286745	*B. subtilis* JRS9	53 279	pLS20
pBsuJRS2‡ (F5)	7	GCA_001286845	*B. subtilis* JRS2	53 276	pLS20
pBspX2† (F6)	1	GCA_002278635	*Bacillus sp*. X2	56 719	pLS20
pBamB1895‡ (F7) #	10	GCA_000696285	*B. amyloliquefaciens* B1895	52 280	pDSYZ
pBamJRS8‡ (F8)	9	GCF_001286965	*B. amyloliquefaciens* JRS8	54 931	pDSYZ
pBatNRS213† (F9) $	1	GCF_001584335	*B. atrophaeus* NRS213	66 742	pLS20 or pDSYZ
pBatNBRC15407‡ (F10)	6	GCA_006539905	*B. atrophaeus* NBRC15407	64 276	pLS20 or pDSYZ
pBatNBRC15539† (F11)	1	GCA_001591925	*B. atrophaeus* NBRC15539	66 706	pLS20 or pDSYZ
pBliB4092‡ (F12)	8	GCA_001587195	*B. licheniformis* B4092	57 450	pBglySRCM10574
pBliYNP2† (F13)	1	GCA_001896285	*B. licheniformis* YNP2	67 839	pBglySRCM10574
pBalRIT380‡ (F14)	2	GCA_001029865	*B. altitudinis* RIT380	41 563	p576
pBpuPS115‡ (F15)	5	GCA_003985025	*B. pumilus* Ps115	37 502	p576
pBpuSAFR032‡ (F16)	2	GCA_007679675	*B. pumilus* SAFR-032	39 331	p576
pBpuDE0471† (F17)	1	GCA_007667475	*B. pumilus* DE0471	40 281	p576
pBsa7783‡ (F18)	7	GCA_002276315	*B. safensis* 7783	32 019	p576
pBspLLTC93‡ (F19)	2	GCA_002993125	*Bacillus sp*. LLTC93	38 911	p576
pBspNMCC4† (F20)	1	GCF_002998555	*Bacillus sp*. NMCC4	82 879	§

*Identified hits were arbitrarily numbered F1 to F20 to facilitate their descriptions (F, Fragmented plasmid).
**Reference plasmid corresponds to the query plasmid for which the highest score was obtained in TblastX searches.
§Reference plasmid pBglySRCM130574 has the best score (563) for *Bacillus* sp. NMCC4. However, the score obtained for this comparison was considerably lower than scores obtained for the other hits.
#, *B. amyloliquefaciens* strain B1895 was kindly provided by Dr Vladimir A. Chistyakov and Dr Michael L. Chikindas.
$, *B. atrophaeus* NRS213 strain was obtained from the *Bacillus* Genetic Stock Center (BGSC, stock 11A2).

whose members are harboured by different *Bacillus* species spread over very different global locations and environmental niches.

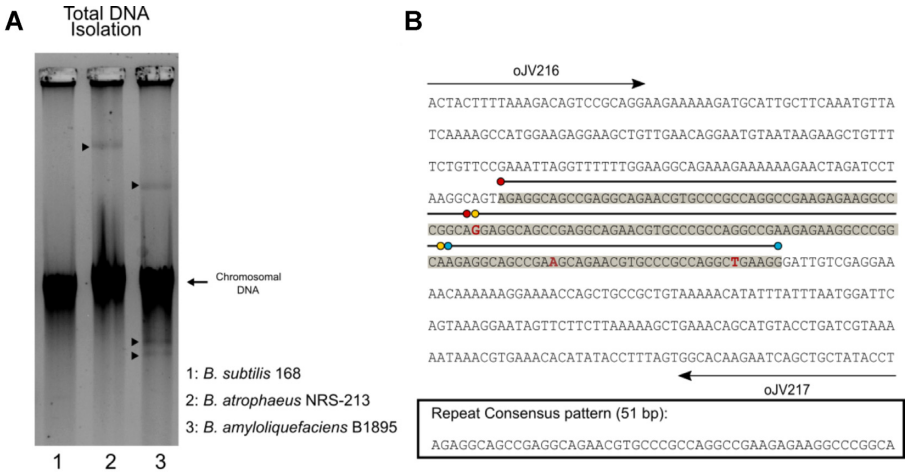
Confirmation that *B. amyloliquefaciens* B1895 and *B. atrophaeus* NRS213 each contain a pLS20-related plasmid

Twenty of the 35 members of the pLS20 family of plasmids were partially reconstructed from one or more contigs of their corresponding NGS projects. To confirm that the sequences of these contigs correspond to plasmids, we selected the putative plasmids pBamB1895‡ and pBatNRS213† from *Bacillus amyloliquefaciens* B1895 and *B. atrophaeus* NRS213, respectively, for further analysis. pBatNRS213† was identified from a single contig, suggesting that its sequence corresponds to almost the entire plasmid. On the other hand, plasmid pBamB1895‡ was reconstructed from 10 contigs (see Table 2). These two plasmids were selected also because their hosts were interesting from a biotechnology point of view. *B. amyloliquefaciens* B1895 is used as a probiotic agent in fish industry or a poultry dietary supplement (58,59); and *B. atrophaeus* NRS213 is considered the type strain of *B. atrophaeus* (60), an organism that is frequently used as a surrogate organism in *B. anthracis* studies (61).

We first tested whether these two strains contained plasmids by isolating their total DNA contents and analysing them by agarose gel electrophoresis. For both strains, but not for the plasmid-free *B. subtilis* strain 168, an additional, discrete DNA band was observed that migrated above the bulk of chromosomal DNA, and could correspond to a large plasmid (Figure 1A). In the case of *B. amyloliquefaciens* B1895, but not *B. atrophaeus* NRS213, additional DNA bands were also detected that migrated below the chromosomal DNA, suggesting that this strain contains more than one plasmid.

To sequence the gap of plasmid pBatNRS213†, two outward-oriented primers that hybridized near the ends of the single long contig were applied as primers in PCR reactions using total DNA of *B. atrophaeus* NRS213 as template. The PCR reaction generated a DNA fragment of about 480 bp, which was cloned in an *E. coli* vector and then sequenced. Analysis of the sequence revealed that the gap has a size of 124 bp, hence pBatNRS213 has a size of 66,829 bp. The gap region contains 2.7 times a direct repeated sequence of 51 bp (see Figure 1B). Probably, these repeated sequences caused problems in the assembly process and have prevented circularization of the plasmid. This region of pBatNRS213 forms part of a gene that is a homologue of pLS20 gene *46*, which contains similar repeats. The annotation of pBatNRS213 (previously contig NZLSBB01000022) was then manually revised, and genes and features were named using the nomenclature of pLS20. The complete pBatNRS213 sequence was deposited in the NCBI database and given accession number OK210094. pBatNRS213 has 83 putative genes, most of them sharing homology with a pLS20 gene (see Supplemental Figure S7 and Supplemental Table S6).

pBamB1895‡, which was reconstructed from 10 contigs, is substantially smaller than the reference plasmid pDSYZ, raising the possibility that its sequence was not complete (see Supplementary Figure S4). Therefore, we performed a new NGS sequencing project of *B. amyloliquefaciens* strain B1895 using special conditions resulting in an extremely high coverage number (see Materials and Methods). A summary of the sequencing results is provided as Supplementary Data. This sequencing approach resulted in the identification of a 63 409 bp extrachromosomal



C

B. amyloliquefaciens B1895 assembly
(contigs > 500bp)

Component	nº Contigs	size	GC%
Bacterial chromosome	13	3,930,816	46,4
Putative phage	1	129,832	35,6
Plasmid pBamB1895	1	63,409	40,9
Plasmid pBamB1895.2	1	12,453	37,8

Figure 1. *B. atrophaeus* NRS213 and *B. amyloliquefaciens* B1895 contain a large plasmid. (**A**) Bacterial DNA was isolated from the plasmid-free control strain *B. subtilis 168*, and from the *B. atrophaeus* NRS213 and *B. amyloliquefaciens* B1895 strains using a total DNA isolation procedure, and subjected to agarose gel electrophoresis. The migration position of chromosomal DNA is indicated with an arrow, and arrowheads indicate the migration position of additional DNA in the *B. atrophaeus* NRS213 and *B. amyloliquefaciens* B1895 strains. DNA was separated on an agarose gel lacking EtBr, which was stained after electrophoresis. Under these conditions, the DNA of large covalently closed plasmids migrates slower than the bulk of fragmented chromosomal DNA. (**B**) Upper panel, gap sequence of plasmid pBatNRS213. Primers oJV216 and oJV217 (arrows) were used to amplify and clone the gap sequence of pBatNRS213 into vector pBluescript II KS (+). Next, the recombinant plasmid was sequenced. This gap region, which is indicated with a grey background, contains 2.7 repeats of a 51 bp sequence. These repeated sequences are over lined with black lines starting and ending with coloured spheres (red and yellow labelled lines correspond to 51 bp, and the blue-labelled line corresponds to 39 bp. Base pairs deviating from the consensus 51 bp repeated sequence [shown in the lower panel] are indicated in red). The direct repeated sequence may have caused assembly problems. (**C**) Characteristics of the re-sequenced *B. amyloliquefaciens* strain B1895.

element that corresponded to the pLS20-related plasmid pBamB1895 (deposited in the NCBI database and given accession number OK210095). All the 10 contigs of the previously published *B. amyloliquefaciens* B1895 sequencing project ASM69628v1 form part of the pBamB1895 plasmid determined here, implying that they were correctly identified from the pool of contigs present in the published NGS project. This validated our strategy for identifying pLS20-like sequences present in NGS projects containing multiple contigs. The size of pBamB1895 obtained in our sequence analysis is about 11 kb larger than that of the plasmid reconstructed from the 10 contigs present in the previously published NGS project. This result shows therefore that a high sequence coverage is an important feature for obtaining high quality sequencing data in general and particularly for identifying plasmid sequences. Both automatic and manual annotation showed that pBamB1895 contains 79 putative genes (see Supplementary Figure S8 and Supplementary Table S7). Plasmid pBamB1895 contains a bacteriocin cassette that is very similar to the lactococcal 972 family of bacteriocin cassettes (62,63). The reference plasmid pDSYZ does not contain such a bacteriocin cassette, and hence this demonstrates, as we predicted, that regions corresponding to DNA sequences that are not present on the reference plasmids, would not be identified as sequences of a plasmid by the strategy applied.

The pLS20 family of plasmids can be divided into four clades

We were interested in gaining insights into the evolutionary relationship between the plasmids identified. However, because plasmids reconstructed from multiple contigs are likely to be incomplete (see foregoing section), these insights could not be obtained by constructing a maximum likelihood tree based on comparison of the complete nucleotide sequences. As an alternative, we constructed, instead, two other maximum likelihood trees. One of these was based on the replication regions (Figure 2A) and the other on a concatenated sequence of the following selected conserved orthologous genes (COGs): *24c* (*alp7A*), *25* (*rap*), *28*, *36*, *38*, *39*, *40*, *58* (*rel*) and *60* (Figure 2B). The strategy used to select the orthologous pLS20 CDSs for comparison and the settings of the programs involved are outlined in Ma-

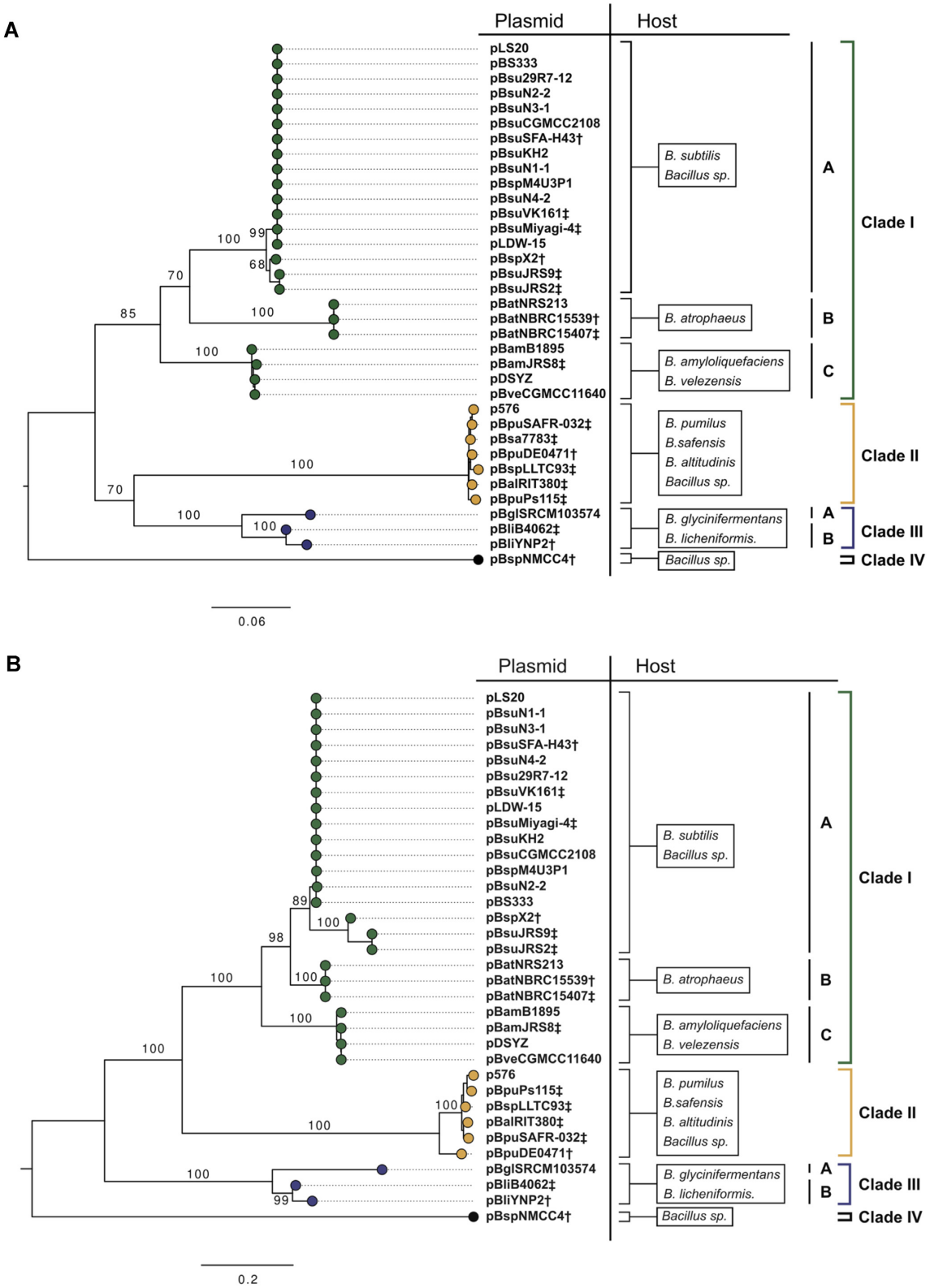


Figure 2. Phylogenetic relatedness of the pLS20 family of plasmids inferred from maximum likelihoods. Phylogenetic trees were constructed for the 35 members of the pLS20 family of plasmids. The tree in panel (**A**) was constructed from the nucleotide sequence corresponding to the replication modules, and the tree in panel (**B**) was built from the nucleotide sequences of ten concatenated CDS encoding orthologous proteins present in all analysed plasmids (see also text and see Materials and Methods section for settings). Statistical evidence for each branch is provided by bootstraps analysis (1000 replicates, indicated as percentages). The roots are located at the midpoint. According to both analyses, the plasmids can be grouped into four clades indicated by Roman numbers I to IV and labelled with colours: clade I, green, clade II, yellow, clade III, blue and clade IV, black. Names of the *Bacillus* strains harbouring the plasmids are given in Tables 1 and 2.

terials and Methods section. The replication module- and the COG-based trees are very similar to each other (compare Figure 2A and B). This indicates that the replication, partitioning and conjugation modules (see below) have not been shuffled between them. Both trees revealed the presence of four main clades (I to IV) using the TreeCluster algorithm (see Materials and Methods section) whose main branches are supported by a >95% of bootstrap replicates. In addition, clade I can be divided into three subclades, and clade III into two subclades. Remarkably, plasmids of different (sub)clades are present in the same or very closely related *Bacillus* species. This strongly suggests that plasmids of each (sub)clade have evolved within a narrow range of closely related species, which might be a major reason for differentiation of the plasmids into the four different clades.

Modular organization of pLS20 family plasmids

Figure 3A shows a genetic map of pLS20. The inner circle of the map shows that pLS20 contains, besides a number of genes with unknown function, at least four modules: the essential replication module, the partition module, the conjugation module that occupies more than half the size of the plasmid, and a DNA repair module that had not been shown to be present on *Bacillus* plasmids previously. Comparisons were made to determine which of these modules/genes were conserved among the pLS20 related plasmids identified, using the following plasmids as references: pLS20 (clade IA), pBatNRS213 (clade IB), pBamB1895 (clade IC), p576 (clade II), pBglSRCM103574 (clade IIIA), pBliYNP2† (clade IIIB) and pBspNMCC4† (clade IV). Based on *in silico* and manual analyses and criteria (see Materials and Methods section), the four modules present on at least four of the seven representative plasmids share 34 conserved genes (see Table 3 for general features of these genes; the accession numbers of the orthologues are given in Supplementary Table S8). Among the 34 conserved genes, two belonged to the partitioning module, two to the DNA repair module, and 30 to the conjugation module (three of which are responsible for regulation of the conjugation operon). This demonstrates that (i) three of the four modules (replication, partitioning and conjugation) identified on pLS20 are conserved among the pLS20 family members, (ii) the DNA repair module is conserved in plasmids of clade I and IV (see below additional information on this module regarding plasmids of clade II and III), and (iii) each plasmid contains a variable number of non-conserved genes of unknown function. Moreover, alignments of the seven reference plasmids showed that the structural organization of the modules is also conserved (Figure 3B). Thus, the replication module is flanked at its left and right by the partition and conjugation modules, respectively, and the DNA repair module, present on four of the seven reference plasmids, is located in between the conjugation and the partitioning modules together with poorly conserved genes. In the following sections, the different modules, and the (dis)similarities between the plasmids of different clades are described in more detail, with special emphasis on the conjugation module.

The partitioning module

The partitioning module of pLS20 is located adjacent to the replication module and is composed of a bicistronic operon and a cis-acting locus located upstream of the operon. The two genes involved are named *alp7A* and *alp7R*. The first gene encodes an actin-like protein that is similar to other partitioning proteins (16). The second gene, *alp7R*, encodes a DNA binding protein that binds to the repeated sequence motifs located upstream of *alp7A* and thereby regulates the expression of the operon. In addition, Alp7R can interact with Alp7A and probably regulates the activity of Alp7A (17). All pLS20 family plasmids analysed contained an *alp7A/R*-type partitioning module. The Alp7R proteins are more diverse than the Alp7A proteins (see Supplementary Table S9). The regions containing the Alp7R binding sites located upstream of the operon, also display considerable diversity (not shown). Possibly, this reflects the specificity of the Alp7R protein for its cognate binding sites to avoid cross-talk between partitioning systems of related plasmids.

The DNA repair module

Genes *12* and *13* of pLS20cat, which form a bicistronic operon, are conserved among pLS20 family members of clade I and IV. The proteins encoded by pLS20 genes *12* and *13* share similarities with the *B. subtilis* UvrX and YolD proteins (see Supplemental Figure S9), repectively. UvrX and YolD are thought to be the homologs of the UmuC and UmuD′ (a processed form of UmuD) subunits of the *E. coli* error prone DNA polymerase (Pol V); UmuC being the catalytic subunit and UmuD′ the auxiliary subunit. The Pol V type error prone DNA polymerases carry out DNA translesion synthesis (TLS) and form part of the SOS response (64). The presence of a possible error prone DNA polymerase based DNA repair system in two clades of pLS20 family members is interesting because, although *uvrX/yolD* modules are present on bacterial genomes and some bacteriophages and homologs of gene *13* are present on several *Listeria* plasmids (65–67), it has not been reported before on *Bacillus* plasmids. Its presence probably provides an advantage for the plasmid, and possibly for the host, to overcome prolonged DNA replication problems

The conjugation module

The quorum sensing-based regulation of conjugation. In pLS20, the large conjugation operon is preceded by three genes, *rco*, *rap* and *phr*, whose encoded products regulate the activity of the main conjugation promoter P_c located upstream of the conjugation operon (for recent review see, 68). A schematic view of how the conjugation promoter is regulated is shown in Supplementary Figure S10). The transcriptional regulator Rco_{pLS20} binds to two operators, one of which coincides with the overlapping divergently oriented promoters P_r and P_c that drive expression of *rco* and the conjugation operon, respectively (21,22,25). Binding of Rco_{pLS20} to its operators results in DNA looping, which is required for the proper regulation of the two promoters (21,22,25). *rap* and *phr* form a bicistronic operon.

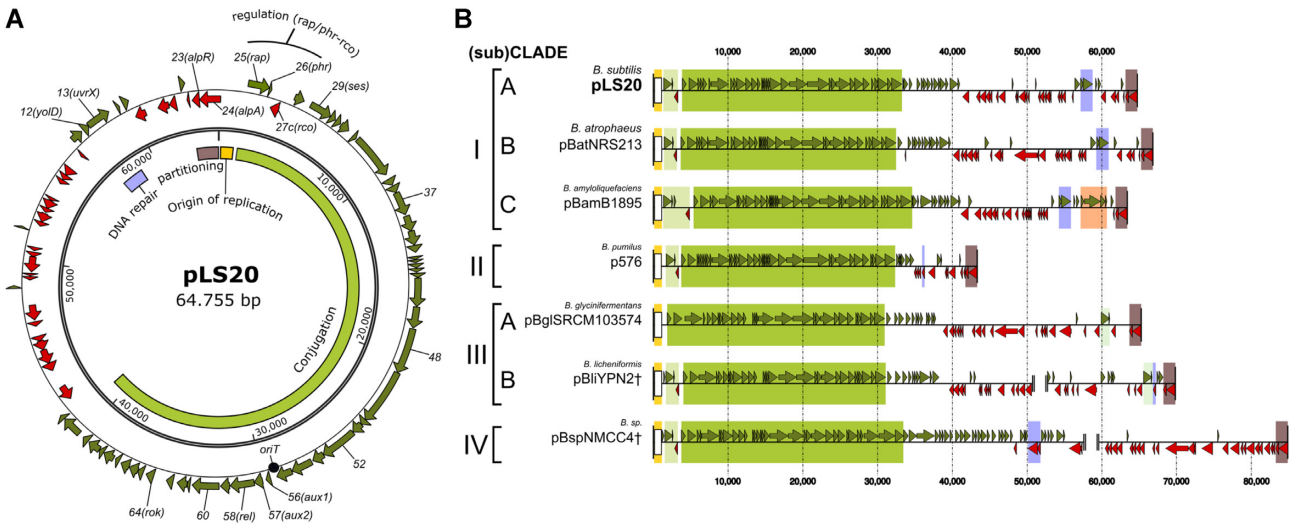


Figure 3. pLS20 family plasmids have a conserved modular organization. (**A**) Genetic map of pLS20. Different modules are indicated on the inner circle of the pLS20 map: replication module (yellow), DNA repair module (purple), partition system (brown) and conjugation module (green). The middle circle indicates base pair positions, ticked every 10 000 bp. All (putative) genes are labelled on the outer circle. Clockwise/counter-clockwise-oriented genes are shown in green and red, respectively. The black circle represents the origin of transfer region (oriT). (**B**) Alignment of linear genetic maps of representative plasmids of the pLS20 family. Alignment of reference plasmids pLS20 (clade IA), pBatNRS213 (clade IB), pBamB1895 (clade IC), p576 (clade II), pBglSRCM103574 (clade IIIA), pBliYNP2† (clade IIIB) and pBspNMCC4† (clade IV). The plasmids are aligned starting at the left side with the replication module. Modules are indicated with rectangles using the same colour code as in panel (A). The bacteriocin cassette present on pBamB1895 is indicated with an orange box. Base pair numbering are shown at the top and bottom; every 10 000 bp are indicated with a vertical interrupted line. Right and left-ward oriented genes are indicated with green and red arrows, respectively.

rap encodes an anti-repressor and *phr* encodes a small pre-proprotein that, after secretion and processing generates the mature signalling peptide Phr*, which inhibits the anti-repressor activity of Rap (24,25).

The location and the genetic organization of the cassette containing these three genes are conserved in most members of the pLS20 family of plasmids. This region is somewhat differently organised in the clade I plasmids that are hosted by *B. velezensis* and *B. amyloliquefaciens* species, and the clade III plasmid pBglSRCM103574. The four clade I plasmids hosted by *B. velezensis* and *B. amyloliquefaciens* contain an insert of (∼1 to 1.5 kb in between the convergently oriented *rap*/*phr* and *rco* genes (Figure 4) that is characterised by several multiple long (up to >200 bp) direct repeated sequences. The function and the origin of this insert are unknown. The clade III plasmids pBglSRCM103574 and pBliYPN2† (as well as the almost identical plasmid pBli4092‡) contain a second distantly related *rap*/*phr* operon (referred to as *rap2*/*phr2* to distinguish them from *rap1*/*phr1*) upstream of the replication module (Figure 4). The two Rap2 proteins share 96% similarity whereas similarity between the Rap1 proteins varies between 60 to 80% (see Supplementary Figure S11). Interestingly, while pBliYNP2† contains both *rap*/*phr* cassettes, pBglSRCM103574 lacks the *rap*/*phr* cassette 1 and the flanking *rco* gene (see Figure 4). The conjugation genes of pBglSRCM103574 are also expected to be regulated, because constitutive expression of the conjugation genes is thought to impose a high metabolic burden on the host cell and has been shown to provoke plasmid instability of pLS20 (21). Interestingly, a convergently oriented gene follows the *rap*/*phr* cassette 2 on pBglSRCM103574. Although the protein encoded by this gene does not share significant homology with Rco_{pLS20}, it is predicted to be a DNA binding protein. It is therefore possible that the *rap2*/*phr2* genes and its flanking gene together regulate expression of the conjugation operon of pBglSRCM103574.

Finally, it is worth mentioning that the orthologues of the regulatory proteins Rap, Phr and Rco are more divergent than most of the other orthologues (see Supplementary Table S9). Possibly, these divergences reflect differences in protein–DNA and protein–protein interactions, thereby ensuring that each plasmid regulates its conjugation genes with high specificity. The *B. subtilis* genome contains eight *rap*/*phr* cassettes and each Rap protein interacts with its cognate mature Phr peptide (69). In this sense, it is interesting to note that the Phr2 peptides encoded by pBglSRCM103574 and pBliYPN2† are identical (see Figure 4).

The conjugation operon. The start of the conjugation operon has been well defined empirically in the case of pLS20. Gene *28* constitutes the first gene of the conjugation operon and the upstream P_c promoter is responsible for the expression of the conjugation operon (21,22,25). The overall genetic organization and synteny of conjugation genes are conserved in all the other members of the pLS20 family of plasmids: they all contain an orthologous gene of pLS20 gene *28*, which in all cases constitutes the first gene of the conjugation operon. The end of the conjugation operon is less clear though. In pLS20, the conjugation operon has been described to span genes *28* to gene *74* (21). However, the last gene that is conserved in the conjugation operon among all the pLS20 family of plasmids is gene *60*, encoding the DNA primase. Therefore, for the studies performed here, we consider gene *60* the last gene of the conjugation

Table 3. Genes conserved in at least four of the seven reference plasmids

Module	Conserved gene	Size*	TMSD#	Signal peptide	Protein family	Function
Quorum sensing	*25*	370 (366–372)	No	No	IPR011990	QS anti-repressor Rap
	26	42 (34–44)	1 (central)	Yes	-	QS peptide Phr
	*27c***	129 (127–166)	No	No	IPR010982	Conjugation repressor Rco
Conjugation operon	*28*	172 (170–177)	No	No	IPR020270	Unknown
	29	362 (342–389)	1 (N-terminal)	Yes	-	Surface exclusion protein Ses
	34	768 (765–778)	1 (N-terminal)	Yes	IPR013552	Adhesion protein Tie
	36	126 (125–133)	1 (N-terminal)	Yes	-	Unknown
	37	402 (388–433)	No	No	IPR001482	Putative VirB11
	38	268 (267–286)	1/2 (N-terminal) 2 (C-terminal)	No	-	Unknown
	39	243 (234–247)	1 (N-terminal) 1 (C-terminal)	No	-	Unknown
	40	120 (120–123)	1 (N-terminal)	No	-	Unknown
	41	47 (45–54)	1 (central)	No	-	Unknown
	43	71 (62–72)	No	No	-	Unknown
	44	63 (58–67)	No	No	-	Unknown
	*45***	76 (76–78)	1 (N-terminal) 1 (C-terminal)	No	-	Unknown
	*46***	439 (363–691)	1 (C-terminal)	No	-	Unknown
	47§	361 (344–363)	No	No	-	Unknown
	48	786 (761–820)	2 (N-terminal)	No	IPR003688	T4CP, putative VirD4
	49	906 (690–1135)	5 (N-terminal)	No	-	Putative extended-VirB6
	50	103 (94–108)	2 (central)	No	IPR020275	unknown
	51	192 (189–226)	No	No	-	unknown
	52	631 (624–643)	No	No	IPR027417	Putative VirB4
	53	205 (190–212)	1 (N-terminal)	No	IPR032710	NTF2-like, putative VirB8
	54	369 (345–369)	1 (N-terminal)	No	IPR023346	Putative lysozyme, putative VirB1
	55	269 (265–298)	1 (N-terminal)	No	-	Unknown
	56	79 (79–87)	No	No	IPR010985	Auxiliar protein Aux1
	57	147 (135–148)	No	No	-	Auxiliar protein Aux2
	58	410 (387–410)	No	No	IPR041073	Relaxase Rel (Mob_L)
	59§	155 (155–156)	No	No	-	Unknown
	60	455 (359–459)	No	No	IPR036977	Putative DNA primase
DNA repair	*12***	117 (110–121)	No	No	IPR014962	YolD-like protein
	13	421 (421–426)	No	No	IPR043502	UvrX-like protein
Partitioning	*23c*	129 (128–130)	No	No	-	Partitioning regulator Alp7R
	24c	391 (389–400)	No	No	-	Partitioning Alp7A

*, The median sizes of the proteins included in the COG are indicated. In parentheses, the minimum and maximum value.
**, Genes 45 and 46 are only conserved in three different clades.
§, Genes 47 and 59 are conserved in two different clades.
#, TMSD, transmembrane spanning domain.

operon. A schematic comparative map of the seven reference plasmids is shown in Figure 5.

A remarkable feature of the pLS20 family of plasmids is that the order of the conjugation genes reflects the sequential steps involved in the conjugation process. The genes for which a function is known will be briefly described following the distinct steps in the conjugation process.

The conAn antitermination system required for proper expression of conjugation genes. The first gene of the pLS20 conjugation operon, gene *28*, which is present in all pLS20 family members (Figure 5), encodes protein ConAn1, the protein component of a bipartite processive antitermination (P-AT) system. We recently characterized the pLS20 P-AT system, which is located at the start of the conjugation operon, and show that ConAn1 acts as a processivity factor while the second component, *conAn2* that exerts antitermination, corresponds to the ∼300 bp region downstream of *conAn1* (29). The system modifies the transcription elongation complex (TEC) such that it is able to antiterminate the multiple terminators inside the conjugation operon, so transcripts initiated at the P_c promoter will not terminate at the first terminator in the operon. The P-AT system also serves to minimize the negative effects of spurious transcription, while allowing differential expression of subsets of genes within the conjugation operon (29). The conjugation oper-

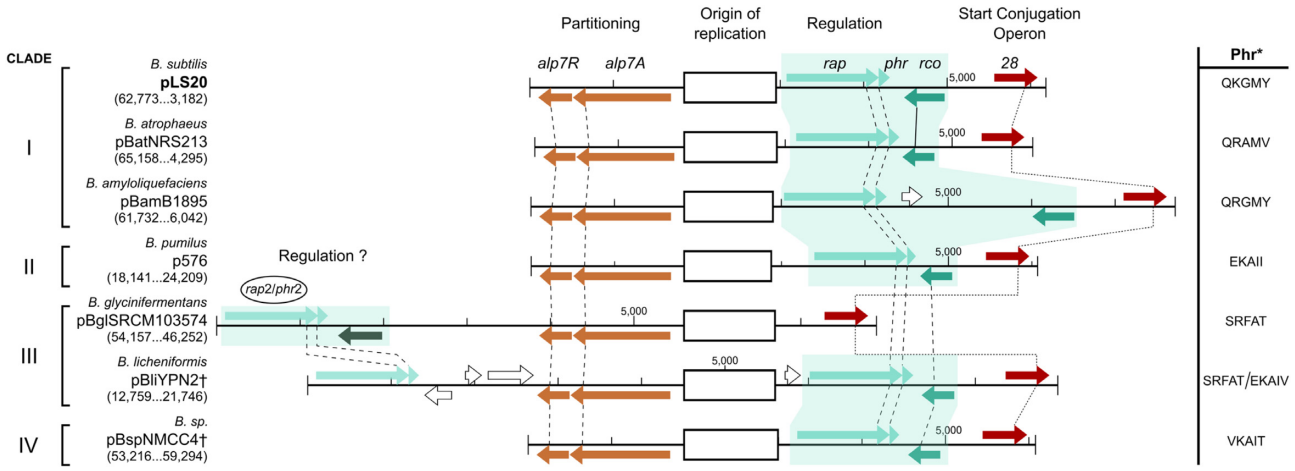


Figure 4. Comparisons between the regions of the seven representative pLS20 family plasmids covering the partitioning and replication modules, and the first gene of the conjugation operon. Schematic representation of the structural organization of the continuous regions encompassing the partitioning module (*alp7R*, *alp7A*), replication module (origin of replication), the three-gene regulatory cassette of conjugation genes (*rap*, *phr*, *rco*) and the first gene of the conjugation operon (*28*) of the seven representative pLS20 family plasmids. In the case of the clade III plasmids, an upstream region containing another *rap*/*phr* cassette is also shown (*rap2*/*phr2*). The names of the plasmids and the bacterial species from which they were isolated are shown on the left, together with the size of the region given in bp. At the right, the sequence is given of each of the (predicted) mature signalling peptides.

ons of the pLS20 family members are all very large, and pLS20 is known to contain >20 putative transcriptional terminators. Therefore, it is probably not surprising that all the members require a PA-T system

Step 1 of the conjugation process: selection of and attachment to the recipient cell. The first step of conjugation process involves selection of and attachment to a suitable recipient cell. Most if not all conjugative elements of G− bacteria produce so-called conjugative pili; organelles that extend from the donor cell surface into the extracellular space and that play an important role in the attachment to recipient cells (70). Conjugative elements of G+ bacteria do not form conjugative pili. Recently, we have shown that pLS20 gene *34*, which is essential for efficient conjugation, particularly in liquid medium, encodes an adhesion protein that contains a class II thioester domain permitting covalent attachment to a recipient cell, and was named tie_{pLS20} (71). All members of the pLS20 family of plasmids contain a homolog of tie_{pLS20} near their start of the conjugation operon. It thus appears that pLS20 family members, and probably G+ bacteria in general, depend on adhesins for attachment to recipient cells.

Transfer of the plasmid between two donor cells negatively affect the conjugation efficiency and so-called exclusion mechanisms have evolved that inhibit the transfer between two donor cells. pLS20 contains an exclusion mechanism that is based on the surface protein Ses_{pLS20} (surface exclusion system), which is encoded by gene *29* (28). Homologs of ses_{pLS20} are present on all pLS20 family plasmids where they are always located immediately downstream of the antitermination genes. Exclusion is not due to the interaction of Ses_{pLS20} proteins present on the surface of both mating cells (28). So far, the molecular mechanism of Ses-mediated exclusion is unknown. Possibly, Ses acts at the initial step of conjugation by affecting proper functioning of Tie; alternatively, Ses may act at a later stage for instance by interfering with a component of the transferosome. In summary, genes encoding proteins involved in the initial steps of the conjugation process are located near the beginning of the conjugation operon.

Step 2 of the conjugation process: the putative T4SS. The ssDNA and a few proteins are transferred from the donor into the recipient through a membrane-localized T4 secretion system (T4SS). T4SSs of G− bacteria are composed of 12 core subunits that elaborate a 'minimal' T4SS spanning both the inner and outer membranes. Some G− plasmids encode a minimal T4SS, but others form 'expanded' T4SSs that comprise besides the core subunits multiple additional, system-specific, subunits that presumably provide additional functions. Major advances have been made in unravelling the structure function relationship of several T4SSs in G− bacteria in the last two decades (for recent review see, 10,72,73). The minimal T4SS of the *Agrobacterium tumefaciens* plasmid pTi is among the best-studied T4SSs of G− bacteria. The 12 T4SS proteins of pTi are encoded by genes *virB1* to *virB11* and *virD4,* and *virB*/*D* has become the unifying nomenclature for describing T4SSs (10,72,74). A schematic view of the 'minimal' T4SS of pTi and a speculative and probably incomplete model of the T4SS formed by pLS20 family plasmids is shown in Figure 6.

A 'minimal' T4SS system of G− plasmids can be divided into a central translocation channel that is coupled at the cytoplasmic side to an energy centre and at its extracellular side to a conjugative pilus. The cytoplasmic energy centre is composed of three ATPases (VirB4, VirB11 and VirD4). The VirD4 protein—aka the T4 coupling protein (T4CP)—recruits DNA and protein substrates to the channel. The central translocating channel is composed of two subassemblies, one of which spans the inner membrane (IM) and the other the periplasm and the outer membrane (OM). The IM complex (IMC) is composed of proteins VirB3, VirB6 and VirB8, and the OM complex (OMC)

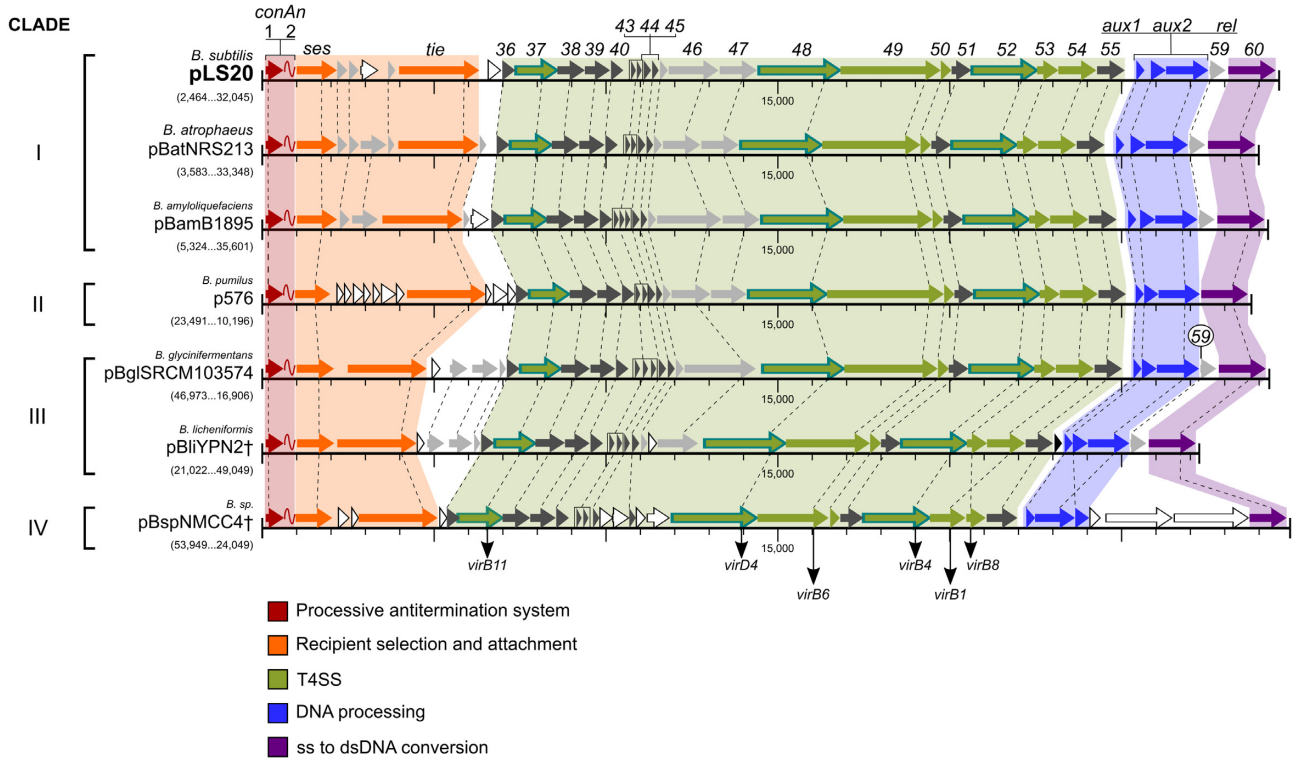


Figure 5. Comparative map of the conjugation operon present on the seven representative plasmids of the pLS20 family. Alignment of the conjugation operon of pLS20 with those present on six other plasmids of the pLS20 family. The names of the plasmids, the bacterial species harbouring them, and the region involved (in bp) are given on the left, as well as the (sub)clade to which they belong. The positions and sizes of genes are indicated with arrows. Dashed lines indicate corresponding positions in the different plasmids. Colour codes are used to indicate individual genes, and clusters of genes that (probably) are involved in a related step in the conjugation process. The background colour codes to indicate clusters of genes are: red, *conAn* P-AT system; orange, recipient cell selection and attachment; green, formation of translocation channel; blue, DNA processing; and purple, conversion of ss to dsDNA. For individual genes, two sets of colour codes are used. White and grey coloured arrows are used to indicate genes for which no function could be attributed. White genes correspond to genes that are present on one or several plasmids but not all the seven reference plasmids. Dark grey coloured genes reflect genes that are conserved on all seven reference plasmids. pLS20 genes *46* and 47, which are conserved on 6 and 4 of the reference plasmids, respectively, are shown in light grey. Genes for which a function has been established experimentally or could be attributed based on sequence similarity are indicated with the following colours: red, P-AT system; orange, genes involved in step 1 of the conjugation process; selection of and attachment to recipient cell. Green, genes involved in step 2 of the conjugation process; formation of the translocation channel (T4SS). Dark blue, genes involved in step 3 of the conjugation process; relaxosome proteins responsible for DNA processing. Purple, genes involved in step 4 of the conjugation process; conversion of transferred ssDNA into double-stranded plasmid DNA. The three genes encoding a putative ATPase are blue outlined. In the pBspNMCC4 plasmid, several genes share <30% identity with the corresponding pLS20 genes but are most likely orthologues based on other criteria (see text).

consists of lipoprotein VirB7 and VirB9. The VirB10 protein forms part of both the IMC and OMC and thereby plays a crucial role in connecting the inner and outer membrane complexes. The conjugative pilus is attached to the OMC, and is composed of multiple VirB2 shaft subunits and a VirB5 pilus-tip adhesion protein. All conjugative elements encode a protein that has cell wall hydrolytic activities. Most probably, these proteins—named VirB1 in the pTi system—facilitate insertion of the T4SS in the cell envelope by local degradation of the cell wall (75).

Compared to those of G− bacteria, little information is available for T4SSs in G+ bacteria, and structural data in particular, is limited. So far, structures are only available for several individual proteins encoded by plasmids pIP501 and pCW3 from *Streptococcus agalactiae* and *Clostridium perfringens*, respectively (for review see 13,74). It seems that the T4SSs are simpler in G+ bacteria than those in G− bacteria. In part, this is probably due to the differences in cell envelopes of G+ and G− bacteria. G+ bacteria lack an outer membrane; instead, a thick peptidoglycan layer containing cell wall surrounds the inner membrane. Consequently, T4SSs in G+ bacteria do not contain an OMC. The presence of a thick peptidoglycan layer is probably also the reason why inactivation of the cell wall hydrolase gene has more drastic effects on conjugation in G+ than in G− bacteria (76–79).

Several genes whose encoded proteins share significant similarity with proteins forming part of the T4SS encoded by other conjugative elements could be identified on the pLS20 family plasmids. These genes are all located in the central region of the conjugation operon. Thus, pLS20 genes *48* and *52* encode the putative ATPases VirD4 (T4CP) and VirB4, respectively. Interestingly, although it has been reported that the T4SSs of G+ bacteria studied so far lack the third ATPase (VirB11) (74), the pLS20 family of plasmids encode a homolog of the VirB11 protein (pLS20 gene *37*). The central region of the conjugation operon also encodes homologs of VirB6 (pLS20 gene *49*), VirB8 (pLS20 gene *53*) and VirB1 (pLS20 gene *54*). Although most of the other genes located in the central region of the con-

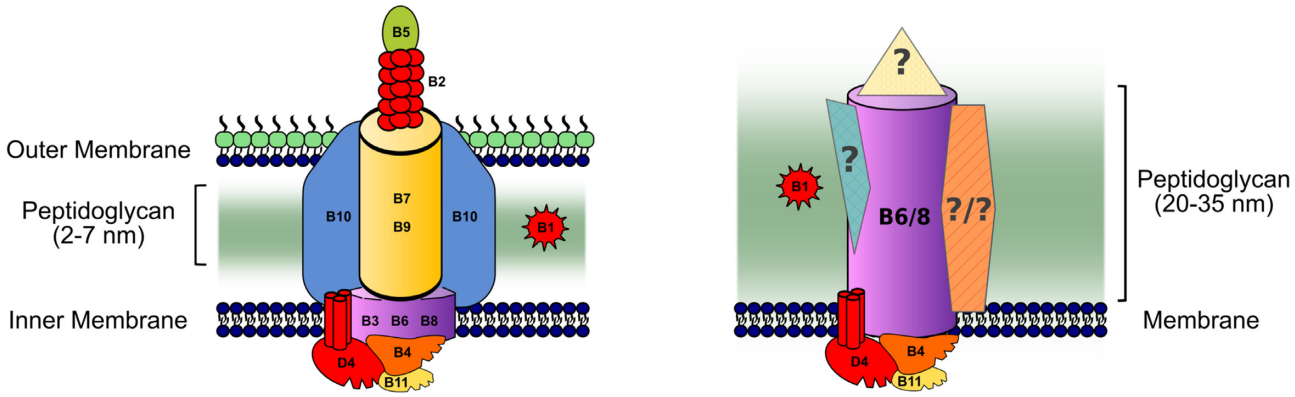


Figure 6. Speculative model of the T4SS encoded by pLS20 family plasmids. Cartoons shown on the left and right present the minimal T4SS encoded by plasmid pTi of the G− bacterium *A. tumefaciens*, and a speculative model of the T4SSs encoded by pLS20 family plasmids, respectively. The minimal T4SS encoded by pTi is formed by the VirB subunits VirB1-VirB11 and the VirD4 receptor (aka T4CP). The central translocation channel is composed of two subassemblies that cross the inner and outer membrane (IM and OM, respectively), with the associated VirB proteins indicated. The VirB10 protein forms part of both the IM and OM subassemblies. Three ATPases, VirD4, VirB4/11, are located at the entry side of the translocation channel. At the cell surface, a conjugative pilus is connected to the translocation channel. The VirB1 protein having cell wall degradation activity is also indicated. The figure is adapted from (72). The conjugation operons of pLS20 family plasmids do not encode homologs of the pilus components (VirB2, VirB5), the OM subassembly (VirB7, VirB9) nor VirB10. They do encode homologs of at least the IM subassembly proteins VirB6 and VirB8, the ATPases VirB4/11 and VirD4, and the cell wall hydrolysing protein VirB1. These VirB/D homologs have been used to generate a speculative model of the T4SS encoded by the pLS20 family plasmids based on the demonstrated function and structural context of these proteins in the minimal T4SS of pTi. Fourteen to fifteen other genes are clustered together with the six *virB/D* homologs in the central region of the conjugation operons of the pLS20 family of plasmids. Probably, several or all of the proteins encoded by these genes form part of the T4SS of pLS20 family plasmids. These proteins are indicated schematically and speculatively with the blue, orange and yellow multi angle boxes labelled with question marks. In addition, it has to be taken into account that the cell wall of G+ bacteria is much thicker than that of G− bacteria suggesting that there may be major differences between the T4SSs formed in G− and G+ bacteria.

jugation operon of pLS20 do not share similarity to *virB* genes they are conserved in all or the majority of the seven reference plasmids (see Figure 5). As the T4SS genes in pTi and on other conjugative elements are clustered together, it is plausible that the other conserved genes in the central region of the conjugation operons of pLS20 family plasmids encode proteins forming part of the T4SS. If correct, this would imply that the T4SS encoded by pLS20 family of plasmids is composed of about 20 proteins and hence would be of the 'expanded' type of T4SSs (72).

Step 3 of the conjugation process: DNA processing. The DNA of the conjugative element has to be processed in order to generate the single-stranded DNA (ssDNA) that is transferred into the recipient cell via the translocating channel. The DNA processing reaction initiates with the formation of a nucleoprotein complex, named relaxosome, at a particular region of ∼400 bp, called *oriT*. The crucial component of the relaxosome is a relaxase that recognizes and binds to specific sequences within the *oriT*. Often, the relaxosome involves other auxiliary proteins, which are generally encoded by the plasmid. The relaxase introduces a strand- and site-specific nick in the *oriT* region and the generated hydroxyl group at 3′-end of the nick site acts as a primer for DNA synthesis. After nicking, the relaxase remains covalently attached to the 5′-end of the nicked DNA strand, and VirD4 (T4CP) recruits this complex to the cytoplasmic site of the T4SS after which the relaxase pilots the transfer of the ssDNA into the recipient cell. In pLS20 the approximately 300 bp intergenic region between genes *55* and *56* corresponds to the *oriT* region, and genes *56* to *58* encode the relaxosome proteins. Proteins p56 and p57 are two auxiliary relaxosome proteins, named Aux1 and Aux2, respectively, and protein p58 is the relaxase, named Rel (also named Mob_L) (23,26). Both auxiliary proteins, as well as the relaxase, are the founding members of relaxosome proteins that are conserved within the Firmicutes phylum of G + bacteria (26). The relative position and organization of the DNA processing cassette is conserved for all the pLS20 family plasmids with one exception; in the case of clade IV plasmid pBspNMCC4† the second auxiliary gene and the relaxase gene have swapped positions.

Step 4 of the conjugation process: conversion of ssDNA into double-stranded plasmid DNA. Only one strand of the conjugative DNA is transferred into the recipient cell. Several conjugative elements encode a primase, which sometimes is fused to the relaxase, and that synthesize short RNAs on the transferred ssDNA that are used as primers for the conversion of ssDNA to dsDNA. Whereas the best-studied conjugative elements of G+ bacteria (pCF10, pIP501, pCW3) do not encode a primase. Gene *60* of pLS20 is predicted to contain a DnaG-like DNA primase and a topoisomerase-primase (Toprim) domain, and this gene is conserved on all seven reference plasmids of the pLS20 family of plasmids. Like the relaxase, primases are produced in the donor cell and are transported into the recipient cell through the translocation channel. The few proteins that are transported in this way possess translocation signals that are used for recruitment to the translocation channel. At least in some cases, a positively charged C-terminus functions as a translocation signal (72,80). The putative DNA primase of pLS20 has a positively charged C-terminus suggesting that this region might be important for the protein to be recruited and transferred to the recipient cell.

Clade/plasmid specific genes in the conjugation operons of pLS20 family plasmids. Three regions in the conjugation operons of the pLS20 family plasmids are variable. The functions of the genes in these regions are not known. The first variable region is located inside the region containing genes involved in selection of and attachment to the recipient cell; i.e. the exclusion and adhesin genes. In plasmids belonging to clade I, II and IV these genes are separated by two to seven genes whereas in the clade III plasmids there are none. The second variable region is located in between the clusters of genes involved in recipient cell selection and attachment, and the T4SS genes; i.e. between pLS20 genes *34* and *36*. This variable region ranges from one to four genes. The third variable region is located in between the genes for relaxosome and the primase, and ranges from zero to three genes. Possibly, these non-conserved genes provide the conjugative system with specific features or functions. Of all the pLS20 family plasmids, clade IV plasmid pBspNMCC4† is the most divergent plasmid. Most of the deduced protein sequences of pBspNMCC4† share a clearly lower level of similarity. This plasmid contains three unique genes in its presumed T4SS region.

CONCLUDING REMARKS

Here we show that pLS20 constitutes the prototype of a family of related plasmids that are harboured by different *Bacillus* species distributed at very different geographical locations around the world. Although the plasmids share a conserved structural organization, they could be divided into four clades. Besides a variable region located in between the partition and conjugation modules, all the plasmids contain three conserved modules: a replication, a partitioning and a conjugation module. In addition, plasmids belonging to clades I and IV contain a DNA repair module that is comprised by genes *uvrX* and *yolD*, while plasmids belonging to clade II and two of the three clade III plasmids only contain a *yolD* gene. The very large conjugation module can be divided into two subregions. Except for one, all the other plasmids contain a 5′-located region of three genes that are responsible for controlling the activation of the conjugation pathway. This region is followed by a very large operon encoding all the genes required for the different steps of the conjugation pathway. Remarkably, the sequential steps of the conjugation process are reflected by the order of genes within the conjugation operon. Thus, the first two genes of the conjugation operon constitute a bipartite P-AT system. These two genes are followed by genes involved in the first steps of the conjugation process: selection of and attachment to a recipient cell. The regions downstream are occupied respectively by genes encoding proteins involved in steps two (generation of the T4SS), three (DNA processing) and four (conversion of single-stranded to double-stranded plasmid DNA). Comparative analysis of representative plasmids of the different clades revealed that most of the genes located in the conjugation operons are conserved in all plasmids, indicating that these perform important functions. In addition, some plasmids contain a few or several unique genes that may contribute additional features. The results of the comparative analyses are useful for future studies, not only allowing conserved and unique genes between plasmids of the pLS20 family to be differentiated, and conserved residues or regions within homologous protein to be identified, our analyses also help better understanding of the relationship between the pLS20 family plasmids and other conjugative plasmids of G+ bacteria.

DATA AVAILABILITY

Sequences of plasmids pBatNRS213 and pBamB1895 were deposited in the NCBI database and given accession numbers OK210094 and OK210095, respectively.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We thank members of the different labs for useful discussions and Dr Jeff Errington for his support, Dr Ricardo Ramos of the 'Parque Científico de Madrid' (Madrid, Spain) for sequencing *B. amyloliquefaciens* strains B1895, and Daniel Zeigler of the Bacillus Genetic Stock Centre for providing plasmids and strains. The *B. atrophaeus* NRS213 strain was obtained from the Bacillus Genetic Stock Center (BGSC, stock 11A2), and *B. amyloliquefaciens* strain B1895 was kindly provided by Dr Vladimir A. Chistyakov and Dr Michael L. Chikindas.
Author contributions: J.V.C., D.A., A.M.A. and W.J.J.M. performed (*in silico*) experiments. W.J.J.M. and L.J.W. wrote the paper. J.V.C. and W.J.J.M. prepared Figures.

FUNDING

Ministry of Science and Innovation of the Spanish Government [bio2016-77883-C2-1-P, PID2019 108778GB C21 (AEI/FEDER, EU) to W.J.J.M.]; Wellcome Trust [209500 to Prof. Jeff Errington supported L.J.W.]; 'Fundación Ramón Areces'; Funding for open access charge: Ministry of Science and Innovation of the Spanish Government [PID2019 108778GB C21 (AEI/FEDER, EU) to W.J.J.M.]; CSIC.
Conflict of interest statement. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCES

1. Nolling,J., van Eeden,F.J., Eggen,R.I. and de Vos,W.M. (1992) Modular organization of related archaeal plasmids encoding different restriction-modification systems in *Methanobacterium thermoformicicum*. *Nucleic Acids Res.*, **20**, 6501–6507.
2. Reuther,J., Wohlleben,W. and Muth,G. (2006) Modular architecture of the conjugative plasmid pSVH1 from *Streptomyces venezuelae*. *Plasmid*, **55**, 201–209.
3. Kulinska,A., Czeredys,M., Hayes,F. and Jagura-Burdzy,G. (2008) Genomic and functional characterization of the modular broad-host-range RA3 plasmid, the archetype of the IncU group. *Appl. Environ. Microbiol.*, **74**, 4119–4132.
4. Dadgostar,P. (2019) Antimicrobial resistance: implications and costs. *Infect Drug Resist*, **12**, 3903–3910.
5. Partridge,S.R., Kwong,S.M., Firth,N. and Jensen,S.O. (2018) Mobile genetic elements associated with antimicrobial resistance. *Clin. Microbiol. Rev.*, **31**, e00088–17.

6. Virolle,C., Goldlust,K., Djermoun,S., Bigot,S. and Lesterlin,C. (2020) Plasmid transfer by conjugation in gram-negative bacteria: from the cellular to the community level. *Genes (Basel)*, **11**, 1239.
7. Cabezon,E., Ripoll-Rozada,J., Pena,A., De la Cruz,F. and Arechaga,I. (2014) Towards an integrated model of bacterial conjugation. *FEMS Microbiol. Rev.*, **39**, 81–95.
8. Guglielmini,J., Neron,B., Abby,S.S., Garcillan-Barcia,M.P., de la Cruz,F. and Rocha,E.P. (2014) Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion. *Nucleic Acids Res.*, **42**, 5715–5727.
9. Christie,P.J. (2016) The mosaic type IV secretion systems. *EcoSal Plus*, **7**, 1–22.
10. Waksman,G. (2019) From conjugation to T4S systems in Gram-negative bacteria: a mechanistic biology perspective. *EMBO Rep.*, **20**, e47012.
11. Arutyunov,D. and Frost,L.S. (2013) F conjugation: back to the beginning. *Plasmid*, **70**, 18–32.
12. Stingl,K. and Koraimann,G. (2017) Prokaryotic information games: how and when to take up and secrete DNA. *Curr. Top. Microbiol. Immunol.*, **413**, 61–92.
13. Goessweiner-Mohr,N., Arends,K., Keller,W. and Grohmann,E. (2014) Conjugation in gram-positive bacteria. *Microbiol Spectr*, **2**, PLAS-0004–2013.
14. Tanaka,T. and Koshikawa,T. (1977) Isolation and characterization of four types of plasmids from *Bacillussubtilis (natto)*. *J. Bacteriol*, **131**, 699–701.
15. Meijer,W.J.J., de Boer,A., van Tongeren,S., Venema,G. and Bron,S. (1995) Characterization of the replication region of the *Bacillussubtilis* plasmid pLS20: a novel type of replicon. *Nucleic Acids Res.*, **23**, 3214–3223.
16. Derman,A.I., Becker,E.C., Truong,B.D., Fujioka,A., Tucey,T.M., Erb,M.L., Patterson,P.C. and Pogliano,J. (2009) Phylogenetic analysis identifies many uncharacterized actin-like proteins (Alps) in bacteria: regulated polymerization, dynamic instability and treadmilling in Alp7A. *Mol. Microbiol*, **73**, 534–552.
17. Derman,A.I., Nonejuie,P., Michel,B.C., Truong,B.D., Fujioka,A., Erb,M.L. and Pogliano,J. (2012) Alp7R regulates expression of the actin-like protein Alp7A in *Bacillus subtilis*. *J. Bacteriol.*, **194**, 2715–2724.
18. Koehler,T.M. and Thorne,C.B. (1987) *Bacillus subtilis (natto)* plasmid pLS20 mediates interspecies plasmid transfer. *J. Bacteriol*, **169**, 5271–5278.
19. Meijer,W.J.J., Wisman,G.B.A., Terpstra,P., Thorsted,P.B., Thomas,C.M., Holsappel,S., Venema,G. and Bron,S. (1998) Rolling-circle plasmids from *bacillussubtilis*: complete nucleotide sequences and analyses of genes of pTA1015, pTA1040, pTA1050 and pTA1060, and comparisons with related plasmids from Gram-positive bacteria. *FEMS Microbiol. Rev*, **21**, 337–368.
20. Itaya,M., Sakaya,N., Matsunaga,S., Fujita,K. and Kaneko,S. (2006) Conjugational transfer kinetics of pLS20 between *Bacillus subtilis* in liquid medium. *Biosci. Biotechnol. Biochem*, **70**, 740–742.
21. Singh,P.K., Ramachandran,G., Ramos-Ruiz,R., Peiro-Pastor,R., Abia,D., Wu,L.J. and Meijer,W.J. (2013) Mobility of the native *Bacillus subtilis* conjugative plasmid pLS20 is regulated by intercellular signaling. *PLos Genet.*, **9**, e1003892.
22. Ramachandran,G., Singh,P.K., Luque-Ortega,J.R., Yuste,L., Alfonso,C., Rojo,F., Wu,L.J. and Meijer,W.J. (2014) A complex genetic switch involving overlapping divergent promoters and DNA looping regulates expression of conjugation genes of a Gram-positive plasmid. *PLos Genet.*, **10**, e1004733.
23. Ramachandran,G., Miguel-Arribas,A., Abia,D., Singh,P.K., Crespo,I., Gago-Cordoba,C., Hao,J.A., Luque-Ortega,J.R., Alfonso,C., Wu,L.J. *et al.* (2017) Discovery of a new family of relaxases in firmicutes bacteria. *PLos Genet.*, **13**, e1006586.
24. Crespo,I., Bernardo,N., Miguel-Arribas,A., Singh,P.K., Luque-Ortega,J.R., Alfonso,C., Malfois,M., Meijer,W.J.J. and Boer,D.R. (2020) Inactivation of the dimeric Rap_{pls20} anti-repressor of the conjugation operon is mediated by peptide-induced tetramerization. *Nucleic Acids Res.*, **48**, 8113–8127.
25. Singh,P.K., Serrano,E., Ramachandran,G., Miguel-Arribas,A., Gago-Cordoba,C., Val-Calvo,J., Lopez-Perez,A., Alfonso,C., Wu,L.J., Luque-Ortega,J.R. *et al.* (2020) Reversible regulation of conjugation of *Bacillus subtilis* plasmid pLS20 by the quorum sensing peptide responsive anti-repressor Rap_{pls20}. *Nucleic Acids Res.*, **48**, 10785–10801.
26. Miguel-Arribas,A., Hao,J.A., Luque-Ortega,J.R., Ramachandran,G., Val-Calvo,J., Gago-Cordoba,C., Gonzalez-Alvarez,D., Abia,D., Alfonso,C., Wu,L.J. *et al.* (2017) The *Bacillus subtilis* conjugative plasmid pLS20 encodes two ribbon-helix-helix type auxiliary relaxosome proteins that are essential for conjugation. *Front Microbiol.*, **8**, 2138.
27. Singh,P.K., Ramachandran,G., Duran-Alcalde,L., Alonso,C., Wu,L.J. and Meijer,W.J. (2012) Inhibition of *Bacillus subtilis* natural competence by a native, conjugative plasmid-encoded ComK repressor protein. *Environ. Microbiol.*, **14**, 2812–2825.
28. Gago-Cordoba,C., Val-Calvo,J., Miguel-Arribas,A., Serrano,E., Singh,P.K., Abia,D., Wu,L.J. and Meijer,W.J.J. (2019) Surface exclusion revisited: function related to differential expression of the surface exclusion system of *Bacillus subtilis* plasmid pLS20. *Front Microbiol.*, **10**, 1502.
29. Miguel-Arribas,A., Val-Calvo,J., Gago-Cordoba,C., Izquierdo,J.M., Abia,D., Wu,L.J., Errington,J. and Meijer,W.J.J. (2021) A novel bipartite antitermination system widespread in conjugative elements of Gram-positive bacteria. *Nucleic Acids Res.*, **49**, 5553–5567.
30. Singh,P.K., Ballestero-Beltran,S., Ramachandran,G. and Meijer,W.J. (2010) Complete nucleotide sequence and determination of the replication region of the sporulation inhibiting plasmid p576 from *Bacillus pumilus* NRS576. *Res. Microbiol*, **161**, 772–782.
31. Val-Calvo,J., Miguel-Arribas,A., Gago-Cordoba,C., Lopez-Perez,A., Ramachandran,G., Singh,P.K., Ramos-Ruiz,R. and Meijer,W.J.J. (2019) Draft genome sequences of sporulation-impaired *Bacillus pumilus* strain NRS576 and its native plasmid p576. *Microbiol. Resour. Announc.*, **8**, 1–2.
32. Lovett,P.S. (1973) Plasmid in *Bcilluspumilus* and the enhanced sporulation of plasmid-negative variants. *J. Bacteriol*, **115**, 291–298.
33. Tanaka,T., Kuroda,M. and Sakaguchi,K. (1977) Isolation and characterization of four plasmids from *Bacillussubtilis*. *J. Bacteriol*, **129**, 1487–1494.
34. Bertani,G. (1951) Studies on lysogenesis. I. The mode of phage liberation by lysogenic *Escherichia coli*. *J. Bacteriol.*, **62**, 293–300.
35. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10–12.
36. Joshi,N.A.F. and J.N. (2011) Sickle: a sliding window, adaptive, quality-based trimming tool for FastQ files. Computer program, https://github.com/najoshi/sickle.
37. Bankevich,A., Nurk,S., Antipov,D., Gurevich,A.A., Dvorkin,M., Kulikov,A.S., Lesin,V.M., Nikolenko,S.I., Pham,S., Prjibelski,A.D. *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, **19**, 455–477.
38. Prjibelski,A.D., Vasilinetc,I., Bankevich,A., Gurevich,A., Krivosheeva,T., Nurk,S., Pham,S., Korobeynikov,A., Lapidus,A. and Pevzner,P.A. (2014) ExSPAnder: a universal repeat resolver for DNA fragment assembly. *Bioinformatics*, **30**, i293–301.
39. Antipov,D., Hartwick,N., Shen,M., Raiko,M., Lapidus,A. and Pevzner,P.A. (2016) plasmidSPAdes: assembling plasmids from whole genome sequencing data. *Bioinformatics*, **32**, 3380–3387.
40. Gurevich,A., Saveliev,V., Vyahhi,N. and Tesler,G. (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, **29**, 1072–1075.
41. Aguirre de Carcer,D., Lopez-Bueno,A., Alonso-Lobo,J.M., Quesada,A. and Alcami,A. (2016) Metagenomic analysis of lacustrine viral diversity along a latitudinal transect of the antarctic peninsula. *FEMS Microbiol. Ecol.*, **92**, fiw074.
42. Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.
43. Madeira,F., Park,Y.M., Lee,J., Buso,N., Gur,T., Madhusoodanan,N., Basutkar,P., Tivey,A.R.N., Potter,S.C., Finn,R.D. *et al.* (2019) The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.*, **47**, W636–W641.
44. Kurtz,S., Phillippy,A., Delcher,A.L., Smoot,M., Shumway,M., Antonescu,C. and Salzberg,S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biol.*, **5**, R12.
45. Darling,A.E., Mau,B. and Perna,N.T. (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One*, **5**, e11147.

46. Contreras-Moreira,B. and Vinuesa,P. (2013) GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl. Environ. Microbiol.*, **79**, 7696–7701.
47. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinform.*, **10**, 421.
48. Nguyen,L.T., Schmidt,H.A., Haeseler,A. and Minh,B.Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, **32**, 268–274.
49. Kalyaanamoorthy,S., Minh,B.Q., Wong,T.K.F., Haeseler,A. and Jermiin,L.S. (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods*, **14**, 587–589.
50. Vinuesa,P., Ochoa-Sanchez,L.E. and Contreras-Moreira,B. (2018) GET_PHYLOMARKERS, a software package to select optimal orthologous clusters for phylogenomics and inferring pan-genome phylogenies, used for a critical geno-taxonomic revision of the genus stenotrophomonas. *Front. Microbiol.*, **9**, 771.
51. Balaban,M., Moshiri,N., Mai,U., Jia,X. and Mirarab,S. (2019) TreeCluster: clustering biological sequences using phylogenetic trees. *PLoS One*, **14**, e0221068.
52. Weaver,K.E., Kwong,S.M., Firth,N. and Francia,M.V. (2009) The RepA_N replicons of Gram-positive bacteria: a family of broadly distributed but narrow host range plasmids. *Plasmid*, **61**, 94–109.
53. Jensen,L.B., Garcia-Migura,L., Valenzuela,A.J., Lohr,M., Hasman,H. and Aarestrup,F.M. (2010) A classification system for plasmids from enterococci and other Gram-positive bacteria. *J. Microbiol. Methods*, **80**, 25–43.
54. Lanza,V.F., Tedim,A.P., Martinez,J.L., Baquero,F. and Coque,T.M. (2015) The plasmidome of firmicutes: impact on the emergence and the spread of resistance to antimicrobials. *Microbiol. Spectr.*, **3**, PLAS-0039–2014.
55. Kwong,S.M., Ramsay,J.P., Jensen,S.O. and Firth,N. (2017) Replication of staphylococcal resistance plasmids. *Front. Microbiol.*, **8**, 2279.
56. Meijer,W.J.J., Smith,M., Wake,R.G., de Boer,A.L., Venema,G. and Bron,S. (1996) Identification and characterization of a novel type of replication terminator with bidirectional activity on the *Bacillussubtilis* theta plasmid pLS20. *Mol. Microbiol.*, **19**, 1295–1306.
57. Zou,X., Caufield,P.W., Li,Y. and Qi,F. (2001) Complete nucleotide sequence and characterization of pUA140, a cryptic plasmid from *Streptococcus mutans*. *Plasmid*, **46**, 77–85.
58. Karlyshev,A.V., Melnikov,V.G. and Chistyakov,V.A. (2014) Draft genome sequence of *Bacillus amyloliquefaciens* B-1895. *Genome Announc.*, **2**, e00633–14.
59. Chistyakov,V., Melnikov,V., Chikindas,M.L., Khutsishvili,M., Chagelishvili,A., Bren,A., Kostina,N., Cavera,V. and Elisashvili,V. (2015) Poultry-beneficial solid-state *Bacillus amyloliquefaciens* B-1895 fermented soybean formulation. *Biosci. Microb. Food Health*, **34**, 25–28.
60. Nakamura,L.K. (1989) Taxonomic relationship of black-pigmented *B.subtilis* strains and a proposal for *Bacillusatrophaeus*. *Int. J. Syst. Bacteriol.*, **39**, 295–300.
61. Greenberg,D.L., Busch,J.D., Keim,P. and Wagner,D.M. (2010) Identifying experimental surrogates for *Bacillus anthracis* spores: a review. *Investig Genet*, **1**, 4.
62. Martinez,B., Suarez,J.E. and Rodriguez,A. (1996) Lactococcin 972 : a homodimeric lactococcal bacteriocin whose primary target is not the plasma membrane. *Microbiology (Reading)*, **142** (Pt 9), 2393–2398.
63. Campelo,A.B., Roces,C., Mohedano,M.L., Lopez,P., Rodriguez,A. and Martinez,B. (2014) A bacteriocin gene cluster able to enhance plasmid maintenance in *Lactococcus lactis*. *Microb. Cell Fact.*, **13**, 77.
64. Timinskas,K. and Venclovas,C. (2019) New insights into the structures and interactions of bacterial Y-family DNA polymerases. *Nucleic Acids Res.*, **47**, 4393–4405.
65. Kuenne,C., Voget,S., Pischimarov,J., Oehm,S., Goesmann,A., Daniel,R., Hain,T. and Chakraborty,T. (2010) Comparative analysis of plasmids in the genus listeria. *PLoS One*, **5**, e12511.
66. Anast,J.M. and Schmitz-Esser,S. (2020) The transcriptome of . *PLoS One*, **15**, e0233945.
67. Permina,E.A., Mironov,A.A. and Gelfand,M.S. (2002) Damage-repair error-prone polymerases of eubacteria: association with mobile genome elements. *Gene*, **293**, 133–140.
68. Meijer,W.J.J., Boer,D.R., Ares,S., Alfonso,C., Rojo,F., Luque-Ortega,J.R. and Wu,L.J. (2021) Multiple layered control of the conjugation process of the *Bacillus subtilis* plasmid pLS20. *Front. Mol. Biosci.*, **8**, 648468.
69. Pottathil,M. and Lazazzera,B.A. (2003) The extracellular Phr peptide-Rap phosphatase signaling circuit of *Bacillus subtilis*. *Front. Biosci.*, **8**, d32–d45.
70. Thompson,M.A., Onyeziri,M.C. and Fuqua,C. (2018) Function and regulation of *Agrobacterium tumefaciens* cell surface structures that promote attachment. *Curr. Top. Microbiol. Immunol.*, **418**, 143–184.
71. Gago-Cordoba,C., Val-Calvo,J., Abia,D., Diaz-Talavera,A., Miguel-Arribas,A., Aguilar Suarez,R., van Dijl,J.M., Wu,L.J. and Meijer,W.J.J. (2021) A conserved class II type thioester domain-containing adhesin is required for efficient conjugation in *Bacillussubtilis*. *mBio*, **12**, 1–15.
72. Costa,T.R.D., Harb,L., Khara,P., Zeng,L., Hu,B. and Christie,P.J. (2020) Type IV secretion systems: advances in structure, function, and activation. *Mol. Microbiol*, **115**, 436–452.
73. Bhatty,M., Laverde Gomez,J.A. and Christie,P.J. (2013) The expanding bacterial type IV secretion lexicon. *Res. Microbiol.*, **164**, 620–639.
74. Grohmann,E., Christie,P.J., Waksman,G. and Backert,S. (2018) Type IV secretion in Gram-negative and Gram-positive bacteria. *Mol. Microbiol.*, **107**, 455–471.
75. Zahrl,D., Wagner,M., Bischof,K., Bayer,M., Zavecz,B., Beranek,A., Ruckenstuhl,C., Zarfel,G.E. and Koraimann,G. (2005) Peptidoglycan degradation by specialized lytic transglycosylases associated with type III and type IV secretion systems. *Microbiology (Reading)*, **151**, 3455–3467.
76. Berger,B.R. and Christie,P.J. (1994) Genetic complementation analysis of the *Agrobacterium tumefaciens* virB operon: virB2 through virB11 are essential virulence genes. *J. Bacteriol.*, **176**, 3646–3660.
77. DeWitt,T. and Grossman,A.D. (2014) The bifunctional cell wall hydrolase CwlT is needed for conjugation of the integrative and conjugative element ICEBs1 in *Bacillussubtilis* and *B. anthracis*. *J. Bacteriol.*, **196**, 1588–1596.
78. Laverde Gomez,J.A., Bhatty,M. and Christie,P.J. (2014) PrgK, a multidomain peptidoglycan hydrolase, is essential for conjugative transfer of the pheromone-responsive plasmid pCF10. *J. Bacteriol.*, **196**, 527–539.
79. Arends,K., Celik,E.K., Probst,I., Goessweiner-Mohr,N., Fercher,C., Grumet,L., Soellue,C., Abajy,M.Y., Sakinc,T., Broszat,M. *et al.* (2013) TraG encoded by the pIP501 type IV secretion system is a two-domain peptidoglycan-degrading enzyme essential for conjugative transfer. *J. Bacteriol.*, **195**, 4436–4444.
80. Zechner,E.L., Lang,S. and Schildbach,J.F. (2012) Assembly and mechanisms of bacterial type IV secretion machines. *Philos. Trans. R. Soc. Lond B Biol. Sci*, **367**, 1073–1087.
81. Dong,H., Chang,J., He,X., Hou,Q. and Long,W. (2017) Complete genome sequence of *Bacillus subtilis* strain CGMCC 12426, an efficient Poly-gamma-Glutamate producer. *Genome Announc.*, **5**, e01163–17.
82. Wei,Y., Cao,J., Fang,J., Kato,C. and Cui,W. (2017) Complete genome sequence of *Bacillus subtilis* strain 29R7-12, a piezophilic bacterium isolated from coal-bearing sediment 2.4 kilometers below the seafloor. *Genome Announc.*, **5**, e01621–16.
83. Tan,S., Meng,Y., Su,A., Zhang,C. and Ren,Y. (2016) Draft genome sequence of *Bacillus subtilis* subsp. natto strain CGMCC 2108, a high producer of Poly-gamma-Glutamic acid. *Genome Announc.*, **4**, e00426–16.
84. Zhao,J., Liu,H., Liu,K., Li,H., Peng,Y., Liu,J., Han,X., Liu,X., Yao,L., Hou,Q. *et al.* (2019) Complete genome sequence of *Bacillus velezensis* DSYZ, a plant growth-promoting rhizobacterium with antifungal properties. *Microbiol. Resour. Announc.*, **8**, 1–13.
85. Xu,S., Yao,J., Wu,F., Mei,L. and Wang,Y. (2018) Evaluation of *Paenibacilluspolymyxa* carboxymethylcellulose/poly (vinyl alcohol) formulation for control of carya cathayen-sis canker caused by *Botryosphaeriadothidea*. *Forest Pathol.*, **48**, e12464.