Genome sequence of *B. amyloliquefaciens* type strain DSM7^T reveals differences to plant-associated *B. amyloliquefaciens* FZB42

Christian Rückert[a], Jochen Blom[a], XiaoHua Chen[b], Oleg Reva[c], Alfred Pühler[a], and Rainer Borriss[b]*

Center for Biotechnology (CeBiTec) Universität Bielefeld, D-33594 Bielefeld, Germany

[b] *Institut für Biologie, Humboldt Universität Berlin, Chausseestrasse 117, D10115 Berlin, Germany*

[c]*Bioinformatics and Computational Biology Unit, Biochemistry Department, University of Pretoria, Lynnwood Road, Hillcrest, 0002 Pretoria, South Africa*

e-mail addresses:

Christian.Rueckert@CeBiTec.Uni-Bielefeld.DE
jblom@cebitec.uni-bielefeld.de
xiaohuac@yahoo.com
oleg.reva@up.ac.za
Puehler@Genetik.Uni-Bielefeld.DE
rainer.borriss@rz.hu.berlin.de

*Corresponding author. Tel +49 30 2093 8137; fax +49 30 2093 8127. *E-mail address:*

rainer.borriss@rz.hu-berlin.de

A B S T R A C T

The complete genome sequence of *Bacillus amyloliquefaciens* type strain DSM7^T is presented. A comparative analysis between the genome sequences of the plant associated strain FZB42 (Chen *et al.,* 2007) with the genome of *B. amyloliquefaciens* DSM7^T revealed obvious differences in the variable part of the genomes, whilst the core genomes were found to be very similar. The strains FZB42 and DSM7^T have in common 3345 genes (CDS) in their core genomes; while 547 and 344 CDS were found to be unique in DSM7^T and FZB42, respectively. The core genome shared by both strains exhibited 97.89 % identity on amino acid level. The number of genes representing the core genome from the strains FZB42, DSM7^T, and *B. subtilis* DSM10^T was calculated as being 3098 and their identity was 92.25 %. The 3,980,199 bp genome of DSM7^T contains numerous genomic islands (GI) detected by different methods. Many of them were located in vicinity of tRNA, *glnA*, and *glmS* gene copies. In contrast to FZB42, but similar to *B. subtilis* DSM10^T, the GI were enriched in prophage sequences and often harbored transposases, integrases and recombinases. Compared to FZB42, *B. amyloliquefaciens* DSM7^T possessed a reduced potential to non-ribosomally synthesize secondary metabolites with antibacterial and/or antifungal action. *B. amyloliquefaciens* DSM7^T did not produce the polyketides difficidin and macrolactin and was impaired in its ability to produce lipopeptides other than surfactin. Differences established within the variable part of the genomes, justify our proposal to discriminate the plant-associated ecotype represented by FZB42 from the group of type strain related *B. amyloliquefaciens* soil bacteria.

Keywords: *Bacillus amyloliquefaciens,* Genome

1. Introduction

A group of Gram positive aerobic endospore-forming bacteria (AEFB), called in the vernacular the "*Bacillus subtilis*" group, is traditionally of outstanding importance in basic and applied microbiology (Fritze, 2004). All members of the group, which originally consisted of *B. subtilis*, *B. licheniformis*, and *B. pumilus*, are placed in 16S rRNA/DNA group 1. The first species added to this group was soil-borne *B. amyloliquefaciens*, originally described as potent producer of liquefying amylase and other extracellular enzymes of industrial importance (Fukumoto, 1943). By 1987, *B. amyloliquefaciens* became accepted as a species of its own (Priest *et al.,* 1987). Plant-associated *Bacillus* strains closely related to *B. amyloliquefaciens* are widely distributed (Idriss *et al.* 2002). Reva *et al.* (2004) reported that seven out of 17 *Bacillus* strains isolated from plants and soil, formed a cluster distinct from *B. amyloliquefaciens* type strain DSM7^T. These strains were generally better adapted to colonization of the rhizosphere than other members of the *Bacillus subtilis* group and were considered as a distinct ecotype of *B. amyloliquefaciens*.

The genome sequence of the plant-associated *Bacillus amyloliquefaciens* FZB42 (Chen *et al.,* 2007) revealed several features distinct to the genome of the closely related soil bacterium *B. subtilis* 168. *B. amyloliquefaciens* FZB42 dedicates about 340 kb, corresponding to 8.5% of its total genetic capacity, to synthesis of secondary metabolites, whilst in *B. subtilis* only 180 kb are devoted to non-ribosomal synthesis of secondary metabolites (Kunst *et al.* 1997). Here, we present the complete genome sequence of *B. amyloliquefaciens* type strain DSM7^T. Both strains are distinguished by their life-styles. *B. amyloliquefaciens* DSM7^T is a soil bacterium characterized by its enormous potential to produce extracellular enzymes of industrials importance including amylases and proteases (Priest, 1985), whilst plant-associated *B. amyloliquefaciens* FZB42 can promote plant growth and is able to suppress

competitive organisms within the plant rhizosphere. Our genome analysis revealed that DSM7^T possesses a comparable low capability to synthesize non-ribosomally lipopeptides and polyketides and to utilize specific plant derived macromolecules as cellulose and hemicelluloses indicated by absence of endo β-1,4 glucanase (EglS) and endo β-1,4 xylanase (XynA).

2. Materials and Methods

Genome sequencing, assembly and annotation. Genomic DNA prepared from DSM7^T was used for construction of a 2 kb long paired end library with a GS FLX library preparation kit in combination with GS FLX paired end adaptors (both Roche, Mannheim, Germany) according to the manufacturers protocol. Sequencing was performed on a Genome Sequencer FLX (Roche, Mannheim, Germany) delivering 851,700 reads with a total of 108,926,617 base pairs. The reads were assembled using the GS de novo Assembler resulting in 13 scaffolds containing 49 contigs (with 55 contigs larger than 500 bp in total) and indicating a genome size of approximately 3,913,433 bp. The scaffolds were oriented based on a comparison to the genome of *B. amyloliquefaciens* FZB42 and the remaining gaps were closed by long range PCR (using Phusion polymerase, New England Biolabs, Frankfurt(Main), Germany) and subsequent Sanger sequencing, resulting in 177 reads in total (IIT Biotech, Bielefeld, Germany). Long repeat structures (the ten copies of the *rrn* operon) were resolved by introducing fake reads based on the consensus sequence. Prediction of protein-encoding sequences was initially accomplished with REGANOR (Linke *et al.*, 2006). Automatic annotation was done using the annotation software GenDB 2.4 (Meyer *et al.,* 2003, https://www.cebitec.uni-bielefeld.de/groups/brf/software/gendb-2.2/cgi-bin/login.cgi). 1473

ORFs out of a total of 3923 were manually curated using comparative annotation with the genomes of FZB42 (Chen *et al.,* 2007) and *B. subtilis* 168 (Barbe *et al.,* 2009, http://subtiwiki.uni-goettingen.de/wiki/index.php/Main_Page).

The genome sequence of *B. amyloliquefaciens* was deposited as FN597644.

Comparative genome analysis was performed using the EDGAR software framework (Blom *et al.*, 2009, http://edgar.cebitec.uni-bielefeld.de/cgi-bin/edgar.cgi). A private project was constructed comprising of *B. amyloliquefaciens* DSM7^{T} and five other selected *Bacillus* strains. To construct a phylogenetic tree for this project, the deduced amino acid sequences of the core genes of the genomes were computed. In a following step multiple alignments of the core genes were generated using MUSCLE (Edgar, 2004), non matching parts of the alignment were masked by GBLOCKS and subsequently removed. The remaining parts of all alignments were concatenated to one large alignment. The PHYLIP package (Felsenstein, 1989) was used to create a phylogenetic tree of this alignment, represented in newick format.

Gene per gene comparison in two genomes of *B. amyloliquefaciens* was performed using BLASTp algorithm implementation in the blastall.exe NCBI executable file (Madden, 2002). It was assumed that the genes in two *B. amyloliquefaciens* genomes showing the best alignment score and e-value below 0.0001 were orthologs (Blom *et al*., 2009*)*. Pairs of orthologous genes then were codon aligned using MUSCLE (Edgar, 2004) with the gap open penalty discarded. The alignment length, nucleotide substitutions, number of indels, identities (numbers of exactly matched amino acids), and positives (numbers of matches of amino acids with positive alignment scores) were calculated for every alignment. The dynamics of accumulation of mutations in pairs of orthologous gene was analyzed by several statistical parameters. Frequencies of nucleotide substitutions including indels per 100 bp of sequence alignments were calculated for every pair of orthologous genes. Percentage of non-

synonymous mutations was calculated as a ratio of amino acid substitutions including indels to the number of nucleotide substitutions including indels. Sequence difference was calculated as *1−(identities / alignment_length)*. A non-specificity coefficient was calculated as a ratio *(positives − identities) / identities*. An in-house Python program was used to calculate the statistical parameters of alignments of orthologous genes and for data visualization.

The global alignment of whole chromosomal sequences was performed by M-GCAT software (Treangen and Messeguer, 2006). The Pathways Tools package (Karp, 2001) was used for prediction of metabolic pathways based on the genome annotation. Horizontally transferred genomic islands were predicted by the SeqWord Genome Browser tool (Ganesan *et al.*, 2008) and SeqWord Sniffer (Bezuidt *et al.*, 2009, both are available at www.bi.up.ac.za/SeqWord/), and using the IslandViewer online tools (www.pathogenomics.sfu.ca/islandviewer/query.php) that combines the prediction results of three algorithms of genomic island identification: IslanPick (Langille *et al.*, 2008), SIGI-HMM (Waack *et al.*, 2006) and IslandPath-DIMOB (Hsiao *et al.*, 2003).

3. Results and Discussion

3.1. Genome sequencing and assembly

Applying next generation sequencing (NGS) technologies has facilitated the rapid determination of whole genome sequences. To determine the genome sequences of *C. urealyticum* (Tauch et al., 2008a) and *C. kroppenstedtii* (Tauch et al., 2008b) it was sufficient to carry out NGS-based whole genome shotgun sequencing. In the meantime it turned out that genomes with even small numbers of repetitive elements can often only be completely sequenced by use of a large insert library for sorting the assembled contigs and for closing existing gaps. In the case of *B. amyloliquefaciens* DSM7^T an alternative method, namely a long paired-end run, was carried out. After sequencing and assembly, the distance of paired-end fragments was determined to be on average 2092 ± 523 bp. In total 851,700 reads were assembled including 254,445 paired reads, with a total of 108,926,617 base pairs. Utilization of the paired-end information allowed scaffolding of the 55 contigs larger than 500 bp into 13 scaffolds containing 49 contigs, indicating a genome size of approximately 3,913,433 bp with a 27.8 fold-coverage.

One of the scaffolds with a size of 3011 bp was found to be a plasmid contamination caused by incomplete removal of pUC19 DNA used as carrier DNA during nebulization. The remaining 12 scaffolds were oriented based on a comparison to the genome of *B. amyloliquefaciens* FZB42, revealing that 10 scaffolds were unique whilst the other two were present in 2 and 10 copies. The latter consists of two contigs corresponding to the 16S and 23S rDNA, respectively.

Gaps between scaffolds as well as the 34 gaps between scaffolded contigs were closed by long range PCR and subsequent Sanger sequencing, resulting in 177 reads in total. Analysis of the gaps revealed that most of the gaps (25 of 34) between scaffolded contigs were relatively small (3 ± 10 bp). They seem to be due to inverted repeats forming stable secondary structures which might result in no or poor amplification during the various PCR steps involved in library preparation and NGS sequencing (data not shown). Similar structures are also the cause of three larger gaps (172 bp, 200 bp, and 451 bp) whilst the remaining 6 intra-scaffold gaps are due to two small and four large repetitive elements. In contrast, the gaps between scaffolds were all caused by extremely large repetitive elements. The 10 rDNA copies were found to be responsible for 8 gaps whilst the remaining two contained a single copy of the second large repetitive element: This element has a size of 12,622 bp and is not present in *B. amyloliquefaciens* FZB42. These repeats could not be properly resolved by long-range PCR due to the formation of chimerical products. Therefore, the ten copies of the *rrn* operon were resolved by introducing simulated reads based on the consensus sequence.

For future projects, two main lessons can be learned from the current project. On one hand, it might be useful to use PCR additives, e.g. betaine, DMSO, and/or trehalose, that reduce the formation of secondary structures to avoid/reduce number of gaps caused by insufficient PCR amplification. This should reduce the number of PCR products and Sanger reads necessary to close intra-scaffold gaps. On the other hand, the size of the paired-end library should be chosen to be at least as large as the largest expected repetitive element (usually >6000 bp, the size of a typical ribosomal RNA operon).

3.2. The genome of B. amyloliquefaciens *DSM7^T*

The principal features of the *B. amyloliquefaciens* DSM7^T genome are summarized in Table 1

and Fig. 1. The size of the circular chromosome (3,980,199 bp) is in a similar range of the closely related *B. amyloliquefaciens* FZB42 (Chen *et al.,* 2007), *B. subtilis* DSM10^T (Barbe *et al.* 2009), *Bacillus licheniformis* DSM13^T (Veith *et al.* 2004, Rey *et al.* 2004), and *Bacillus pumilus* (Gioia *et al.* 2007). A comparative analysis between the genome sequences of the plant associated strain FZB42 with the genome of *B. amyloliquefaciens* DSM7^T revealed that the core genomes of these two strains are very similar. *B. amyloliquefaciens* FZB42 and DSM7^T have in common 3345 genes (CDS) in their core genomes with an average 97.89 % identity on the amino acid level; 547 coding sequences of DSM7^T and 344 CDS of FZB42 were singletons. . The number of genes representing the core genome of the strains FZB42, DSM7^T, and *B. subtilis* DSM10^T was calculated as being 3098 (±85) and their average identity was 92.25% (Fig. 2). The pan genome formed by the three strains consists of 5,284 CDS, whilst 476 genes were determined to be unique for *B. amyloliquefaciens* DSM7^T , i.e. they are not shared with *B. amyloliquefaciens* FZB42 and *B. subtilis* 168 (Fig. 2). A phylogram based on computing of the *Bacillus* core genomes (Blom *et al.,* 2009) suggested close relationship between both *B. amyloliquefaciens* genomes (Fig. 3).

These species share large blocks of almost identical sequences with other closely related organisms: *B. licheniformis* (see M-GCAT genome alignments in Supplementary Fig. 1) and *B. pumilus* (data not shown). Another common characteristic of all these organisms is that multiple copies of ribosomal RNA operons are found in their genomes. There are 7 *rrn* alleles in the genomes of *B. pumilus* and *B. licheniformis* and 10 alleles in *B. subtilis* and *B. amyloliquefaciens*. The alleles *rrnD*, *rrnF* and *rrnG* originated in the *B. subtilis* / *B. amyloliquefaciens* common ancestor after its division from the *B. pumilus* and *B. licheniformis* lineages. In the plant-associated strain *B. amyloliquefaciens* FZB42, the allele *rrnF* is truncated. Locations of other *rrn* operons on the chromosome remain quite conserved

in all these four organisms of the *B. subtilis* group.

3.3 Genomic plasticity and mobile elements

Besides the core genome, the DSM7^T genome contains sites of plasticity in which exchanges of DNA often occur. Genomic islands and large indels in DSM7^T and FZB42 genomes were identified by SeqWord Sniffer, IslandViewer and M-GCAT. Genome comparison revealed 27 large regions (longer than 5 kb) of genomic plasticity (GP; see Fig. 4 and Table 2 for details). Some of them are insertions adjacent to t-RNAs (GP 3, 9, 11, 12, 16), *glmS* (GP2), and *glnA* (GP14) containing transposases and integrases known as being typical for DNA-islands in Gram-negative and Gram-positive bacteria. For instance, three copies of DNA sequences similar to Tn1546 and its corresponding resolvase were found in GP14, GP15-17, GP21, and GP27. Sequences of phage related integrase/recombinase proteins were present in GP5, GP10, GP11, GP18, and GP20. FZB42 and *B. subtilis* genomes differ remarkably in their content of prophage or prophage remnants. The *B. subtilis* genome harbors 10 well documented regions of phage origin (Kunst *et al.* 1997) and four GI were identified in this study (Fig. 4), which were found mainly absent in FZB42 (Chen *et al.,* 2007). DSM7^T contained clearly more phage remnants than FZB42 (Fig. 4). Sequences, similar to the skin element (GP11, GP18), phage PBSX (GP3, GP12, GP20), SPBc2 (GP3, GP9-11, GP14, GP18, GP24), and several other prophages (GP23, GP27) were detected in the genome of *B. amyloliquefaciens* DSM7^T (Table 2 and Supplementary Table 1).

An extended analysis revealed that some of the genomic islands present in DSM7^T have counterparts in the genomes of other related *Bacillus* species. GP2 (489189-491604) is an old island present in *B. amyloliquefaciens* FZB42, *B. subtilis, B. licheniformis*, and *B. clausii*. Genome rearrangements, deletions, and insertions happened here quite often. Prophage

containing GP3 (573646-617807) corresponds to genomic islands in *B. amyloliquefaciens* FZB42, *B. subtilis,* and *B. licheniformis*. GP10 (917647-991120) is similar to a genomic island from *Streptococcus sanguinis* . Except its 5´region, GP14 (1876213-1919259) is shared with genomic islands in *B. amyloliquefaciens* FZB42, *B. subtilis*, and *B. licheniformis*.

Large regions of genome plasticity which are specific for the plant related *B. amyloliquefaciens* FZB42 contain several large operons of genes encoding none-ribosomal polipetide synthases, which are absent or rudimentary in the strain DSM7^T. Those include GP13 with the macrolactin synthase, GP14 with the bacillomycin synthase, GP15 with the fengycin synthase, GP19 with the difficidin synthase, and GP22 containing a 10235 bp long synthetase of an unknown secondary metabolite. Some of these genome plasticity regions are associated with phage related genes (GP14 and GP15) which implicates a possible involvement of phages in the distribution of polypeptide antibiotic encoding genes. Besides acquisition events in *B. amyloliquefaciens* FZB42, at least one gene loss event in progress was detected in the strain DSM7^T. In the chromosomal locus that corresponds to the fengycin synthase operon of *B. amyloliquefaciens* FZB42, two genes, *fenE* and *fend*, are located in the DSM7^T chromosome. The latter one is in fact a fusion of the *fenA* and *fenD* genes of the fengycin operon of *B. amyloliquefaciens* FZB42 while the genes *fenB* - *fenD* are missing in strain DSM7^T. Thus, the scarcity of polypeptide synthases in *B. amyloliquefaciens* DSM7^T may result equally likely from both: the acquisition of new polypeptide synthases by the strain FZB42 and deletion of these genes in the strain DSM7^T.

The comparison of metabolic pathways predicted for *B. amyloliquefaciens* DSM7^T and FZB42 by the Pathway Tools showed that horizontal exchange of genes did not affect the main metabolic pathways. Several differences in bacterial metabolism depicted by the Pathway Tools will be discussed in the next section. However, must be noted that the role of

horizontal gene exchange may be obscure as many of these genes are annotated as conserved hypotheticals. Among the functional genes that very likely have been transferred by lateral exchange there are genes for transcriptional regulators, ABC-transport proteins, UV-damage repair protein and several others (Table 2).

3.4 Protein secretion and extracellular enzymes

Environmental members of the *B. subtilis* group secrete different hydrolases, enabling them to use external cellulosic and hemicellulosic substrates present in plant cell walls. Two of these enzymes, endo-1,4-β-glucanase or carboxymethylcellulase (CMCase, cellulase, EC 3.2.1.4) and endo-1,4-β-xylanase (xylanase, 1,4-β-xylan xylanohydrolase, EC 3.2.1.8) are encoded in *B. subtilis* 168 by the genes *eglS* and *xynA*, respectively (Wolf *et al.* 1995). The same genes were also present in the genome of FZB42, but found absent in *B. amyloliquefaciens* DSM7^T. Similarly, the genes *xylA*, involved in xylose degradation (EC 5.3.1.5), *xynP*, encoding an oligosaccharide transporter, *xynB*, encoding 1,4-β-xylan xylosidase (EC 3.2.1.37), *xylR*, encoding the xylose operon repressor, and *bglC*, encoding an endo-1,4-□-glucanase (EC 3.2.1.4) are present in *B. subtilis* 168 and *B. amyloliquefaciens* FZB42 but missing in the *B. amyloliquefaciens* DSM7^T genome.

The members of the *B. amyloliquefaciens* DSM7^T related clade secrete a starch-liquefying alpha-amylase (*amyA*) with high industrial potential, whilst *B. subtilis* secretes a saccharifying enzyme (*amyE*). Sequences of both genes are not similar. Unlike DSM7^T, the genome of FZB42 does not contain the *amy*A sequence, corroborating an earlier finding of Reva *et al.* (2004), who were unable to amplify *amyA*-like sequences in plant-associated *B. amyloliquefaciens* strains. Instead, FZB42 possessed an *amyE*-like gene, resembling that of *B. subtilis* (Chen et al., 2007).

The known *B. amyloliquefaciens* genomes also lack several genes encoding enzymes of

the Entner-Doudoroff pathway. The Entner-Doudoroff pathway does not function in *B. subtilis* as well, due to the absence of a corresponding 6-P-gluconate dehydratase homologue (Zamboni *et al.*, 2004). Degradation of the Entner-Doudoroff pathway was found to have progressed further in both *B. amyloliquefaciens* strains with loss of the *ykgB* gene that encodes a 6-phosphogluconolactonase catalyzing the upstream conversion of D-glucono-δ-lactone-6-phosphate to 6-phospho-D-gluconate (EC 3.1.1.3.1). In the plant associated *B. amyloliquefaciens* strain FZB42, additionally *kdgA*, encoding 2-keto-3-deoxygluconate-6-phosphate aldolase (EC 4.1.2.14), was found to be absent. Interestingly, the *zwf* gene encoding glucose-6-phosphate 1-dehydrogenase, the first enzyme of the Entner-Doudoroff pathway that converts □-D-glucose-6-phosphate to D-glucono-δ-lactone-6-phosphate (EC 1.1.1.49), is quite conserved in all these organisms. It is possible that D-glucono-δ-lactone-6-phosphate is used in *Bacillus* in other metabolic pathways. In this case the loss of the enzymes of the Entner-Doudoroff pathway as possible substrate competitors has an evolutionary implication.

3.5 Restriction and modification

The strains *B. amyloliquefaciens* DSM7^T, FZB42, and *B. subtilis* DSM10^T evolved independent restriction modification (RM) systems without sequence similarity. Remarkably, type II restriction modification genes present in DSM7^T shared 98 % (methyl transferase) and 100 % (restrictase) identity, respectively, with the *Bam*HI system known for *B. amyloliquefaciens* H (Roberts *et al.,* 1977).

3.6. Positive selection of amino acid substitutions

The organism specification on the genomic level involves structural and compositional modifications of proteins adjusting their kinetic parameters and the substrate specificity to

new needs. The process of retention of beneficial mutations in a population is known as positive Darwinian selection. Loci or genes undergoing positive selection usually are determined by a higher frequency of non-synonymous versus synonymous nucleotide substitution known as Ka/Ks-ratio (Li, 1993). In this work, we analyzed the dynamics of accumulation of synonymous and non-synonymous mutations in orthologous genes of the *B. amyloliquefaciens* strains DSM7^{T} and FZB42 by calculating several statistical parameters: frequency of nucleotide substitutions per alignment, percentage of non-synonymous mutations, amino acid sequence difference, and non-specificity of mutation accumulation (i.e. *(positives − identities) / identities*). (all pairs of orthologous genes and their statistical parameters are listed in Supplementary Table 2.) Most genes in these two closely related organisms are conserved with the frequency of mutations around 2-3 nucleotides per 100 bp of the alignment. However, there are two groups of genes showing significantly higher level of mutation affecting both the nucleotide sequences of the genes and the amino acid sequences of the encoded proteins. These two groups may be distinguished by the non-specificity parameter (Fig. 5). It is assumed that in the case of a random accumulation of mutations in two orthologous genes, the number of identities will drop faster than the number of positives (Monzoorul *et al.*, 2009). In one group of genes, the rate of mutations is in the range from 30 to 50 nucleotide substitutions per 100 bp and the number of identities deviates from the number of positives exponentially with the sequence difference growing (Fig. 5; green pins elevated over the plot). These genes are most likely no longer used in one strain, therefore nonspecific mutations are accumulatingin one of the orthologs.

The other group of variable genes show an alternative pattern with an average rate of mutations from 10 to 20 substitutions per 100 bp and an almost precise equality between identities and positives meaning that non-conserved replacements of amino acids are

evolutionary favourable in these proteins (Fig. 5; yellow dots in the shaded area; Table 3). Based on a visual analysis of parameter distributions to separate an outlying group of orthologous genes, the threshold values of parameters were set for the percentage of non-synonymous mutations >= 33%, difference >= 0.32 and non-specificity <= 0.1.

Among the genes under positive selection (Table 3) there are 16 metabolic enzymes forming almost complete pathways of galacturonate, glutamine and several other compounds degradation; however, a possible involvement of these pathways in the adaptation to plant colonization remains obscure. γ-Glutamyltranspeptidase is known to be upregulated by plant exudate in plant-associated strains (Chen *et al.*, 2007). One gene is involved in antibiotic biosynthesis and transport. Five genes are transcriptional regulators and two of them are spore germination proteins trigging the spore germination by environmental signals. The haem-based aerotactic transducer is a chemotaxis protein sensing oxygen. Plant colonizing bacilli are under permanent oxidation stress (Reva *et al.*, 2004) and sensing of oxygen concentration and alternative environmental signals of spore germination probably is the factor of success for plant colonizers. Another stress response protein CsbD and 4 restriction/repair proteins are under positive evolutionary selection. Trans-membrane transport system of these bacteria is under the evolutionary pressure as well. Interesting, that three phage related skin elements showed the same pattern of accumulation of mutations that may indicate their involvement in bacterial metabolism and evolution.

3.7 Non-ribosomal synthesis of secondary metabolites

B. amyloliquefaciens strains are distinguished by their potential to non-ribosomally synthesize a huge spectrum of different secondary metabolites, many of them with antibacterial and/or antifungal action (Schneider *et al.* 2007). A survey of the genomes of four other *B. subtilis*

strains (http://www.bacillusgenomics.org/bsubtilis/) and of *B. amyloliquefaciens* DSM7^T corroborated that strains belonging to *B. subtilis sensu stricto* and *B. amyloliquefaciens* type strain did not produce the polyketides difficidin and macrolactin and are often impaired in their ability to produce lipopeptides other than surfactin.

Whilst plant-associated FZB42 dedicates about 340 kb, corresponding to 8.5% of the whole genome, to non-ribosomal synthesis of lipopeptides and polyketides, the genome of DSM7^T lacks several of the giant gene cluster involved in synthesis of antimicrobial compounds and siderophores (Chen *et al.,* 2007). No gene clusters for the polyketides macrolactin, difficidin, and the unknown *nrs* gene product were detected. Moreover, DSM7^T harbors only a rudimentary)gene set *fenDE* responsible for non-ribosomal synthesis of the antifungal lipopeptide fengycin. By contrast, DSM7^T contains a complete gene cluster for non-ribosomal synthesis of the lipopeptide iturin A.

Synthesis of the lipopeptides and polyketides mentioned above is dependent on Sfp, an enzyme that transfers 4´-phosphopantetheine from coenzyme A to the carrier proteins of nascent peptide or polyketide chains. An exception is the antibacterial dipeptide bacilysin (Chen *et al.,* 2009). The 6.3 kb gene cluster encoding bacilysin synthesis was found conserved in *B. subtilis* and both *B. amyloliquefaciens* genomes.

A comparative survey of nonribosomal polyketides, lipopeptides and siderophores (bacillibactin) in plant-associated (related to FZB42) and non-associated (related to DSM7^T) *B. amyloliquefaciens* strains corroborated our genomic findings. *B. amyloliquefaciens* DSM7^T (“F”), and related strains S23 (“N”), and ATCC15841 did not produce difficidin, macrolactin, and fengycin, whilst plant-associated strains including FZB42 did. Presence of bacillaene, surfactin, iturin A, and bacillibactin as the only products of non-ribosomal synthesis of polyketides and peptides was corroborated by MALDI-TOF mass spectrometry in DSM7^T

(Borriss *et al.,* IJSEM, manuscript under revision).

3.8 Ribosomal synthesis of small peptides

Ribosomal synthesis of small peptides with antimicrobial action (bacteriocins) is rather scarce in *B. amyloliquefaciens* FZB42 due to absence of genomic prophage insertions. Only operon fragments directing immunity against subtilin (*spaKREGF*) and mersacidin (*mrsKRFGE*) were detected, whilst their respective synthesis genes were missing (Chen *et al.,* 2007). DSM7^T contained also *spaKREGF*, but did not harbor remnants of the *mrs* gene cluster.

Recently, two gene clusters for ribosomal synthesis of antimicrobial peptides were identified in FZB42. One of them, designated *pzn* (plantazolicin), was found responsible for synthesis, processing, and modification of a highly modified peptide with a molecular mass of 1335 Da. The compound exerted only weak antagonistic action against closely related *Bacillus* strains, and might have some function in plant-bacteria interactions (Scholz *et al.* manuscript submitted). Another gene cluster in FZB42, *acn*, was shown to encompass synthesis of a class I cyclic bacteriocin with (m/z) 6362.5 and 6382.2, respectively, named amylocyclicin. Amylocyclicin was highly efficient against Gram-positive bacteria, especially against a *sigW* mutant of *B. subtilis* (Butcher and Helmann 2006, Scholz *et al.* manuscript in prep.). Genes corresponding to the *acn* genes, RBAM 02190 – 22040, were detected in DSM7^T, too. All of them were more than 90 % identical on the level of amino acids. DSM7^T did not harbor elements of the *pzn* gene cluster including the genes mediating immunity against *pzn*.

4. Conclusions

Due to deviations in DNA core genomes, changes in the variable portion of the genomes, in specific marker gene sequences (amylase, endo-1,4 β-glucanase, restriction modification), and

occurrence of gene clusters for non-ribosomal synthesis of secondary metabolites, we propose that *B. amyloliquefaciens* FZB42 and DSM7^T belong to taxonomically related but distinct units. This proposal has been justified by extended taxonomical studies, recently performed with a group of plant-associated and non-associated *B. amyloliquefaciens* strains (Borriss *et al.* IJSEM, under revision).

Acknowledgements

Technical assistance of Christiane Müller and Yvonne Kutter and financial support in frame of the competence network Genome Research on Bacteria (GenoMikPlus) and the Chinese-German collaboration program by the German Ministry for Education and Research is gratefully acknowledged. Oleg Reva acknowledges funding from the National Research Foundation of South Africa for computer program development.

References

Blom, J., Albaum, S., Doppmeier, D., Pühler, A., Vorhölter, F.-J., Zakrzewski, M., Goesmann, A. 2009. EDGAR: A software framework fort the comparative analysis of prokaryotic genomes. *BMC Bioinformatics* 10, 154.

Barbe, V., Cruveiller, S., Kunst, F., Lenoble, P., Meurice, G., Sekowska, A., Vallenet, D., Wang, T., Moszer, I., Medigue, C., Danchin, A. 2009. From a consortium sequence to a unified sequence: the *Bacillus subtilis* 168 reference genome a decade later. *Microbiology* 155, 1758-1775.

Bezuidt, O., Lima-Mendez, G., Reva, O. N. 2009. SeqWord Gene Island Sniffer: a program to study the lateral genetic exchange among bacteria. *World Academy of Science, Engineering and Technology*, 58, 1169-1174.

Butcher, B. G., Helmann, J. D. 2006. Identification of *Bacillus subtilis* sigma-dependent genes that provide intrinsic resistance to antimicrobial compounds produced by Bacilli. *Mol. Microbiol*. 60, 765-782.

Chen, X.H., Koumoutsi, A., Scholz, R., Eisenreich, A., Schneider, K., Heinemeyer, I., Morgenstern, B., Voss, B., Hess, W.R., Reva, O., Junge, H., Voigt, B., Jungblut, P.R., Vater, J., Süssmuth, R., Liesegang, H., Strittmatter, A., Gottschalk, G., Borriss, R. 2007. Comparative analysis of the complete genome sequence of the plant growth-promoting bacterium *Bacillus amyloliquefaciens* FZB42. *Nat. Biotechnol*. 25, 1007-1014.

Chen, X.H., Koumoutsi, A., Scholz, R., Borriss, R. 2009. More than anticipated – production of antibiotics and other secondary metabolites by *Bacillus amyloliquefaciens* FZB42. *J. Mol. Microbiol. Biotechnol*. 16, 14-24.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792-1797.

Felsenstein, J. 1989. PHYLIP: phylogeny interference package. *Cladistics* 5, 154-166.

Fukumoto, J. 1943. Studies on the production of bacterial amylase. I. Isolation of bacteria secreting potent amylases and their distribution (in Japanese). *J. Agr. Chem. Soc. Jpn.* 19, 487-503.

Fritze, D. 2004. Taxonomy of the genus *Bacillus* and related genera: the aerobic endospore-forming bacteria. *Phytopathology* 94, 1245-1248.

Ganesan, H., Rakitianskaia, A. S., Davenport, C. F., Tümmler, B., Reva, O. N. 2008. The SeqWord Genome Browser: an online tool for the identification and visualization of atypical regions of bacterial genomes through oligonucleotide usage. *BMC Bioinformatics*, 9, 333.

Gioia, J., Yerrapragada, S., Xiang, Q., Jiang, H., Igboeli, O. C., Muzny, D., Dugan-Rocha, S., Ding, Y., Hawes, A. *et al.* 2007. Paradoxical DNA repair and peroxide resistance gene conservation in *Bacillus pumilus* SAFR-032. *PloS ONE* 2, e928.

Hsiao, W., Wan, I., Jones, S. J., Brinkman, F. S. L. 2003. IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinformatics*, 19, 418-420.

Idriss, E. E., Makarewicz, O., Farouk, A., Rosner, K., Greiner, R., Bochow, H., Richter, T., Borriss, R. 2002. Extracellular phytase activity of *Bacillus amyloliquefaciens* FZB45 contributes to its plant-growth-promoting effect. *Microbiology* 148, 2097-2109.

Karp, P. D. 2001. Pathway Database: a case study in computational symbiotic theories. *Science*, 293, 2040-2044.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V., Bertero, M. G., Bessières, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S. C., Bron, S., Brouillet, S., Bruschi, C.V,, Caldwell, B., Capuano, V., Carter, N. M., Choi, S. K., Codani, J. J., Connerton, I. F., Danchin, A., *et al.* 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* 390, 249-256.

Langille, M. G. I., Hsiao, W. W. L., Brinkman, F. S. L. 2008. Evaluation of genomic island predictors using a comparative genomics approach. *BMC Bioinformatics*, 9, 329.

Li, W. H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol*., 36, 96-99.

Linke, B., McHardy, A.C., Neuweger, H., Krause, L., Meyer, F. 2006. REGANOR: a gene prediction server for prokaryotic genomes and a database of high quality gene predictions for prokaryotes. *Appl. Bioinformatics* 5, 193-198.

Madden, T. 2002. The BLAST sequence analysis tool. In McEntyre, J. and Ostell, J. Eds. The NCBI Handbook, National Library of Medicine, Bethesda (MD), chapter 16, pp. 1-18

Meyer, F., Goesmann, A. McHardy, A. C. Bartels, D. Bekel, T. Clausen, J. Kalinowski, J. Linke, B. Rupp, O. Giegerich R., Pühler A. 2003. GenDB - an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res*. 31, 2187-2195.

Monzoorul, H. M., Ghosh, T. S., Komanduri, D., Mande, S. S. 2009. Sort-ITEMS: sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics*, 25, 1722-1730.

Priest, F. G., Goodfellow, M., Shute, L. A. Berkeley, W. 1987. *Bacillus amyloliquefaciens* sp. nov., nom. rev. *Int. J. Syst. Bact.* 37, 69-71.

Priest F. G. 1985. Synthesis and secretion of extracellular enzymes by bacilli. *Microbiol. Sci*, 2, 278-282.

Reva, O. N., Dixelius, C., Meijer, J. Priest, F. G. 2004. Taxonomic characterization and plant colonizing abilities of some bacteria related to *Bacillus amyloliquefaciens* and *Bacillus subtilis*. *FEMS Microbiol. Ecol*. 48, 249-259.

Rey, M. W., Ramaiya, P., Nelson, B., Brody-Karpin, S. D., Zaretsky, E. J., Tang, M., Lopez de Leon, A., Xiang, H. *et al.* 2004. Complete genome sequence of the industrial bacterium *Bacillus licheniformis* and comparison with closely related *Bacillus* species. *Genome Biology* 5, R77.

Roberts, R. J., Wilson, G. A., Young, F. E. 1977. Recognition sequence of specific endonuclease *Bam*HI from *Bacillus amyloliquefaciens* H. *Nature* 265, 82-84.

Schneider, K., Chen, X.-H., Vater, J., Franke, P., Nicholson, G., Borriss, R., Süssmuth, R. 2007. Macrolactin is the polyketide biosynthesis product of the pks2 cluster of *B. amyloliquefaciens* FZB42. *J. Nat. Prod*. 70, 1417-1423.

Treangen, T. J., Messeguer, X. 2006. M-GCAT: interactively and efficiently constructing large-scale multiple genome comparison frameworks in closely related species. *BMC Bioinformatics*, 7, 433.

Veith, B., Herzberg, C., Steckel, S., Feesche, J., Maurer, K.H., Ehrenreich, P., Bäumer, S., Henne, A., Liesegang, H., Merkl, R., Ehrenreich, A., Gottschalk, G. 2004. The complete genome sequence of *Bacillus licheniformis* DSM13, an organism with great industrial potential. *J. Mol. Microbiol. Biotechnol*. 7, 204-211.

Waack, S., Keller, O., Asper, R., Brodag, T., Damm, C., Fricke, W. F., Surovcik, K.,

Meinicke, P., Merkl, R. 2006. Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics* 16, 142.

Wolf, M., Geczi, A., Simon, O., Borriss, R. 1995. Genes encoding xylan and β-glucan hydrolsing enzymes in *Bacillus subtilis*: characterization, mapping and construction of strains deficient in lichenase, cellulose and xylanase. *Microbiology* 141, 281-290.

Zamboni, N., Fisher, E., Laudert, D., Aymerich, S., Hohmann, H.-P., Sauer, U. 2004. The *Bacillus subtilis yqjI* gene encodes the $NADP^+$-dependent 6-P-gluconate dehydrogenase in the pentose phosphate pathway. *J. Bacteriol*. 186, 4528-4534.

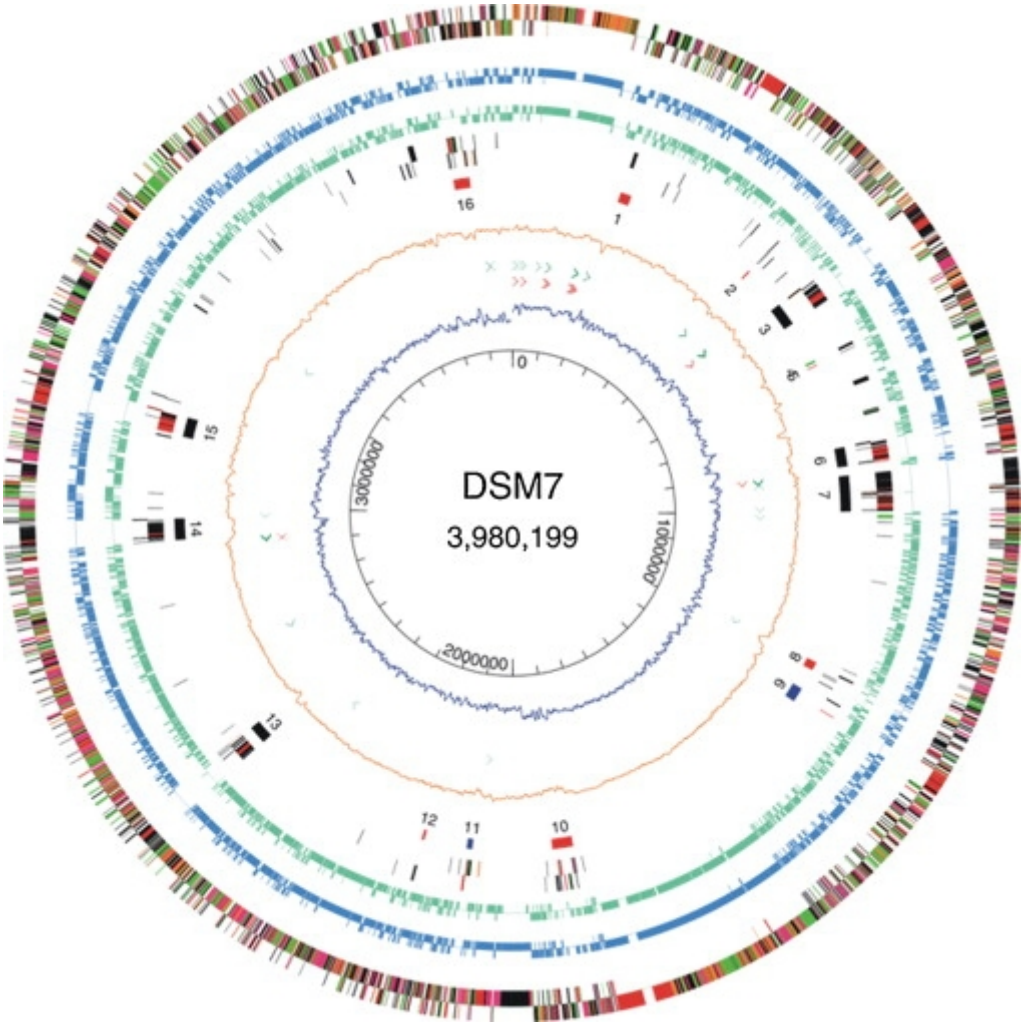


Fig. 1. Circular representation of the DSM7^T genome for several specific genomic features. The likely origin of replication was defined by similarities to the corresponding region in other *Bacillus* genomes. Outmost circle (1st): all genes in colour code according to their functions: cell envelope and cellular processes, green; information pathways, orange; intermediary metabolism, pink, other functions, red; unknown, black. 2nd circle: Genes having a presumed orthologue in *B. amyloliquefaciens*, blue; 3rd circle: Genes having a presumed orthologue in *B.subtilis*, green; 4th circle: Unique genes (singletons), only present in DSM7, in colour code according to their functions; 5th circle: Genomic islands of DSM7^T (1-16, see supplementary Table 1). DNA islands, unique in DSM7, black; DNA islands also present in B. subtilis, blue; DNA islands also present in FZB42, green; DNA islands also present in FZB42 and *B. subtilis;* red; 6th circle: GC deviation profile; 7th circle: rRNAs, green; 8th circle: tRNA, red; 9th circle: GC skew (G+C/G-C using a 1 kb sliding window); 10th circle: scale

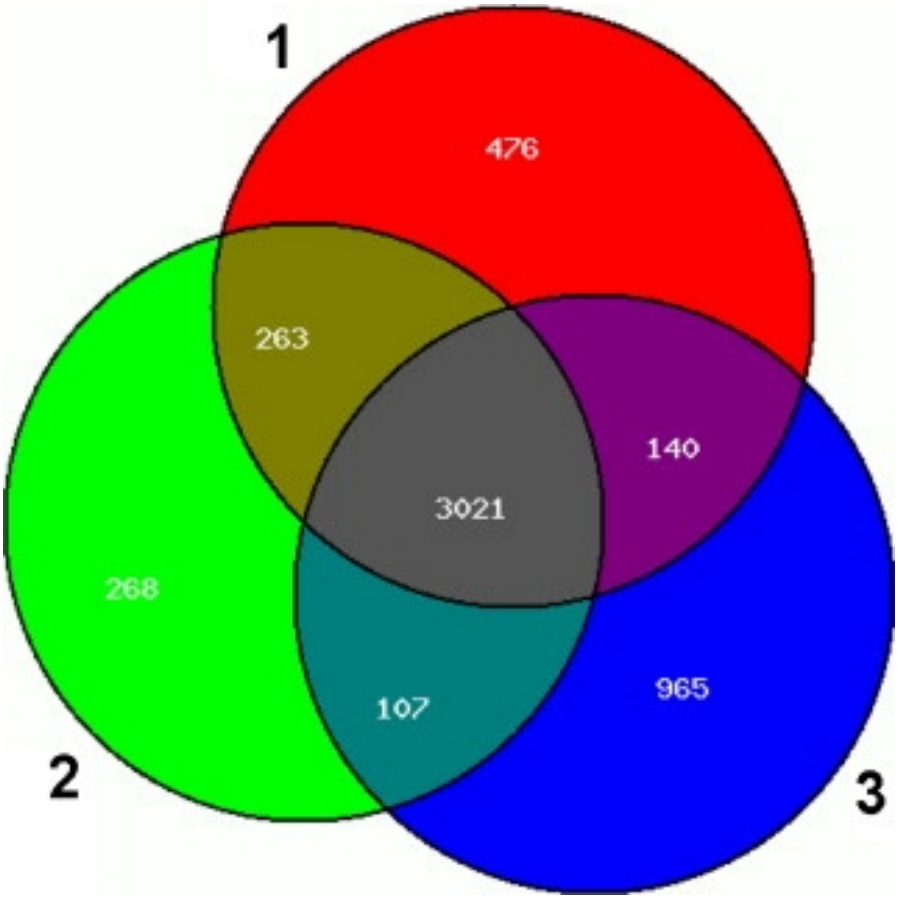


Fig. 2. Distribution of orthologous genes in the genomes of *Bacillus amyloliquefaciens* DSM7^T (1), FZB42 (2), and *B. subtilis* 168 (3). The Venn diagram was prepared by the EDGAR software (Blom et al., 2009). This analysis exploits all CDS of the genomes and is not restricted to the core genome.

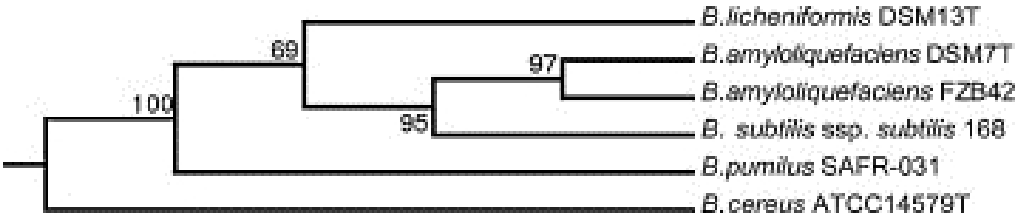


Fig. 3. Phylogenetic tree drawn from the core genomes of FZB42, DSM7^T, *B. subtilis* 168 (DSM10^T), *B. licheniformis* DSM13^T, *B. pumilus* SAFR-032 and *B. cereus* ATCC14579^T (see Section 2). Bootstrap values are indicated in % of repetitions.

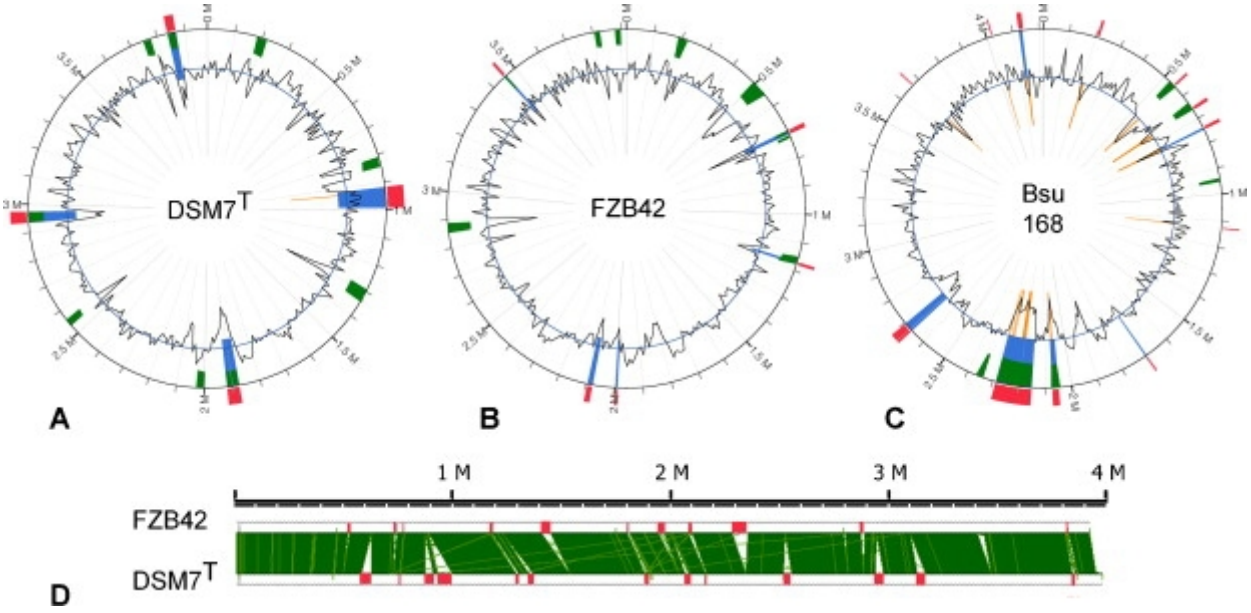


Fig. 4. Chromosomal locations of genome plasticity regions identified by different methods. On the chromosomal maps A, B and C constructed for *B. amyloliquefaciens* DSM7^T, FZB42 and *B. subtilis* 168 genomes correspondingly the locations of putative horizontally transferred genomic islands are depicted as predicted by IslanViewer (red spots), IslandPath (blue spots), SIGI-HMM (yellow spots) and SeqWord Sniffer (green spots). D graph shows a global alignment of bacterial chromosomes built by the M-GCAT program. Indels are depicted by red rectangles.

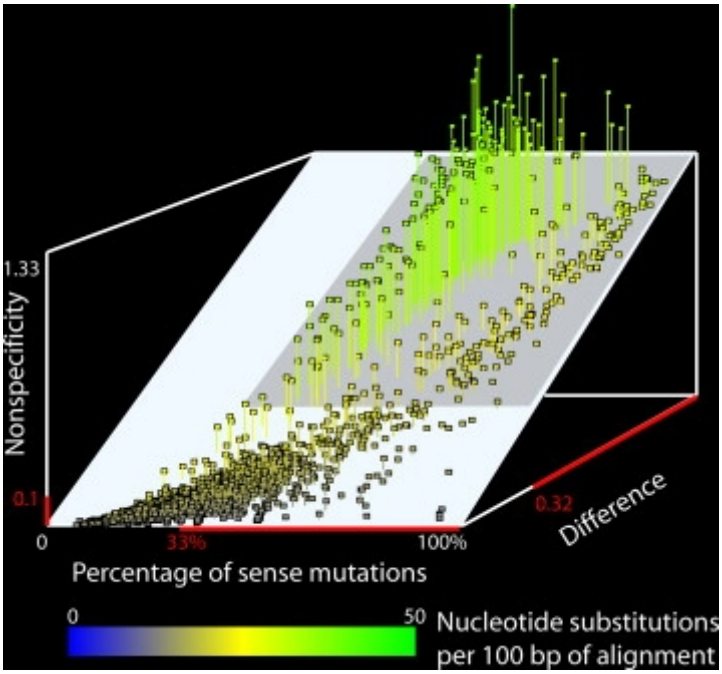


Fig. 5. Distribution of statistical parameters of accumulation of mutations in pairs of orthologues genes of *B. amyloliquefaciens* DSM7^{T} and FZB42. Each pair of orthologous genes is represented by a pin. The length of the pin corresponds to the non-specificity value calculated for this pair of genes. Frequency of nucleotide substitutions per 100 bp of alignment is depicted by the colour code. The area of distribution of genes undergoing evolutionary positive selection is shaded.Sequence difference was calculated as 1 − (identities/alignment_length). A non-specificity coefficient was calculated as a ratio (positives − identities)/identities.

Table 1. Genomic features of the *B. amyloliquefaciens* DSM7^T genome and comparison with genomes of other *Bacillus* spp. belonging to the B. subtilis group

Features	*B. amyloliquefaciens*		*B. subtilis*	*B. licheniformis*	*B. pumilus*
	DSM7^T	FZB42	168^T	DSM13^T	SAFR-013
Genome size (bp)	3,980,199	3,918,591	4,215,606	4,222,645	3,704,465
G+C content (mol%)	46.08	46.49	43.51	46.19	41.29
Protein-coding sequences (CDS)	3923	3660	4245	4196	3681
Number of CDS shared with DSM7	3923	3284	3161	2728	2503
Unique CDS (not shared with DSM7)	0	527	606 (375)[1]	986 (331)[2]	1221 (327)[3]
Number of CDS of the Pan-genome	3,923	4,302	5,284	6,447	7,298
Ribosomal RNA operons	30	38	30	21	21
Number of t-RNAs	94	101	86	72	69

[1] Number in brackets are singletons not shared with DSM7^T and FZB42
[2] Number in brackets are singletons not shared with DSM7^T, FZB42, and *B. subtilis*
[3] Number in brackets are singletons not shared with DSM7^T, FZB42, *B. subtilis,* and DSM13

Table 2. Genome plasticity regions found in *B. amyloliquefaciens* DSM7^T and FZB42 genomes.

GP	DSM 7	FZB42	Prediction Program	Annotation
1	198048..220858	198702..222440	SeqWord	GlmS and other metabolic proteins
2	489189..491604	495372..532439	SeqWord and M-GCAT	transcriptional regulators, acyl carrier protein
3	573646..617807	No	M-GCAT	PBSX, SPP1 and SPBc2 phage proteins, tRNA
4	No	697568..712055	Predicted by multiple methods	hypothetical proteins
5	No	701003..713073	M-GCAT	integrase/recombinase, phi-105 phage proteins
6	718747..720716	724191..740699	SeqWord and M-GCAT	ABC-transporters
7	749407..755745	765371..768886	M-GCAT	surface adhesine
8	807068..833743	No	SeqWord	ABC-transporter, protein kinase
9	869019..906335	No	M-GCAT	p27 and SPBc2 phage proteins, tRNA
10	917647..991120	No	Predicted by multiple methods	integrase/recombinase, SPBc2 phage proteins
11	1277390..1305801	1159500..1177276	SeqWord	skin elements, integrase/recombinase, SPBc2 phage proteins, tRNA
12	1343976..1364385	No	M-GCAT	complete PBSX prophage, tRNA
13	No	1402380..1445564	M-GCAT	macrolactin synthase
14	1876213..1919259	1794841..1803302	Predicted by multiple methods	bacillomycin synthase, resolvase, glnA, SPBc2 phage proteins
15	No	1939781..1967431	M-GCAT	fengycin synthase, resolvase
16	2058682..2077387	No	SeqWord	Tn1546 transposon, resolvase, tRNA
17	No	1990181..1996951	Predicted by multiple methods	hypothetical proteins
18	2156713..2159759	2078463..2093011	Predicted by multiple methods	UV-damage repair, integrase/recombinase and skin related proteins, SPBc2 phage proteins

19	No	2276734..2347685	M-GCAT	difficidin synthase
20	2515764..2551722	No	SeqWord	integrase/recombinase, resolvase, PBSX phage proteins, hypothetical proteins
21	No	2789121..2796777	M-GCAT	hypothetical proteins, resolvase
22	No	2868278..2887889	SeqWord	large (10235 bp) synthetase of unknown secondary metabolite
23	2936272..2974039	No	Predicted by multiple methods	p7 and HK97 phage proteins
24	3128111..3165599	No	M-GCAT	*xerC* integrase and other SPBc2 phage proteins
25	No	3451377..3460293	Predicted by multiple methods	hypothetical proteins
26	3764279..3785774	3809403..3835271	SeqWord and M-GCAT	phage proteins, Tn1546 transposase
27	3838361..3867587	3878265..3899599	Predicted by multiple methods	resolvase, hypothetical proteins

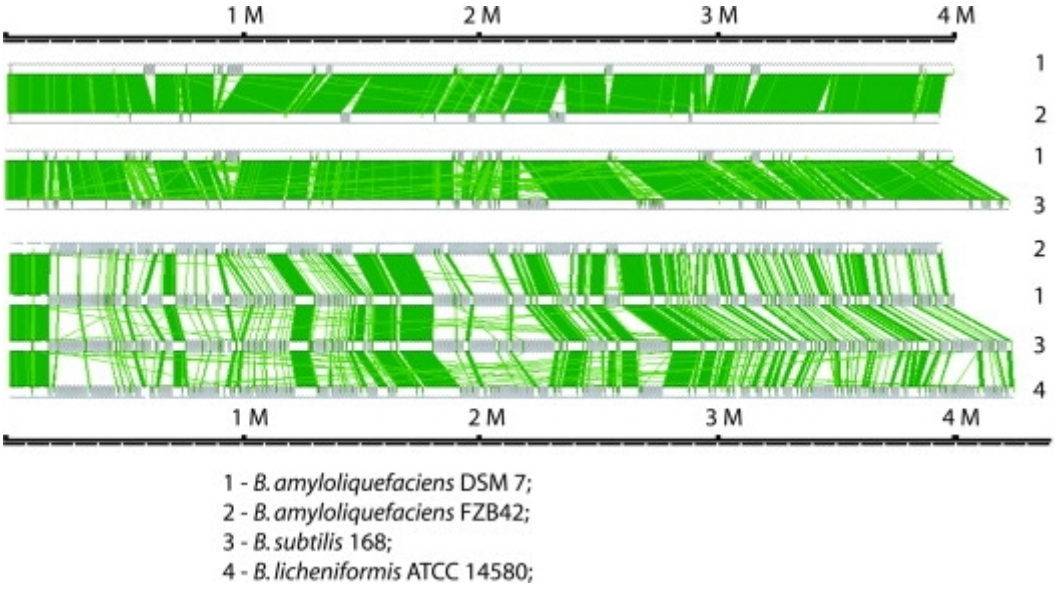
Table 3. Pairs of orthologous genes of DSM7^T and FZB42 genomes under positive evolutionary selection.

Pairs of *B. amyloliquefaciens* genes DSM7^T	FZB42	Function	Category
ydfJ; [541418-542177]	*ydfJ*; [569531-571685]	antibiotic transport associated protein	antibiotic synthesis
hemAT; [1151215-1152145]	*hemAT*; [1036806-1038099]	haem-based aerotactic transducer	chemotaxis
yheE; [1152081-1152501]	*ydhE*; [588055-589249]	macrolide glycosyltransferase	
cotV; [1285258-1285726]	*cotV*; [1152132-1152519]	spore coat protein	membrane
yplQ; [2169159-2169411]	*yplQ*; [2101629-2102268]	hemolysin III homolog	membrane
yrrR; [2584335-2584800]	*yrrR*; [2560287-2562042]	penicillin-binding protein	membrane
ybgJ; [249351-249873]	*ybgJ*; [253751-254735]	glutaminase, glutamine degradation pathway	metabolic
ycbJ; [253105-254026]	[257254-257461]	aminoglycoside phosphotransferase, glutamine degradation pathway	metabolic
lip; [268054-268213]	*lip*; [271862-272507]	triacylglycerol lipase, triacylglycerol degradation pathway	metabolic
yfmJ; [743143-743458]	*yfmJ*; [758636-759656]	NADH-dependent oxidoreductase	metabolic
yfjR; [846506-846941]	*yfjR*; [811041-811902]	oxidoreductase	metabolic
ctaO; [845994-846510]	*ctaO*; [838892-839840]	protoheme IX farnesyltransferase	metabolic
yjgC; [1316513-1318859]	*yjgC*; [1194669-1197660]	formate dehydrogenase alpha subunit, purine degradation pathway	metabolic
uxaA; [1330979-1331897]	*uxaA*; [1212796-1214290]	altronate hydrolase, D-galacturonate degradation pathway	metabolic
[1330812-1330947]	*uxaB*; [1211354-1212800]	tagaturonate reductase, D-galacturonate degradation pathway	metabolic
uxaC; [1330812-1330947]	*uxaC*; [1207396-1208794]	uronate isomerase, D-galacturonate degradation pathway	metabolic
yfnB; [1887548-1888799]	*yfnB*; [219937-220408]	2-haloacid dehalogenase, 1,2-dichlorethane	metabolic

		degradation	
ggt; [2079852-2079963]	*ggt*; [1975809-1977585]	gamma-glutamyltranspeptidase	metabolic
yfnB; [220144-220858]	*yfnB*; [219937-220408]	1,2-dichloroethane hydrolase	metabolic
ndk; [2253471-2253696]	*ndk*; [2187816-2188263]	nucleoside diphosphate kinase, pyrimidine catabolism	metabolic
yqjM; [2343482-2344073]	*yqjM*; [2349352-2350369]	NADH-dependent flavin oxidoreductase	metabolic
ybjI; [2557997-2558396]	[2533523-2534336]	HAD hydrolase	metabolic
rbsK; [3470371-3470635]	*rbsK*; [3433169-3434051]	ribokinase, ribose degradation pathway	metabolic
lmrA; [268054-268213]	*lmrA*; [269806-270373]	transcription repressor	regulation
ykoM; [1501633-1501786]	*ykoM*; [1277908-1278373]	transcriptional regulator	regulation
ccpB; [3928707-3929253]	*ccpB*; [3898630-3899569]	catabolic control protein	regulation
gerKA1; [3974623-3976093]	*gerKA1*; [3429146-3429890]	spore germination protein	regulation
gerKC1; [3977591-3978296]	*gerKC1*; [3430960-3432094]	spore germination protein	regulation
yrvE; [249351-249873]	*yrvE*; [2595066-2596452]	single-strand DNA-specific exonuclease	restriction/reparation
uvrX; [1887548-1888799]	*uvrX*; [2082876-2083506]	UV-damage repair protein	restriction/reparation
yrvE; [2618134-2620492]	*yrvE*; [2595066-2596452]	single-strand DNA-specific exonuclease	restriction/reparation
rnr; [3271923-3273225]	*rnr*; [3205126-3207460]	ribonuclease R	restriction/reparation
[3533471-3533750]	*csbD*; [3498834-3499023]	stress response protein	stress response
ycbE; [309192-309915]	*ycbE*; [312976-314287]	sugar transport permease	transport
treP; [784393-784633]	*treP*; [785158-786574]	phosphotransferase system (PTS) trehalose-specific enzyme IIBC	transport
yjbQ; [1262462-1263077]	*yjbQ*; [1140999-1142844]	Na+/H+ antiporter	transport
yjkB; [1325079-1325346]	*yjkB*; [1202814-1203564]	amino acid ABC transporter	transport
ykkC; [1394345-1394837]	*ykkC*; [1253673-1254012]	multidrug resistance protein	transport
yvrC; [3215053-	*yvrC*; [3145355-	iron-binding protein	transport

3215683]	3146300]		
nasA; [323905-324187]	*nasA*; [332804-334010]	nitrate transporter	transport
yvbW; [3318470-3319811]	*yvbW*; [3252880-3253672]	amino acid permease	transport

Supplementary materials:



Supplementary Figure 1. Global alignments of chromosomes calculated by the M-GCAT program.

Supplementary Table 1. Genes, unique in the genome of DSM7T (not present in the genome of FZB42). GI means Genomic island. Orthologous genes of B.

Supplementary Table 2. Statistical parameters of accumulation of mutations in pairs of orthologous genes of *B. amyloliquefaciens* DSM7^{T} and FZB42