



Phylogenetic Distribution of Secondary Metabolites in the *Bacillus subtilis* Species Complex

 Kat Steinke,^{a,b}  Omkar S. Mohite,^b  Tilmann Weber,^b  Ákos T. Kovács^a

^aBacterial Interactions and Evolution Group, DTU Bioengineering, Technical University of Denmark, Kongens Lyngby, Denmark

^bThe Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark

Kat Steinke and Omkar S. Mohite contributed equally to this work. Author order was determined based on contributions to the initial creation of the data set.

ABSTRACT Microbes produce a plethora of secondary (or specialized) metabolites that, although not essential for primary metabolism, benefit them to survive in the environment, communicate, and influence cell differentiation. Biosynthetic gene clusters (BGCs), responsible for the production of these secondary metabolites, are readily identifiable on bacterial genome sequences. Understanding the phylogeny and distribution of BGCs helps us to predict the natural product synthesis ability of new isolates. Here, we examined 310 genomes from the *Bacillus subtilis* group, determined the inter- and intraspecies patterns of absence/presence for all BGCs, and assigned them to defined gene cluster families (GCFs). This allowed us to establish patterns in the distribution of both known and unknown products. Further, we analyzed variations in the BGC structures of particular families encoding natural products, such as plipastatin, fengycin, iturin, mycosubtilin, and bacillomycin. Our detailed analysis revealed multiple GCFs that are species or clade specific and a few others that are scattered within or between species, which will guide exploration of the chemodiversity within the *B. subtilis* group. Surprisingly, we discovered that partial deletion of BGCs and frameshift mutations in selected biosynthetic genes are conserved within phylogenetically related isolates, although isolated from around the globe. Our results highlight the importance of detailed genomic analysis of BGCs and the remarkable phylogenetically conserved erosion of secondary metabolite biosynthetic potential in the *B. subtilis* group.

IMPORTANCE Members of the *B. subtilis* species complex are commonly recognized producers of secondary metabolites, among those, the production of antifungals, which makes them promising biocontrol strains. While there are studies examining the distribution of well-known secondary metabolites in *Bacilli*, intraspecies clade-specific distribution has not been systematically reported for the *B. subtilis* group. Here, we report the complete biosynthetic potential within the *B. subtilis* group to explore the distribution of the biosynthetic gene clusters and to reveal an exhaustive phylogenetic conservation of secondary metabolite production within *Bacillus* that supports the chemodiversity within this species complex. We identify that certain gene clusters acquired deletions of genes and particular frameshift mutations, rendering them inactive for secondary metabolite biosynthesis, a conserved genetic trait within phylogenetically conserved clades of certain species. The overview guides the assignment of the secondary metabolite production potential of newly isolated *Bacillus* strains based on genome sequence and phylogenetic relatedness.

KEYWORDS *Bacillus*, biosynthetic gene clusters, fengycin, iturin, phylogeny, plipastatin, secondary metabolite

B *acilli* can be isolated from various environments, such as the plant rhizosphere and the animal and human digestive systems, where secondary (or specialized)

Citation Steinke K, Mohite OS, Weber T, Kovács ÁT. 2021. Phylogenetic distribution of secondary metabolites in the *Bacillus subtilis* species complex. *mSystems* 6:e00057-21. <https://doi.org/10.1128/mSystems.00057-21>.


Editor Marnix Medema, Wageningen University

Ad Hoc Peer Reviewer Gajender Aleti, University of California San Diego

The review history of this article can be read [here](#).

Copyright © 2021 Steinke et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](#).

Address correspondence to Tilmann Weber, tive@biosustain.dtu.dk, or Ákos T. Kovács, atkovacs@dtu.dk.

 The biosynthetic gene clusters for secondary (or specialised) metabolite production in the *Bacillus subtilis* species complex show remarkable phylogenetic distribution, including clade-specific partial gene cluster deletions and frame shift mutations

Received 16 January 2021

Accepted 19 February 2021

Published 9 March 2021

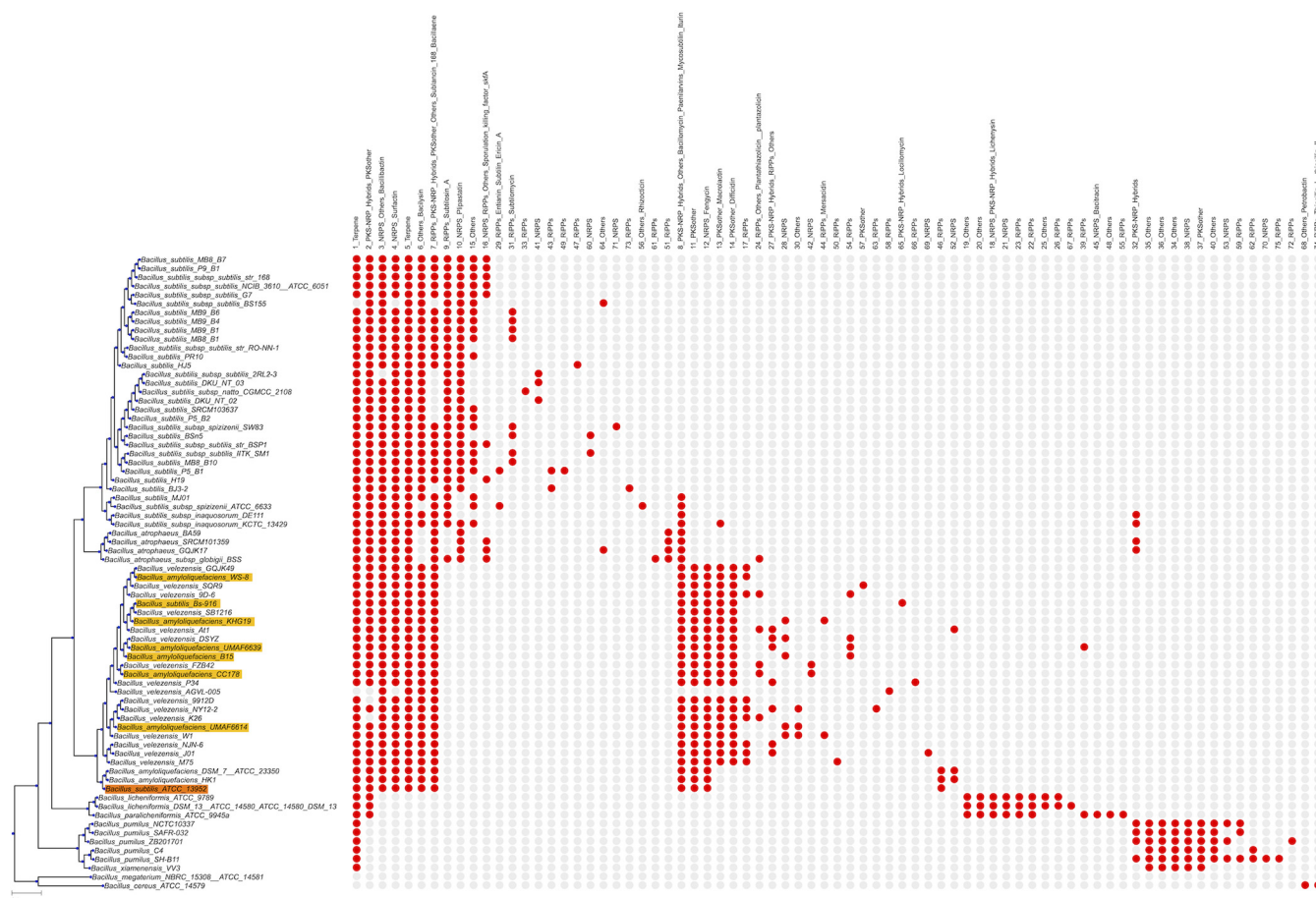


FIG 1 Phylogenetic tree reconstructed based on a multilocus sequence alignment of 30 genes with a modified version of autoMLST, using IQ-TREE and ultrafast bootstrapping with 1,000 replicates. *B. cereus* ATCC 14579 and *B. megaterium* NBRC 15308 were used as an outgroup. The presence-absence matrix of GCFs is visualized with a red dot indicating presence and a gray dot indicating absence. Figure S1 includes the complete tree. Strains with disagreements in NCBI and GTDB taxonomy are highlighted.

metabolites (SMs) play a pivotal role. The *Bacillus subtilis* group, which includes *B. subtilis* and its closely related species (Fig. 1), comprises common producers of bioactive SMs, such as antimicrobials and cytotoxic substances, empowering them for a range of industrial applications, including plant pathogen biocontrol (1, 2). Members of the *B. subtilis* group are producers of numerous well-known natural products, such as iturin, mycosubtilin, fengycin (FEN)/plipastatin (PPS), or bacillaene. Previous studies have globally reported the presence of known and novel biosynthetic gene clusters (BGCs) in *Bacillus* and related genera, highlighting the diverse potential of SM production in these bacteria (3–5). Additional reviews provide an overview of various SMs produced by these *Bacilli* (1, 6). Only recently, the species-level distribution of the corresponding BGCs in numerous coisolates from the *B. subtilis* group has been experimentally investigated (7).

Here, we specifically expand previous studies by investigating patterns in all complete genomes of *B. subtilis* group as of July 2019 to dissect inter- and intraspecies diversity. Therefore, we examined the phylogenetic distribution of BGC families across 310 *B. subtilis* group genomes (see Data Set S1 in the supplemental material) by predicting BGCs with a modified version of antiSMASH v5.0 (8), clustering these into gene cluster families (GCFs) with BiG-SCAPE (9), and visualizing GCF distributions across a phylogenetic tree generated with autoMLST-derived scripts (10) (Fig. S1 and S2).

The phylogenetic tree was reconstructed based on a multilocus sequence alignment of 30 conserved single-copy genes (Fig. 1; Fig. S1), generally reflecting NCBI

taxonomy, but with certain disagreements in the *B. velezensis* and *B. amyloliquefaciens* clades (highlighted in Fig. 1 and below).

The 3,643 BGCs identified using antiSMASH v5.0 (8) were assigned into 75 GCFs and 62 singletons with BiG-SCAPE (9) (Fig. S2); GCFs were subsequently mapped to the tree (Fig. 1; Fig. S1). Only one predicted GCF, coding a terpene (sesquiterpene), was found in nearly all strains, while another, a predicted nonribosomal peptide synthetase (NRPS)/polyketide synthase (PKS) hybrid, was found in most species except *B. pumilus* and *B. xiamenensis*. Other widespread GCFs are bacillibactin, surfactin, and bacilysin. Bacillaene and sublancin 168 families were found in most species, except *B. licheniformis*, *B. paralicheniformis*, *B. pumilus*, and *B. xiamenensis*; however, there are two gaps seemingly following clade boundaries in *B. subtilis*. A similar gap in distribution occurs in bacilysin, which is absent in *B. spizizenii* and *B. atrophaeus*. No correlation was identified between the determined BGC number of each strain and the source of isolation (e. g., rhizosphere, soil, food, or environment) (Data Set S1).

Such apparently clade-linked patterns in the absence or presence of GCFs were common, and many GCFs were distributed according to phylogeny. This occurred in both individual species and clades spanning multiple species. For example, a clade-specific GCF, lichenysin, was identified only in *B. licheniformis* and *B. paralicheniformis*. The distribution of the highly similar lipopeptides fengycin and plipastatin also followed clade boundaries, with fengycin in *B. velezensis* and *B. amyloliquefaciens*, whereas plipastatin was found in *B. subtilis* and *B. atrophaeus*. As previously reported (11), GCFs for rhizocticin but not plipastatin were found in *B. spizizenii*, supporting the biosynthetic distinctness of this clade.

Other examples of clusters almost or entirely limited to one species in the tree included bacitracin, which was present in all examined *B. paralicheniformis* genomes, as well as diffridin and macrolactin, both found in most *B. velezensis* strains (though macrolactin was also present in single isolates of other species). Certain species-specific GCFs were found dispersedly; for instance, the *B. subtilis*-specific subtilomycin was apparently linked to particular clades within the species or the ribosomally synthesized and posttranslationally modified peptide-coding 33_RiPP and 49_RiPP families, which were also species specific. Additionally, some families appeared in multiple clades but in a clade-linked pattern (17 ribosomally synthesized and posttranslationally modified peptides [RiPPs] in *B. velezensis*), while others were missing in one or more clades (15_Others in *B. subtilis*).

Finally, other GCFs appeared more scattered within a species, with no evident link to an individual clade, such as the 42_NRPS GCF in *B. velezensis* (Fig. S1). Only a few GCFs (e.g., 39_RiPPs), appeared scattered across the entire tree without a noticeable link to particular clades. Horizontal gene transfer (HGT), in accordance with the natural competence of *B. subtilis* (6), might drive the scattered patterns and random occurrences of GCFs outside key species. For instance, the 42_NRPS family contains the *nrs* cluster of *B. velezensis* FZB42, previously suggested to be acquired via HGT (12).

Next, we compared the genetic variations within particular GCFs and investigated the phylogenetic relationship among these variants, selecting families that code for important *Bacillus* SMs, namely fengycin, plipastatin, iturin, bacillomycin, and mycosubtilin.

A total of 127 BGCs were part of the similarity network with BGCs for plipastatin, a biodegradable fungicide (1). Based on the similarity network, these BGCs were placed into 6 groups: PPS, PPS groups B to E, and PPS_others (Fig. 2; Data Set S2). Plipastatins are mostly observed in the *B. subtilis* strains, with the exception of group B BGCs found in *B. atrophaeus*. We found that 71 BGCs from group PPS and 7 BGCs from group B had the complete BGC for plipastatin (*ppsA* to *ppsE*). In contrast, groups C, D, E, and "others" had BGCs missing up to three biosynthetic genes (BGs) (Fig. S3), consistent with experimental data demonstrating a lack of plipastatin production in *B. subtilis* natto BEST195 (13) and *B. subtilis* P5_B2 (7) (Fig. 2). A similar deletion of BGs was found in several other strains. Interestingly, these strains are phylogenetically close to each other, suggesting such deletions being conserved within a single clade (Fig. 2F; Fig. S4).

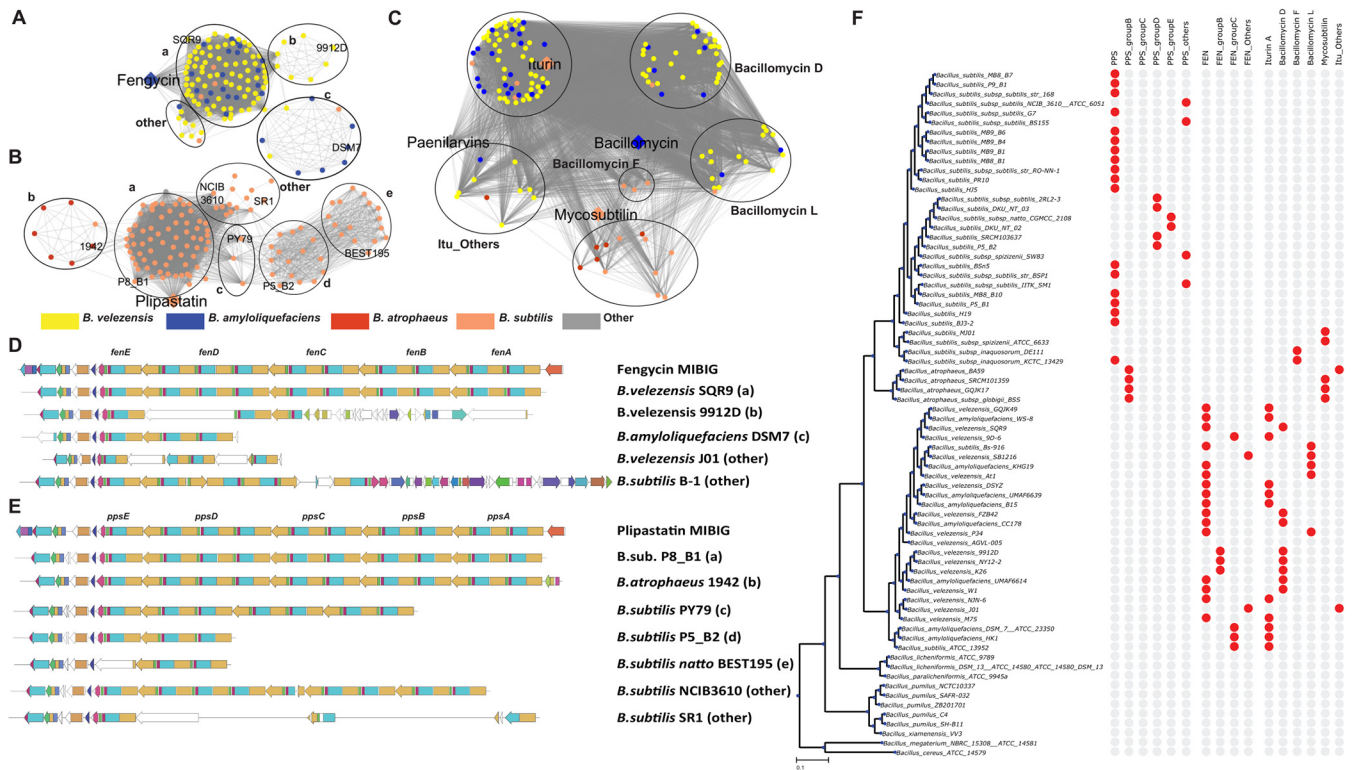


FIG 2 Comparison of plipastatin/fengycin/iturin families of BGCs. (A to C) Similarity networks representing the plipastatins, fengycins, and iturinic lipopeptide GCFs, respectively. The different colors represent different species of *Bacillus*. Iturinic lipopeptides are grouped based on amino acid specificity predictions instead of BiG-SCAPE-generated similarity index (Table S1). (D to E) Selected clusters are shown from different groups of plipastatins and fengycins, respectively. The detailed genetic structures of all incomplete BGC families can be found in Fig. S3. (F) The phylogenetic distribution of different groups of BGCs is presented across selected genomes. For a complete list of all genomes and different groups of BGCs, see Fig. S4.

Additionally, in plipastatin BGCs of group E, gene *ppsE* appeared to have missing domains (Fig. S5). Investigation of the nucleotide sequences of *ppsE* gene homologs revealed a deletion at position 232 of the reference *ppsE* gene across all 16 members of group E, leading to a frameshift mutation causing alternative protein sequence translation that lacks the respective functional domains (Fig. S5). This frameshift was present in multiple strains isolated from distinct geographic locations (Data Set S2) but belonging to the same phylogenetic clade, suggesting an evolutionarily conserved frameshift in the *ppsE* gene that may lead to loss of function.

The fengycin family network contained 123 BGCs from *B. velezensis* and *B. amyloliquefaciens* species, in addition to 5 isolates likely misclassified as "*B. subtilis*." Based on our multilocus sequence assay (MLSA) data, these should be assigned as *B. velezensis* (Data Set S2; Fig. S4). The fengycin BGCs could be divided into four groups, with 96 BGCs containing all BGs (*fenA* to *fenE*). BGCs from groups B, C, and others contained incomplete BGCs, with up to three of the BGs missing (Fig. S3). The strains harboring these incomplete fengycin BGCs were also phylogenetically close, similar to plipastatins, suggesting that these deletions were conserved within a single clade (Fig. 2F; Fig. S4). As noted above for the *ppsE* gene, many phylogenetically close strains harboring group B of fengycin contained a frameshift at positions 3126 to 3127 of the *fenD* gene, suggesting a possible evolutionary trait of the clade (Fig. S6). Again, these frameshift mutations result in the translation of an alternative protein sequence lacking the functional domains of FenD.

Unlike with the above, sequence similarity alone could not divide the 141 iturin-like BGCs into distinct groups due to conserved BG sequences. These BGs differ only in the individual amino acid substrate specificities leading to the production of diverse lipopeptides, like iturin A, bacillomycin D-F-L, and mycosubtilin (14), with different levels of bioactivity (6). Therefore, the antiSMASH-predicted amino acid substrate specificities

for all NRPS adenylation domains were used to group the BGCs into iturin A, bacillomycin D-F-L, mycosubtilin, and “others” that have less than the 7 amino acid substrates that are typical for iturins (Table S1; Data Set S2). Mapping these data onto the phylogenetic tree revealed that each group is conserved in closely related strains. The mycosubtilin group was detected in *B. atrophaeus* and in some *B. subtilis* and *B. spizizenii* strains, and bacillomycin was detected in three strains of *B. inaquosorum*, whereas iturin A, bacillomycin D, and bacillomycin F were spread across different *B. velezensis* and *B. amyloliquefaciens* isolates, confirming the previously proposed species- and strain-level presence of iturinic lipopeptides (14).

The lack of SM production, specifically surfactin, in domesticated strains of *B. subtilis* has previously been connected to a frameshift mutation in the *sfp* gene, coding for a 4-phosphopantetheinyl transferase, which transfers the essential phosphopantetheine prosthetic group to the surfactin NRPS (15). This frameshift mutation in the *sfp* gene (all identical to previously reported positions) was detected in only 13 of 260 genes, all belonging to closely related laboratory strains and an additional *B. subtilis* isolate (Data Set S2; Fig. S7), suggesting that the inactivation of lipopeptide production in natural *Bacillus* isolates is not as common as might have been expected based on the laboratory observations.

Our detailed BGC comparison identified variations in particular GCFs to be phylogenetically conserved but also revealed that particular GCFs were clade rather than species specific. Therefore, our study improves upon the previous systematic descriptions of BGCs within the *Bacillus* genus (3–5) by providing a more detailed, species-level examination of the distributions and features of BGCs within strains belonging to the environmentally important *B. subtilis* group. Such phylogenetic correlation of different BGC groups and particular frameshifts suggest evolutionary relationships among production capabilities of *Bacillus* strains. Therefore, our workflow, combining comparative analysis of BGCs and phylogenetic relationships, revealed how a particular BGC evolves within a species. This knowledge, and closer examination of the exceptions, may guide the selection of specific strains as antimicrobial producers within underexplored groups of SM producers.

Genome selection. Initially, all genomes of *B. amyloliquefaciens*, *B. atrophaeus*, *B. licheniformis*, *B. paralicheniformis*, *B. pumilus*, *B. subtilis*, *B. velezensis*, *B. xiamenensis*, and a few related *Bacillus* sp. strains with assembly status “complete” or “chromosome” publicly available from the NCBI in July 2019 were selected. Additionally, the type strains of *B. cereus* and *B. megaterium* were included as outgroups. The strain list was then curated to remove duplicates. Further, the genomes of engineered *B. subtilis* and strains were removed (*B. subtilis* BEST7613, *B. subtilis* delta6, *B. subtilis* IIG-Bs27-47-24, *B. subtilis* PS38, and *B. subtilis* PG10, as described in reference 16, as well as *B. subtilis* BEST7003, *B. subtilis* QB5413, *B. subtilis* QB5412, *B. subtilis* QB928, and *B. subtilis* WB800N). After preliminary tree reconstruction, *B. subtilis* HDZK-BYSB7 was found to group with *B. cereus* rather than the other *B. subtilis* strains and was therefore removed; it has since been reclassified as *B. anthracis*. Initial examination of results also found BGCs to be split by the origin in *B. velezensis* Hx05; for ease of analysis, this strain was therefore dropped. Subsequently, *B. velezensis* AGVL-005 was found to contain many frameshifted proteins; however, it was retained. A further 13 in-house genomes of *B. subtilis* and one of *B. licheniformis* (17) were included. This led to a final count of 310 genomes.

Genome acquisition and strain name annotation. Genomes were downloaded in the NCBI GenBank format with the ncbi-acc-download tool (<https://github.com/kblin/ncbi-acc-download>). As many of the GenBank entries did not contain strain information in the “Source” or “Organism” features, which are required by the autoMLST and BiG-SCAPE tools to distinguish the individual strains, the Python script `rename_strainless_organisms.py` (found in the tree and matrix construction pipeline [see below]) was employed to transfer strain information from the “strain” field to these fields.

Genome mining. In a first step, all downloaded genomes were initially mined for SMs with antiSMASH 5.0 (8). antiSMASH collapses gene clusters that are in close proximity, such as the iturin and fengycin clusters in *Bacilli*, into a single biosynthetic “region.” A modified version of antiSMASH (<https://github.com/KatSteinke/dmz-antismash>) that contains the additional functionality to split known clusters at a user-defined gene, resulting in two independent “regions,” was developed. In all other respects, this version of antiSMASH is identical to antiSMASH 5.0.0. The modified version of antiSMASH was run as an antiSMASH fast run with the default parameters. The genes selected to split between adjacent clusters were *dacC* and *yngH* for the plipastatin/fengycin clusters and *yxjF* and *xynD* for the iturin clusters. For assigning the plipastatin/fengycin boundary genes, homologs from several species were selected to reflect species variations: *dacC* homologs from *B. velezensis*, *B. subtilis*, *B. amyloliquefaciens*, and *B. atrophaeus* and *yngH* homologs from *B. subtilis* and *B. atrophaeus*. These were selected so that the cut would yield the intersection of both clusters as found on MIBiG, from *dacC* to *yngH*, as other boundaries led to incorrect splits, either failing to cut the cluster or cutting it twice. The genes are identified by a BLAST search in the examined genome, with coverage and identity of at least 90% each needed for identification. During this step, errors in the GenBank file of *B. licheniformis* PB3 (accession no. [NZ_CP025226.1](https://ncbi.nlm.nih.gov/nucl/NZ_CP025226.1)) were detected, as they caused subsequent errors in antiSMASH; the erroneous portions, CXG95_RS00005 and CXG95_RS00010, were consequently deleted.

GCF identification and clustering. We used BiG-SCAPE (9) at default settings to identify families of homologous gene clusters present in multiple species (gene cluster families [GCFs]). In order to automatically identify any known compounds, reference clusters from the MIBiG database (18) were included in the networking analysis. Singleton clusters were not returned. As this produced almost exclusively GCFs split along species lines, even for compounds known to be found in all species, connected components were identified with the NetworkX library (19) using an approach similar to that in reference 20. However, as BiG-SCAPE was left at default options, duplicated entries were later merged.

Tree building. For getting a highly resolved phylogeny of the closely related *Bacillus* strains, maximum-likelihood trees were constructed with a pipeline based on autoMLST (10) and by using autoMLST defaults to the greatest extent. We introduced a modification to autoMLST that skipped the automated search and inclusion of similar genomes and thus processed only the supplied genomes. Subsequently, the pipeline identified all conserved single-copy genes from these genomes. Additionally, the *gbk2sql*.py script in autoMLST, which was employed in the pipeline, was patched to use the same *hmm* database (*reducedcore.hmm*) as the main *automl*.py script. The modified version, including *reducedcore.hmm*, is available at <https://github.com/KatSteinke/automl-simplified-wrapper>.

Both for the short tree shown in Fig. 1 and the full tree (Fig. S1), analysis with this pipeline yielded 30 single-copy/housekeeping genes for each tree; however, not all of these were identical between the trees. For generating the multilocus alignment, each individual gene was aligned with MAFFT (21) and the alignment trimmed using trimAl (22); then, all alignments were concatenated. As in autoMLST, the tree was generated with IQ-TREE (23), using Ultrafast Bootstrap (24) with 1,000 replicates.

The resulting tree was rerooted in ETE3 (25) during the visualization step, using *B. megaterium* NBRC 15308 and *B. cereus* ATCC 14579 as an outgroup. During this process, it was found that the GenBank file of *B. subtilis* subsp. *subtilis* NCD-2 had been excluded from the tree because it lacked gene annotations; thus, it was annotated with Prokka.

In our global analysis of all 310 genomes, we identified a total of 28 strains whose genome-based taxonomy conflicts with their assigned species names (Fig. S1; Data Set S1). Based on our analysis, in line with the recently released genome-based taxonomy in GTDB (26, 27), these strains should be designated *B. velezensis* or *B. amyloliquefaciens*, respectively. Additionally, strains designated *B. subtilis* subsp. *inaquosorum* and

B. subtilis subsp. *spizizenii* by NCBI form their own clades, consistent with their recent promotion to species status (11). The tree thus appears to reflect genome-based taxonomy well.

Absence/presence matrix. We established an automated pipeline for the tree and matrix construction pipeline that combined the individual steps of the analysis. The script for this pipeline takes as arguments the location of a base directory in which analysis results are to be placed, the location of a file listing accession numbers to be downloaded, the name of the final tree to be output, and optional outgroups to be used. It creates all the files and directories necessary for the subsequent analysis (see below). The script can be downloaded at <https://github.com/KatSteinke/AbsPresTree>.

From the connected component GCFs, a matrix-counting occurrence of each GCF in each strain was computed. GCFs were subsequently clustered according to their occurrence in each strain using SciPy's clustering package (28); hierarchical clustering was performed. Subsequently, the absence/presence matrix was reordered to reflect the clustering of GCFs.

It must be noted, however, that the connected-component GCFs are based on the placement of gene clusters in a network, and even incomplete or inactive clusters may be included if they pass the threshold for clustering. The tree and matrix were visualized in ETE3 using ETE3's clustering module. Subsequently, matrix columns were manually arranged to follow the phylogeny of the strains primarily represented per column.

Variations within particular GCFs. Based on the similarity networks of the plipastatin and fengycin GCFs, we created groups within a GCF. The fengycin GCF was split into four groups, and the plipastatin GCF was split into six groups. The genetic structure variations among groups with few missing BGs are shown in Fig. S3. The genes *ppsE* and *fenD* from plipastatin group E and fengycin group B, respectively, are further selected for multiple-sequence alignment (Fig. S5 and S6). For the iturinic lipopeptide GCF, substrate specificities of the A domain were collected from antiSMASH annotations. Based on the individual amino acid specificities, the BGCs from this GCF are further classified into iturin A, bacillomycin D, F, and L, and mycosubtilin (Table S1). In addition to analyzing GCF variation, we aligned nucleotide sequences of the *sfp* gene, coding for 4-phosphopantetheinyl transferase, from 260 BGCs of the surfactin family (Data Set S2). A frameshift mutation previously known to disrupt *sfp* function was detected across 14 of the 260 genes (Fig. S7). We generated a presence-absence matrix where the rows represent 310 genomes and the columns represent groups of plipastatins (PPS, PPS groups B to E, PPS others), fengycins (FEN, FEN groups B and C, and FEN others), iturin A, bacillomycin D, F, and L, mycosubtilin, and *sfp* gene frameshift mutation (Fig. S4). The presence-absence matrix is visualized against the tree to understand the evolutionary aspects of GCF variation. The scripts used to analyze the variations in GCF can be downloaded at https://github.com/OmkarSaMo/GCF_variation_Bacillus.

Data availability. The data with NCBI accession IDs and information on all detected gene clusters are available in Data Sets S1 and S2. Code used to generate the data is available at <https://github.com/KatSteinke/AbsPresTree>. The script used to analyze the variations in GCF is available at https://github.com/OmkarSaMo/GCF_variation_Bacillus.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

DATA SET S1, XLSX file, 0.3 MB.

DATA SET S2, XLSX file, 0.1 MB.

FIG S1, PDF file, 2.1 MB.

FIG S2, PDF file, 1.7 MB.

FIG S3, PDF file, 2.4 MB.

FIG S4, PDF file, 1 MB.

FIG S5, PDF file, 1.8 MB.

FIG S6, PDF file, 1 MB.

FIG S7, PDF file, 2.3 MB.

TABLE S1, PDF file, 0.02 MB.

ACKNOWLEDGMENTS

This work was funded by the Danish National Research Foundation (grant DNRF137) for the Center for Microbial Secondary Metabolites. T.W. and O.S.M., furthermore, acknowledge funding from the Novo Nordisk Foundation (grants NNF20CC0035580 and NNF16OC0021746).

We declare that we have no competing interests.

K.S. and O.S.M. performed the bioinformatic analysis, K.S., O.S.M., T.W., and A.T.K. interpreted the data, and K.S., O.S.M., T.W., and A.T.K. wrote the manuscript.

REFERENCES

- Harwood CR, Mouillon J-M, Pohl S, Arnau J. 2018. Secondary metabolite production and the safety of industrially important members of the *Bacillus subtilis* group. *FEMS Microbiol Rev* 42:721–738. <https://doi.org/10.1093/femsre/fuy028>.
- Fira D, Dimkić I, Berić T, Lozo J, Stanković S. 2018. Biological control of plant pathogens by *Bacillus* species. *J Biotechnol* 285:44–55. <https://doi.org/10.1016/j.jbiotec.2018.07.044>.
- Aleti G, Sessitsch A, Brader G. 2015. Genome mining: prediction of lipopeptides and polyketides from *Bacillus* and related Firmicutes. *Comput Struct Biotechnol J* 13:192–203. <https://doi.org/10.1016/j.csbj.2015.03.003>.
- Zhao X, Kuipers OP. 2016. Identification and classification of known and putative antimicrobial compounds produced by a wide variety of *Bacillales* species. *BMC Genomics* 17:882. <https://doi.org/10.1186/s12864-016-3224-y>.
- Grubbs KJ, Bleich RM, Santa Maria KC, Allen SE, Farag S, Shank EA, Bowers AA. 2017. Large-scale bioinformatics analysis of *Bacillus* genomes uncovers conserved roles of natural products in bacterial physiology. *mSystems* 2:e00040-17. <https://doi.org/10.1128/mSystems.00040-17>.
- Kaspar F, Neubauer P, Gimpel M. 2019. Bioactive secondary metabolites from *Bacillus subtilis*: a comprehensive review. *J Nat Prod* 82:2038–2053. <https://doi.org/10.1021/acs.jnatprod.9b00110>.
- Kiesewalter HT, Lozano-Andrade CN, Wibowo M, Strube ML, Maróti G, Snyder D, Jørgensen TS, Larsen TO, Cooper VS, Weber T, Kovács ÁT. 2021. Genomic and chemical diversity of *Bacillus subtilis* secondary metabolites against plant pathogenic fungi. *mSystems* 6:e00770-20. <https://doi.org/10.1128/mSystems.00770-20>.
- Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, Medema MH, Weber T. 2019. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res* 47:W81–W87. <https://doi.org/10.1093/nar/gkz310>.
- Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, De Los Santos ELC, Yeong M, Cruz-Morales P, Abubucker S, Roeters A, Lokhorst W, Fernandez-Guerra A, Cappelini LTD, Goering AW, Thomson RJ, Metcalf WW, Kelleher NL, Barona-Gomez F, Medema MH. 2020. A computational framework to explore large-scale biosynthetic diversity. *Nat Chem Biol* 16:60–68. <https://doi.org/10.1038/s41589-019-0400-9>.
- Alanjary M, Steinke K, Ziemert N. 2019. AutoMLST: an automated web server for generating multi-locus species trees highlighting natural product potential. *Nucleic Acids Res* 47:W276–W282. <https://doi.org/10.1093/nar/gkz282>.
- Dunlap CA, Bowman MJ, Zeigler DR. 2020. Promotion of *Bacillus subtilis* subsp. *inaquosorum*, *Bacillus subtilis* subsp. *spizizenii* and *Bacillus subtilis* subsp. *stercoris* to species status. *Antonie Van Leeuwenhoek* 113:1–12. <https://doi.org/10.1007/s10482-019-01354-9>.
- Chen XH, Koumoutsis A, Scholz R, Eisenreich A, Schneider K, Heinemeyer I, Morgenstern B, Voss B, Hess WR, Reva O, Junge H, Voigt B, Jungblut PR, 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. <https://doi.org/10.1038/nbt1325>.
- Nishito Y, Osana Y, Hachiya T, Popendorf K, Toyoda A, Fujiyama A, Itaya M, Sakakibara Y. 2010. Whole genome assembly of a natto production strain *Bacillus subtilis* natto from very short read data. *BMC Genomics* 11:243. <https://doi.org/10.1186/1471-2164-11-243>.
- Dunlap CA, Bowman MJ, Rooney AP. 2019. Iturinic lipopeptide diversity in the *Bacillus subtilis* species group—important antifungals for plant disease biocontrol applications. *Front Microbiol* 10:1794. <https://doi.org/10.3389/fmicb.2019.01794>.
- Kearns DB, Chu F, Rudner R, Losick R. 2004. Genes governing swarming in *Bacillus subtilis* and evidence for a phase variation mechanism controlling surface motility. *Mol Microbiol* 52:357–369. <https://doi.org/10.1111/j.1365-2958.2004.03996.x>.
- Wu H, Wang D, Gao F. 17 February 2020. Toward a high-quality pan-genome landscape of *Bacillus subtilis* by removal of confounding strains. *Brief Bioinform* <https://doi.org/10.1093/bib/bbaa013>.
- Kiesewalter HT, Lozano-Andrade CN, Maróti G, Snyder D, Cooper VS, Jørgensen TS, Weber T, Kovács ÁT. 2020. Complete genome sequences of 13 *Bacillus subtilis* soil isolates for studying secondary metabolite diversity. *Microbiol Resour Announc* 9:e01406-19. <https://doi.org/10.1128/MRA.01406-19>.
- Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, de Bruijn I, Chooi YH, Claesen J, Coates RC, Cruz-Morales P, Duddela S, Dusterhaus S, Edwards DJ, Fewer DP, Garg N, Geiger C, Gomez-Escribano JP, Greule A, Hadjithomas M, Haines AS, Helfrich EJM, Hillwig ML, Ishida K, Jones AC, Jones CS, Jungmann K, Kegler C, Kim HU, Kötter P, Krug D, Masschelein J, Melnik AV, Mantovani SM, Monroe EA, Moore M, Moss N, Nützmann H-W, Pan G, Pati A, Petras D, Reen FJ, Rosconi F, Rui Z, Tian Z, Tobias NJ, Tsunematsu Y, Wiemann P, Wyckoff E, Yan X, et al. 2015. Minimum information about a biosynthetic gene cluster. *Nat Chem Biol* 11:625–631. <https://doi.org/10.1038/nchembio.1890>.
- Hagberg AA, Schult DA, Swart PJ. 2008. Exploring network structure, dynamics, and function using NetworkX, p 11–15. In Varoquaux G, Vaught T, Millman J (ed), 7th Python in Science Conference (SciPy 2008), Pasadena, CA.
- Mohite OS, Lloyd CJ, Monk JM, Weber T, Palsson BO. 2019. Pangenome analysis of Enterobacteria reveals richness of secondary metabolite gene clusters and their associated gene sets. *bioRxiv* <https://doi.org/10.1101/781328>.
- Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30:772–780. <https://doi.org/10.1093/molbev/mst010>.
- Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–1973. <https://doi.org/10.1093/bioinformatics/btp348>.
- Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 32:268–274. <https://doi.org/10.1093/molbev/msu300>.
- Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol* 35:518–522. <https://doi.org/10.1093/molbev/msx281>.
- Huerta-Cepas J, Dopazo J, Gabaldón T. 2010. ETE: a Python environment for tree exploration. *BMC Bioinformatics* 11:24. <https://doi.org/10.1186/1471-2105-11-24>.
- Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* 36:996–1004. <https://doi.org/10.1038/nbt.4229>.

27. Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. 2020. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat Biotechnol* 38:1079–1086. <https://doi.org/10.1038/s41587-020-0501-8>.
28. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, SciPy 1.0 Contributors. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 17:261–272. <https://doi.org/10.1038/s41592-019-0686-2>.