scientific data

OPEN

DATA DESCRIPTOR

Expanding the genome information on *Bacillales* for biosynthetic gene cluster discovery

Lijie Song[1,10], Lasse Johan Dyrbye Nielsen[1,10], Xinming Xu[1,2], Omkar Satyavan Mohite[3], Matin Nuhamunada[3], Zhihui Xu[4], Rob Murphy[5], Kasun Bodawatta[6,7], Michael Poulsen[5], Mohamed Hatha Abdulla[8], Eva C. Sonnenschein[1,9], Tilmann Weber[3] ✉ & Ákos T. Kovács[1,2] ✉

This study showcases 121 new genomes of spore-forming *Bacillales* from strains collected globally from a variety of habitats, assembled using Oxford Nanopore long-read and MGI short-read sequences. *Bacilli* are renowned for their capacity to produce diverse secondary metabolites with use in agriculture, biotechnology, and medicine. These secondary metabolites are encoded within biosynthetic gene clusters (smBGCs). smBGCs have significant research interest due to their potential as sources of new bioactivate compounds. Our dataset includes 62 complete genomes, 2 at chromosome level, and 57 at contig level, covering a genomic size range from 3.50 Mb to 7.15 Mb. Phylotaxonomic analysis revealed that these genomes span 16 genera, with 69 of them belonging to *Bacillus*. A total of 1,176 predicted BGCs were identified by *in silico* genome mining. We anticipate that the open-access data presented here will expand the reported genomic information of spore-forming *Bacillales* and facilitate a deeper understanding of the genetic basis of *Bacillales*' potential for secondary metabolite production.

Background & Summary

Bacillus is a genus of Gram-positive, rod-shaped bacteria that are widely distributed in soil, water, and other diverse environments. *Bacillus* species have been extensively studied for their potential to produce secondary metabolites (SMs), which have a wide range of functions and activities, and are being harnessed in various fields, such as agriculture, biotechnology, and medicine[1,2]. Several studies have reported that *Bacillus* and related genera produce secondary metabolites, an ability conferred by the presence of biosynthetic gene clusters[3–5].

Secondary metabolite biosynthetic gene clusters (smBGCs) are genomic regions containing two or more genes involved in the biosynthetic pathway of secondary metabolites. These genes encode enzymes, transport proteins, regulatory factors, and other accessory proteins that contribute to the secondary metabolite biosynthetic process[6]. The composition and structures of smBGCs can vary widely across and even within the same species. The importance and feasibility of exploring species-specific BGCs have been recently highlighted[7,8]. Many bioinformatics tools have been developed to predict, identify, and characterize smBGCs[9], which require high quality genome sequences[10]. The development of sequencing technologies has made whole genome sequencing simpler and faster. In particular, the integration of high throughput sequencing (short-read) and long-read sequencing data, can lead to high quality assemblies of genomes, including complete genomes[11].

In this study, we performed whole genome sequencing for strains collected from different countries and regions spanning four different continents (Fig. 1), based on an integrated approach, including Oxford Nanopore

[1]DTU Bioengineering, Technical University of Denmark, 2800 Kgs, Lyngby, Denmark. [2]Institute of Biology, Leiden University, 2333BE, Leiden, Netherlands. [3]The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, 2800 Kgs, Lyngby, Denmark. [4]Key laboratory of organic-based fertilizers of China and Jiangsu Provincial Key Laboratory for solid organic waste utilization, Nanjing Agricultural University, Nanjing, Jiangsu, People's Republic of China. [5]Section for Ecology and Evolution, Department of Biology, University of Copenhagen, 2100, Copenhagen, Denmark. [6]Natural History Museum of Denmark, University of Copenhagen, 2100, Copenhagen, Denmark. [7]Section for Molecular Ecology and Evolution, Globe Institute, University of Copenhagen, 2100, Copenhagen, Denmark. [8]Department of Marine Biology, Microbiology and Biochemistry, School of Marine Sciences, Cochin University of Science and Technology, Kerala, 682 016, India. [9]Department of Biosciences, Faculty of Science and Engineering, Swansea University, Swansea, SA2 8PP, United Kingdom. [10]These authors contributed equally: Lijie Song, Lasse Johan Dyrbye Nielsen. ✉e-mail: tiwe@biosustain.dtu.dk; a.t.kovacs@biology.leidenuniv.nl

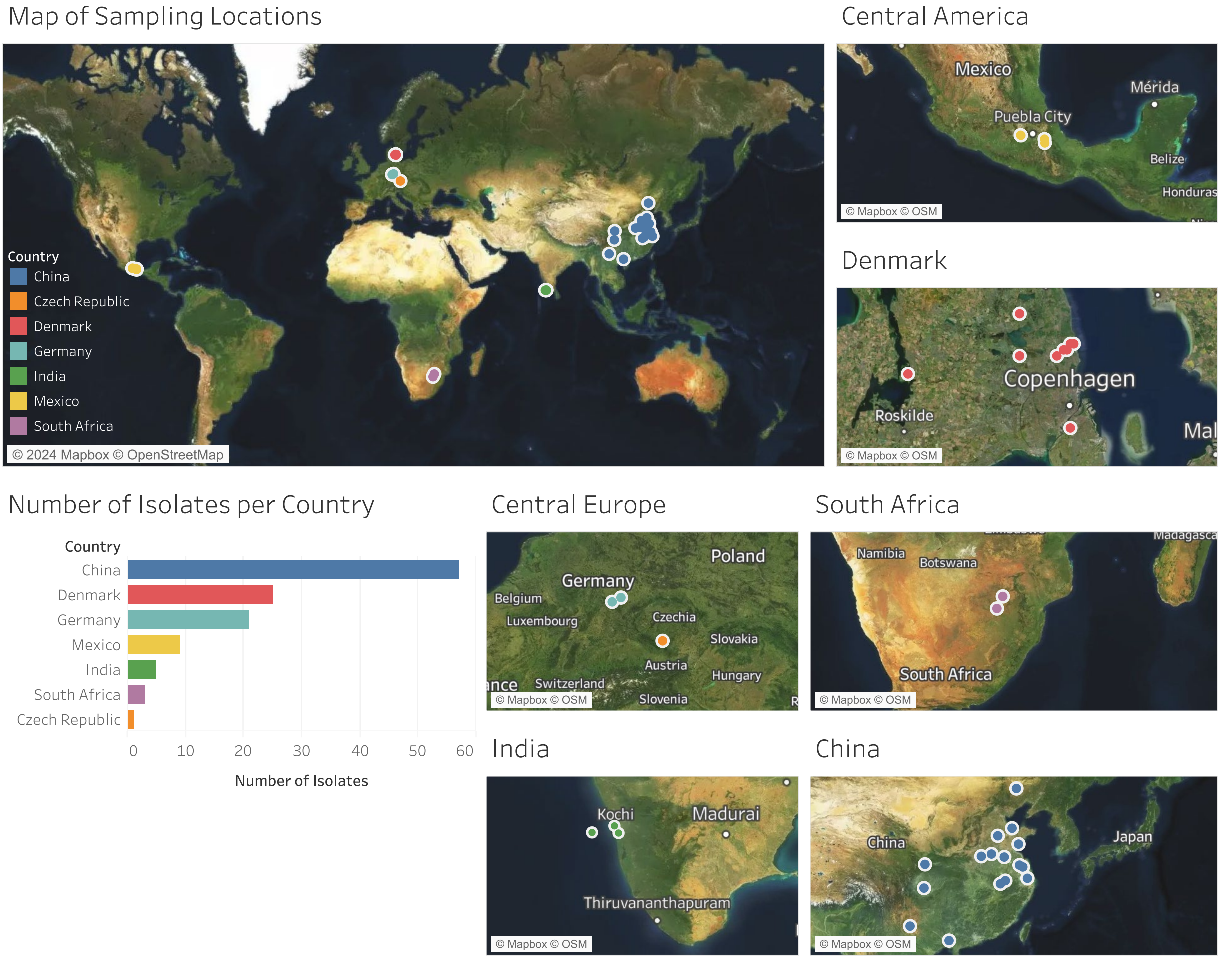

Fig. 1 Distribution of sample collection site coordinates depicted using OpenStreetMap.

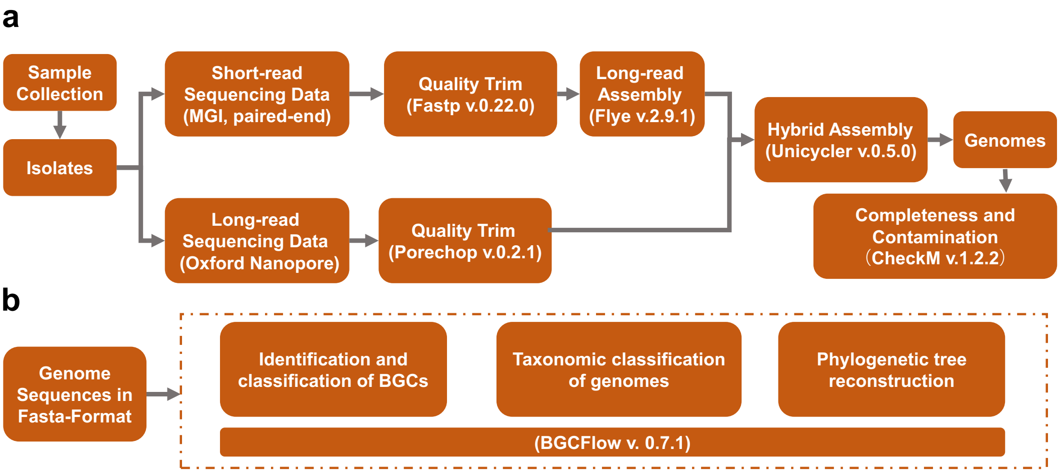
a
Sample Collection
Isolates
Short-read Sequencing Data (MGI, paired-end)
Quality Trim (Fastp v.0.22.0)
Long-read Assembly (Flye v.2.9.1)
Long-read Sequencing Data (Oxford Nanopore)
Quality Trim (Porechop v.0.2.1)
Hybrid Assembly (Unicycler v.0.5.0)
Genomes
Completeness and Contamination（CheckM v.1.2.2）
b
Genome Sequences in Fasta-Format
Identification and classification of BGCs
Taxonomic classification of genomes
Phylogenetic tree reconstruction
(BGCFlow v. 0.7.1)

Fig. 2 Illustration of Genome assembly and BGC analysis. (**a**) Strategy for sequencing and genome assembly, (**b**) the BGC analysis pipeline.

long-read sequencing and MGI short-read sequencing. Here, we sequenced and assembled 121 genomes using this approach. An outline of the study's experimental and analysis design is presented in Fig. 2, and detailed descriptions of the workflow are provided in the methodology sections. According to the completeness criteria of the National Center for Biotechnology Information (NCBI), we produced, in total, 62 assemblies at a complete genome level, 2 at chromosome level, and a remaining 57 at contig level (Supplementary Table 1 for details). Overall, the genome sizes range from 3.50 Mb to 7.15 Mb (5.09 Mb on average), with a GC content ranging

		Assembly level			Average genome size (min and max) (Mbps)	Average GC (min and max) (%)	Average gene number (min and max)
Genus	# of isolates	complete genome	chromosome	contig			
Bacillus	69	35	2	32	4.86 (3.50–6.33)	39.46 (34.40–46.50)	4,998 (3,616–6,554)
Peribacillus	16	7	0	9	5.59 (5.35–6.02)	40.25 (39.50–40.50)	5,452 (5,245–5,817)
Neobacillus	8	6	0	2	6.05 (5.59–6.29)	38.63 (38.00–40.00)	5,917 (5,383–6,233)
Paenibacillus	5	3	0	2	6.40 (5.29–7.15)	46.60 (44.00–50.00)	5,724 (4,881–6,560)
Lysinibacillus	5	2	0	3	5.17 (4.56–5.54)	36.80 (36.50–37.50)	5,180 (4,575–5,605)
Cytobacillus	4	2	0	2	5.03 (4.73–5.30)	40.50 (37.00–42.00)	5,034 (4,752–5,251)
Fictibacillus	4	1	0	3	5.90 (3.94–5.40)	42.88 (39.50–44.00)	5,070 (4,156–5,551)
Brevibacillus	2	1	0	1	5.68 and 6.08	52.00 and 54.00	5,498 and 5,699
Mesobacillus	1	1	0	0	4.80	43.00	4,782
Siminovitchia	1	1	0	0	3.74	44.00	3,727
Halobacillus	1	1	0	0	3.68	47.00	3,841
Virgibacillus	1	1	0	0	4.01	37.50	3,880
Priestia	1	1	0	0	3.82	38.00	3,985
Rossellomorea	1	0	0	1	4.33	48.00	4,467
Ureibacillus	1	0	0	1	4.45	36.00	4,240
Ferdinandcohnia	1	0	0	1	4.77	37.50	4,746
In total	121	62	2	57	5.09 (3.50–7.15)	40.19 (34.50–54.00)	5,119 (3,616–6,560)

Table 1. Summary of general genome information for each genus.

from 34.50% to 54.00% (40.19% on average). The base accuracy of each assembly was assessed using yak[12], and the quality value (QV) ranges from 41.13 to 69.52 (56.41 on average). Based on NCBI PGAP[13], an average of 5,119 genes, including 4,851 protein-coding genes were annotated in the genomes (Table 1). Taxonomic analysis showed that these 121 genomes could be classified into 16 genera within the *Bacillales* order, most of which were species from the *Bacillus* genus (Fig. 3). (Supplementary Table 2).

To assess the potential for secondary metabolite production in these isolates, the genome mining tool BGCFlow[14] was applied for BGC identification and annotation, resulting in a total of 1,176 BGCs predicted. The BGCs were categorized into seven classes through BiG-SCAPE[15], part of the BGCFlow executable, which showed that RiPPs have the greatest count of 381 and comprise the highest percentage at 32.4% (Supplementary Table 3). The distribution of BGC counts per genus highlights the uneven abundance of BGCs between the distinct genera (Fig. 4). Notably, the genera *Bacillus* and *Paenibacillus* harbor the highest number of BGCs among the genomes presented here.

To assess the novelty of the BGCs found in the dataset, the sequence similarity network of the BGCs constructed using BiG-SCAPE was further enriched with the top KnownClusterBlast hit to the known entries from the MIBiG database[16], resulting in 283 connected components of gene cluster families (GCFs). Using this analysis, we can group 310 BGCs (26%) into 27 GCFs with high similarity to 37 known compounds produced by *Bacillus* or related genera, such as bacilysin, surfactin, and subtilosin (Figure S1A, Supplementary Table 4a,b). Meanwhile, 435 BGCs (37%) can be grouped into 55 GCFs with medium to low similarity to 59 known compounds (Figure S1B, Supplementary Table 4c,b). Almost half of the known compound hits in this category (29 hits) are also produced by *Bacillus* and related genera, hinting that these GCFs might produce analogs or compounds of similar types. The remaining 30 compounds are known to be produced by other distantly related genera. This is likely due to the limitation of the database and suggest further detailed BGC comparison to identify possible products. Finally, the remaining 431 BGCs (36%) can be grouped into 201 GCFs without any hits to known BGCs in the database (Figure S1C, Supplementary Table 4e,f). While some of the GCFs in this category are quite conserved in some clades, a third of the GCFs are singletons. These unknown categories hold the potential for further experimentation to find novel compounds from the *Bacillales* group. A detailed list of the GCF assignments can be found in Supplementary Table 4.

The datasets and genomic analysis results described here greatly expand the reported genomic information of spore-forming *Bacillales* and will also strengthen studies advancing our understanding of the secondary metabolite potential of the *Bacillales* order.

Method

Sample collection and isolation. Sample collection was dependent on the isolating laboratory. Using soil samples collected at diverse locations in Germany, Denmark, China, and Mexico, spore-forming soil bacteria were isolated after heat treatment at 80 °C for 10 minutes and spreading the soil suspension on lysogeny broth (LB) or tryptic soy broth (TSB) plates with 1.5% agar that were incubated at 37 °C for 2 days.

Bacillus altitudinis J6-1 and J6-2 were isolated from a biofilm sample obtained from the pier at Jyllinge Harbour (55.744923; 12.094888). Biofilm samples were incubated at 80 °C for 15 mins and subsequently plated on LB agar and incubated at 25 °C.

Other marine samples were collected from the Cochin estuary and adjacent coastal waters (South-west coast of India), during pre-monsoon (March), monsoon (August) and post-monsoon (December) periods of the year 2012 and 2013. Water samples were serially diluted and spread on Norris Glucose Nitrogen free medium

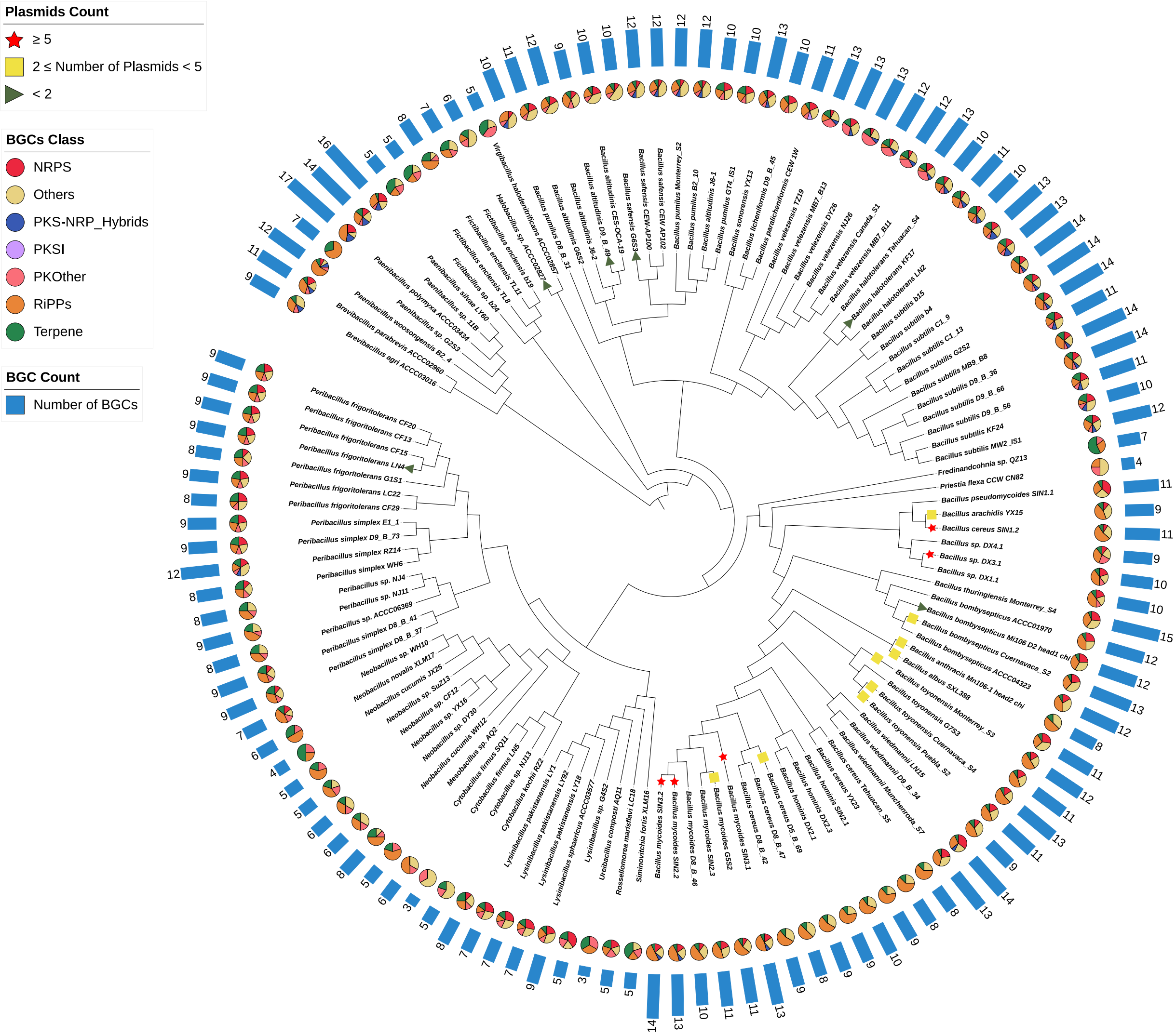

Fig. 3 A phylogenetic tree of the 121 genomes with plasmid content and BGC class and count indicated.

(NGNF medium, HIMEDIA-M712) with 1.5% agar (Himedia GRM 666) and incubated at 28 ± 1 °C for 7–14 days. Separated colonies with different morphologies were picked using a sterile inoculation loop, re-streaked and maintained on the slants of fresh nitrogen free culture medium at 4 °C. Cell morphology and presence of endospore was analyzed by light microscopy (Olympus CX21i). Rod shaped endospore forming isolates were selected for this study.

Isolate Mi106 D2 head1 chi was obtained from the head of a worker termite from a colony of *Microtermes* sp. and Mn106-1 head2 chi was obtained from the head of a worker from a colony of *Macrotermes natalensis* in Mookgophong, South Africa (S24 40 30.5 E28 47 50.4) in 2010. In both cases, the surface of a worker termite was rinsed using phosphate buffer saline (PBS). Subsequently, the head of the termite was crushed in 200 µl PBS, which was subsequently spread onto chitin medium (4 g chitin, 0.7 g K_2HPO_4, 0.3 g KH_2PO_4, 0.5 g $MgSO_4 \times 5H_2O$, 0.01 g $FeSO_4 \times 7H_20$, 0.001 g $ZnSO_4$, 0.001 g $MnCl_2$, and 20 g of agar per liter). Growing colonies on plates were streaked onto Yeast Malt Extract Agar medium (4 g yeast extract, 10 g malt extract, 4 g D-glucose and 20 g bacteriological agar per liter), and once in pure culture, stored in 10% glycerol at −20 °C. Isolate 11B was obtained using the same approach on a fragment from a fungus garden of a *Macrotermes natalensis* colony collected in Rietondale, South Africa (S25 43 45.6 E28 14 09.9) in 2010.

Strains GT4_IS1 and MW2_IS1 were previously isolated from the uropygial glands of Great tits (*Parus major*) from Denmark and Czechia respectively[17].

In each case, observed colonies were re-streaked to obtain single colonies, and subsequently stored at −80 °C with 28% glycerol added. To obtain primary information about these strains, colony PCR was employed to amplify the 16S rRNA gene. Strains that exhibited low similarity and distant branches in the 16S rRNA phylogenetic tree were selected for further study.

Genomic DNA (gDNA) extraction. For genomic DNA (gDNA) extraction, a pure single colony of each isolate was inoculated in 5 ml of LB and incubated at 37 °C for more than 12 hours. Then gDNA was extracted

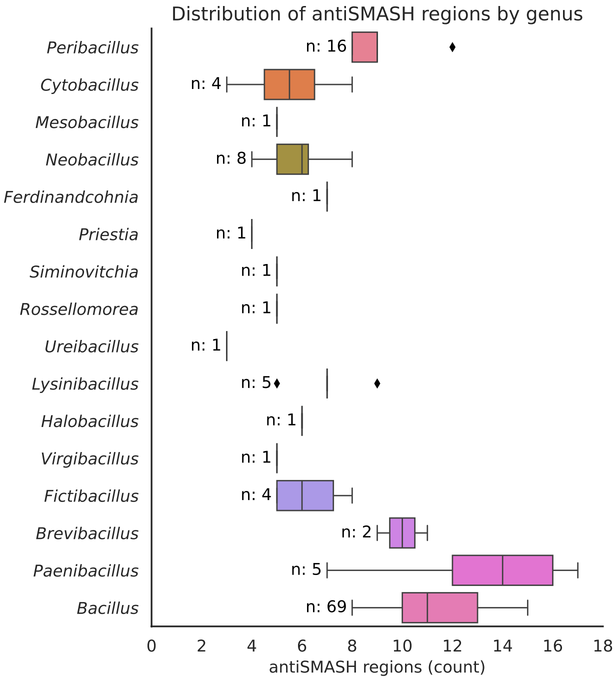

Fig. 4 The number of BGCs in each genus of the 121 genomes.

using E.Z.N.A. DNA extraction kits (OMEGA Bio-Tek Inc., Norcross, GA, USA) following the manufacturer's instructions. The quality and quantity of gDNA were assessed using agarose gel electrophoresis and Nanodrop (Thermo Fisher Scientific, MA, USA), to guarantee that the integrity, concentration, and purity met the requirements for library construction and sequencing.

Short-read sequencing on MGI platform. For each strain, 300 ng gDNA was used for short-read sequencing library construction according to MGI paired-end libraries construction protocol[18]. Briefly, gDNA was fragmented to 200–300 bp using segmentase followed by fragment selection with VAHTS™ DNA Clean Beads (Vazyme, Nanjing, Jiangsu, China). Subsequently, end repair, A-tailing reactions and adapter ligation were implemented. After PCR and purification, the concentration of each library was determined using Qubit® dsDNA HS Assay Kit (Thermo Fisher Scientific) as quality control. The qualified libraries were sequenced on the DNBSEQ-G400 (MGI Tech Co., Ltd.) platform according to the manufacturer's instructions to generate paired end reads (150 bps at each end).

Long-read sequencing on Oxford nanopore platform. For Oxford Nanopore sequencing, the libraries were prepared using the SQK-RBK110.96 barcoding kit (Oxford Nanopore Technologies, Oxford, UK) starting from 50 ng DNA for each strain. In brief, each sample was fragmented and ligated by a unique rapid barcode with incubation at 30 °C for 2 minutes and then at 80 °C for 2 minutes, then all barcoded samples were pooled together in a 1:1 ratio and purified by SPRI beads. After ligation of 1 μl of Rapid Adapter F (RAP F) to 11 μl of pooled DNA, the final library was quantified using Nanodrop. The ONT library was loaded into the MinION spot-on Flow Cell (R9 Version) and sequenced on a MinION Mk1B device according to standard protocol. The resulting reads were basecalled and demultiplexed with MinKNOW UI v.4.1.22.

Genome assembly. For *de novo* assembly, the MGISEQ paired end short reads were adapter and quality trimmed using fastp v.0.22.0 and the Nanopore long reads were adapter trimmed using porechop v.0.2.1, using standard settings[19,20]. The trimmed long reads from Nanopore were assembled with flye v.2.9.1-b1780, and subsequently the trimmed reads from both platforms and the long-read assembly were hybrid assembled with Unicycler v.0.5.0 using the *–existing_long_read_assembly* option[21,22]. The completeness and contamination level of each genome were assessed using CheckM v.1.2.2[23] with the command 'checkm lineage_wf <genomes folder> <output folder>,' which places each genome phylogenetically before choosing the set of single-copy conserved genes to evaluate by the completeness defined as the proportion of markers present, and the contamination defined as the proportion of markers present in multiple copies (see details in Supplementary Table 5). Yak[12] was used to assess quality values of each genome following the protocol[24]. To account for the *Bacillales* genome size of ~6 Mbp, the K value was changed to 6 M.

Genome annotation, taxonomic analysis and BGC prediction. The genomes of the 121 isolates were taxonomically classified and gene-annotated in a two-step process. Initially, we employed GTDB-Tk v2.11[25], using the 'classify_wf' command, to preliminarily assign taxonomic classifications to the FASTA format genomes. Subsequently, these genomes were uploaded to the NCBI GenBank database, where they were annotated using the NCBI Prokaryotic Genome Annotation Pipeline (PGAP). Following this, we conducted a comprehensive analysis of the annotated genomes using BGCFlow v0.7.1. This tool integrates multiple genome mining and phylogenetic tools into one pipeline[14]. To set up the analysis, we created a folder containing the project configuration structure as defined by BGCFlow Portable Encapsulated Project (PEP) specification[26]. The designated project folder contains a comma separated sample file which contains the NCBI-assigned GenBank accession numbers of the 121 *de novo* assembled genomes and the PEP configuration file for the BGCFlow run. The YAML configuration file for the project was configured to enable GTDB-Tk and autoMLST wrapper for phylogenetic tree construction, antiSMASH[27] for BGC annotation, and BiG-SCAPE[15] for BGC dereplication into gene cluster families (GCFs) and generating summary tables. The resulting GCFs were then visualized using Cytoscape version 3.10.2[28]. BGCFlow was executed using standard settings, which include KnownClusterBlast search against the MIBiG database[16] for known BGCs.

We conducted a non-exhaustive search for plasmids within our *de novo* assembled genomes by identifying contiguous sequences (contigs) as plasmids if they were circular and if RFPlasmid[29] (v.0.0.18), an open-source software that classifies contigs as plasmid or chromosomal based on the presence of marker genes and k-mers, classified them as plasmids. Due to the incomplete assembly of several genomes, which resulted in the presence of linear fragments, the absence of any plasmid identified by this method does not necessarily indicate their true absence.

Data Records

The sample information and assembled genomes were deposited in NCBI BioProject under PRJNA960711[30], and raw reads of long-read sequencing on Nanopore and short-read sequencing on MGISEQ have been deposited in NCBI Sequence Read Archive (SRA) under SRP485167[31] (Supplementary Tables 1, 6 for accession and other details).

Technical Validation

In this study, the main steps of experimental procedures and data analysis have been validated. For short-read sequencing on MGI, the libraries were quantified with a minimum of 10 ng/μl. For *de novo* assembly, default parameters were used for quality trimming. In brief, after filtering, an average of 2.69 G MGI reads (0.66 G-6.52 G, PE150) and 76,507 Nanopore reads with mean N50 of 6,709 bp (1,777bp-13,698 bp) for each sample were generated (Supplementary Table 7). CheckM was used for validation of the genome completeness and contamination.

Code availability

The software versions and parameters used for sequence filtering, assembly, and genome mining in this work are described in Methods. Custom code for setting up the BGCFlow run, processing the output, and producing figures, as well as for downloading the genomes, is available at https://github.com/ljdnielsen/bacillales_genomes_figures; https://doi.org/10.5281/zenodo.10907189[32].

Received: 24 April 2024; Accepted: 11 November 2024;
Published online: 21 November 2024

References

1. Caulier, S. *et al*. Overview of the antimicrobial compounds produced by members of the *Bacillus subtilis* group. *Front Microbiol* **10**, 302 (2019).
2. Stein, T. *Bacillus subtilis* antibiotics: Structures, syntheses and specific functions. *Mol Microbiol* **56**, 845–857 (2005).
3. Kaspar, F., Neubauer, P. & Gimpel, M. Bioactive secondary metabolites from *Bacillus subtilis*: A comprehensive review. *J Nat Prod* **82**, 2038–2053 (2019).
4. Kiesewalter, H. T. *et al*. Genomic and chemical diversity of *Bacillus subtilis* secondary metabolites against plant pathogenic fungi. *mSystems* **6**, e00770-20 (2021).
5. Xia, L. *et al*. Biosynthetic gene cluster profiling predicts the positive association between antagonism and phylogeny in *Bacillus*. *Nat Commun* **13**, 1023 (2022).
6. Harwood, C. R., Mouillon, J. M., Pohl, S. & Arnau, J. Secondary metabolite production and the safety of industrially important members of the *Bacillus subtilis* group. *FEMS Microbiol Rev* **42**, 721–738 (2018).
7. Steinke, K., Mohite, O. S., Weber, T. & Kovács, Á. T. Phylogenetic distribution of secondary metabolites in the *Bacillus subtilis* species complex. *mSystems* **6**, e00057-21 (2021).
8. Yin, Q. J. *et al*. Species-specificity of the secondary biosynthetic potential in *Bacillus*. *Front Microbiol* **14**, 1271418 (2023).
9. Blin, K. *et al*. antiSMASH 6.0: improving cluster detection and comparison capabilities. **49**, 29–35 (2021).
10. Blin, K., Kim, H. U., Medema, M. H. & Weber, T. Recent development of antiSMASH and other computational approaches to mine secondary metabolite biosynthetic gene clusters. *Brief Bioinform* **20**, 1103–1113 (2019).
11. Tyler, A. D. *et al*. Evaluation of Oxford Nanopore's MinION sequencing device for microbial whole genome sequencing applications. *Sci Rep* **8**, 10931 (2018).
12. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175 (2021).
13. Tatusova, T. *et al*. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res* **44**, 6614–6624 (2016).
14. Nuhamunada, M., Mohite, O. S., Phaneuf, P. V., Palsson, B. O. & Weber, T. BGCFlow: systematic pangenome workflow for the analysis of biosynthetic gene clusters across large genomic datasets. *Nucleic Acids Res* **52**, 5478–5495 (2024).
15. Navarro-Muñoz, J. C. *et al*. A computational framework to explore large-scale biosynthetic diversity. *Nat Chem Biol* **16**, 60–68 (2020).
16. Terlouw, B. R. *et al*. MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Res* **51**, D603–D610 (2023).
17. Bodawatta, K. H. *et al*. Great Tit (Parus major) uropygial gland microbiomes and their potential defensive roles. *Front Microbiol* **11**, 1735 (2020).
18. Yang, Z. *et al*. Test development, optimization and validation of a WGS pipeline for genetic disorders. *BMC Med Genomics* **16**, 74 (2023).
19. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

20. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb Genom* **3**, e000132 (2017).
21. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37**, 540–546 (2019).
22. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* **13**, e1005595 (2017).
23. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**, 1043–1055 (2015).
24. Liao, W. W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
25. Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* **38**, 5315–5316 (2022).
26. Sheffield, N. C., Stolarczyk, M., Reuter, V. P. & Rendeiro, A. F. Linking big biomedical datasets to modular analysis with Portable Encapsulated Projects. *Gigascience* **10**, giab077 (2021).
27. Blin, K. *et al.* AntiSMASH 7.0: New and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res* **51**, W46–W50 (2023).
28. Shannon, P. *et al.* Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498–2504 (2003).
29. van der Graaf-Van Bloois, L., Wagenaar, J. A. & Zomer, A. L. RFPlasmid: predicting plasmid sequences from short-read assembly data using machine learning. *Microb Genom* **7**, 000683 (2021).
30. *NCBI BioProject.* https://identifiers.org/ncbi/bioproject:PRJNA960711 (2023).
31. *NCBI Sequence Read Archive.* https://identifiers.org/ncbi/insdc.sra:SRP485167 (2023).
32. *Zenodo.* https://doi.org/10.5281/zenodo.10907189.

Acknowledgements

This project was supported by the Danish National Research Foundation (DNRF137) for the Center for Microbial Secondary Metabolites, and Novo Nordisk Foundation within the INTERACT project of the Collaborative Crop Resiliency Program (NNF19SA0059360). TW acknowledges funding from the Novo Nordisk Foundation (NNF20CC0035580).

Author contributions

L.S. performed MGISEQ and Nanopore sequencing, analysis of genomes, interpreted the data, and wrote the manuscript. L.J.D.N. performed Nanopore sequencing, assembly, and analysis of genomes, interpreted the data, and wrote the manuscript. X.X. provided bacterial isolates, performed 16S rRNA gene sequencing, preliminary 16S rRNA-based phylotaxonomics, data visualization and helped to write the manuscript. O.S.M. helped with data analysis, contributed with BGCFlow, and helped to write the manuscript. M.N. contributed with BGCFlow and helped to write the manuscript. Z.X. provided new bacterial isolates and helped to write the manuscript. R.M. provided new bacterial isolates and helped to write the manuscript. K.B. provided new bacterial isolates and helped to write the manuscript. M.P. provided new bacterial isolates and helped to write the manuscript. M.H.A. provided new bacterial isolates and helped to write the manuscript. E.C.S. provided new bacterial isolates and helped to write the manuscript. T.W. conceived and supervised the project, contributed with BGCFlow, and wrote the manuscript. Á.T.K., conceived and supervised the project, and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-024-04118-x.

Correspondence and requests for materials should be addressed to T.W. or Á.T.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.