scientific data

OPEN

DATA DESCRIPTOR

De-novo assembly of 82 bacterial genomes using Nanopore sequencing and prediction of biosynthetic capacity

Dimitra Basdani[1,5], Sokratis Zekkas[1,5], Athanasios Kylonis[1], Dimosthenis Tzimotoudis[1], Giannoulis Fakis[1], Tamás Felföldi[2,3], Károly Márialigeti[2] & Sotiria Boukouvala[1,4] ✉

Historically, antibiotic development has been driven by research into secondary metabolism which is responsible for the biosynthesis of so-called natural products. This research has been revolutionized by our growing capacity to rapidly generate and analyze prokaryotic genome sequences, facilitating the identification of biosynthetic gene clusters (BGCs) involved in secondary metabolism. We present the genomic analysis of 82 bacterial isolates from our collection, covering a broad taxonomic spectrum. The bulk of analyzed isolates belong to the actinobacterial class of *Actinomycetes*, known for their remarkable ability to generate antimicrobial compounds with diverse chemistries. Whole genome sequencing was performed using the Oxford Nanopore technology and the majority of assembled sequences exhibit over 90% gene completeness, enabling reliable functional annotation. Genome mining analyses revealed variability in BGC content across different phyla, confirming isolates in the *Streptomyces* genus as preeminent producers of secondary metabolites. Utilization of the current dataset holds potential not only for the elucidation of the bacterial biosynthetic machinery, but also for phylogenetic comparison and biotechnological exploitation of the isolates under study.

Background & Summary

Microbes display tremendous biosynthetic capacities, producing a vast spectrum of secondary metabolites in response to environmental changes or interactions with other organisms[1]. Beyond their ecological significance, these compounds, also known as natural products, serve as a rich source for the discovery of drugs, particularly antibiotics[2,3]. Genes responsible for the biosynthesis of a specific secondary metabolite co-localize together in the microbial genome, forming biosynthetic gene clusters (BGCs)[4]. Actinomycetes, in particular, have been well-documented as prime producers of different antibiotic classes, mainly thanks to bacteria in the *Streptomyces* genus that typically harbor multiple (20 to 70) BGCs per genome[5,6]. Considering the abundance of as yet unknown or uncultivated bacterial strains, the fact that the majority of BGCs remain "cryptic" in the absence of specific stimuli, as well as the appreciable diversity of BGC content even at the individual genome level, it is reasonable to assume that a great number of novel bioactive compounds remains untapped[7–9].

The current availability of numerous prokaryotic genomic sequences has fueled genome mining as a more targeted approach to elucidate bacterial biosynthetic capacity, enabling the identification and exploitation of microbial BGCs and significantly enhancing the potential for discovery of novel natural products[10–13]. As expected, the efficiency of this approach is reliant on the quality of the sequenced genome, underscoring the need for contiguous and accurate genome assemblies[14–17]. Short-read sequencing technologies have displayed some limitations in generating complete microbial genome assemblies, due to their difficulty in resolving repetitive and GC-rich regions[18–20]. Conversely, long-read sequencing technologies are reported to improve bacterial genome assembly continuity by enabling the reconstruction of complex regions and by bridging the gaps between contigs[21–24]. Although long-read technologies exhibit low base accuracy compared with short-read

[1]Department of Molecular Biology and Genetics, Democritus University of Thrace, Alexandroupolis, Greece. [2]Department of Microbiology, ELTE Eötvös Loránd University, Budapest, Hungary. [3]Institute of Aquatic Ecology, HUN-REN Centre for Ecological Research, Budapest, Hungary. [4]Hellenic Pasteur Institute, Athens, Greece. [5]These authors contributed equally: Dimitra Basdani, Sokratis Zekkas. ✉e-mail: sboukouv@mbg.duth.gr

technologies, this disadvantage can be compensated for by production of high-coverage sequencing data[25,26]. Error-prone reads can also be resolved through the implementation of various read correction tools[27,28]. Such tools polish the bacterial genome assembly by performing raw read self-correction, thereby reducing sequencing errors and significantly improving the quality of the final genomic assembly. Another strategy to ameliorate quality of the sequencing output combines the high accuracy of short-read data with the sequence continuity of long-read data, thus producing hybrid genomic assemblies[29–35].

Given the undeniably higher cost and complexity of combining different sequencing technologies to generate hybrid assemblies, efforts have focused on maximizing the capacity of long-read methods to generate low-complexity prokaryotic genome sequences as accurately as possible[15,22,26,36–39]. As reported, recent advances in Oxford Nanopore flowcell R10.4 have enabled the generation of reliable, near-complete, bacterial genome assemblies of 40x coverage without the need for short-read sequencing[40,41]. Furthermore, investigators have reported accurate identification of antimicrobial resistance and virulence traits of pathogens by utilizing Nanopore-only data[42,43].

In an effort to further explore bacterial biosynthetic capacity, this study presents sequencing of 82 new genomes from our in-house collection of isolates spanning a wide taxonomic range, particularly species of the *Streptomyces* genus. Since information about BGCs can be more efficiently extracted from long-read sequencing data[14,15,44], we employed Oxford Nanopore sequencing for this study and the presented *de-novo* assemblies were based solely on data from this technology. To evaluate the completeness and accuracy of our genomic assemblies, various computational tools were employed, while all bacterial genomes were finally computationally screened for their BGC content. The introduced dataset is anticipated to serve as a useful resource to scientists who wish to investigate the broad biosynthetic diversity among these isolates, identifying potentially robust producers of secondary metabolites, including antibiotics and other bioactive molecules. It is further anticipated that the generated genomic sequences will more broadly facilitate research into the biology of these particular microorganisms, available to the community as pure strains from our in-house collection.

Methods

Microbiological material. Bacterial isolates were collected and microbiologically characterized at the Department of Microbiology, ELTE, Budapest, Hungary. They represent the phyla of *Pseudomonadota* (proteobacteria), *Actinomycetota* (actinobacteria) and *Bacillota* (*Firmicutes*), as our dataset aimed to represent a wide taxonomic spectrum to enable comprehensive analyses across different bacterial lineages (Fig. 1). Strains with laboratory serial numbers preceded by the symbol "#" are available as frozen monocultures and have been described previously[45]. *Streptomyces* strains with laboratory serial numbers preceded by the letter "M" were originally archived as lyophilized stocks in the early 1990s, prepared by late ELTE Professor István Mihály Szabó who isolated them from soil in Hungary. All strains are also maintained as frozen (−80 °C, 25% v/v glycerol) stocks in the corresponding author's laboratory in Alexandroupolis. The strains are presented in Fig. 1 and excel files labelled "Strain Genomes File" and "Strain 16S rRNA File" available on Figshare[46].

Starter liquid cultures were prepared from stocks by inoculating in 5 ml of Nutrient Broth No. 4 (NB, Sigma-Aldrich/Merck, Darmstadt, Germany) at 28 °C without shaking until bacterial growth was observed. Following two consecutive passages on the appropriate agar media (see "Strain Genomes File")[46] to ensure purity of each strain, individual colonies were retrieved and used to inoculate 5 ml of fresh NB medium. Following incubation (28 °C, shaking at 180–200 rpm), bacterial cultures were ready for DNA extraction.

DNA extraction and quality assessment. High-molecular-weight genomic DNA was extracted using the MagAttract HMW DNA Kit (Qiagen, Hilden, Germany) following the manufacturer's instructions for lysis of Gram-negative or Gram-positive bacteria. DNA size distribution and minimal obvious degradation were confirmed by gel electrophoresis (0.7% w/v agarose stained with ethidium bromide). DNA concentration and purity were initially assessed using Nanodrop 2000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). To enhance purity and deplete short DNA fragments, size-selection was carried out using 1:1 v/v solid-phase reversible immobilization (SPRI) magnetic beads (Mag-Bind TotalPure NGS by Omega Bio-tek, Norcross, GA, USA). Size-selected DNA preparations were then re-assessed on the Nanodrop and further quantified with a Qubit 3.0 fluorimeter, using the recommended Qubit dsDNA BR assay kit (both from Thermo Fisher Scientific). For quality of individual preparations, see "Strain Genomes File"[46].

Molecular phylogenetic classification. Molecular phylogenetic classification was already available for 37 bacterial isolates (those assigned serial numbers starting with #)[45], and those 16S rRNA gene sequences can be accessed through European Nucleotide Archive (ENA) Project ID: PRJEB47054[47].

Classification by 16S rRNA gene sequencing was also performed for an additional 49 bacterial isolates (see "Strain 16S rRNA File")[46]. PCR amplification with universal primers 27 F (5′-AGAGTTTGATCCTGGCTCAG-3′) and 1492 R (5′-TACGGYTACCTTGTTACGACTT-3′)[48] was carried out as follows: each reaction in a final volume of 25 µl contained 1 U DreamTaq DNA Polymerase in 1x DreamTaq Buffer, 0.2 mM dNTP mixture, 10 µg of bovine serum albumin (all reagents from Thermo Fisher Scientific), 0.325 µM of each primer and 2 µl of genomic DNA. The reactions were carried out under the following conditions: 98 °C for 5 min, followed by 32 amplification cycles (94 °C for 30 s; 48 °C annealing temperature for 30 s; 72 °C for 60 s) and a final extension step at 72 °C for 10 min. Automated Sanger sequencing of purified amplification products was facilitated by LGC Genomics (Berlin, Germany). Sequences were assigned unique accession numbers and deposited to the ENA with Project ID: PRJEB71440[49] (see "Strain 16S rRNA File")[46].

For taxonomic identification of bacterial isolates, the EzBioCloud database[50,51] was used to analyze all generated 16S rRNA gene sequences and determine the closest related species/strain that matched each one of our investigated isolates (see "Strain 16S rRNA File")[46]. The Contamination Estimator by 16S (ContEst16S)

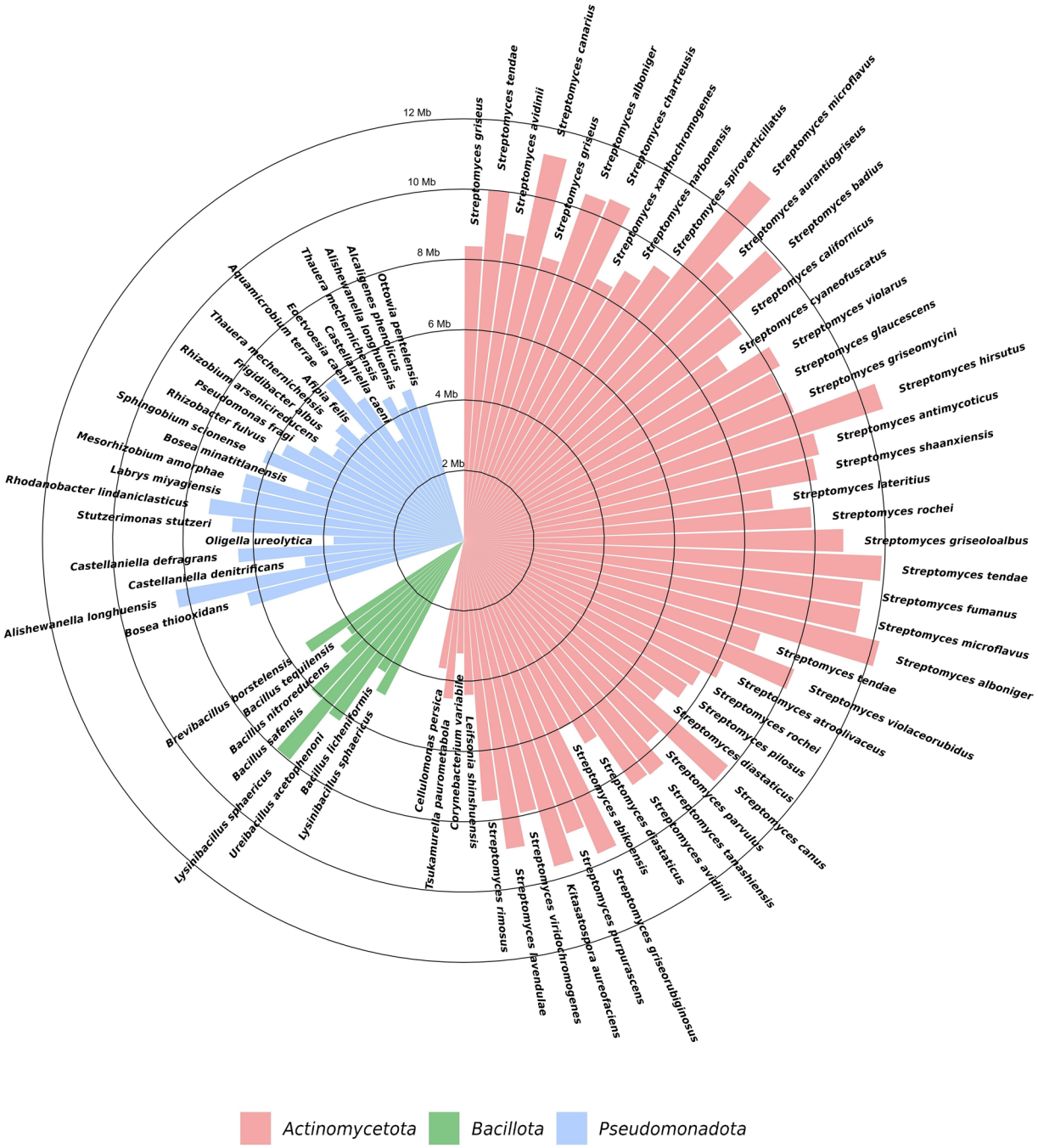

Fig. 1 Phylum-level distribution of sequenced bacterial strains. Each phylum is represented by a different bar color; concentric circles indicate 2 Mb intervals.

tool[52,53] was subsequently utilized to confirm purity of the strains and exclude those where 16S rRNA gene sequences indicated the presence of contamination. Assigned taxonomy identifiers (TAXIDs) are according to the Taxonomy database[54,55], while species names are according to the List of Prokaryotic Names with Standing in Nomenclature[56], as applicable at the time (Nov.–Dec. 2024).

Oxford Nanopore sequencing. Library preparation was performed using the Native Barcoding Kit 24 V14 (SQK-NBD114.24, Oxford Nanopore Technologies, Oxford, UK) according to the manufacturer's recommended protocol. Briefly, for library preparation, DNA samples of similar concentration and purity were grouped together, so as to favor more balanced occupancy of the nanopores by DNA fragments during sequencing. Each run involved multiplexing of 24 barcoded DNA samples, as instructed by the manufacturer. For the clean-up step after adapter ligation, the Long Fragment Buffer was used to enrich for DNA fragments of 3 kb or longer. Libraries were loaded onto R.10.4 flowcells for sequencing using the MinION Mk1B device (Oxford Nanopore Technologies). Four separate sequencing experiments were performed, during which raw data acquisition and

basecalling were carried out simultaneously with sequencing, using the MinKNOW software incorporating Guppy, enabling real-time data processing and demultiplexing. The first two sequencing experiments were controlled by MinKNOW version v23.04.6, while the software was updated to v23.07.12 for the third and fourth run.

Read processing. Each sequencing run generated 24 directories corresponding to each demultiplexed barcode, containing raw reads stored as multiple.fast5 files. To achieve the highest possible raw read accuracy, the Super accurate (SUP) model of Guppy software v6.5.7[57] (Oxford Nanopore Technologies) was utilized, employing the "dna_r10.4.1_e8.2_400bps_sup.cfg" configuration file. Quality control assessments were conducted using MinIONQC v1.4.1[58,59], FastQC v0.12.1[60,61], and Nanoplot v1.41.6[62,63], with default parameters, to evaluate read quality and detect any potential issues. Adapters were trimmed using Porechop v0.2.4[64] with default settings. The reads were filtered with Filtlong v0.2.1[65] to keep 90% of the longest and most accurate reads. Then, an additional round of quality control was performed using the same tools.

***De novo* assembly pipeline and evaluation.** Assembly of all quality-checked Nanopore sequencing reads was performed using Flye v2.9.2[66,67] with the "–nano-raw" option and default parameters. The Flye assemblies were subsequently polished using Medaka v2.0.1[68] (Oxford Nanopore Technologies) with model "r1041_e82_400bps_sup_v5.0.0" and all QC-passed sequencing reads per barcode. Assembly quality assessment was conducted using Quast v5.2.0[69,70], with default parameters, to evaluate assembly quality metrics such as N50, L50, and GC content. Genome coverage statistics were evaluated using bamdst v1.0.9[71]. The completeness of genome assemblies was assessed using Benchmarking Universal Single-copy Orthologs (BUSCO) v5.8.1[72–74] with default parameters and the "–lineage_dataset" option enabled for order-specific analysis.

To streamline the analysis process and ensure reproducibility, all tools were integrated into a single Python script, except for bamdst and BUSCO. To avoid conflicts among different libraries and software dependencies, all software was containerized using Anaconda[75] virtual environments, creating isolated environments for each tool. This approach ensured robust analysis pipelines, allowing seamless integration and execution of the various bioinformatics tools used in this study.

Genome annotation and investigation of bacterial secondary metabolism. Genome assemblies were annotated using Prokka v1.14.6[76,77], employing a protein dataset for each sample derived from the most closely related reference proteome (obtained from UniProt and NCBI databases). Prokka annotation process was performed using default parameters along with the "–proteins" and "–addgenes" options.

The bacterial genome assemblies were computationally analyzed using antiSMASH v7.1.0[78,79] to assess the biosynthetic capacity of investigated bacterial strains via prediction of putative BGCs. The analyses were conducted enabling features ClusterBlast and Minimum Information about a Biosynthetic Gene cluster (MIBiG) cluster comparison.

Data Records

The whole genome sequences of 82 bacterial isolates were deposited in the ENA database with Project ID: PRJEB71438[80] (see "Strain Genomes File" for individual accession numbers per strain)[46]. The 16S rRNA gene sequences for streptomyces strains reported in this study were also deposited in the ENA database with Project ID: PRJEB71440[49] (see "Strain 16S rRNA File" for individual accession numbers per strain)[46]. Details of bacterial isolates and their molecular classification are provided in the "Strain Genomes File" and "Strain 16S rRNA File"[46]. The results of antiSMASH analyses are presented in "Strain BGCs File"[46]. All accompanying material associated with this manuscript is available on Figshare[46]. All other data and results are included in the main manuscript.

Technical Validation

The reconstruction and functional characterization of bacterial genomic sequences is pivotal for understanding the bacterial biosynthetic machinery. Such information can be exploited to advance scientific progress, including pharmaceutical development. Since our primary motivation for this study was to explore bacterial BGCs involved in secondary metabolism, a subset of our in-house collection of bacterial isolates was subjected to whole-genome sequencing analysis using Nanopore technology. The collection displays diversity in the phylogeny of selected isolates, but we prioritized the sequencing of *Streptomyces* species considering their enhanced biosynthetic capabilities. Initially, the sequenced subset included a total of 96 bacterial isolates from our collection, however, 14 of those were abandoned (and are not reported here), either because the collected data was insufficient, or because the analysis of 16S rRNA gene sequences with the ContEst16S tool on EzBioCloud demonstrated that the sequencing output contained genomic segments of different bacterial origin, indicating contamination of the original culture. This highlights the importance of incorporating a contamination check in the sequencing workflow, ensuring the purity of starting cultures and DNA samples. The 82 genomic assemblies that passed those initial tests are reported here and encompass 50 *Actinomycetota*, 24 *Pseudomonadota* and 8 *Bacillota* (Fig. 1 and "Strain Genomes File"[46]).

We performed *de-novo* assembly of bacterial genomes, facilitated by the use of Nanopore long-read sequencing technology, as this circumvents some limitations of reference-based approaches. Specifically, reference-based methods may exclude sequence reads from the strain under study, if these do not align with the available reference genome, potentially overlooking genomic regions important for capturing microbial diversity[22]. The assemblies were generated using Flye, a high-performance assembler for Nanopore long-read data. While we did not perform an exhaustive benchmark of all available assembly tools, our pipeline followed the established best practices for the technology platform at the time of the study. Moreover, depositing of the raw reads to the ENA database ensures that, as bioinformatics tools and assembly algorithms continue to evolve, the community

Barcoded samples	Species name	Assembly length (bp)	Number of contigs	GC (%)	N50 (bp)	Average depth (x)
1_nb01	*Bosea thiooxidans*	6,349,271	256	67.37	120,849	24.65
1_nb03	*Lysinibacillus sphaericus*	4,935,994	140	37.55	227,665	19.14
1_nb05	*Bacillus licheniformis*	4,364,058	33	45.84	1,326,042	48.26
1_nb06	*Ureibacillus acetophenoni*	6,229,690	195	34.99	62,021	43.67
1_nb07	*Alishewanella longhuensis*	8,329,063	343	47.22	42,005	21.32
1_nb09	*Castellaniella denitrificans*	4,565,760	148	67.84	96,483	68.25
1_nb10	*Castellaniella defragrans*	6,436,768	249	63.73	61,759	34.17
1_nb11	*Oligella ureolytica*	3,710,395	137	43.70	63,017	9.94
1_nb12	*Stutzerimonas stutzeri*	6,609,220	256	64.11	67,114	14.68
1_nb13	*Rhodanobacter lindaniclasticus*	7,309,924	304	67.16	46,451	18.15
1_nb14	*Streptomyces griseus*	8,377,519	118	72.30	1,533,214	28.94
1_nb16	*Streptomyces tendae*	9,986,999	387	72.10	120,464	18.63
1_nb17	*Streptomyces avidinii*	8,824,473	260	71.80	282,487	18.47
1_nb18	*Streptomyces canarius*	11,234,897	396	71.92	142,725	16.67
1_nb19	*Streptomyces griseus*	8,381,104	133	71.71	527,544	37.80
1_nb20	*Streptomyces alboniger*	10,458,132	443	71.76	74,823	13.48
1_nb21	*Streptomyces chartreusis*	10,572,294	45	70.79	2,955,546	45.64
1_nb22	*Streptomyces xanthochromogenes*	8,355,972	65	71.28	2,428,440	46.65
1_nb23	*Streptomyces narbonensis*	8,940,517	338	72.06	143,075	23.00
1_nb24	*Streptomyces spiroverticillatus*	9,516,703	98	71.20	544,295	40.23
2_nb01	*Lysinibacillus sphaericus*	7,961,604	396	37.25	36,231	11.57
2_nb02	*Bacillus safensis*	6,058,678	321	41.83	36,550	17.97
2_nb04	*Streptomyces microflavus*	13,075,107	672	71.23	50,866	13.67
2_nb05	*Streptomyces aurantiogriseus*	10,721,168	232	70.93	2,089,405	42.30
2_nb06	*Streptomyces badius*	11,908,868	585	71.65	43,706	13.03
2_nb07	*Streptomyces californicus*	9,863,829	432	72.65	43,000	6.99
2_nb08	*Streptomyces cyaneofuscatus*	8,933,790	480	71.53	60,290	15.37
2_nb09	*Streptomyces violarus*	10,285,877	633	71.18	42,177	9.19
2_nb10	*Streptomyces glaucescens*	9,965,271	557	72.27	44,522	8.11
2_nb11	*Streptomyces griseomycini*	10,083,991	509	72.65	99,455	17.31
2_nb12	*Streptomyces hirsutus*	12,515,227	545	71.13	49,492	8.52
2_nb13	*Streptomyces antimycoticus*	10,408,690	408	70.95	56,074	6.83
2_nb14	*Streptomyces shaanxiensis*	10,208,071	363	71.12	213,402	26.39
2_nb15	*Streptomyces lateritius*	8,855,071	420	71.33	39,910	7.23
2_nb16	*Streptomyces rochei*	9,912,515	573	72.56	64,306	13.57
2_nb17	*Streptomyces griseoloalbus*	10,807,963	569	72.59	41,114	12.55
2_nb18	*Streptomyces tendae*	11,908,553	681	72.47	42,905	13.14
2_nb19	*Streptomyces fumanus*	11,427,027	562	73.01	38,469	9.16
2_nb20	*Streptomyces microflavus*	11,451,493	448	71.28	43,483	7.62
2_nb22	*Streptomyces alboniger*	12,178,729	638	70.12	30,785	7.84
2_nb23	*Streptomyces tendae*	8,847,876	335	72.04	45,050	7.22
2_nb24	*Streptomyces violaceorubidus*	10,097,966	441	72.34	39,630	7.98

Table 1. Quality assessment metrics of 42 barcoded *de novo* genome assemblies generated from sequencing runs 1 and 2. As discussed in the Technical Validation section, assemblies generated from sequencing runs 1 and 2 were more fragmented, due to technical issues of MinKNOW v23.04.6 which were resolved after updating to version v23.07.12 of the software. Lower sequencing coverage was also attributed to inconsistent performance of the MinION and pore decline in reused flowcells.

will be able to perform more sophisticated re-analyses or hybrid assemblies to further improve these results in the future.

The output of the sequencing process (Tables 1, 2; Figs. 2, 3) was subjected to quality assessment per DNA sample, examining parameters such as the number of contigs, N50, Q-score and genome coverage for all sequencing runs. Regarding the number of contigs, 42.6% of the bacterial genome assemblies contain less than 20 contigs, while approximately 24% were assembled into 1 to 3 contigs. A greater number of contigs was observed for assemblies generated from the first and second sequencing runs (Table 1). This was caused by multiple interruptions of the sequencing device during prolonged (up to 36 h) continuous usage, an inherent flaw that MinKNOW v23.04.6 software reports attributed to device errors, such as failed script or data transfer from the device, even though hardware check was successfully passed during experiments. Those issues were

Barcoded samples	Species name	Assembly length (bp)	Number of contigs	GC (%)	N50 (bp)	Average depth (x)
3_nb01	*Labrys miyagiensis*	6,451,965	1	63.68	6,451,965	61.31
3_nb02	*Mesorhizobium amorphae*	6,475,598	1	62.07	6,475,598	82.44
3_nb03	*Bosea minatitlanensis*	4,732,811	1	68.57	4,732,811	102.87
3_nb05	*Sphingobium scionense*	6,142,515	9	64.00	4,819,090	53.97
3_nb06	*Rhizobacter fulvus*	5,725,747	1	68.59	5,725,747	62.13
3_nb07	*Pseudomonas fragi*	5,054,805	12	59.32	2,861,027	48.50
3_nb08	*Bacillus nitroreducens*	4,617,405	4	37.45	3,812,000	109.94
3_nb09	*Rhizobium arsenicireducens*	4,439,158	3	62.15	4,285,843	167.69
3_nb10	*Frigidibacter albus*	4,556,185	6	66.91	4,350,424	36.23
3_nb11	*Thauera mechernichensis*	4,805,311	5	65.21	4,283,565	88.78
3_nb12	*Afipia felis*	4,130,270	4	62.84	3,880,815	118.52
3_nb13	*Aquamicrobium terrae*	5,911,167	10	65.23	2,273,974	78.54
3_nb14	*Eoetvoesia caeni*	4,953,590	2	58.42	4,715,377	30.79
3_nb15	*Castellaniella caeni*	3,389,029	1	64.70	3,389,029	181.01
3_nb17	*Streptomyces atroolivaceus*	8,170,221	2	70.70	8,046,083	37.09
3_nb18	*Streptomyces rochei*	7,700,368	17	72.57	5,336,130	21.76
3_nb19	*Streptomyces pilosus*	7,567,575	2	72.44	7,309,760	47.28
3_nb20	*Streptomyces diastaticus*	7,070,005	4	73.38	6,616,679	64.87
3_nb21	*Streptomyces canus*	9,876,360	2	70.30	8,721,601	31.71
3_nb22	*Streptomyces parvulus*	7,765,051	2	72.74	7,149,483	18.74
3_nb23	*Streptomyces tanashiensis*	8,510,880	2	72.34	7,250,060	35.36
3_nb24	*Streptomyces avidinii*	8,455,439	2	72.31	8,451,836	57.30
4_nb01	*Streptomyces diastaticus*	6,729,465	8	73.51	4,488,046	14.16
4_nb02	*Streptomyces abikoensis*	6,029,806	5	72.72	5,371,170	97.29
4_nb04	*Streptomyces griseorubiginosus*	9,693,545	74	70.89	311,467	12.14
4_nb05	*Streptomyces purpurascens*	8,834,081	162	71.06	88,743	6.86
4_nb06	*Kitasatospora aureofaciens*	9,623,598	66	72.52	548,223	15.72
4_nb07	*Streptomyces viridochromogenes*	7,910,144	6	72.02	7,030,946	30.52
4_nb08	*Streptomyces lavendulae*	8,858,801	20	72.66	1,269,958	23.57
4_nb09	*Streptomyces rimosus*	7,436,011	140	71.84	85,432	5.20
4_nb11	*Leifsonia shinshuensis*	4,389,968	3	71.05	4,235,312	14.01
4_nb13	*Corynebacterium variabile*	3,207,583	19	67.33	240,792	9.87
4_nb14	*Tsukamurella paurometabola*	4,510,532	27	68.46	448,140	9.83
4_nb15	*Cellulomonas persica*	3,677,918	1	73.54	3,677,918	31.99
4_nb16	*Bacillus tequilensis*	4,149,298	2	43.76	4,145,740	328.68
4_nb17	*Brevibacillus borstelensis*	5,366,844	1	51.69	5,366,844	79.49
4_nb18	*Thauera mechernichensis*	4,599,847	2	65.55	4,426,618	66.74
4_nb19	*Alishewanella longhuensis*	4,147,163	2	47.17	4,143,592	42.72
4_nb20	*Alcaligenes phenolicus*	4,568,329	5	56.87	4,414,899	90.76
4_nb21	*Ottowia pentelensis*	4,007,106	1	68.43	4,007,106	142.74

Table 2. Quality assessment metrics of 40 barcoded *de novo* genome assemblies generated from sequencing runs 3 and 4. Samples generating genome assemblies with <20 contigs and/or >30x coverage are shown in bold.

resolved with the updated version v23.07.12 of the MinKNOW software that was used during the third and fourth sequencing runs (Table 2).

Similar observations, i.e. regarding the continuity of assemblies, were made when evaluating the N50 metric. Specifically, the third sequencing run yields an overall N50 value of 2 to 8 Mbp, which (considering the small size of bacterial genomes) is indicative of highly contiguous assemblies. A greater diversity in N50 values is observed for the fourth run, with the majority of barcodes hovering at around 4 Mbp (Table 2). In contrast, substantially lower N50 values are reported for the first two runs. Specifically, in the first sequencing run, N50 values fall within the range of hundreds of thousands base pairs, suggesting a minor yet not concerning decline in assembly quality. However, in the case of the second run, most N50 values range around 50 Kbp, indicating a significant drop in assembly continuity (Table 1). Such limitations of the technology are important to keep in mind, for example, when interpreting experiments conducted with older MinKNOW versions.

To assess basecalling accuracy, Q-score metrics were examined (Fig. 2). Approximately 90% of the total reads of each barcode had a quality score greater than 15, signifying an accuracy of over 96.84%. A rather lower base accuracy relative to short-read sequencing was anticipated, considering that Nanopore is an error-prone technology and no short-read polishing was applied.

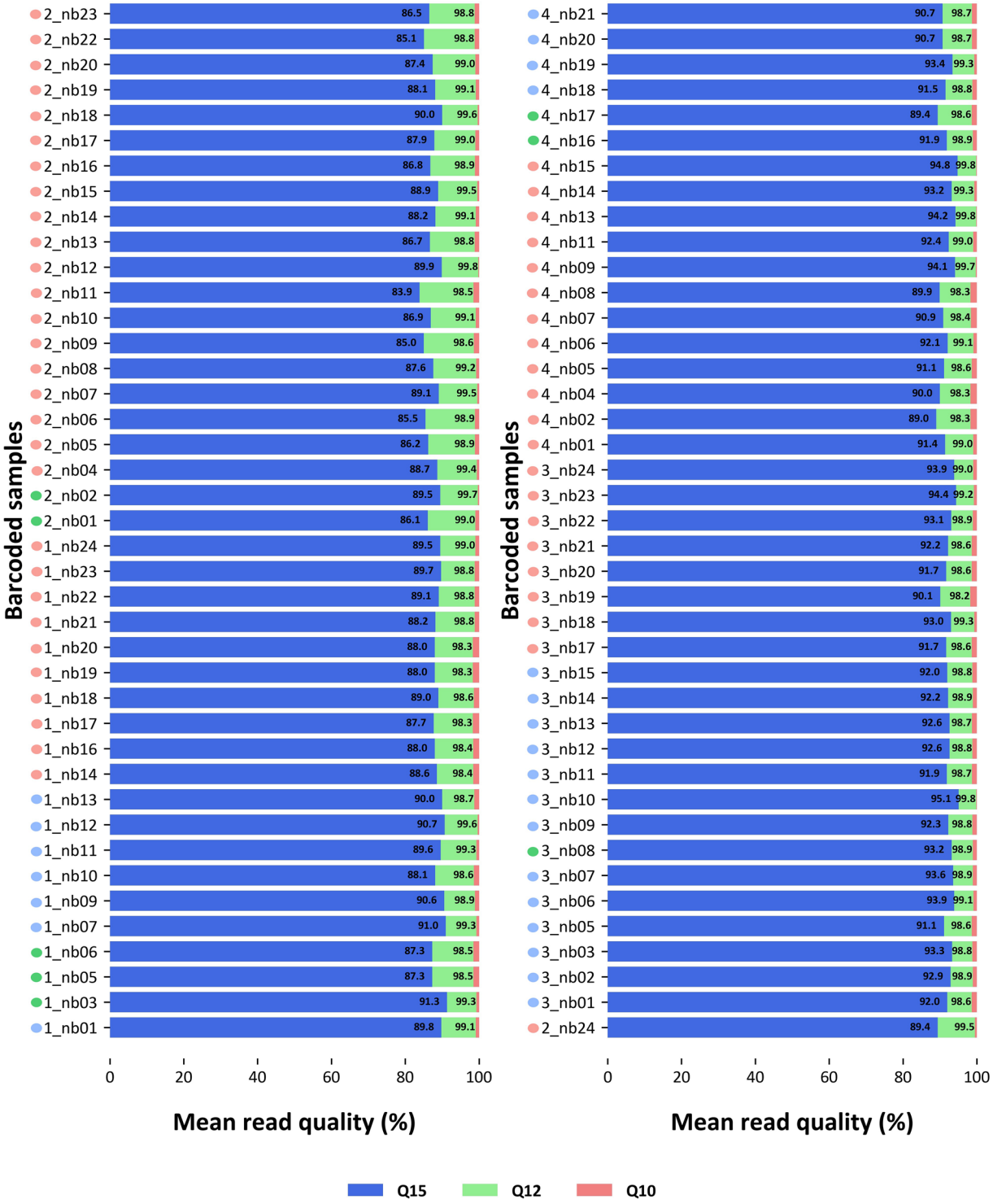

Fig. 2 Distribution of mean read quality across 82 bacterial genome assemblies. Each bar is divided into three segments representing the proportion of reads that satisfy the Q15 (blue), Q12 (green), or Q10 (coral) quality threshold; colored dots next to the barcodes indicate phyla (as in Fig. 1).

Another important quality metric is the average sequencing depth, which could mitigate the low-per-base accuracy that is an inherent limitation of Nanopore sequencing. In this study, around 46% of the genome assemblies have a coverage exceeding 30x, with 7 assemblies surpassing 100x coverage. No correlation between the sequencing coverage and the length or GC content of genome assemblies was observed (Tables 1, 2).

The completeness of the bacterial sequences was evaluated with BUSCO analysis, using the corresponding order-level dataset for every species (Fig. 3). For non-model genomes, the accepted BUSCO score ranges from 50% to 95% completeness, contingent upon the biological complexity of each species. In the current study, 63

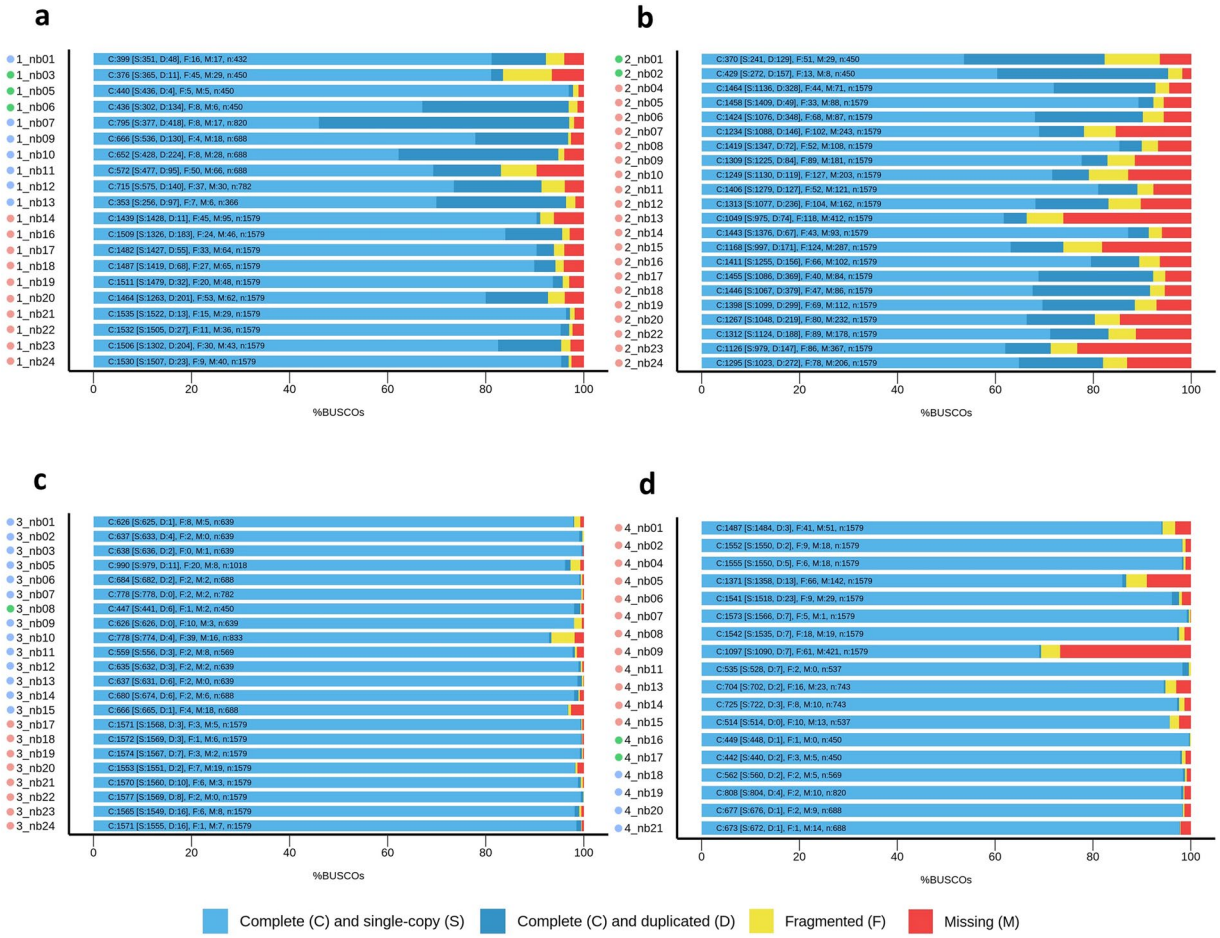

Fig. 3 Evaluation of genome assembly completeness by BUSCO analysis. Assemblies were analyzed employing the closest available lineage for each species. Panels a-d represent barcoded samples across four sequencing runs; colored dots next to the barcodes indicate phyla (as in Fig. 1).

assembled genomes exceeded the threshold of 90% completeness, with the highest levels recorded during the third sequencing run. No sample ranked below 50% completeness. On the other hand, the relatively higher rate of missing, fragmented, or complete and duplicated BUSCO genes in the first two runs, can serve as an indicator of issues during the sequencing or assembly processes[74], discussed above.

Through the Prokka whole genome annotation process, it was possible to determine the number of coding (CDS), rRNA and tRNA genomic sequences per assembly (Table 3). Out of 82 genome assemblies, annotation was successfully applied to 79, as a reference genome, proteome or both were not available for the remaining 3 samples. The 79 annotated genomes contain an average of 7,417 CDS, 16 rRNA, and 87 tRNA sequences. A surge in the number of predicted genes observed in the second sequencing run could be due to possible overestimation, considering the high fragmentation of the genome assembly leading to the scattering of genes across multiple contigs[81].

Overviewing the quality assessment metrics mentioned above, several factors should be taken into account when troubleshooting Nanopore sequencing which remains a less reliable technology compared with more established short-read platforms. Firstly, sequencing performance is largely dependent on the quantity and quality of the starting genetic material used[82]. Given that the four sequencing runs differed in the concentration and purity of the bacterial DNA preparations pooled together in the same library, variations in the sequencing throughput were anticipated across runs. Specifically, the first and third runs were performed using DNA of highest quality in terms of quantity and purity, while the second and fourth runs were carried out with DNA preparations of less optimal quality (see "Strain Genomes File")[46]. It must be stressed, however, that variability in DNA preparation quality was largely due to inherent biological properties of different bacterial strains, affecting the concentration and purity of recovered DNA, rather than its integrity which remained high (see "Strain Genomes File")[46]. For example, gamma-proteobacteria typically reach high cell densities quickly, allowing for higher prep concentrations compared to alpha-proteobacteria that grow more slowly and to lower densities; Gram-negative cells are typically easier to lyse compared with Gram-positive cells; and the streptomycetes are Gram-positive bacteria, typically growing in tight clumps that are difficult to disperse during lysis.

Secondly, it should be noted that, in line with the manufacturer's specifications, two flowcells were used, each one for two 36-hour runs. Consequently, it is reasonable to expect a decline in sequencing throughput, as

Barcoded sample	CDS	rRNA	tRNA	Barcoded sample	CDS	rRNA	tRNA
1_nb01	6,264	6	62	2_nb24	10,505	16	109
1_nb03	5,410	27	101	3_nb01	5,951	6	50
1_nb05	4,536	15	80	3_nb02	6,227	6	48
1_nb06	6,729	64	125	3_nb03	4,509	6	58
1_nb07	8,671	29	151	3_nb05	5,932	12	63
1_nb09	4,168	6	64	3_nb06	5,290	3	53
1_nb10	6,316	10	71	3_nb07	4,563	4	69
1_nb11	4,591	7	54	3_nb10	4,430	6	53
1_nb12	7,008	15	92	3_nb11	4,381	12	64
1_nb13	7,440	10	75	3_nb12	3,910	3	51
1_nb14	7,262	18	78	3_nb13	5,756	3	50
1_nb16	9,512	21	95	3_nb14	4,669	6	49
1_nb17	8,205	19	88	3_nb15	3,125	9	56
1_nb18	10,220	10	99	3_nb17	7,051	18	75
1_nb19	7,412	23	101	3_nb18	6,771	18	89
1_nb20	10,079	17	100	3_nb19	6,660	18	79
1_nb21	9,238	16	86	3_nb20	5,926	21	95
1_nb22	7,356	25	82	3_nb21	8,853	18	83
1_nb23	8,268	26	93	3_nb22	6,867	18	78
1_nb24	8,443	24	91	3_nb23	7,497	21	85
2_nb01	10,138	35	165	3_nb24	7,369	21	93
2_nb02	7,075	38	115	4_nb01	5,770	21	93
2_nb04	13,149	22	138	4_nb02	5,277	21	85
2_nb05	9,833	17	91	4_nb04	8,658	15	82
2_nb06	12,193	20	158	4_nb05	8,447	6	82
2_nb07	10,240	20	102	4_nb06	8,415	21	97
2_nb08	8,586	13	88	4_nb07	7,029	18	79
2_nb09	10,447	15	108	4_nb08	8,012	21	95
2_nb10	10,580	15	96	4_nb09	7,319	14	80
2_nb11	9,672	14	105	4_nb11	4,190	3	54
2_nb12	13,273	20	106	4_nb13	3,029	3	60
2_nb13	10,960	16	91	4_nb14	4,416	3	54
2_nb15	9,651	20	91	4_nb15	3,267	6	52
2_nb16	9,437	23	89	4_nb16	4,133	29	86
2_nb17	10,780	15	107	4_nb17	5,168	32	126
2_nb18	11,844	19	142	4_nb18	4,115	12	62
2_nb19	11,352	35	146	4_nb19	3,600	15	75
2_nb20	12,100	30	121	4_nb20	4,228	9	57
2_nb22	12,986	12	102	4_nb21	3,817	6	48
2_nb23	9,418	10	89				

Table 3. Number of coding (CDS), rRNA, and tRNA sequences per barcoded bacterial assembly, as predicted by Prokka genome annotation.

available flowcell pores progressively deteriorated during the second and fourth runs. Indeed, lower sequencing coverage was more likely linked to inconsistent performance of the MinION and pore decline in reused flowcells, although the low concentration of DNA preparations in runs 2 and 4 also contributed. Lastly, as already discussed, it is important to keep in mind that the output of Nanopore sequencing is still heavily reliant on the software performance and its stable connectivity to the sequencing device, factors that may be beyond user's ability to fully control.

To gain insight into the biosynthetic potential of the 82 sequenced bacterial isolates, their genomic assemblies were subjected to antiSMASH analysis, so as to predict the presence of BGCs organized in so-called "Regions" defined by the antiSMASH expert community (Fig. 4 and "Strain BGCs File"[46]). Moderate correlation was observed between the number of contigs per assembly and the corresponding number of antiSMASH predicted Regions (Fig. 5), supporting that, even though assembly fragmentation may influence BGC prediction to some extent, antiSMASH annotations can be considered as reasonably reliable with regard to possible BGC fragmentation during genome assembly. Indeed, upon visual inspection of antiSMASH results for possible BGC disruption, only a limited number of clusters were found to localize at the edges of contigs. An increase in BGC

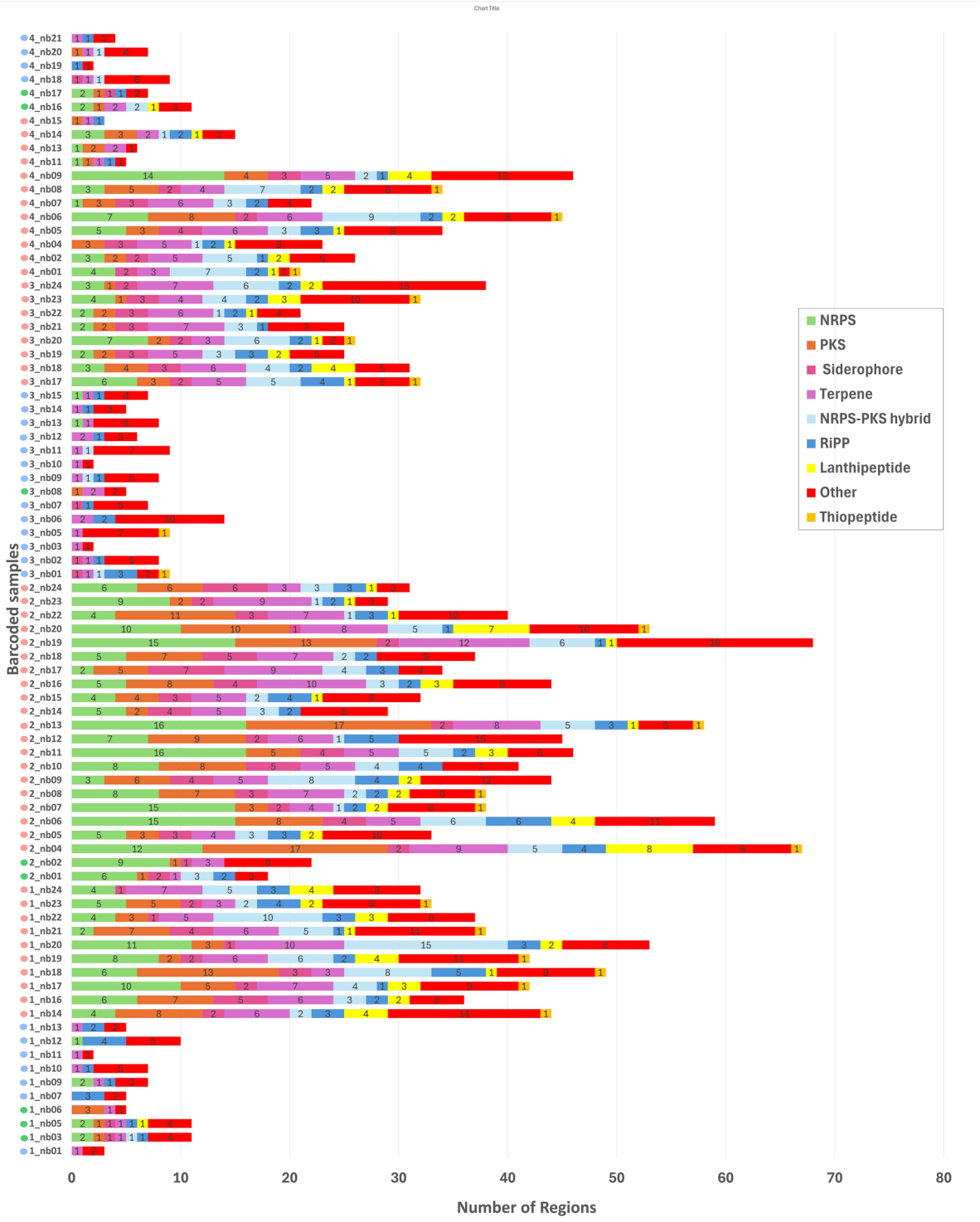

Fig. 4 BGC type distribution per barcoded bacterial sample relative to the total number of predicted BGC Regions. Each BGC type is color-coded and the following abbreviations were used: NRPS for non-ribosomal peptide synthetase, PKS for polyketide synthase and RiPP for ribosomally synthesised and post-translationally modified peptides. Colored dots next to the barcodes indicate phyla (as in Fig. 1).

fragmentation was nevertheless detected in the second sequencing run, and this can be attributed to the high degree of contig fragmentation in the corresponding genome assemblies, as discussed above.

As expected, the BGC content varies according to the taxonomic classification of each bacterial isolate. In line with existing literature, isolates in the *Streptomyces* genus exhibited the greatest number of BGCs compared to all other bacteria, harboring 21–68 BGCs per genome which are predicted to produce a whole range of secondary metabolites belonging to diverse chemical classes. Search of the MIBiG database further demonstrates the

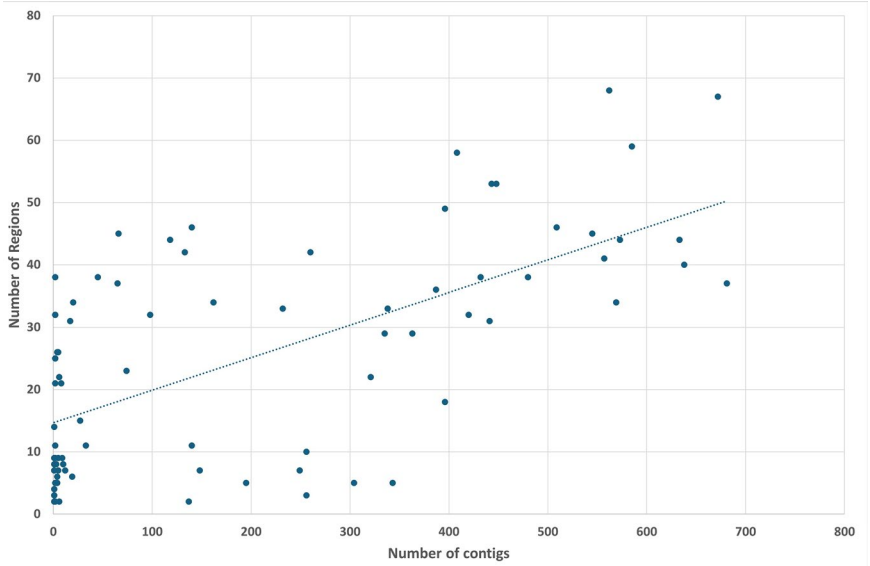

Fig. 5 Correlation between the number of predicted BGC Regions and the number of contigs per bacterial genome assembly ($y = 0.0522x + 14.67$, $R^2 = 0.4186$).

potential medicinal value of the predicted secondary metabolites, many of which are likely to have antimicrobial or antitumor activity. Identifying specific isolates as proficient producers of secondary metabolites through genome mining of putative BGCs may advance their targeted investigation and biotechnological exploitation towards the discovery of new compounds with unprecedented functional and medicinal properties.

Furthermore, the generated dataset is anticipated to more broadly support investigations into the biology of microbes in our collection, advancing academic collaboration and scientific progress. To this end, even sub-optimal assemblies remain valuable as genomic maps for the design of "wet-lab" research involving unique non-model strains in our collection. For example, our group is currently using this genomic dataset to guide whole BGC cloning and heterologous expression of antibiotic secondary metabolites in genetically engineered hosts. We believe that providing genomic assemblies alongside transparent quality measures – and making raw reads available for re-analyses as computational tools evolve – is balancing bioinformatics rigor with "wet-lab" utility, encouraging future reuse of these data associated with our collection of strains.

Data availability

ENA Project ID: PRJEB71438[80] includes the whole-genome sequences of investigated strains. ENA Project ID: PRJEB71440[49] includes the 16S rRNA gene sequences of *Streptomyces* strains reported in this study. Details of bacterial isolates and their molecular classification, as well as the results of antiSMASH analyses are available on Figshare[46]. All other data is presented in the main manuscript.

Code availability

The complete bioinformatic pipeline, including steps for data acquisition, software requirements, and the final Python code, is available through Figshare and GitHub[83,84].

Received: 17 September 2025; Accepted: 31 March 2026;
Published online: 08 April 2026

References

1. Newman, D. J. & Cragg, G. M. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J. Nat. Prod.* **83**, 770–803, https://doi.org/10.1021/acs.jnatprod.9b01285 (2020).
2. Genilloud, O. The re-emerging role of microbial natural products in antibiotic discovery. *Antonie van Leeuwenhoek* **106**, 173–188, https://doi.org/10.1007/s10482-014-0204-6 (2014).
3. Katz, L. & Baltz, R. H. Natural product discovery: past, present, and future. *J. Ind. Microbiol. Biotechnol.* **43**, 155–176, https://doi.org/10.1007/s10295-015-1723-5 (2016).
4. Osbourn, A. Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation. *Trends Genet.* **26**, 449–457, https://doi.org/10.1016/j.tig.2010.07.001 (2010).
5. De Simeis, D. & Serra, S. Actinomycetes: a never-ending source of bioactive compounds – An overview on antibiotics production. *Antibiotics (Basel)* **10**, 483, https://doi.org/10.3390/antibiotics10050483 (2021).
6. Genilloud, O. Actinomycetes: still a source of novel antibiotics. *Nat. Prod. Rep.* **34**, 1203–1232, https://doi.org/10.1039/c7np00026j (2017).
7. Nett, M., Ikeda, H. & Moore, B. S. Genomic basis for natural product biosynthetic diversity in the actinomycetes. *Nat. Prod. Rep.* **26**, 1362–1384, https://doi.org/10.1039/b817069j (2009).
8. Seipke, R. F. Strain-level diversity of secondary metabolism in *Streptomyces albus*. *PLoS One* **10**, e0116457, https://doi.org/10.1371/journal.pone.0116457 (2015).

9. Belknap, K. C., Park, C. J., Barth, B. M. & Andam, C. P. Genome mining of biosynthetic and chemotherapeutic gene clusters in *Streptomyces* bacteria. *Sci. Rep.* **10**, 2003, https://doi.org/10.1038/s41598-020-58904-9 (2020).
10. Bachmann, B. O., Van Lanen, S. G. & Baltz, R. H. Microbial genome mining for accelerated natural products discovery: is a renaissance in the making? *J. Ind. Microbiol. Biotechnol.* **41**, 175–184, https://doi.org/10.1007/s10295-013-1389-9 (2014).
11. Baltz, R. H. Gifted microbes for genome mining and natural product discovery. *J. Ind. Microbiol. Biotechnol.* **44**, 573–588, https://doi.org/10.1007/s10295-016-1815-x (2017).
12. Baral, B., Akhgari, A. & Metsä-Ketelä, M. Activation of microbial secondary metabolic pathways: avenues and challenges. *Synth. Syst. Biotechnol.* **3**, 163–178, https://doi.org/10.1016/j.synbio.2018.09.001 (2018).
13. Baltz, R. H. Natural product drug discovery in the genomic era: realities, conjectures, misconceptions, and opportunities. *J. Ind. Microbiol. Biotechnol.* **46**, 281–299, https://doi.org/10.1007/s10295-018-2115-4 (2019).
14. Klassen, J. L. & Currie, C. R. Gene fragmentation in bacterial draft genomes: extent, consequences and mitigation. *BMC Genomics* **13**, 14, https://doi.org/10.1186/1471-2164-13-14 (2012).
15. Goldstein, S., Beka, L., Graf, J. & Klassen, J. L. Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing. *BMC Genomics* **20**, 23, https://doi.org/10.1186/s12864-018-5381-7 (2019).
16. Salzberg, S. L. Next-generation genome annotation: we still struggle to get it right. *Genome Biol.* **20**, 92, https://doi.org/10.1186/s13059-019-1715-2 (2019).
17. Van Der Hooft, J. J. J. *et al*. Linking genomics and metabolomics to chart specialized metabolic diversity. *Chem. Soc. Rev.* **49**, 3297–3314, https://doi.org/10.1039/d0cs00162g (2020).
18. Chaisson, M., Pevzner, P. & Tang, H. Fragment assembly with short reads. *Bioinformatics* **20**, 2067–2074, https://doi.org/10.1093/bioinformatics/bth205 (2004).
19. Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36**, e105, https://doi.org/10.1093/nar/gkn425 (2008).
20. Cahill, M. J., Köser, C. U., Ross, N. E. & Archer, J. A. C. Read length and repeat resolution: exploring prokaryote genomes using next-generation sequencing technologies. *PLoS One* **5**, e11518, https://doi.org/10.1371/journal.pone.0011518 (2010).
21. Koren, S. *et al*. Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **30**, 693–700, https://doi.org/10.1038/nbt.2280 (2012).
22. Sohn, J. Il. & Nam, J. W. The present and future of *de novo* whole-genome assembly. *Brief. Bioinform.* **19**, 23–40, https://doi.org/10.1093/bib/bbw096 (2018).
23. Acuña-Amador, L., Primot, A., Cadieu, E., Roulet, A. & Barloy-Hubler, F. Genomic repeats, misassembly and reannotation: a case study with long-read resequencing of *Porphyromonas gingivalis* reference strains. *BMC Genomics* **19**, 54, https://doi.org/10.1186/s12864-017-4429-4 (2018).
24. Segerman, B. The most frequently used sequencing technologies and assembly methods in different time segments of the bacterial surveillance and RefSeq genome databases. *Front. Cell. Infect. Microbiol.* **10**, 527102, https://doi.org/10.3389/fcimb.2020.527102 (2020).
25. Baker, M. *De novo* genome assembly: what every biologist should know. *Nat. Methods* **9**, 333–337, https://doi.org/10.1038/nmeth.1935 (2012).
26. Koren, S. *et al*. Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol.* **14**, R101, https://doi.org/10.1186/gb-2013-14-9-r101 (2013).
27. Safar, H. A. *et al*. The impact of applying various *de novo* assembly and correction tools on the identification of genome characterization, drug resistance, and virulence factors of clinical isolates using ONT sequencing. *BMC Biotechnol.* **23**, 26, https://doi.org/10.1186/s12896-023-00797-3 (2023).
28. Watson, M. & Warr, A. Errors in long-read assemblies can critically affect protein prediction. *Nat. Biotechnol.* **37**, 124–126, https://doi.org/10.1038/s41587-018-0004-z (2019).
29. Risse, J. *et al*. A single chromosome assembly of *Bacteroides fragilis* strain BE1 from Illumina and MinION nanopore sequencing data. *Gigascience* **4**, 60, https://doi.org/10.1186/s13742-015-0101-6 (2015).
30. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb. Genom.* **3**, e000132, https://doi.org/10.1099/mgen.0.000132 (2017).
31. Arredondo-Alonso, S. *et al*. A high-throughput multiplexing and selection strategy to complete bacterial genomes. *Gigascience* **10**, giab079, https://doi.org/10.1093/gigascience/giab079 (2021).
32. Lee, N. *et al*. Thirty complete *Streptomyces* genome sequences for mining novel secondary metabolite biosynthetic gene clusters. *Sci. Data* **7**, 55, https://doi.org/10.1038/s41597-020-0395-9 (2020).
33. Lipworth, S. *et al*. Optimized use of Oxford Nanopore flowcells for hybrid assemblies. *Microb. Genom.* **6**, mgen000453, https://doi.org/10.1099/mgen.0.000453 (2020).
34. Chen, Z., Erickson, D. L. & Meng, J. Polishing the Oxford Nanopore long-read assemblies of bacterial pathogens with Illumina short reads to improve genomic analyses. *Genomics* **113**, 1366–1377, https://doi.org/10.1016/j.ygeno.2021.03.018 (2021).
35. Sanderson, N. D. *et al*. Comparison of R9.4.1/Kit10 and R10/Kit12 Oxford Nanopore flowcells and chemistries in bacterial genome reconstruction. *Microb. Genom.* **9**, mgen000910, https://doi.org/10.1099/mgen.0.000910 (2023).
36. Quick, J., Quinlan, A. R. & Loman, N. J. A reference bacterial genome dataset generated on the MinION™ portable single-molecule Nanopore sequencer. *Gigascience* **3**, 22, https://doi.org/10.1186/2047-217X-3-22 (2014).
37. Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled *de novo* using only Nanopore sequencing data. *Nat. Methods* **12**, 733–735, https://doi.org/10.1038/nmeth.3444 (2015).
38. Chakraborty, M., Baldwin-Brown, J. G., Long, A. D. & Emerson, J. J. Contiguous and accurate *de novo* assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res.* **44**, e147, https://doi.org/10.1093/nar/gkw654 (2016).
39. Wick, R. R. & Holt, K. E. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Res.* **8**, 2138, https://doi.org/10.12688/f1000research.21782.4 (2019).
40. Sereika, M. *et al*. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nat. Methods* **19**, 823–826, https://doi.org/10.1038/s41592-022-01539-7 (2022).
41. Sanderson, N. D. *et al*. Evaluation of the accuracy of bacterial genome reconstruction with Oxford Nanopore R10.4.1 long-read-only sequencing. *Microb. Genom.* **10**, 001246, https://doi.org/10.1099/mgen.0.001246 (2024).
42. Zhang, C. *et al*. Determining antimicrobial resistance profiles and identifying novel mutations of *Neisseria gonorrhoeae* genomes obtained by multiplexed MinION sequencing. *Sci. China Life Sci.* **63**, 1063–1070, https://doi.org/10.1007/s11427-019-1558-8 (2020).
43. Foster-Nyarko, E. *et al*. Nanopore-only assemblies for genomic surveillance of the global priority drug-resistant pathogen, *Klebsiella pneumoniae*. *Microb. Genom.* **9**, mgen000936, https://doi.org/10.1099/mgen.0.000936 (2023).
44. Harrison, J. & Studholme, D. J. Recently published *Streptomyces* genome sequences. *Microb. Biotechnol.* **7**, 373–380, https://doi.org/10.1111/1751-7915.12143 (2014).
45. Kontomina, E. *et al*. A taxonomically representative strain collection to explore xenobiotic and secondary metabolism in bacteria. *PLoS One* **17**, e0271125, https://doi.org/10.1371/journal.pone.0271125 (2022).
46. Basdani, D. *et al*. *De-novo* assembly of 82 bacterial genomes using Nanopore sequencing and prediction of biosynthetic capacity (manuscript datasets). *figshare* https://doi.org/10.6084/m9.figshare.30120808 (2026).
47. *European Nucleotide Archive (ENA) Project: PRJEB47054.* https://www.ebi.ac.uk/ena/browser/view/PRJEB47054 (2022).

48. Lane, D.J. 16S/23S rRNA Sequencing. In: Stackebrandt, E. and Goodfellow, M., Eds., *Nucleic Acid Techniques in Bacterial Systematics*, John Wiley and Sons, New York, 115–175 (1991).
49. *European Nucleotide Archive (ENA) Project: PRJEB71440*. https://www.ebi.ac.uk/ena/browser/view/PRJEB71440 (2025).
50. Chalita, M. *et al*. EzBioCloud: a genome-driven database and platform for microbiome identification and discovery. *Int. J. Syst. Evol. Microbiol.* **74**, 006421, https://doi.org/10.1099/ijsem.0.006421 (2024).
51. *EzBioCloud Database*. https://www.ezbiocloud.net/.
52. Lee, I. *et al*. ContEst16S: an algorithm that identifies contaminated prokaryotic genomes using 16S RNA gene sequences. *Int. J. Syst. Evol. Microbiol.* **67**, 2053–2057, https://doi.org/10.1099/ijsem.0.001872 (2017).
53. *Contamination Estimator by 16S (ContEst16S)*. https://www.ezbiocloud.net/tools/contest16s.
54. Schoch, C. L. *et al*. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)* **2020**, baaa062, https://doi.org/10.1093/database/baaa062 (2020).
55. *Taxonomy Database*. https://www.ncbi.nlm.nih.gov/taxonomy.
56. *LPSN Database*. https://lpsn.dsmz.de/.
57. *Guppy software, Oxford Nanopore Technologies*. https://github.com/nanoporetech/pyguppyclient.
58. *MinIONQC software*. https://github.com/roblanf/minion_qc.
59. Lanfear, R., Schalamun, M., Kainer, D., Wang, W. & Schwessinger, B. MinIONQC: fast and simple quality control for MinION sequencing data. *Bioinformatics* **35**, 523–525, https://doi.org/10.1093/bioinformatics/bty654 (2019).
60. Andrews, S. *FastQC: a quality control tool for high throughput sequence data*. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).
61. *FastQC software*. https://github.com/s-andrews/FastQC.
62. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: Visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669, https://doi.org/10.1093/bioinformatics/bty149 (2018).
63. *Nanoplot software*. https://github.com/wdecoster/NanoPlot.
64. *Porechop*. https://github.com/rrwick/Porechop.
65. *Filtlong software*. https://github.com/rrwick/Filtlong.
66. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546, https://doi.org/10.1038/s41587-019-0072-8 (2019).
67. *Flye software*. https://github.com/mikolmogorov/Flye.
68. *Medaka software, Oxford Nanopore Technologies*. https://github.com/nanoporetech/medaka.
69. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075, https://doi.org/10.1093/bioinformatics/btt086 (2013).
70. *Quast software*. https://github.com/ablab/quast.
71. *bamdst software*. https://github.com/shiquan/bamdst.
72. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654, https://doi.org/10.1093/molbev/msab199 (2021).
73. *Benchmarking Universal Single-copy Orthologs (BUSCO)*. https://busco.ezlab.org/.
74. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness. *In: Methods in Molecular Biology* **1962**, 227–245, https://doi.org/10.1007/978-1-4939-9173-0_14 (2019).
75. Anaconda Software Distribution. Computer software. Vers. 2-2.4.0. Anaconda, Nov. 2016. https://anaconda.com.
76. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069, https://doi.org/10.1093/bioinformatics/btu153 (2014).
77. *Prokka software*. https://github.com/tseemann/prokka.
78. Blin, K. *et al*. AntiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res.* **51**, W46–W50, https://doi.org/10.1093/nar/gkad344 (2023).
79. *antiSMASH*. https://antismash.secondarymetabolites.org/.
80. *European Nucleotide Archive (ENA) Project: PRJEB71438*. https://www.ebi.ac.uk/ena/browser/view/PRJEB71438 (2025).
81. Denton, J. F. *et al*. Extensive error in the number of genes inferred from draft genome assemblies. *PLoS Comput. Biol.* **10**, e1003998, https://doi.org/10.1371/journal.pcbi.1003998 (2014).
82. Oxford Nanopore Technologies. Input DNA/RNA QC. *Oxford Nanopore Technologies Community* https://nanoporetech.com/document/input-dna-rna-qc (2016).
83. Basdani, D. *et al*. *De-novo* assembly of 82 bacterial genomes using Nanopore sequencing and prediction of biosynthetic capacity (manuscript custom code). *figshare* https://doi.org/10.6084/m9.figshare.31378534 (2026).
84. *Manuscript custom code*. https://github.com/SBGF-LAB/Bacterial-genome-assembly-using-Nanopore-sequencing-data.

Acknowledgements

The corresponding author thanks Dr. Evanthia Kontomina for helpful early discussions and Professor George Patrinos for providing a welcoming academic environment at the Hellenic Pasteur Institute during completion of this manuscript. The authors also thank Evgenia Marangou for valuable assistance during grant administration and Chen Xiao Yi for technical support. Finally, we gratefully acknowledge the contribution of late ELTE Professor István Mihály Szabó, whose painstaking isolation and characterization of streptomycetes continues to benefit science 35 years on, making this study possible. The research project was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "2nd Call for H.F.R.I. Research Projects to support Faculty Members & Researchers" (Project Number: 3712). Additional support was provided by research grant 82418 with Medicon Hellas. T.F. was supported by the Sustainable Development and Technologies National Programme of the Hungarian Academy of Sciences (FFT NP FTA). The funders hold no ownership of the data and had no involvement in its collection, analysis, interpretation, or the preparation of the manuscript.

Author contributions

D.B.: Microbiological work, Nanopore sequencing (wet-lab), BGC analyses; S.Z.: 16S rRNA gene sequencing, bioinformatics (MinION operation and advanced genomic analyses); A.K. & D.T.: bioinformatics (MinION operation, software pipeline and initial genomic analyses); G.F.: Resources, co-supervision (D.B.) and expertise; T.F.: Resources, co-supervision (S.Z.) and expertise; K.M.: Resources and expertise; S.B.: Conceptualization, team supervision, resources, investigation, project administration, funding acquisition. D.B., S.Z., A.K., D.T. and S.B. wrote the manuscript which was reviewed and approved by all authors.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to S.B.

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.