Cost-effective hybrid long-short read assembly delineates alternative GC-rich *Streptomyces* chassis for natural product discovery

Elena Heng[a], Lee Ling Tan[a], Dillon W. P. Tay[b], Yee Hwee Lim[b], Lay-Kien Yang[c], Deborah C.S. Seow[c], Chung Yan Leong[c], Veronica Ng[c], Siew Bee Ng[c], Yoganathan Kanagasundaram[c], Fong Tian Wong[a,#], Lokanand Koduru[a,#]

Affiliations:

[a] Molecular Engineering Laboratory, Institute of Molecular and Cell Biology, A*STAR, Singapore.

[b] Organic and Biomolecular Chemistry, Institute of Sustainability for Chemicals, Energy and Environment, A*STAR, Singapore.

[c] Singapore Institute of Food and Biotechnology Innovation, A*STAR, Singapore.

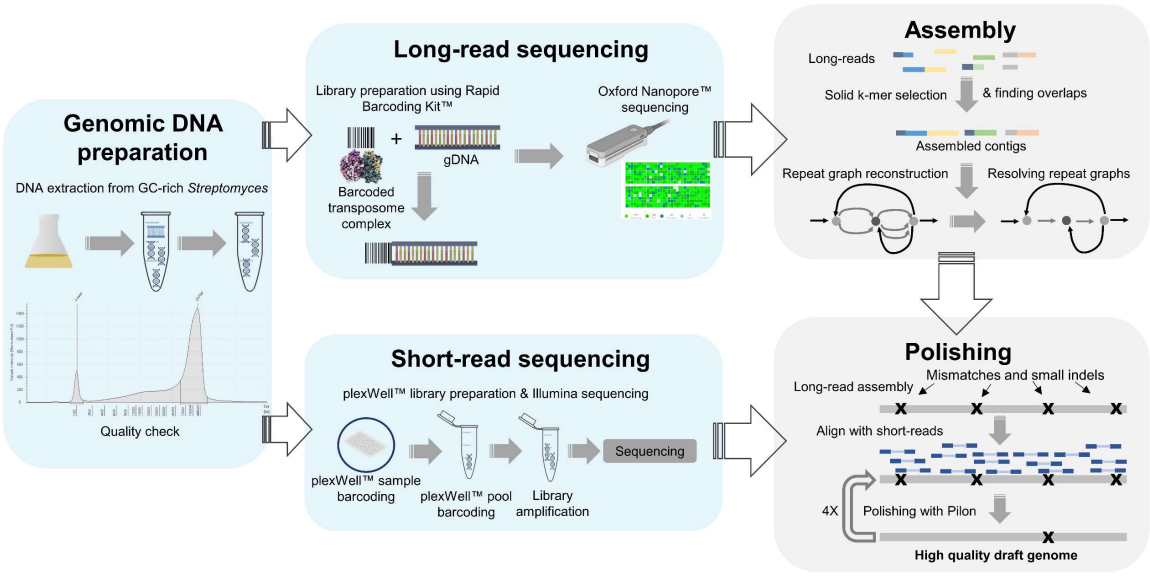
[#]Corresponding authors:

F.T.W.: wongft@imcb.a.star.edu.sg, Level 7, 61 Biopolis Drive, Proteos, Singapore 138673

L.K.: kodurul@imcb.a-star.edu.sg, Level 7, 61 Biopolis Drive, Proteos, Singapore 138673

Keywords: *Streptomyces*, GC-rich, Hybrid assembly, Biosynthetic gene cluster (BGC), Natural product (NP) chassis, Type I polyketide synthase (PKS)

Funding: This work was supported by National Research Foundation, Singapore (NRF-CRP19-2017-05-00) and Agency for Science, Technology and Research (A*STAR), Singapore (#21719).



Graphic abstract. Schematic of hybrid long- and short read assembly workflow for genome sequencing of GC-rich *Streptomyces*. Boxes shaded blue and grey correspond to experimental and *in silico* workflows, respectively.

Highlights

- A cost-effective genome sequencing approach for GC-rich *Streptomyces* is presented
- Hybrid assembly improves BGC annotation and identification
- A new species, *Streptomyces sydneybrenneri,* identified by taxonomic analysis
- Genomes of 8 *Streptomyces* species are reported and analysed in this study

Abstract

With the advent of rapid automated *in silico* identification of biosynthetic gene clusters (BGCs), genomics presents vast opportunities to accelerate natural product (NP) discovery. However, prolific NP producers, *Streptomyces*, are exceptionally GC-rich (>80%) and highly repetitive within BGCs. These pose challenges in sequencing and high-quality genome assembly which are currently circumvented *via* intensive sequencing. Here, we outline a more cost-effective workflow using multiplex Illumina and Oxford Nanopore sequencing with hybrid long-short read assembly algorithms to generate high quality genomes. Our protocol involves subjecting long read-derived assemblies to up to 4 rounds of polishing with short reads to yield accurate BGC predictions. We successfully sequenced and assembled 8 GC-rich *Streptomyces* genomes whose lengths range from 7.1 to 12.1 Mb at an average N50 of 5.9 Mb. Taxonomic analysis revealed previous misrepresentation among these strains and allowed us to propose a potentially new species, *Streptomyces sydneybrenneri*. Further comprehensive characterization of their biosynthetic, pan-genomic and antibiotic resistance features especially for molecules derived from type I polyketide synthase (PKS) BGCs reflected their potential as NP chassis. Thus, the genome assemblies and insights presented here are envisioned to serve as gateway for the scientific community to expand their avenues in NP discovery.

Introduction

Natural products (NPs) are secondary metabolites with myriad applications in therapeutics, agriculture, and food. After decades of technological advancement in deciphering their biosynthetic pathways, we now have highly predictive bioinformatics tools capable of translating genes to biosynthetic information and molecular structures [1], [2]. This has become vital in the new era of computationally driven antibiotics discovery [3]. Due to inactivated gene clusters or low yields of microbial NP under lab conditions, traditional screening and assay-based drug discovery methods are unable to keep up with the growing crisis of antibiotics resistance and medical needs [3]. However, advances in high throughput genomic and bioinformatic tools present a potential reservoir of untapped novel biosynthetic gene clusters (BGCs) in silent strains or in uncultivated microbiomes as sources of new drugs [4].

Streptomyces are highly prolific producers and a critical source of NPs, however they are characterized by high-GC content and abundant repetitive sequences, leading to problems in assembling high quality genomes, especially with short-read sequencing [5]. Instead, long-read sequencing provided by PacBio is greatly preferred to circumvent these limitations [5], [6]. More recently, Oxford Nanopore sequencing, along with optimized DNA isolation methods, has also been used to obtain long reads for whole-genome sequencing (WGS) of *Streptomyces* [7][8]. While high expense hinders the use of PacBio, Oxford Nanopore sequencing is not exploited to its full potential due to its higher error rate [9]. A plausible alternative to this is to employ a hybrid sequencing approach, where assemblies obtained using long reads are subsequently polished with short reads to limit errors in the final assemblies. Here, we describe a simple cost-efficient multiplex sequencing pipeline for the assembly and generation of GC-rich *Streptomyces* genomes. Furthermore, no optimization for multiplex DNA preparation was required beyond commercial protocols for sequencing.

In this study, we were interested in lesser studied *Streptomyces*, including *Streptomyces albulus, Streptomyces diastatochromogenes, Streptomyces lydicus, Streptomyces noursei,* and *Streptomyces* sp. MOE7. With these strains, we demonstrate that standard multiplex Illumina and Oxford Nanopore sequencing protocols can be used in tandem with hybrid long-short read assembly to generate high quality assemblies of lengths ranging from 7.1 to 12.1 Mb and with % completeness varying from 84% for one genome to >93% for the remaining 7 genomes. We further subjected the genome assemblies to taxonomic, pan-genomic and biosynthetic pathway analyses, to resolve misrepresentation in species assignment and explore their potential as alternative NP chassis.

Results

A total of 8 *Streptomyces* strains were obtained from various sources for this study. 6 *Streptomyces* strains initially annotated as *Streptomyces albulus, Streptomyces diastatochromogenes, Streptomyces lydicus* and *Streptomyces noursei*, were obtained from ATCC (Table 1). *Streptomyces* sp. A44034 and A41733 were obtained from the National Organism Library, Singapore [10] and were annotated as *Streptomyces* sp. MOE7 (NRRL ISP-5461) and *Streptomyces albulus* (ATCC 12757) respectively.

Table 1. Whole genome-based taxonomy

Strain	Closest Placement Taxonomy	Originally Assigned Taxonomy (Source[a])	Closest Placement ANI	MSA Percent
A41733 (ATCC 12757)	*Streptomyces noursei*	*Streptomyces albulus* (16S)	97.68	92.73
A44034 (NRRL ISP-5461)	*Streptomyces lydicus*	*Streptomyces* sp. MOE7 (16S)	99.91	89.16
ATCC11455a	*Streptomyces noursei*	*Streptomyces noursei* (ATCC)	99.88	87.87
ATCC21481	*Streptomyces vietnamensis*[b]	*Streptomyces diastatochromogenes* (ATCC)	92.93	78.49
ATCC23862	*Streptomyces diastatochromogenes*	*Streptomyces diastatochromogenes* (ATCC)	99.88	85.05
ATCC31561	*Streptomyces mirabilis*	*Streptomyces diastatochromogenes* (ATCC)	96.6	87.65
ATCC31713	*Streptomyces albulus*	*Streptomyces albulus* (ATCC)	99.14	81.29
ATCC31975	*Streptomyces lydicus*	*Streptomyces lydicus* (ATCC)	99.91	88.74

[a]Source refers to how the original taxonomy was assigned; ATCC refers to the assignment as on ATCC (accessed Nov 2022) and 16S refers to assignment by 16S alone.
[b]This is a “low confidence” species assignment by GTDB-Tk (v2.1.0) [11] as indicated by average nucleotide identity (ANI) and multiple sequence alignment (MSA) percent score; the strain is likely to be of different/new species.

Cost-effective hybrid genome sequencing of GC-rich *Streptomyces*

Long-read sequencing has been especially valuable in sequencing *Streptomyces* whose genomes are characterized by high-GC content and numerous repetitive sequences [5]. Although Nanopore-sequenced assemblies are high in contiguity, low per read sequence quality is also common [9]. In the context of BGC annotation, this resulting low quality could lead to missing gene or protein domain annotations, thereby resulting in fragmented or mis-annotated gene clusters with incomplete

biosynthetic information. Consequently, a combination of long and short reads is required to assemble the vastly repetitive GC-rich sequences of these prolific *Streptomyces*. Prior studies sequencing *Streptomyces* whole genomes have mostly employed PacBio [12] or use one MinION™ flow cell per genome [13][14], to obtain sufficient coverage and consequently high quality genomes. Here, we demonstrate a simple and convenient multiplexed workflow which does not required further optimization of the genomic preparation for the GC-rich *Streptomyces*. By using standard phenol:chloroform library preparation [15] protocols for these GC-rich genomes; protocols can be performed as described in the commercial kits for multiplex Oxford Nanopore and Illumina sequencing. For Oxford Nanopore, we used the rapid barcoding kit (Oxford Nanopore Technologies, SQK-RBK004) to multiplex 12 genomes on one MinION™ flow cell. For Illumina sequencing preparation, we used plexWell™ 96 (PW096) kit to generate 96 plex samples. The plexWell™ kit uses a reagent limited transposition to ensure uniform read count across samples with various GC and DNA concentrations. Overall cost comparison of Illumina-PacBio sequencing versus our multiplex Nanopore-Illumina protocol estimated that our method of genomic sequencing (including library preparation) has reduced the sequencing costs by >50% (Table S1-2).

Hybrid assemblies improve genome sequence quality

Nanopore reads are first assembled with the long read assembler, Flye (v2.8) [16], and the resulting assemblies (N50 > 1Mb, Table S4) were polished with short reads obtained from Illumina sequencing (see Methods). Subsequently, we subjected the unpolished and polished assemblies to two different assembly completeness tests: BUSCO [17] and CheckM [18]. In our analysis, we observed that the number of polishing rounds was important for assembly quality. Both BUSCO and CheckM analyses revealed a clear improvement in assembly quality following each round of polishing (Figure 1). Maximum improvement occurred after the first round of polishing with subsequent quality improvement almost saturating after 4 rounds.

In the 8 high quality genomes assembled for the various *Streptomyces* species, final assemblies observed genome lengths ranging from 7.1 to 12.1 Mb, with an average N50 of 5.9 Mb, and CheckM completeness of 84 to 99% (Table S3). In comparison, the previous genome version of the strain A44034 available on Genbank (accession: GCA_000717345.1) has an N50 of 182 kb. Annotation of the genomes also identified an average of 9850 genes per genome. All except one strain (ATCC 31713) had the completeness values >93 %. The lower CheckM completeness of ATCC 31713 is likely due to its lower quality Nanopore reads (N50 ~70Kb).

Here, we also demonstrate the importance of polishing for sequence quality improvement and its translation to a functional output. Molecular networking of tandem mass spectra (MS/MS) from fermentation extracts of *Streptomyces* sp. A44034 revealed a compound cluster annotated by DEREPLICATOR+ [19] as anti-tumour cyclic siderophore, desferrioxamine E and its rarer analogs [20] (Figure 2A, Table S4). However, to our surprise, antiSMASH (v6.0) of long read assembled (unpolished) and one round of polishing (Figure 2B, Pilon 1) did not uncover a desferrioxamine BGC. Instead, multiple rounds of polishing A44034's Nanopore assembly with the corresponding short reads were required to achieve sufficient quality to allow accurate BGC annotations. The putative BGC corresponding to desferrioxamine E [21] was observed only after at least two rounds of polishing (Figure 2B, Pilon 2). Also, at Pilon 3, full length genes are now observed for the BGC (Figure 2B). Similarly, we observed an increase in the number and quality of BGCs across the strains with subsequent rounds of polishing (Figure 1D). We therefore highlight the importance of "polishing" to ensure correct and high-resolution analysis of highly GC-rich genomes.

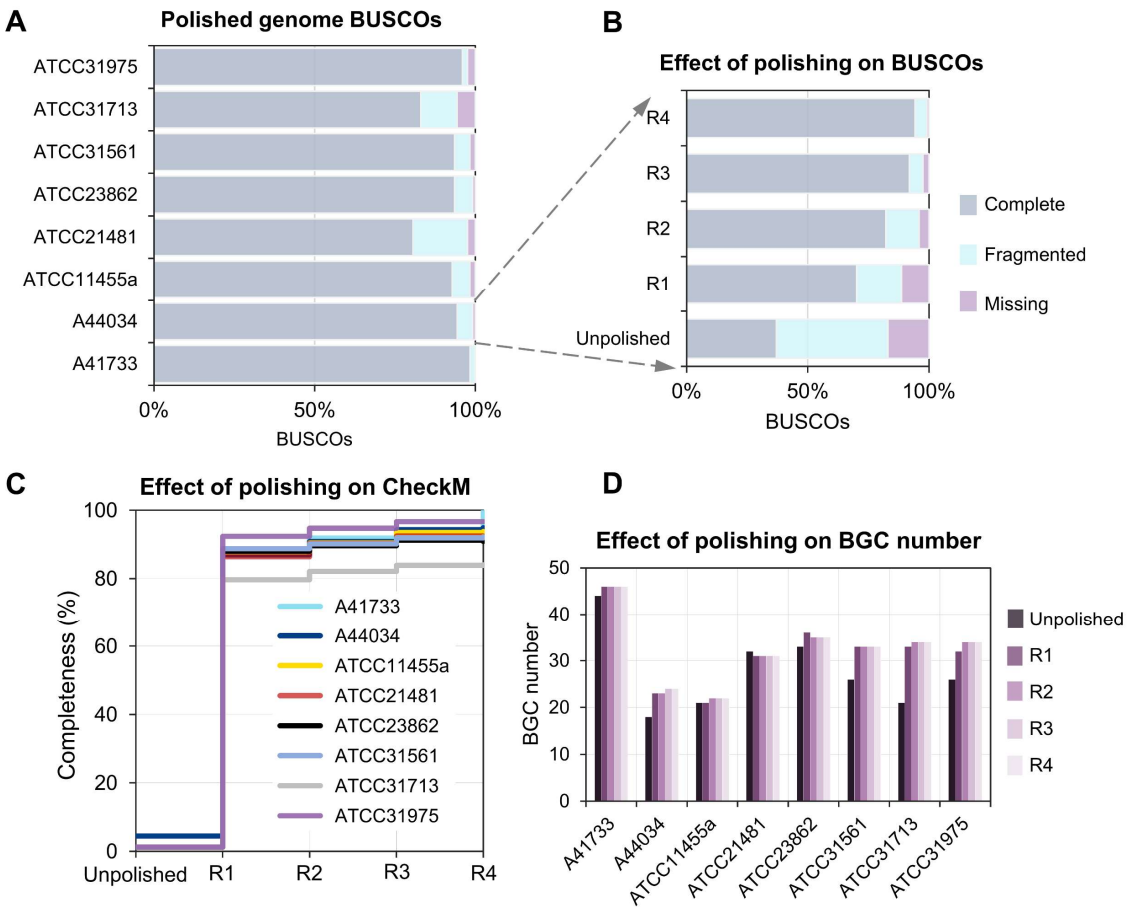


Figure 1. Effect of polishing long read assemblies on genome quality. (A) Percentage lineage-specific completeness and contamination levels in polished assemblies as assessed by BUSCO [17]. **(B)** Effect of polishing on completeness and contamination. **(C)** Effect of polishing on completeness in polished assemblies as assessed by CheckM [18]. **(D)** Effect of polishing on number of BGC identified in assemblies. R1, R2, R3 and R4 in (B-D) correspond to the number of rounds of polishing the long read assemblies with short reads using Pilon (v1.24) [22].

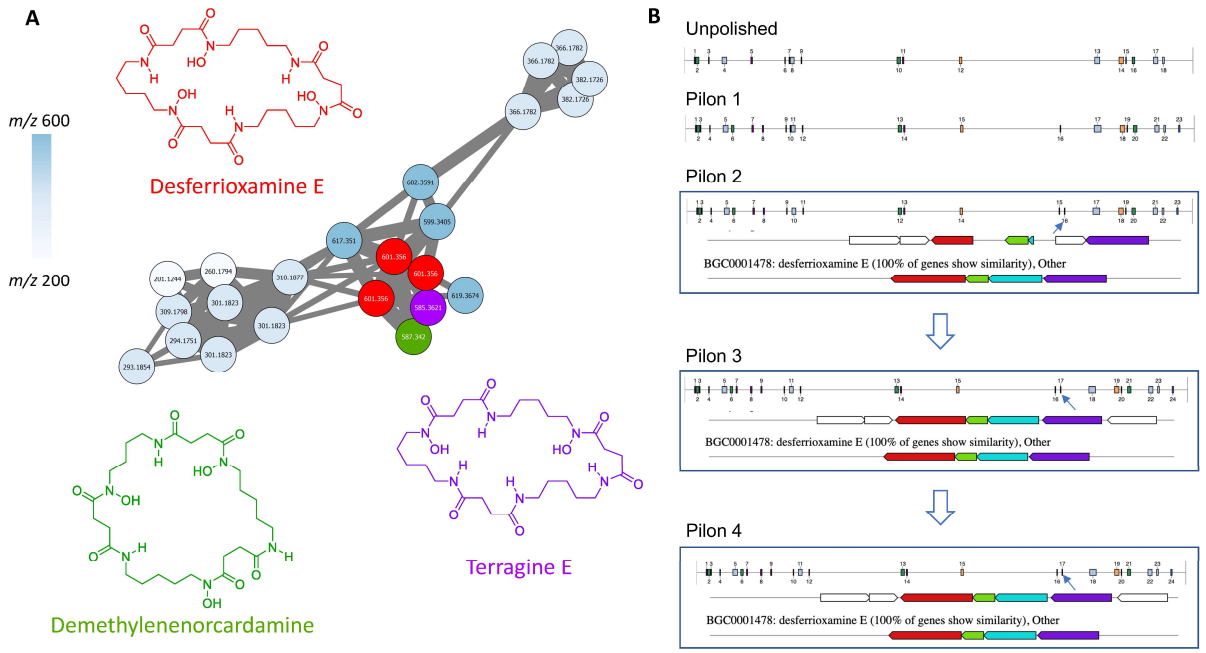


Figure 2. Molecular networking cluster of Desferrioxamine E, and its putative biosynthetic gene cluster (BGC) in *Streptomyces* sp. A44034. **(A)** Desferrioxamine E and its analogs as determined by their MS/MS spectra and molecular networking [23] (Figure S1, S2) **(B)** Revolution of annotated BGCs with increasing polishing rounds. Overview of clusters and their relative positions on the genome, along with multi-gene alignment as analysed using antiSMASH (v6.0) [24] are shown. Blue arrows are used to annotate location of desferrioxamine BGC. Gene cluster table of the putative BGC is available in Table S5. Pilon 1-4 refers to the assembly after 1-4 rounds of polishing, respectively.

Taxonomy and pan-genomic analysis reveal conserved unique features

All assemblies were subjected to GTDB-Tk based taxonomy [11] (Table 1). Average nucleotide identity (ANI) was used to assign species to the genomes. All except one of the strains (ATCC21481) are classified into known 'species' of *Streptomyces*. The current standard species boundary is proposed to be ANI>94 [25]. ATCC21481 was originally assigned as *Streptomyces diastachromogenes*. However, our analysis indicates that it has the closest taxonomic proximity to *Streptomyces vietnamensis* (ANI=92.93) with low multiple sequence alignment (MSA) percent (78.49), rendering it likely to be a potential candidate for new species within the known Genbank collection. We further extracted the 16S rRNA sequence from ATCC21481 genome and subjected it to BLASTn against the NCBI database. The closest 16S homolog at 99.6% similarity was to an unclassified strain *Streptomyces*, named *Streptomyces sp.* 21-4. Moreover, we assigned *Streptomyces noursei* to A41733 and *Streptomyces mirabilis* to ATCC31561, which were previously *Streptomyces albulus* [26] and *Streptomyces diastatochromogenes* [27] , respectively. The closest 16S homologs to A41733 and ATCC31561 were *Streptomyces noursei* ATCC 11455 at 99.9% similarity and *Streptomyces* sp. RLB1-8 at 100% similarity, respectively.

We then performed pan-genomic analysis of the 8 assemblies along with 4 model *Streptomyces* species using the Anvio pipeline (see Methods). The pan-genomic map consists of a few distinct regions: the dense region toward the top-left corresponds to the conserved regions most likely harboring genes involved in housekeeping and other core functions (Figure 3). We mapped the presence of the 5 most commonly studied housekeeping genes, rpoB, gyrB, atpD, recA and trpB [28]. These were all positioned inside this dense 'Core' pan-genomic region. Significant segments contributing to the biosynthetic diversity in different strains also are observed in regions that are unevenly spread across the strain genomes. Interestingly, the accessory regions ('Cloud' and 'Shell' pan-genome) of the ATCC genomes were enriched with type I polyketide synthases (PKSI) (bottom left of Figure 3). The regions annotated with PKSI were comparatively less in the prototypical *Streptomyces*. A detailed analysis of different BGC classes is presented in the next section.

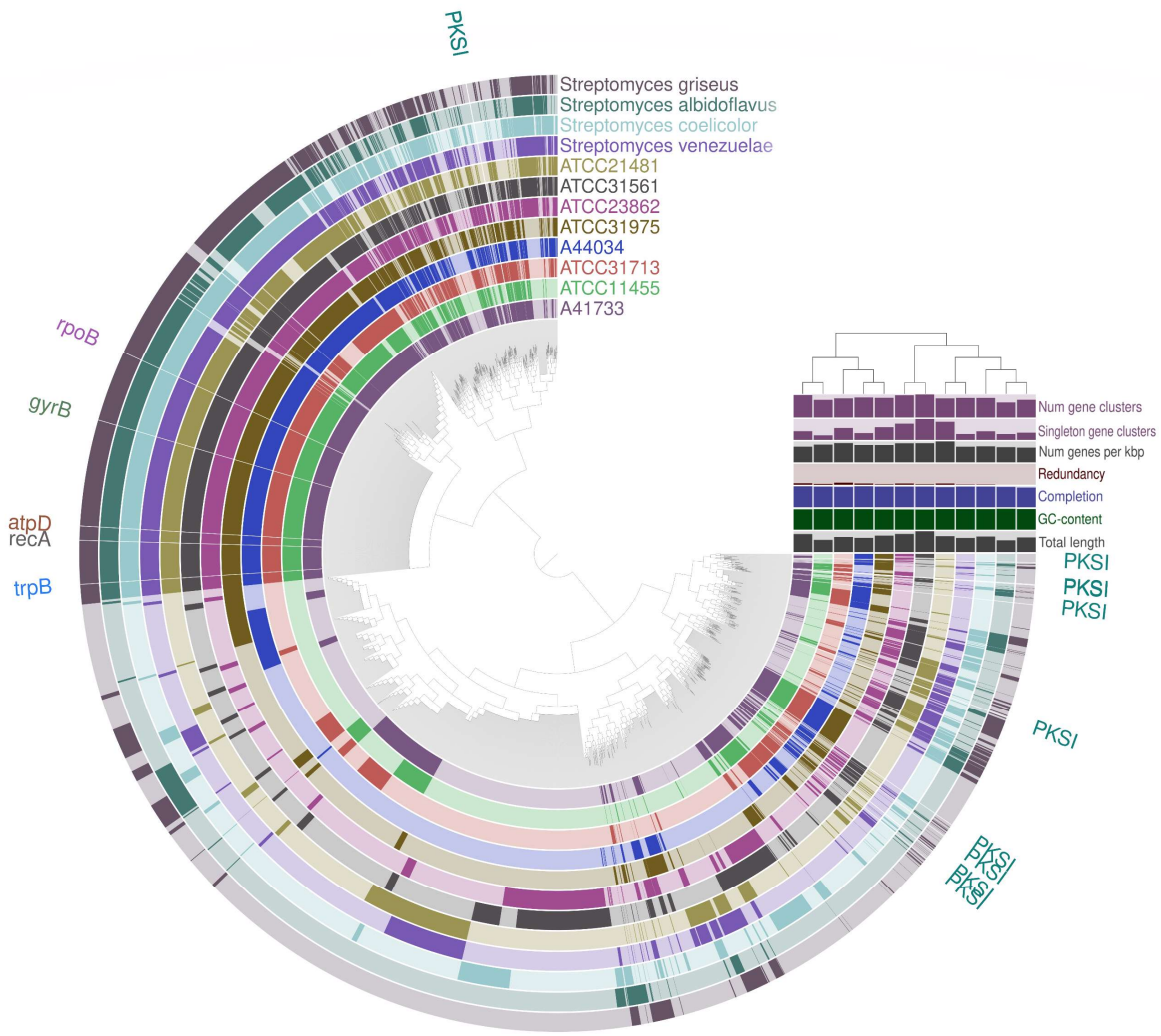


Figure 3. **Pan-genomic analysis of the *Streptomyces* strains**. The circular strip represents the presence/absence of gene clusters in each of the strains. PKSI: type I PKS; rpoB, gyrB, atpD, recA and trpB are housekeeping genes of *Streptomyces*. Bar plots at the end of each strip represents additional layers summarizing pan-genomic statistics of strains. 'Num gene clusters' represents the total number of gene clusters in strains, 'Singleton gene clusters' are orphan clusters which do not have homologs in other strains. Dendrogram is constructed using gene cluster presence/absence and Ward clustering. Tree over the layers is constructed based on Euclidean distance between the corresponding values for the strains using hierarchical clustering. Numbers corresponding to different layers are provided in detail in Table S6.

BGC repertoire and gene cluster family diversity

With genomes of 8 strains assembled, we also sought to explore the BGC diversity among them. To compare if these strains encode any unique BGCs compared to prototypical *Streptomyces* species, we also included *Streptomyces coelicolor* (Accession: AL645882.2) [29], *Streptomyces griseus* (Accession: NC_010572.1) [30], *Streptomyces albidoflavus* (Accession: NC_020990.1) [31], and *Streptomyces venezuelae* (Accession: NZ_CP018074.1) [32] to our analysis on BGC diversity comparisons. First, we constructed a whole genome-based phylogenetic tree using M1CR0B1AL1Z3R [33] and iTOL [34] (see Methods). AntiSMASH [24] based BGC annotation was then performed on all strains. The identified BGCs were then subjected to BiG-SCAPE (v1.1.0) [35] based annotation of classes and gene cluster families (GCFs) while including BGCs from MiBIG 3.0 [36] in the analysis. Finally, the BiG-SCAPE-based BGC classes were aligned to the phylogenetic tree (Figure 4). Clearly, we observe that these strains encode more type I PKS BGCs compared the model *Streptomyces*, possibly suggesting that they could be better hosts than the model *Streptomyces to* produce PKSI-derived natural products. Within A41733, A44034, ATCC 23862, ATCC 31975, we

observed desferrioxamine E and a derivative of streptolydigin (Table S4). Although this is only a fraction out of the gene clusters identified *in silico* including pathways towards curamycin, quinomycin, lantipeptide planosporicin, citrulassin D and nystatin (Table S8, S9, S12, S15), it is not unexpected and can be contributed to factors such as unfavourable fermentation conditions [37] or silent gene expression [38]. Within these 8 genomes, 27 clusters with less than 30% homology to known MiBiG 3.0 clusters were also observed as potentially new unique BGCs (Table S8-S15). Furthermore, we used antibiotic resistance target seeker (ARTS) [39] to identify BGCs with potential bioactivity. We were able to identify the presence of two categories of resistance, namely the 'known-resistance mechanisms' and 'core duplications' among multiple BGCs (Figure 4A and 4B). Most strains, except for ATCC 21481 and ATCC 31713, show multiple resistance mechanisms as part of their BGCs. We observed increased presence of ABC efflux transporters among the known resistance mechanisms, indicating potential bactericidal activity among the BGCs (Figure S3). It is interesting to find 1-deoxy-D-xylulose-5-phosphate synthase (*dxs*, [40]) duplicated in the terpene BGCs, as 1-deoxy-D-xylulose-5-phosphate is also intermediate in the non-mevalonate pathway required for terpenoid biosynthesis in *Streptomyces.* However, further experimental validation, and perhaps even strain engineering to activate these clusters, is required to determine if these are indeed critical duplications to combat bioactivity of the corresponding biosynthetic product in the gene clusters.

In addition, we carried out a biochemical network-based pathway enrichment analysis to assess the strain metabolic capacities. The four prototypical *Streptomyces* species and an outgroup heterologous host, *Escherichia coli*, were included in the analysis to facilitate a thorough comparison. The procedure has been detailed in Methods section. Briefly, we first performed protein annotations on the genome assemblies of all strains using Bakta (v1.4.0) [41], the annotations were then used to build draft metabolic networks using Chimera pipeline [42]. Finally, KEGG pathway information was extracted from files containing the genome-wide metabolic mapping information (Figure 5). It is interesting to note that two distinct clades emerge from the dendrogram: clade-2 is considerably more enriched in biosynthetic pathways related to several NPs compared to clade-1 (p-value = 1.2055e-04, chi-squared = 14.78, Kruskal-Wallis test). Clade-2 also include commonly used *Streptomyces* chassis, golden standard *Streptomyces coelicolor* [29] and fast growing *Streptomyces venezuelae* [32]. Also, 5 out of the 8 strains examined in this study, including A41733, ATCC23862, ATCC31561, ATCC31713 and A44034, belong to clade-2 [43][44][45], hinting at the potential of these strains as alternative chassis for NP production.

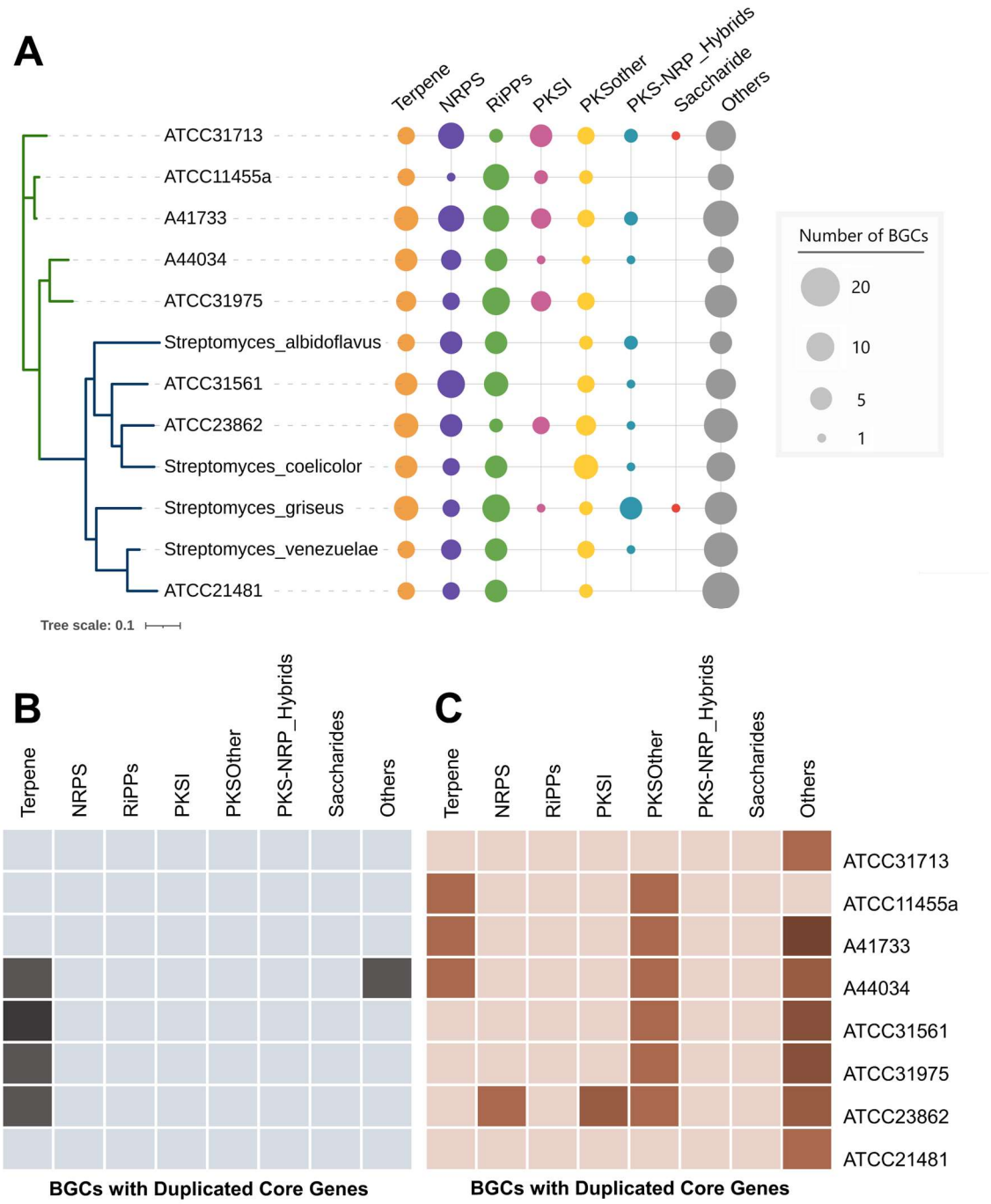


Figure 4. Mining BGCs and resistance mechanisms (A) Enrichment of BGC classes in different *Streptomyces* strains. The size of the circle is proportional to number of BGCs harboured by the corresponding strain. The tree is constructed based on whole-genome phylogeny (bootstrap=100). The clade overrepresenting model *Streptomyces* is coloured blue, while the strains in this study mostly fall under the green clade. More details are provided in Table S7. **(B)** Presence of known-resistance mechanisms identified by ARTS analysis (v2.0) [39] . **(C)** Presence of duplicated core genes in BGCs as identified by ARTS. NRPS: Non ribosomal peptide synthase, RiPPs: Ribosomally synthesized and post-translationally modified peptides, PKSI: Type I PKS, PKSother: PKS other than type I PKS, PKS-NRP_hybrids: Hybrid PKS-NRPS, Others: Cluster that does not fit into any of categories shown. [35]

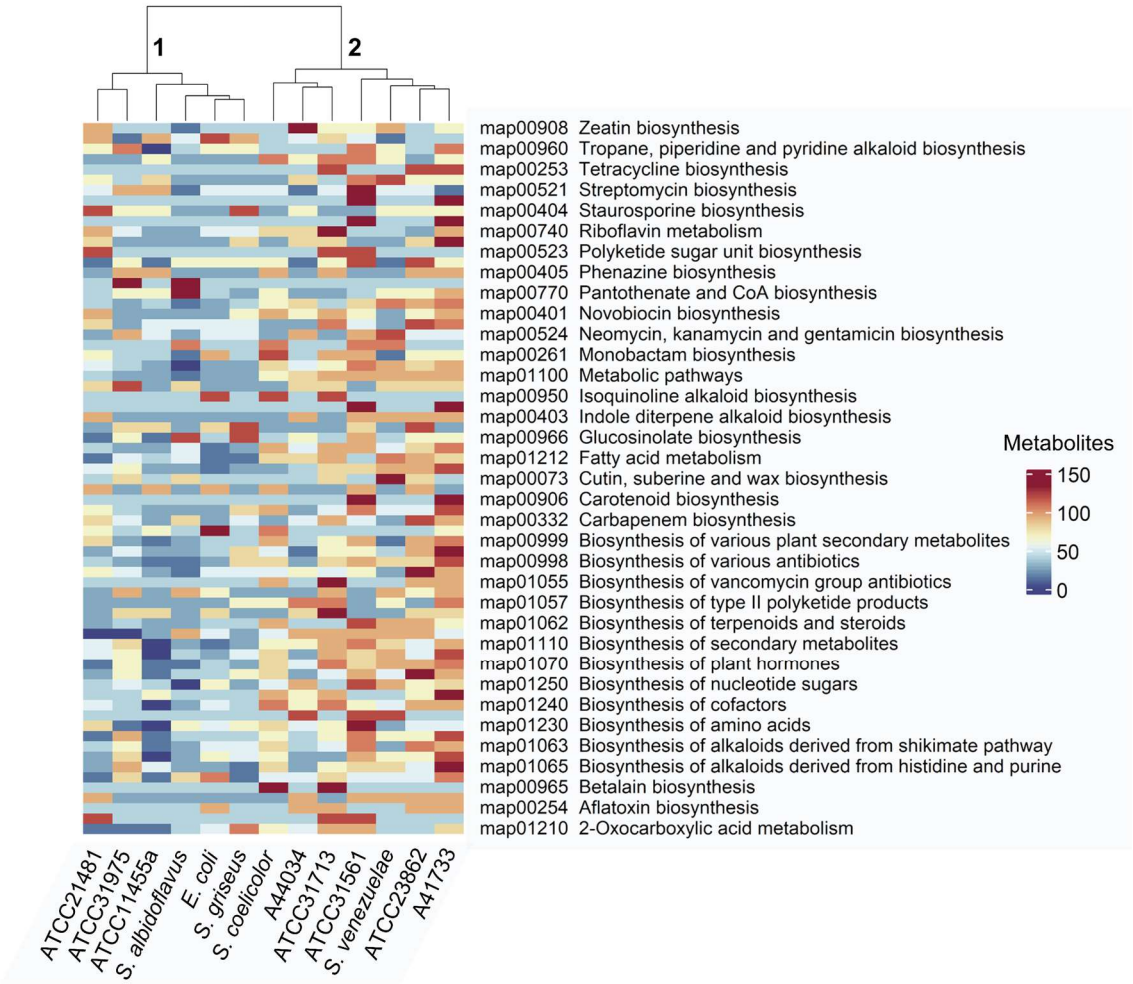


Figure 5. Biochemical network-guided metabolic pathway enrichment. The heatmap showcases enrichment of different KEGG pathway maps related to NP biosynthesis in the 8 strains, 4 prototypical *Streptomyces* and an outgroup heterologous host, *E. coli*. The color scale corresponds to number of metabolites in each KEGG pathway, red being 'high' and blue corresponding to 'low' values. Labels '1' and '2' at the top represent two distinct clades showing differential enrichment in NP biosynthetic pathways (p-value = 1.2055e-04, chi-squared = 14.78, Kruskal-Wallis test). The 'map' IDs denote KEGG pathways [46]. The dendrogram was generated using hierarchical clustering. Detailed numbers can be found in Table S16.

Discussion and conclusions

Drastic cost reduction of NGS technologies has greatly accelerated accessibility to genome sequencing, consequently resulting in new applications for exploitation of microbial genomes [47]. However, the field of natural product discovery also relies on extensively curated genome sequences with high contiguity [48], which allows for better annotation of BGCs due to accurate positional context of various genes in the genome. Oftentimes, microbes of interest, such as actinobacterial NP producers, are GC-rich and repetitive and PacBio and Oxford Nanopore have been the critical drivers of long read sequencing approaches required to generate high quality assemblies. Although Oxford Nanopore sequencing is several orders cheaper than PacBio, the resulting reads are typically characterized by low per read accuracy, especially in GC-rich actinobacteria which are known to be major reservoirs of natural products [5]. In this work, we present a multiplexed short- and long-read sequencing protocol to assemble high quality genomes of GC-rich *Streptomyces*. Here, we sequenced 8 *Streptomyces* strains with potential as natural product chassis and assemble them using

a hybrid approach. We observe that polishing Nanopore-assembled genomes with Illumina short read sequences significantly improves BGC annotations. With a substantial reduction (>50%) in sequencing costs compared to the more conventional PacBio-Illumina workflow, this protocol has the potential to considerably advance natural product research. Among the 8 sequenced strains, we have also identified one strain (ATCC 21481) as a presumptive new species, which we have renamed as *Streptomyces sydneybrenneri.* Genomic analyses of these strains also suggest that although most of the BGCs were not expressed under laboratory conditions, these strains have significant potential as NP chassis. Pan-genomic comparation of these strains observed that 6 out of 8 strains possess a higher number of type I PKS BGCs compared with model *Streptomyces*, suggesting their potential as better hosts for type I PKS-derived natural products. Genome-scale metabolic model-KEGG mapping also observed an overall enrichment of metabolic pathways related to the biosynthesis of several NPs in 5 out of 8 strains. Most strains also consist of more than one resistance gene/mechanism, implying the existence of putative bioactivity in them.

In summary, this study describes a cost-efficient multiplex protocol that drastically reduces the cost and complexity of generating high quality GC-rich genome assemblies. Here, we use this protocol to obtain genome assemblies of 8 strains, consisting of *Streptomyces noursei, Streptomyces lydicus, Streptomyces diastatochromogenes, Streptomyces mirabilis,* and *Streptomyces albulus*, including a potentially new *Streptomyces* species. Analysis of these genomes indicate high biosynthetic potential, including possible opportunities for their usage as type I PKS chassis or mining of cryptic gene clusters. We also envision that this protocol will aid in accelerate whole genome sequencing and assembly of these traditionally challenging strains, and in doing so, open further opportunities into NP discovery and applications.

Materials and Methods

Strains

Streptomyces strains (ATCC® 11455a, 21481, 23862,31561, 31713, 31975) were purchased from ATCC Global Bioresource Center (Manassas, VA, USA). *Streptomyces* sp. A44034 and A41733 were obtained from the A*STAR National Organism library [10].

Fermentation and Liquid Chromatography-Mass Spectrometry (LC-MS)

Seed cultures were inoculated into 50 mL of fermentation media, CA02LB, CA07LB, CA08LB, CA09LB or CA10LB. The cultures were incubated at 28°C for 9 days at 200 rpm shaking. Cultures were then collected, freeze dried and extracted with methanol. The extracts were analysed as described in a previous study [49].

Genomic DNA extraction

Strains are seeded in 4mL SV2 (1.5% glucose, 1.5% glycerol, 1.5% soya peptone, 0.1% $CaCO_3$, pH 7.0) and incubated for 4 days in a 30°C shaker. Cell pellets are collected and lysed with lysozyme at 5mg/mL for 1 hour. Genomic DNA is extracted with equal volumes of phenol:chloroform:isoamylalcohol and precipitated with 0.6M ammonium acetate and ethanol. DNA precipitate was washed with 80% ethanol and reconstituted with 10mM Tris-HCl pH 8.0.

Quality characterisation

Quality of genomic DNA was verified with NanoDrop™ 2000/2000c Spectrophotometer and Qubit dsDNA BR Assay Kit. Only samples with ratios of 260:280 and 260:230 more than 1.8 and Nanodrop to Qubit ratio of 1.0 to 1.5 were used for Nanopore sequencing. To achieve a ratio of 1.0 to 1.5 for Nanodrop to Qubit ratio, 0.6x Ampure XP beads were used to clean-up fragments lesser than 500bp. The quality of genomic DNA was further validated using Agilent Genomic DNA ScreenTape on a tapestation.

Genomic library preparation and sequencing

To sequence genomes, a library was generated using 10 ng genomic DNA per strain with plexWell™ 96 kit, as described in its manual. The library was sequenced using 2 x 151 bp paired end protocol on illumina HiSEQ 4000.

400 ng genomic DNA per strain was used for sequencing on MinION™ system (Oxford Nanopore Technologies). Rapid Barcoding Sequencing (SQK-RBK004) kit was used for library preparation. All was done accordingly to provided manual. The library was added to a R9.5 Flowcell for 72 hours run on the MinION™ Sequencer.

Data processing and genome assembly

Hiseq reads were check for quality using FastQC [50], adaptors were trimmed using Trimmomatic (v0.32) [51]. Short reads were assembled using Spades assembler (v3.15.5) [52]. Nanopore raw data (FAST5) was basecalled and quality-assessed using Guppy GPU version (Oxford Nanopore Technologies GitHub: Guppy [https://github.com/nanoporetech].). The base called FASTQ reads were assembled using Flye (v2.8) [16] using expected genome size of 8Mb and coverage of 30. Long reads were polished with short reads using Pilon (v1.24) [22]. Four rounds of polishing were done to ensure better per read quality. BUSCO [17] and CheckM [18] analysis was performed to check assembly completeness.

Phylogenetic analysis and taxonomic inference

Whole genome phylogenetic tree was contructed using M1CR0B1AL1Z3R [33]. The newick trees were visualized using iTOL browser [34]. BGC class bubble plot annotations were added using iTOL's 'create dataset' options. Taxonomic inferences were made based on whole genome assemblies using GTDB-Tk (v2.1.0) [11]. A scratch directory was used to avoid RAM limitations of PPlacer algorithm part of the GTDB-Tk pipeline. 16S-based taxonomy was inferred as follows: 16S sequences were extracted from the genome assemblies FASTA files using Barrnap (v0.9) [53]

Genome mining, BiG-SCAPE, enrichment analysis and resistance guided BGC prioritization pipeline

AntiSMASH (v6.0) [24] in 'full' mode was used for BGC annotation. BGC class and GCF annotations were derived from BiG-SCAPE analysis (v1.1.0) [35] with default options while manually including

MiBiG 3.0 BGCs [36] to the input. Identification of resistance mechanisms was carried out using antibiotic resistance target seeker (ARTS) pipeline (v2.0) [39]. Metabolic pathway enrichment analysis

The genome sequences were first subjected to annotation using Bakta pipeline (v1.4.0) [41]. Strain-specific genome-scale metabolic models were generated from the Bakta protein annotations using CarveMe algorithm (v1.4.1) [54] part of Chimera pipeline [42]. KEGG [46] pathway information for each strain was retrieved from Chimera outputs and merged to generate pathway enrichment heatmaps. Significant difference in KEGG pathway enrichment between the two dendrogram clades was evaluated using Kruskal-Wallis test.

DEREPLICATOR+ Metabolite identification

Metabolite identification from tandem mass spectra (MS/MS) data using the online workflow DEREPLICATOR+ [19] available on the Global Natural Products Social Networking (GNPS) website (http://gnps.ucsd.edu, accessed November 2022). Precursor and fragment ion mass tolerance thresholds were set to ± 0.005 Da and ± 0.01 Da respectively. Maximum charge allowed was set to 2. Fragmentation model used was 2-1-3, which indicates there will be at most two bridges, one 2-cut, and three cut in total. The spectra were searched against the “AllDB” small molecule structure database of about 720K compounds.

Molecular Networking of Desferrioxamine E Cluster

MSConvert v3.0.22198-0867718 from Proteowizard [55] was used for initial processing of raw liquid chromatography-tandem mass spectrometry (LC-MS/MS) data into open-source mascot generic format (.mgf). All tandem mass spectra (MS/MS) signals with intensity values below 1000 signal intensity were removed as background correction. All peaks in a +/- 17 Da around the precursor ion mass were deleted to remove residual precursor ions, and peaks not in the top 6 most intense peaks in a +/- 50Da window were filtered out. MetGem v1.3.6 [56] was used to plot molecular networks with the following filters (m/z tolerance = 0.02, minimum matched peaks = 4, max nearest neighbors = 10, minimal cosine score value = 0.70, max connected component size = 1000).

Data availability:

All genome sequences generated in this study have been deposited to GenBank (Accession: PRJNA907517).

Author contributions:

L.K.: Conceptualization, methodology, software, data curation, formation analysis, visualization, original draft-writing, review and editing, F.T.W.: Conceptualization, investigation, methodology, original draft-writing, review and editing, E.H.: investigation, methodology, data acquisition (genomic sequencing), review and editing, L.L.T.: methodology, data acquisition, V.N., C.Y.L: Data acquisition, strain resources, fermentation, S.B.N.: Data acquisition, strain resources, fermentation, writing-review and editing, L.K.Y., D.S.: Data acquisition, Y.K: Data acquisition, review and editing, D.T., Y.H.L.; Data analysis, writing-review and editing.

Competing interests:

The authors declare no competing interests.

Acknowledgments:

The authors gratefully acknowledged the funding for this work from National Research Foundation, Singapore (NRF-CRP19-2017-05-00) and Agency for Science, Technology and Research (A*STAR), Singapore (#21719). The authors also acknowledge the support of the late Professor Sydney Brenner, founder of Molecular Engineering Laboratory at A*STAR, Singapore.

References

[1] K. Blin *et al.*, “AntiSMASH 5.0: Updates to the secondary metabolite genome mining pipeline,” *Nucleic Acids Res*, vol. 47, no. W1, 2019, doi: 10.1093/nar/gkz310.

[2] M. A. Skinnider, N. J. Merwin, C. W. Johnston, and N. A. Magarvey, “PRISM 3: Expanded prediction of natural product chemical structures from microbial genomes,” *Nucleic Acids Res*, vol. 45, no. W1, 2017, doi: 10.1093/nar/gkx320.

[3] K. Lewis, “The Science of Antibiotic Discovery,” *Cell*, vol. 181, no. 1. 2020. doi: 10.1016/j.cell.2020.02.056.

[4] N. Lee, S. Hwang, J. Kim, S. Cho, B. Palsson, and B. K. Cho, “Mini review: Genome mining approaches for the identification of secondary metabolite biosynthetic gene clusters in Streptomyces,” *Computational and Structural Biotechnology Journal*, vol. 18. 2020. doi: 10.1016/j.csbj.2020.06.024.

[5] J. P. Gomez-Escribano, S. Alt, and M. J. Bibb, “Next generation sequencing of actinobacteria for the discovery of novel natural products,” *Marine Drugs*, vol. 14, no. 4. 2016. doi: 10.3390/md14040078.

[6] A. Rhoads and K. F. Au, “PacBio Sequencing and Its Applications,” *Genomics, Proteomics and Bioinformatics*, vol. 13, no. 5. 2015. doi: 10.1016/j.gpb.2015.08.002.

[7] K. Athanasopoulou, M. A. Boti, P. G. Adamopoulos, P. C. Skourou, and A. Scorilas, “Third-generation sequencing: The spearhead towards the radical transformation of modern genomics,” *Life*, vol. 12, no. 1, 2022, doi: 10.3390/life12010030.

[8] L. Slemc *et al.*, “Reference-Grade Genome and Large Linear Plasmid of Streptomyces rimosus: Pushing the Limits of Nanopore Sequencing.,” *Microbiol Spectr*, vol. 10, no. 2, p. e0243421, Apr. 2022, doi: 10.1128/spectrum.02434-21.

[9] S. L. Amarasinghe, S. Su, X. Dong, L. Zappia, M. E. Ritchie, and Q. Gouil, “Opportunities and challenges in long-read sequencing data analysis,” *Genome Biol*, vol. 21, no. 1, p. 30, Dec. 2020, doi: 10.1186/s13059-020-1935-5.

[10] S. B. Ng, Y. Kanagasundaram, H. Fan, P. Arumugam, B. Eisenhaber, and F. Eisenhaber, “The 160K Natural Organism Library, a unique resource for natural products research,” *Nature Biotechnology*, vol. 36, no. 7. 2018. doi: 10.1038/nbt.4187.

[11] P.-A. Chaumeil, A. J. Mussig, P. Hugenholtz, and D. H. Parks, “GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database,” *Bioinformatics*, Nov. 2019, doi: 10.1093/bioinformatics/btz848.

[12] N. Lee *et al.*, “Thirty complete Streptomyces genome sequences for mining novel secondary metabolite biosynthetic gene clusters,” *Sci Data*, vol. 7, no. 1, p. 55, Dec. 2020, doi: 10.1038/s41597-020-0395-9.

[13] T. Gren *et al.*, “Characterization and engineering of Streptomyces griseofuscus DSM 40191 as a potential host for heterologous expression of biosynthetic gene clusters,” *Sci Rep*, vol. 11, no. 1, p. 18301, Dec. 2021, doi: 10.1038/s41598-021-97571-2.

[14] M. Kibret *et al.*, “Streptomyces spp. From Ethiopia Producing Antimicrobial Compounds: Characterization via Bioassays, Genome Analyses, and Mass Spectrometry,” *Front Microbiol*, vol. 9, Jun. 2018, doi: 10.3389/fmicb.2018.01270.

[15] J. Quick, “Ultra-long read sequencing protocol for RAD004 V.3,” *protocols.io*, Jan. 22, 2018. dx.doi.org/10.17504/protocols.io.mrxc57n (accessed Dec. 06, 2022).

[16] M. Kolmogorov, J. Yuan, Y. Lin, and P. A. Pevzner, “Assembly of long, error-prone reads using repeat graphs,” *Nat Biotechnol*, vol. 37, no. 5, pp. 540–546, May 2019, doi: 10.1038/s41587-019-0072-8.

[17] F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. v. Kriventseva, and E. M. Zdobnov, “BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs,” *Bioinformatics*, vol. 31, no. 19, pp. 3210–3212, Oct. 2015, doi: 10.1093/bioinformatics/btv351.

[18] D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson, “CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes,” *Genome Res*, vol. 25, no. 7, pp. 1043–1055, Jul. 2015, doi: 10.1101/gr.186072.114.

[19] H. Mohimani *et al.*, “Dereplication of microbial metabolites through database search of mass spectra,” *Nat Commun*, vol. 9, no. 1, 2018, doi: 10.1038/s41467-018-06082-8.

[20] G. Y. S. Wang *et al.*, “Novel natural products from soil DNA libraries in a streptomycete host,” *Org Lett*, vol. 2, no. 16, 2000, doi: 10.1021/ol005860z.

[21] F. Barona-Gómez, U. Wong, A. E. Giannakopulos, P. J. Derrick, and G. L. Challis, “Identification of a Cluster of Genes that Directs Desferrioxamine Biosynthesis in *Streptomyces c oelicolor* M145,” *J Am Chem Soc*, vol. 126, no. 50, pp. 16282–16283, Dec. 2004, doi: 10.1021/ja045774k.

[22] B. J. Walker *et al.*, “Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement,” *PLoS One*, vol. 9, no. 11, 2014, doi: 10.1371/journal.pone.0112963.

[23] A. T. Aron *et al.*, “Reproducible molecular networking of untargeted mass spectrometry data using GNPS,” *Nat Protoc*, vol. 15, no. 6, 2020, doi: 10.1038/s41596-020-0317-5.

[24] K. Blin *et al.*, “antiSMASH 6.0: improving cluster detection and comparison capabilities,” *Nucleic Acids Res*, vol. 49, no. W1, pp. W29–W35, Jul. 2021, doi: 10.1093/nar/gkab335.

[25] M. Richter and R. Rosselló-Móra, “Shifting the genomic gold standard for the prokaryotic species definition,” *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 19126–19131, Nov. 2009, doi: 10.1073/pnas.0906412106.

[26] J. W. Wong and M. P. Burns, “Microbial hydroxylation of (Z)-2-benzylidene-1-azabicyclo[2.2.2]octan-3- one,” *Tetrahedron Asymmetry*, vol. 10, no. 22, 1999, doi: 10.1016/S0957-4166(99)00471-1.

[27] M. Kittelmann, R. Lattmann, and O. Ghisalba, “Preparation of 10,11-Epoxy-carbamazepine and 10,11-Dihydro-10-hydroxy-carbamazepine by Microbial Epoxidation and Hydroxylation,” *Biosci Biotechnol Biochem*, vol. 57, no. 9, 1993, doi: 10.1271/bbb.57.1589.

[28] H. Komaki, “Resolution of housekeeping gene sequences used in MLSA for the genus Streptomyces and reclassification of Streptomyces anthocyanicus and Streptomyces tricolor as heterotypic synonyms of Streptomyces violaceoruber,” *Int J Syst Evol Microbiol*, vol. 72, no. 5, May 2022, doi: 10.1099/ijsem.0.005370.

[29] S. D. Bentley *et al.*, “Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2),” *Nature*, vol. 417, no. 6885, pp. 141–147, May 2002, doi: 10.1038/417141a.

[30] Y. Ohnishi *et al.*, “Genome Sequence of the Streptomycin-Producing Microorganism *Streptomyces griseus* IFO 13350,” *J Bacteriol*, vol. 190, no. 11, pp. 4050–4060, Jun. 2008, doi: 10.1128/JB.00204-08.

[31] N. Zaburannyi, M. Rabyk, B. Ostash, V. Fedorenko, and A. Luzhetskyy, “Insights into naturally minimised Streptomyces albus J1074 genome,” *BMC Genomics*, vol. 15, no. 1, p. 97, Dec. 2014, doi: 10.1186/1471-2164-15-97.

[32] J. P. Gomez-Escribano *et al.*, "*Streptomyces venezuelae* NRRL B-65442: genome sequence of a model strain used to study morphological differentiation in filamentous actinobacteria," *J Ind Microbiol Biotechnol*, vol. 48, no. 9–10, Dec. 2021, doi: 10.1093/jimb/kuab035.

[33] O. Avram, D. Rapoport, S. Portugez, and T. Pupko, "M1CR0B1AL1Z3R—a user-friendly web server for the analysis of large-scale microbial genomics data," *Nucleic Acids Res*, vol. 47, no. W1, pp. W88–W92, Jul. 2019, doi: 10.1093/nar/gkz423.

[34] I. Letunic and P. Bork, "Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation," *Nucleic Acids Res*, vol. 49, no. W1, pp. W293–W296, Jul. 2021, doi: 10.1093/nar/gkab301.

[35] J. C. Navarro-Muñoz *et al.*, "A computational framework to explore large-scale biosynthetic diversity," *Nat Chem Biol*, vol. 16, no. 1, pp. 60–68, Jan. 2020, doi: 10.1038/s41589-019-0400-9.

[36] S. A. Kautsar *et al.*, "MIBiG 2.0: a repository for biosynthetic gene clusters of known function," *Nucleic Acids Res*, Oct. 2019, doi: 10.1093/nar/gkz882.

[37] M. Liu, T. Grkovic, X. Liu, J. Han, L. Zhang, and R. J. Quinn, "A systems approach using OSMAC, Log P and NMR fingerprinting: An approach to novelty," *Synth Syst Biotechnol*, vol. 2, no. 4, 2017, doi: 10.1016/j.synbio.2017.10.001.

[38] Z. Liu, Y. Zhao, C. Huang, and Y. Luo, "Recent Advances in Silent Gene Cluster Activation in Streptomyces," *Frontiers in Bioengineering and Biotechnology*, vol. 9. 2021. doi: 10.3389/fbioe.2021.632230.

[39] M. D. Mungan, M. Alanjary, K. Blin, T. Weber, M. H. Medema, and N. Ziemert, "ARTS 2.0: feature updates and expansion of the Antibiotic Resistant Target Seeker for comparative genome mining," *Nucleic Acids Res*, vol. 48, no. W1, pp. W546–W552, Jul. 2020, doi: 10.1093/nar/gkaa374.

[40] T. Kuzuyama, M. Takagi, S. Takahashi, and H. Seto, "Cloning and characterization of 1-deoxy-D-xylulose 5- phosphate synthase from Streptomyces sp. Strain CL190, which uses both the mevalonate and nonmevalonate pathways for isopentenyl diphosphate biosynthesis," *J Bacteriol*, vol. 182, no. 4, 2000, doi: 10.1128/JB.182.4.891-897.2000.

[41] O. Schwengers, L. Jelonek, M. A. Dieckmann, S. Beyvers, J. Blom, and A. Goesmann, "Bakta: Rapid and standardized annotation of bacterial genomes via alignment-free sequence identification," *Microb Genom*, vol. 7, no. 11, 2021, doi: 10.1099/MGEN.0.000685.

[42] G. Tamasco, R. R. da Silva, and R. Silva-Rocha, "ChiMera: An easy to use pipeline for Bacterial Genome Based Metabolic Network Reconstruction, Evaluation and Visualization," *Biorxiv*, Dec. 2021.

[43] W. Wang *et al.*, "Harnessing the intracellular triacylglycerols for titer improvement of polyketides in Streptomyces," *Nat Biotechnol*, vol. 38, no. 1, 2020, doi: 10.1038/s41587-019-0335-4.

[44] R. Chen *et al.*, "Engineering cofactor supply and recycling to drive phenolic acid biosynthesis in yeast," *Nat Chem Biol*, vol. 18, no. 5, pp. 520–529, May 2022, doi: 10.1038/s41589-022-01014-6.

[45] Q. Xu, H. Deng, X. Li, and Z.-S. Quan, "Application of Amino Acids in the Structural Modification of Natural Products: A Review," *Front Chem*, vol. 9, Apr. 2021, doi: 10.3389/fchem.2021.650569.

[46] M. Kanehisa, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Res*, vol. 28, no. 1, pp. 27–30, Jan. 2000, doi: 10.1093/nar/28.1.27.

[47] M. Land *et al.*, “Insights from 20 years of bacterial genome sequencing,” *Funct Integr Genomics*, vol. 15, no. 2, pp. 141–161, Mar. 2015, doi: 10.1007/s10142-015-0433-4.

[48] J. J. J. van der Hooft, H. Mohimani, A. Bauermeister, P. C. Dorrestein, K. R. Duncan, and M. H. Medema, “Linking genomics and metabolomics to chart specialized metabolic diversity,” *Chem Soc Rev*, vol. 49, no. 11, pp. 3297–3314, 2020, doi: 10.1039/D0CS00162G.

[49] D. Tay *et al.*, “Training old dogs to do new tricks: A general multi-pronged activation approach for natural product discovery in Actinomycetes,” *Res Sq*, Sep. 2022.

[50] S. Andrews, “FastQC - A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/,” *Babraham Bioinformatics*, 2010.

[51] A. M. Bolger, M. Lohse, and B. Usadel, “Trimmomatic: a flexible trimmer for Illumina sequence data,” *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, Aug. 2014, doi: 10.1093/bioinformatics/btu170.

[52] A. Bankevich *et al.*, “SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing,” *Journal of Computational Biology*, vol. 19, no. 5, pp. 455–477, May 2012, doi: 10.1089/cmb.2012.0021.

[53] T. Seemann, “Barrnap 0.9: rapid ribosomal RNA prediction,” 2013. https://github.com/tseemann/barrnap (accessed Nov. 29, 2022).

[54] D. Machado, S. Andrejev, M. Tramontano, and K. R. Patil, “Fast automated reconstruction of genome-scale metabolic models for microbial species and communities,” *Nucleic Acids Res*, vol. 46, no. 15, pp. 7542–7553, Sep. 2018, doi: 10.1093/nar/gky537.

[55] M. C. Chambers *et al.*, “A cross-platform toolkit for mass spectrometry and proteomics,” *Nature Biotechnology*, vol. 30, no. 10. 2012. doi: 10.1038/nbt.2377.

[56] F. Olivon, N. Elie, G. Grelier, F. Roussi, M. Litaudon, and D. Touboul, “MetGem Software for the Generation of Molecular Networks Based on the t-SNE Algorithm,” *Anal Chem*, vol. 90, no. 23, 2018, doi: 10.1021/acs.analchem.8b03099.