Data in Brief 36 (2021) 107128

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Draft genome sequence data of *Streptomyces* sp. FH025

Lucky Poh Wah Goh, Fauze Mahmud, Ping-Chin Lee*

Faculty of Science and Natural Resources, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah. Malaysia

ARTICLE INFO

Article history:
Received 16 March 2021
Revised 19 April 2021
Accepted 30 April 2021
Available online 12 May 2021

Keywords:
Streptomyces sp.
Draft genome sequence
FH025
Secondary metabolites
Anti-malarial activity

ABSTRACT

The genome data of *Streptomyces* sp. FH025 comprised of 8,381,474 bp with a high GC content of 72.51%. The genome contains 7035 coding sequences spanning 1261 contigs. *Streptomyces* sp. FH025 contains 57 secondary metabolite gene clusters including polyketide synthase, nonribosomal polyketide synthase and other biosynthetic pathways such as amglyccycl, butyrolactone, terpenes, siderophores, lanthipeptide-class-iv, and ladderane. 16S rRNA analysis of *Streptomyces* sp. FH025 is similar to the *Streptomyces* genus. This whole genome project has been deposited at NCBI under the accession JAFLNG000000000.

* Corresponding author.
E-mail address: leepc@ums.edu.my (P.-C. Lee).

https://doi.org/10.1016/j.dib.2021.107128

Specification Table

Subject	Biology
Specific subject area	Microbiology, Bacterial genomics, Biotechnology
Type of data	Figure, Table, Draft genome sequence data
How data were acquired	Genome sequencing on Miseq
Data format	Raw and analyzed
Parameters for data collection	Genomic DNA was isolated from a pure culture of *Streptomyces* sp. FH025. AntiSMASH software predicted the putative biosynthetic gene clusters.
Description of data collection	Whole-genome sequencing, assembly, and annotation
Data source location	Soil samples used for bacteria isolation were collected at Likas, Sabah, Malaysia. (06°2′18.4″N 116° 7′16.6″E)
Data accessibility	The data is available at NCBI Genbank from the following links: http://www.ncbi.nlm.nih.gov/bioproject/705517 https://www.ncbi.nlm.nih.gov/biosample/18091016 https://www.ncbi.nlm.nih.gov/sra/PRJNA705517

Value of the Data

- The *Streptomyces* strain FH025 draft genome showed that it is unique as compared to other strains and has the potential to produce novel bioactive compounds.
- The secondary metabolite putative genes identified in *Streptomyces* sp. FH025 genome could contribute greatly to the antibiotic and drug discovery for treatment of various human diseases.
- Based on the genome data and previous study, this strain could be a potential strain for study of anti-malarial compounds as well as various enzymes production.

1. Data Description

Streptomyces sp. FH025 was isolated from Likas, Sabah, Malaysia (06°2′18.4″ N 116° 7′16.6″ E). The draft genome characteristics of *Streptomyces* sp. FH025 were summarized in Table 1. There were 1261 number of contigs with a total contig size of 8381,474 bp and N50 contig number of 10,071. The L50 value was 246 and the GC content was 72.51%. Based on genome annotation, there were 1261 number of contigs with protein encoding genes and 406 number of sub systems with 7035 number of coding sequences (Table 1, Fig. 1). There were 74 RNAs.

Additionally, *Streptomyces* sp. FH025 could produce important secondary metabolites when analyzed using antiSMASH. It was estimated that there were 51 secondary metabolites cluster of genes (smCOG) (Table 2). The secondary metabolite genes present were type I and type III polyketides synthase (PKS). There were 9 non-ribosomal polypeptide synthetase (NRPS), 10 NRPS-like and 1 NRPS-Type I PKS identified. Besides, several secondary metabolite biosynthetic pathways were present such as amglyccycl, butyrolactone, terpenes, siderophores, lantipeptide and ladderane.

Table 1
Characteristics of draft genome assembly of *Streptomyces* sp. FH025.

Number of contigs	1261
Total contig size (bp)	8381,474
N50 contig number[a]	10,071
L50	246
GC content (%)	72.51
Number of contigs (with protein encoding genes)	1261
Number of subsystems	406
Number of coding sequences	7035
Number of RNAs	74

[a] Minimum set of contigs that represent at least 50% of total genome sequence.

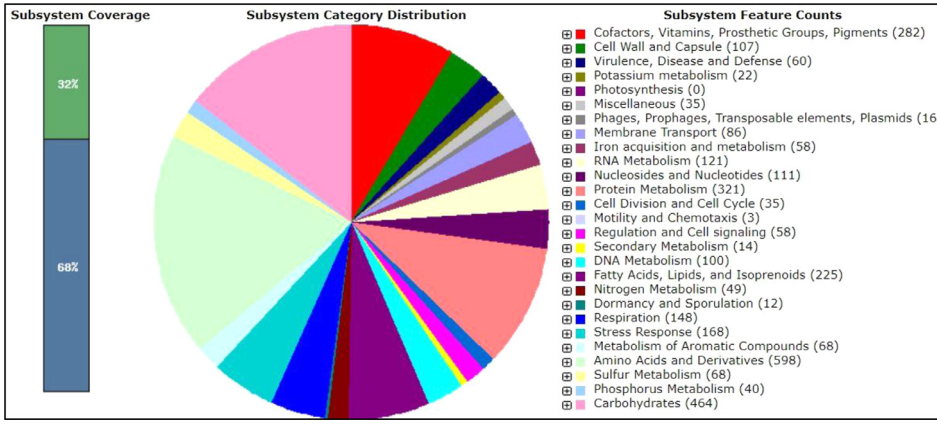

Fig. 1. Subsystem statistics information of FH025 using RAST annotation. The subsystems category and corresponding feature counts were shown in the legend.

Table 2
Putative gene clusters coding for secondary metabolites detected by antiSMASH annotation of *Streptomyces* sp. FH025.

Features	Number of clusters
No of smCOG[1]	57
PKS[2]	
PKS-like	2
Type I	17
Type III	2
NRPS[3]	9
NRPS-like	10
NRPS-Type I PKS	1
Biosynthetic Pathways	
Amglyccycl	1
Butyrolactone	1
Terpenes	3
Siderophores	4
Lanthipeptide-class-iv	1
Ladderane	1
RiPP-like	1
RRE-containing	2
NAPAA	1
Others	1

[1] Secondary metabolism Clusters of Orthologous Groups.
[2] Polyketide synthase.
[3] Nonribosomal polypeptide synthetase.

ContEst16S software analysis indicated that the draft genome assembly did not have contamination of other prokaryotic genome. The 16S rRNA phylogenetic analysis revealed that *Streptomyces* sp. FH025 is closely related to the *Streptomyces* genus (Fig 2). Furthermore, genome-based taxonomy analysis revealed that strain FH025 has the highest average nucleotide identity (ANI) value (89.42%) and highest digital DNA-DNA hybridization (dDDH) value (38.4%) with *Kitasatospora aureofaciens* strain DM-1 (Table 3). However, strain FH025 was not affiliated as *Kintasatospora* because the values of ANI and dDDH were not greater than the established cutoff values on species delimitation for ANI (> 95–96%) [1] and dDDH value (>70%), respectively [2]. The low genome identity of strain FH025 with other strains analyzed indicated that strain FH025 is unique and warrant further investigation.

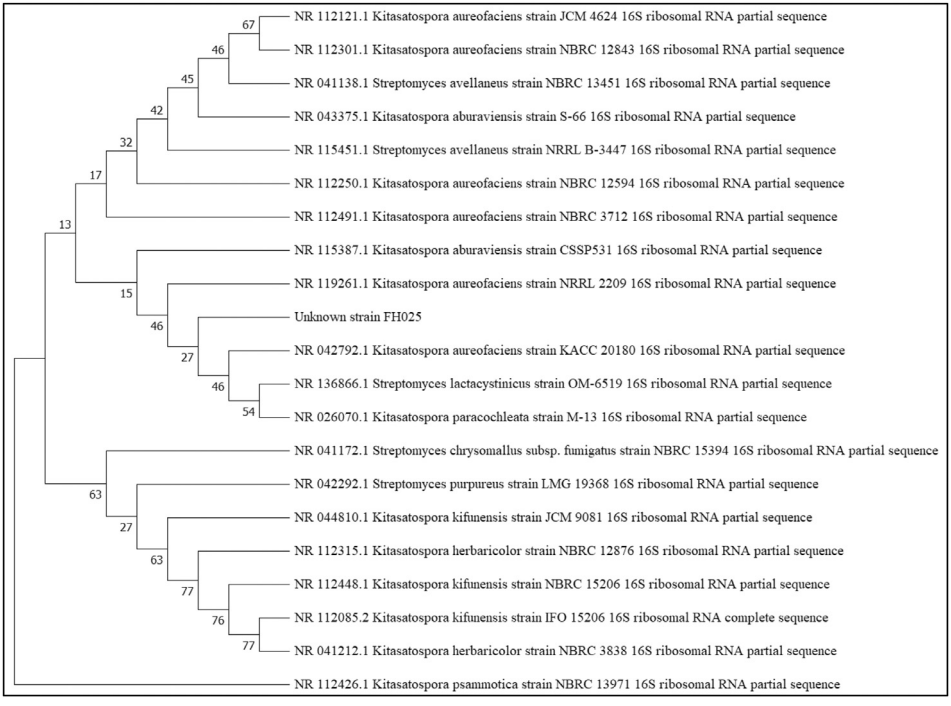

Fig. 2. Phylogenetic tree diagram of FH025 generated using neighbor-joining based on 16S rRNA gene sequence (947 bp) shows that FH025 was closely related with the *Streptomyces* genus. The numbers at branch nodes indicate percentages from 1000 bootstraps.

Table 3
The 16S rRNA sequence similarity, ANI and dDDH values of strain FH025 and its closely related species.

Closely related species	16S rRNA sequence similarity (%)	OrthoANIu value (%)	dDDH value (%)
NC_016109.1 Kitasatospora setae KM 6054, complete sequence	98.17	80.54	24.6
NZ_CP020563.1 Kitasatospora albolonga strain YIM 101,047 chromosome, complete genome	97.46	75.52	21.8
NZ_CP020567.1 Kitasatospora aureofaciens strain DM-1 chromosome, complete genome	99.80	89.42	38.4
NZ_CP025394.1 Kitasatospora sp. MMS16-BH015 chromosome, complete genome	98.67	81.01	25.2
NZ_CP054919.1 Kitasatospora sp. NA04385 chromosome, complete genome	98.57	80.72	24.7
Streptomyces clavuligerus strain ATCC 27,064 chromosome, complete genome	96.64	75.89	21.8
Streptomyces galilaeus strain ATCC 14,969 chromosome, complete genome	96.13	75.59	21.3
Streptomyces nitrosporeus strain ATCC 12,769 chromosome, complete genome	97.25	75.87	21.7
Streptomyces subrutilus strain ATCC 27,467 chromosome, complete genome	96.85	76.23	21.5
Streptomyces tsukubensis strain NRRL 18,488 chromosome, complete genome	96.95	75.59	21.8

2. Experimental Design, Materials and Methods

2.1. Sample collection and isolation of streptomyces

Soil samples covered with dead leaves were collected under a tree, *Shorea parvifolia* from Likas, Sabah, Malaysia and bacteria isolation was performed as previously described [3]. Briefly, serial dilution was performed on the soil samples and bacteria isolation was carried out using modified humic acid agar (with addition of vitamin B). Screening of isolates exhibiting anti-malaria activities was conducted and FH025 was observed to exhibit anti-malarial activities as previously described [3]. The isolate was sub-cultured on oatmeal agar (pH 7.2) at 28 °C to obtain a pure isolate named FH025. The culture was stored in 20% glycerol stock at −80 °C.

2.2. DNA isolation, genome sequencing, assembly, and annotation

Genomic DNA was isolated using Wizard® Genomic DNA Purification Kit according to manufacturer's instructions (Promega, USA). A whole-genome sequencing library was prepared using Nextera XT DNA library preparation kit following manufacturer's instructions (illumina, USA). The libraries were sequenced using the Miseq platform (Illumina, USA) to generate 2 × 250 paired end reads. The raw reads adapters were trimmed. Low quality sequences (<Q30) were trimmed by Trimmomatic version 0.38.0 [4]. Primary genome assembly was performed using Unicycler version 0.4.8.0 [5]. The primary draft genome was analyzed by rapid annotation using subsystems technology (RAST) [6–8]. The secondary metabolites biosynthetic gene clusters of strain FH025 draft genome were identified using antiSMASH version 5.0 [9].

2.3. 16S rRNA phylogenetic analysis

ContEst16S software was used to extract *Streptomyces* sp. FH025 16S rRNA gene sequence (981 bp) and analyze for any contamination of prokaryotic genomes [10]. Basic local alignment search tool (BLAST) analysis was performed against NCBI database and the top 20 near species 16S rRNA gene sequence was retrieved. The sequences were aligned using ClustalW and trimmed to 947 bp [11]. The phylogenetic tree was constructed by neighbor joining method with 1000 bootstraps using MEGA X software [12].

2.4. Average nucleotide identity and digital dna-dna hybridization genome-based taxonomy analysis

The ANI between the genome of strain FH025 and related species with complete genome from NCBI database were determined by OrthoANIu algorithm [13]. Digital DNA-DNA hybridization (dDDH) was performed using genome blast distance phylogeny with 10 closely related species with complete genome sequence obtained from NCBI database [14].

Data Availability

The whole genome project was deposited at NCBI under the accession JAFLNG000000000.

Ethics Statement

This study did not involve any human subjects and animal experiments. No ethical approval was required.

CRediT Author Statement

Lucky Poh Wah Goh: Formal analysis, Data curation, Writing – original draft, Writing – review & editing; **Fauze Mahmud:** Writing – review & editing; **Ping-Chin Lee:** Conceptualization, Resources, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they do not have conflict of interest that could influence the work reported in this paper.

Acknowledgement

This work is partly supported by Universiti Malaysia Sabah (GKP22–2018).

References

[1] N.J. Varghese, S. Mukherjee, N. Ivanova, K.T. Konstantinidis, K. Mavrommatis, N.C. Kyrpides, A. Pati, Microbial species delineation using whole genome sequences, Nucleic. Acid. Res. 43 (2015) 6761–6771, doi:10.1093/nar/gkv657.
[2] J.P. Meier-Kolthoff, H.-P. Klenk, M. Göker, Taxonomic use of DNA G+C content and DNA–DNA hybridization in the genomic age, Int. J. Syst. Evol. Microbiol. 64 (2014) (2014) 352–356, doi:10.1099/ijs.0.056994-0.
[3] D.E. Dahari, R.M. Salleh, F. Mahmud, P.-C. Lee, N. Embi, H.M. Sidek, Anti-malarial activities of two soil actinomycete isolates from sabah via inhibition of glycogen synthase kinase 3β, Trop Life Sci Res. 27 (2016) 53–71, doi:10.21315/tlsr2016.27.2.5.
[4] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinfo. 30 (2014) 2114–2120, doi:10.1093/bioinformatics/btu170.
[5] R.R. Wick, L.M. Judd, C.L. Gorrie, K.E. Holt, Unicycler: resolving bacterial genome assemblies from short and long sequencing reads, PLoS Comput. Biol. 13 (2017) e1005595, doi:10.1371/journal.pcbi.1005595.
[6] R.K. Aziz, D. Bartels, A.A. Best, M. DeJongh, T. Disz, R.A. Edwards, K. Formsma, S. Gerdes, E.M. Glass, M. Kubal, F. Meyer, G.J. Olsen, R. Olson, A.L. Osterman, R.A. Overbeek, L.K. McNeil, D. Paarmann, T. Paczian, B. Parrello, … O. Zagnitko, The RAST server: rapid annotations using subsystems technology, BMC. Genomics. 9 (2008) 75, doi:10.1186/1471-2164-9-75.
[7] R. Overbeek, R. Olson, G.D. Pusch, G.J. Olsen, J.J. Davis, T. Disz, R.A. Edwards, S. Gerdes, B. Parrello, M. Shukla, V. Vonstein, A.R. Wattam, F. Xia, R. Stevens, The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST), Nucleic. Acid. Res. 42 (2013) D206–D214, doi:10.1093/nar/gkt1226.
[8] T. Brettin, J.J. Davis, T. Disz, R.A. Edwards, S. Gerdes, G.J. Olsen, R. Olson, R. Overbeek, B. Parrello, G.D. Pusch, M. Shukla, J.A. Thomason, R. Stevens, V. Vonstein, A.R. Wattam, F. Xia, RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes, Sci. Rep. 5 (2015), doi:10.1038/srep08365.
[9] K. Blin, S. Shaw, K. Steinke, R. Villebro, N. Ziemert, S.Y. Lee, M.H. Medema, T. Weber, antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline, Nucleic. Acid. Res. 47 (2019) W81–W87, doi:10.1093/nar/gkz310.
[10] I. Lee, M. Chalita, S.-M. Ha, S.-I. Na, S.-H. Yoon, J. Chun, ContEst16S: an algorithm that identifies contaminated prokaryotic genomes using 16S RNA gene sequences, Int. J. Syst. Evol. Microbiol. 67 (2017) 2053–2057, doi:10.1099/ijsem.0.001872.
[11] J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic. Acid. Res. 22 (1994) 4673–4680, doi:10.1093/nar/22.22.4673.
[12] S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, X. MEGA, Molecular evolutionary genetics analysis across computing platforms, Mol. Biol. Evol. 35 (2018) 1547–1549, doi:10.1093/molbev/msy096.
[13] S.-H. Yoon, S. Ha, J. Lim, S. Kwon, J. Chun, A large-scale evaluation of algorithms to calculate average nucleotide identity, Antonie Van Leeuwenhoek 110 (2017) 1281–1286, doi:10.1007/s10482-017-0844-4.
[14] J.P. Meier-Kolthoff, A.F. Auch, H.-P. Klenk, M. Göker, Genome sequence-based species delimitation with confidence intervals and improved distance functions, BMC. Bioinfo. 14 (2013) 60, doi:10.1186/1471-2105-14-60.