Data in Brief 35 (2021) 106768

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Data on DNA-seq analysis of Endophytic *Streptomyces* sp. SUK 48

Siti Junaidah Ahmad[a,b], Noraziah Mohamad Zin[a,*]

[a] Center for Diagnostic, Therapeutic and Investigative Studies, Faculty of Health Sciences, Universiti Kebangsaan Malaysia, Jalan Raja Muda Abd Aziz, 50300 Kuala Lumpur, Malaysia

[b] Faculty of Health Sciences, Universiti Sultan Zainal Abidin, 21300, Kuala Nerus, Terengganu, Malaysia

ARTICLE INFO

Article history:
Received 14 October 2020
Revised 11 January 2021
Accepted 14 January 2021
Available online 20 January 2021

Keywords:
Streptomyces sp.
Ethnomedicinal plant
Whole genome sequence
SUK 48

ABSTRACT

The data genome sequence of SUK 48 consists of 8,341,706 bp, comprising of one contig with a high G+C content of 72.33%. The genome sequence encodes for 67 tRNAs and 21 rRNAs in one contig. SUK48 was found to have low similarities with other *Streptomyces* sp. (81–93% ANI indices) indicating that the isolated strain has a unique genome property and is presumably a novel species. This genome includes 34 genetic clusters responsible for the synthesis of secondary metabolites, including two polyketide synthase (PKS) clusters; one PKS type II cluster gene, one PKS gene cluster type III, five NRPS genetic clusters, and five PKS/NRPS hybrid clusters.

Specification Table

Subject	Biology
Specific subject area	Microbiology, Bacteria genomics, Biotechnology
Type of data	Figure, Table, Draft genome sequence, raw sequence data

(*continued on next page*)

[*] Corresponding author.
E-mail address: noraziah.zin@ukm.edu.my (N.M. Zin).

https://doi.org/10.1016/j.dib.2021.106768

How data were acquired	genome sequencing by PacBio RS II
Data format	Raw and analyzed
Parameters for data collection	PACBio RS II carried out genome sequencing of the strain. BUSCO annotated and analysed genes; the average Nucleotide Identity (ANI) analysis was also determined. The AntiSMASH software predicted putative biosynthetic gene clusters.
Description of data collection	Endophytic *Streptomyces* sp. SUK 48 was isolated from fruit of *Brasilia* sp. located at the Universiti Kebangsaan Malaysia reserve forest. *Streptomyces* sp. SUK 48 was cultured on Starch Yeast Casein Agar (SYCA) and maintained on International *Streptomyces* Project 2 agar (ISP2) for 14 days on 28 °C until whitish spore formed. *Streptomyces* SUK 48 sp. genomic DNA was sequenced by PacBio RS II. Using the NCBI Reference Sequence (RefSeq) protein database and the Swiss-Prot protein database, the functional annotation has been determined. AntiSMASH programme predicted biosynthetic clusters, and genomic data were compared with other Streptomyces spp using ANI (Average Nucleotide Index) analysis.
Data source location	Novel Antibiotic Laboratory, Centre of Diagnostics, Therapeutics & Investigations, Faculty of Health Sciences, Universiti Kebangsaan Malaysia
Data accessibility	The data for this draft genome has been deposited in the GenBank under the accession number CP045740 and the raw reads data under the SRA accession number SRP229701. The data described in this paper is the under bioproject; PRJNA587018 and available at http://www.ncbi.nlm.nih.gov/bioproject/587018.

Value of the Data

- Thirty-four secondary metabolites putative genes were identified in *Streptomyces* sp. SUK 48 genome. These candidates' genes can be useful leads in antibiotic discovery.
- The candidates' genes highlighted in this article could be used in further validation studies by using for example genome editing.
- 'Streptomyces genome can be used for further comparative genomics studies.

1. Data Description

Here we represent raw data sequence-reads, an assembled data genome of *Streptomyces* sp. SUK 48 isolated from fruit of *Brasilia* sp. Both the raw data and assembled data genome are available at NCBI's Sequence Read Achive as bioproject PRJNA587018 and available at http://www.ncbi.nlm.nih.gov/bioproject/587018.

The predicted coding sequences ($\geq$99 nucleotide) were used for functional annotation. Diamond v0.9.22 was used to BLAST the mRNA sequences against the RefSeq database, while NCBI-Blast v2.2.28+ was used to BLAST the same gene collection against the Swist-Prot database. The cut-offs were set at the overall estimated value of 1×10^{-5} for both BLAST searches. For the standalone analysis of the Blast2GO pipeline, the BLAST outputs of both databases were used in gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) [1]. Secondary metabolism was analyzed via antiSMASH v.5.0 [2].

The analyses of the assembled genome revealed a genome size of about 8,341,706 bp which consisted of 1 contig and with a high G + C content of 72.33% (Table 1). Prior to gene prediction, 67 tRNAs and 21 rRNAs (7 copies of 5S, 7 copies of 16S and 7 copies of 23S) were identified. A total of 7,354 coding genes were predicted for the masked genome. The gene coding sequences cover approximately 87.7% of the entire draft genome (Table 1). The average nucleotide identity (ANI) analysis of SUK48 with closely related species revealed the following similarity indices: *Streptomyces griseofuscus* 64 (93.85%); *Streptomyces misionensis* DSM 40306 (89.37%); *Streptomyces roseochromogenus* (86.14%); *Streptomyces kebangsaanensis* SUK 12 (84.15%); *Streptomyces*

Table 1
Statistics of assembled sequence length and gene prediction and structural annotation statistics.

Polished contigs	1
Maximum Contig Length	8,341,706
N50 Contig Length	8,341,706
Sum of Contig Length	8,341,706
Number of bases	1,490,720,717
Number of Reads	97,280
N50 Read Length	24,932
Mean Read length	15,324
Mean Read Score	0.84
Reference Consensus Concordance (mean) (%)	99.999976
Reference coverage mean (%)	147.28
G+C content (%)	72.33
Number of Ns	0
Number of predicted protein-coding-genes*	7,358
Total length of coding sequences*	7,312,320
Number of predicted protein-coding-genes (≥99 bp)*	7,354
Total length of coding sequence (≥99 bp)*	7,311,951
tRNA*	67
rRNA*	21

* Gene prediction and structural annotation statistics

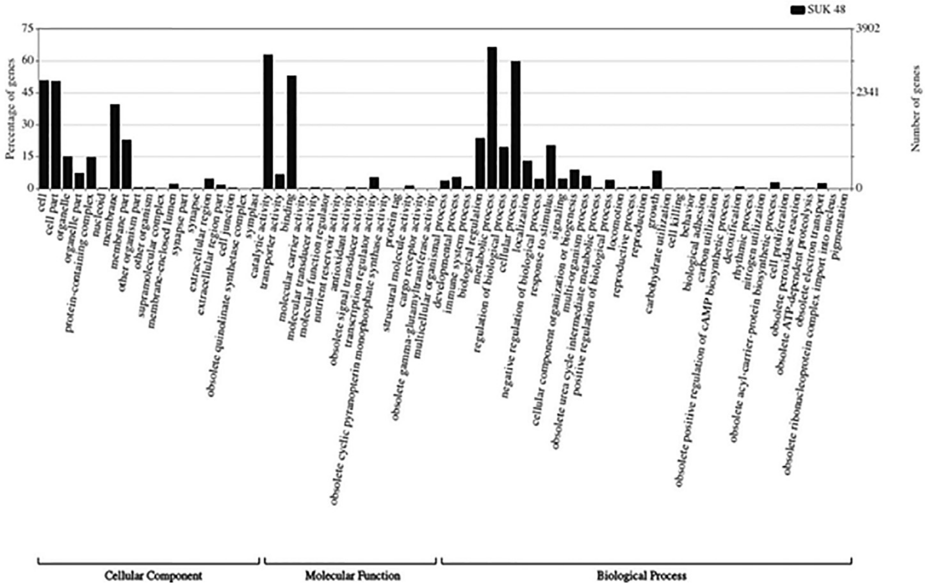


Fig. 1. Gene ontology (level 2) categories distribution.

coelicolor A3(2) (83.64%); *Streptomyces exfoliates* NRRL B-2924 (81.48%) and *Streptomyces griseolus* NRRL B-2925 (81.06%). This result suggests that *Streptomyces* sp. SUK 48 strain has unique genome properties.

A total of 7,354 protein coding genes were predicted, approximately 87.7% of the entire draft genome. Of the 7,354 protein coding genes, 7,261 (98.74%) of the genes were annotated with hits using the RefSeq database, whereas BLAST search against Swiss-Prot returned 4,434 or 60.29% of genes with hits. A total of 1,122 putative enzymes were mapped to 139 KEGG pathway maps. About 70.75% (5,203) of the sequences was annotated with 11,887 unique GO identifications. GO (level 2) categories distribution of SUK 48 is briefly described in Fig. 1.

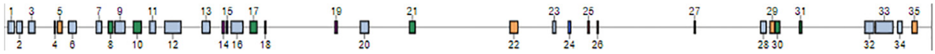
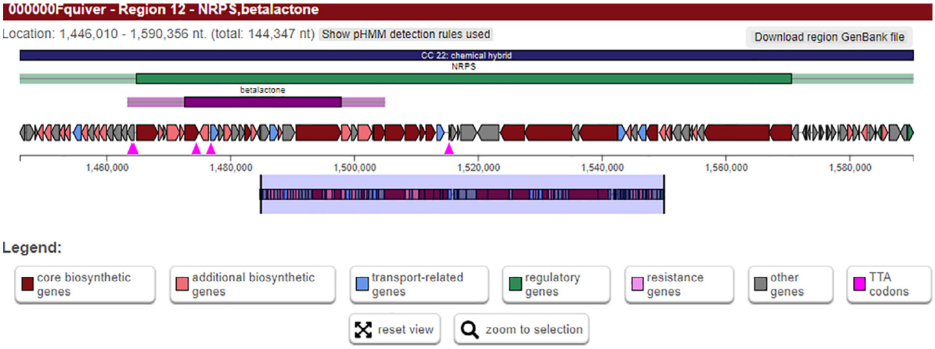


Fig. 2. Overview of BGCs of SUK 48.

Table 2
Putative gene clusters coding for secondary metabolites in SUK 48. Secondary metabolite detected by antiSMASH. T1pks: Type I PKS; T2pks: Type II PKS; T3pks: Type III PKS; Others: other types of PKS cluster; PKS: polyketide synthase; NRPS: Nonribosomal polypeptide synthetase; Bacteriocin: Bacteriocin or other unspecified ribosomally synthesis and post-translationally modified peptide product (RiPP) cluster; Lassopeptide: Lasso peptide cluster.

Region	Type	Length (bp)	Most similar known cluster	% of genes show similarity
1	Nrps-t1pks	56,343	kanamycin	1
2	Lanthipeptide-terpene-others	51,366	Carotenoid	63
3	Nrps-like-terpene	57,417	Ebelactone	8
4	Bacteriocin	7,847	Informatipeptin	28
5	T1pks	44,010	Versipelostatin	5
6	Lanthipeptide-nrps	70,195	Bleomycin	9
7	Nrps-t1pks	48,029	Rakicidin A/ rakicidin B	22
8	Nrps	43,782	Azicemicin	11
9	Nrps-t1pks-transAT-pks-like	90,163	Mirubactin	78
10	Nrps	71,088	Streptothricin	83
11	Thiopeptide-lap-terpene	52,731	Hopene	92
12	Nrps-betalactone	144,346	Kirromycin	16
13	Nrps-siderophore	70,321	Friulimicin	21
14	Terpene	21,836	Geosmin	100
15	Bacteriocin	11,280	-	-
16	Nrps-pks-like-t1pks-others	105,043	Thiolutin	40
17	Nrps	60,577	Ulleungmycin	11
18	Siderophore	10,340	-	-
19	Terpene	21,085	Albaflavenone	100
20	Oligosaccharide-pks-like-t1pks-t2pks	73,615	SF2572	54
21	Nrps	53,189	PM100117/PM100118	8
22	T2pks	72,511	Spore pigment	83
23	Butyrolactone-betalactone	27,548	A201A	6
24	Lassopeptide	22,553	Ikarugamycin	12
25	Siderophore	11,769	Desferrioxamine B	83
26	Melanin	10,599	Melanin	60
27	Ectoine	10,410	Ectoine	100
28	hglE-KS-t1pks	51,554	Cinnamycin	19
29	T3pks	41,064	Herboxidiene	7
30	Nrps	42,519	Stenothricin	13
31	Cdps	20,683	Foxicins A-D	12
32	Nrps-like-t1pks-nrps	80,998	Diisonitrile antibiotic SF2768	55
33	Nrps-t1pks	154,141	Filipin	100
34	Betalactone-nrps-like	41,784	Kirromycin	3
35	T1pks	46,305	Leinamycin	2

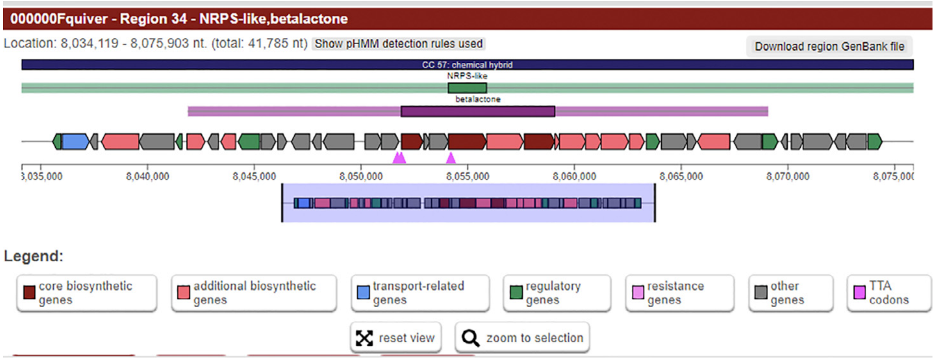
In addition, SUK 48 could produce important secondary metabolites. It was estimated that 34 gene clusters (Fig. 2) will be involved in the secondary metabolism of antiSMASH (Table 2). The present research based on nonribosomal peptides, ectoin and various BGC polyketides (types I, II and III) (biosynthetic gene clusters). While several BGCs were highly homologous with known secondary metabolite synthesis genes such as albaflavenone, ectoine, geosmin and filipine, most of them shared very low similarity with known BGCs. SUK 48 BGCs share the highest gene cluster similarity with the hopene (92 %) of *Streptomyces coelicolor* A3(2). Some NRPS BGCs are preserved and are not closely related to characterised homologs compared to recognised BGCs in the antiSMASH database. For example, kirromycin BGC in regions 12 (Fig. 3) and 34 (Fig. 4) only matched by 3-16% to homolog BGCs in SUK 48. This kirromycin BGC had responsible in production of kirromycin which in turn act as anti-plasmodial agent [3].



000000Fquiver - Region 12 - NRPS,betalactone

Location: 1,446,010 - 1,590,356 nt. (total: 144,347 nt)

Fig. 3. Overview of cluster 12 BGC of SUK 48.



000000Fquiver - Region 34 - NRPS-like,betalactone

Location: 8,034,119 - 8,075,903 nt. (total: 41,785 nt)

Fig. 4. Overview of cluster 34 BGC of SUK 48.

2. Experimental Design, Materials and Methods

Endophytic *Streptomyces* sp. SUK 48 was isolated from the Universiti Kebangsaan Malasia reserve forest [4]. Fruit of Brasilia sp. was cut into small pieces measured between 3 to 5 cm and cleaned under running tap water to cleanse from macroscopic foreign substance. Then, sterilization steps were done to cleanse form epiphyte microorganism. Sterilization begin with dissolved fruit sample withb 99% ethanol (v/l) within 60 s. Then, emersed them in 3.5% sodium hypochoride (NaClO) (v/l) within 6 min. Then the sample dissolved back in 99% ethanol (v/l) within 3 to 5 min. Lastly, the sample was rinse three time using sterile distilled water. Sterilization effectiveness was examined through dropping few drops of final sterile water on nutrient agar and incubated the culture at 37 °C for 5 to 7 days. The sample of fruit was culture on AIA (Actinomycetes Isolation Agar), WA (water agar) and SYCA (Starch Yeast Casein Agar) at 28 °C and monitored for 7 till 21 days. The bacterium of SUK 48 was isolated after two weeks of culture on SYCA. Then, the culture was maintained on International *Streptomyces* Project 2 agar (ISP2) at 28 °C [4–9].

Genomic DNA was extracted by using a Wizard® Genomic DNA Purification Kit as described by the manufacturer (Promega, USA). The sequencing was performed on a PacBio RS II platform (Treecodes, Singapore) generating one SMRT (single-molecule real-time) cell of sequencing data. Briefly, a DNA template consisting of a single molecule bound to a DNA polymerase was immobilized at the bottom of a ZMW (zero mode waveguide). This combined structure was illuminated from below by a laser light. Each of the four DNA bases was connected to one of four different fluorescent colours. When the nucleotide was incorporated into the DNA polymerase, the fluorescent tag was sealed off and diffused out of the ZMW observation field, where its fluorescence was no longer detectable. The detector was used to detect the fluorescent signal of nucleotide incorporation, and the base call was based on the corresponding fluorescence of the dye [10]. The sequencing data were pre-processed, de novo assembled and polished using the command line pbsmrtpipe of SMRT Link v6.0.0. Then, the Hierarchical Genome Assembly Process 4.0 (HGAP 4.0) was used to assemble the whole genome [10]. The assembly was improved by Quiver iteratively for three times using the resequencing pipeline. Through the Benchmarking Universal Single-Copy Orthologs (BUSCO) assessment, the polished genome was analyzed to obtain the complete genome [11].

The RS II data obtained in the h5 format were converted to the subreads.bam format to be fed into the SMRT Link v6.0.0 [10]. All files in the bax.h5 format were used to create the subreadset.xml file required for the SMRT Link analysis using the –type HdfSubreadSet parameter. The pipeline ID of the pbsmrtpipe.pipelines.sa3_hdfsubread_to_subread was used to convert the h5 reads to the analysis-ready subreads in the subreads.bam format. Next, the subreads were assembled using the Hierarchical Genome Assembly Process 4.0 (HGAP 4.0) pipeline with the pbsmrtpipe.pipelines.polished_falcon_fat. The parameters used for the genome assembly included the following settings: falcon_ns.task_options.HGAP_GenomeLenght_str to 8,000,000, input pa_DBsplit_option=-x500 –s100; ovlp_DBsplit_option = -x500 –s100 for falcon_ns.task_options.HGAP_FalconAdvanced_str along with the aggressive assembly mode turned on. The assembled genome from HGAP 4 was improved by Quiver iteratively for three times using the Resequencing pipeline pbsmrtpipepipeline.sa3_ds_resequencing_fat with default parameters. Benchmarking Universal Single-Copy Ortologs (BUSCO) v2 was used to test the completeness of the polished genome [12]. The actinobacteria odb9 profile was selected as the reference profile for this study.

The polished genome was taken as the input for structural annotation. First, tRNA was predicted using tRNAscan-SE v1.3.1 with default parameters [13]. Then, rRNA prediction was carried out using rnammer v1.2 by adding these parameters –S bac and –multi [14]. The polished genome was then masked off for the regions predicted to be tRNAs and rRNAs. The masked polished genome was used for gene prediction using Prodigal v2.6.3 with –c –m turned on. After gene prediction, the full repertoire of peptide sequences ($\geq$ 33 amino acids) was evaluated using BUSCO v2.0. The actinobacteria odb9 profile was selected as the reference profile for this study. The average nucleotide identity (ANI) analysis was calculated according to Goris et al. (2007) [15].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relation-ships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors wish to thanks to grant GP-2019-K006981for financial support assistance in this study.

Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2021.106768.

References

[1] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, M. Hattori, The KEGG resource for deciphering the genome, Nucleic Acids Research 32 (2004), doi:10.1093/nar/gkh063.

[2] K. Blin, V. Pascal Andreu, E.L.C. De Los Santos, F. Del Carratore, S.Y. Lee, M.H. Medema, T. Weber, The antiSMASH database version 2: A comprehensive resource on secondary metabolite biosynthetic gene clusters, Nucleic Acids Research 47 (2019), doi:10.1093/nar/gky1060.

[3] B. Clough, K. Rangachari, M. Strath, P.R. Preiser, R.J.M. Wilson, Antibiotic inhibitors of organellar protein synthesis in Plasmodium falciparum, Protist 150 (1999) 189–195, doi:10.1016/S1434-4610(99)70021-0.

[4] N.M. Zin, A.B. Nur Faizah, I. Aishah, M.S. Baba, N.M. Dan Sidik, Pencirian dan aktiviti antibakteria endofit streptomyces sp. dari hutan simpan penyelidikan UKM Bangi, Malaysian Applied Biology (2015).

[5] M.S. Baba, N.M. Zin, Z.A.A. Hassan, J. Latip, F. Pethick, I.S. Hunter, R.A. Edrada-Ebel, P.R. Herron, In vivo antimalarial activity of the endophytic actinobacteria, Streptomyces SUK 10, Journal of Microbiology 53 (2015) 847–855, doi:10.1007/s12275-015-5076-6.

[6] N.I.M. Sarmin, G.Y.A. Tan, C.M.M. Franco, R. Edrada-Ebel, J. Latip, N.M. Zin, Streptomyces kebangsaanensis sp. nov., an endophytic actinomycete isolated from an ethnomedicinal plant, which produces phenazine-1-carboxylic acid, International Journal of Systematic and Evolutionary Microbiology 63 (2013) 3733–3738, doi:10.1099/ijs.0.047878-0.

[7] A.S. Junaidah, S. Suhaini, H.M. Sidek, D.F. Basri, N.M. Zin, Anti-methicillin resistant staphylococcus aureus activity and optimal culture condition of streptomyces sp. SUK 25, Jundishapur Journal of Microbiology (2015), doi:10.5812/jjm.16784.

[8] M.M. Alshaibani, N. Mohamadzin, J. Jalil, N.M. Sidik, S.J. Ahmad, N. Kamal, R. Edrada-Ebel, Isolation, purification, and characterization of five active diketopiperazine derivatives from endophytic Streptomyces SUK 25 with antimicrobial and cytotoxic activities, Journal of Microbiology and Biotechnology (2017), doi:10.4014/jmb.1608.08032.

[9] E.B. Shirling, D. Gottlieb, Methods for characterization of Streptomyces species, International Journal of Systematic Bacteriology 16 (1966) 313–340, doi:10.1099/00207713-16-3-313.

[10] C.-S. Chin, D.H. Alexander, P. Marks, A.A. Klammer, J. Drake, C. Heiner, A. Clum, A. Copeland, J. Huddleston, E.E. Eichler, S.W. Turner, J. Korlach, Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data, Nature Methods (2013), doi:10.1038/nmeth.2474.

[11] F.A. Simão, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics (2015), doi:10.1093/bioinformatics/btv351.

[12] F.A. Simão, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics 31 (2015), doi:10.1093/bioinformatics/btv351.

[13] T.M. Lowe, P.P. Chan, tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes, Nucleic Acids Research (2016), doi:10.1093/nar/gkw413.

[14] K. Lagesen, P. Hallin, E.A. Rødland, H.H. Stærfeldt, T. Rognes, D.W. Ussery, RNAmmer: Consistent and rapid annotation of ribosomal RNA genes, Nucleic Acids Research (2007), doi:10.1093/nar/gkm160.

[15] J. Goris, K.T. Konstantinidis, J.A. Klappenbach, T. Coenye, P. Vandamme, J.M. Tiedje, DNA-DNA hybridization values and their relationship to whole-genome sequence similarities, International Journal of Systematic and Evolutionary Microbiology (2007), doi:10.1099/ijs.0.64483-0.