Data Article

First draft genome sequence data of TA4-1, the type strain of Gram-positive bacterium *Streptomyces chiangmaiensis*

Montri Yasawong [a,b], A'liyatur Rosyidah [c], Thunwarat Songngamsuk [a], Manassanan Phatcharaharikarn [a], Phongsakorn Ganta [d], Panjamaphon Chanthasena [e], Nuannoi Chudapongse [f], Napatsorn Santapan [d], Wissarut Srisakvarangkool [d], Supavadee Kerdtoob [e], Nawarat Nantapong [d,*]

[a] *Programme on Environmental Toxicology, Chulabhorn Graduate Institute, Chulabhorn Royal Academy, Bangkok 10210, Thailand*
[b] *Center of Excellence on Environmental Health and Toxicology (EHT), OPS, MHESI, Bangkok 10400, Thailand*
[c] *Research Center for Vaccine and Drugs, National Research and Innovation Agency (BRIN), Bogor 16911, Indonesia*
[d] *School of Preclinical Sciences, Institute of Science, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand*
[e] *Faculty of Allied Health Sciences, Nakhonratchasima College, Nakhon Ratchasima 30000, Thailand*
[f] *Department of Basic Medical Sciences, Faculty of Medicine, Siam University, Bangkok 10160, Thailand*

A R T I C L E   I N F O

A B S T R A C T

TA4-1 is the type strain of *Streptomyces chiangmaiensis*. The TA4-1 strain was isolated from a stingless bee (*Tetragonilla collina*). Here we present the draft genome sequence data of *S. chiangmaiensis* TA4-1. The Illumina NextSeq 550 sequencer was used to generate paired-end reads from the genomic DNA of the pure culture of *S. chiangmaiensis* TA4-1. The draft genome sequence of strain TA4-1 consists of 776 contigs with a total size of 9,707,984 base pairs, an N50 of 32,937 base pairs, and a GC content of 69.73 %. Digital DNA-DNA hybridisation (dDDH) and average nucleotide identity (ANI) analysis showed that *S. yaanensis* CGMCC 4.7035 had the highest dDDH value (32.7 %) and ANIm value (88.50 %) when compared with TA4-1. The presented data indicate the

* Corresponding author.
E-mail address: nawarat@sut.ac.th (N. Nantapong).
Social media: @OhooNanta (N. Nantapong).

potential for a reference genome sequence in bacterial taxonomy, comparative genomics, and the investigation of bioactive compound biosynthesis in *S. chiangmaiensis* TA4-1. The draft genome sequence data have been deposited at NCBI under the Bioproject accession number PRJNA680432.

Specifications Table

Subject	Biological sciences
Specific subject area	Omics: Genomics
Data format	Raw and analysed
Type of data	Tables, figures
Data collection	DNA was extracted using the phenol-chloroform method and sequenced on an Illumina NextSeq 550. Fastp v0.23.4 was used for adapter trimming and quality filtering. Unicycler v0.5.0 was employed for genome assembly, and QUAST v5.0.2 was used to determine assembly metrics. CheckM v1.1.2 was used to assess genome quality. The Type (Strain) Genome Server was used for phylogenomic tree analysis. A genomic map was generated using Proksee. Genome annotation was performed using NCBI PGAP. Gene clusters involved in secondary metabolite biosynthesis were identified using antiSMASH v7.0.
Data source location	*Streptomyces chiangmaiensis* TA4-1 (TISTR 1982) was purchased from the Thailand Institute of Scientific and Technological Research (TISTR), Pathum Thani, Thailand (14°02′39.5″N 100°43′08.0″E).
Data accessibility	Sequencing data were deposited in the National Center for Biotechnology Information (NCBI) Genbank database under the accession number JAYWVC000000000. The deposited draft genome sequencing data are available at https://www.ncbi.nlm.nih.gov/nuccore/JAYWVC000000000.

1. Value of the Data

- The draft genome sequence of *S. chiangmaiensis* TA4-1 serves as a reference genome for bacterial taxonomy using average nucleotide identity and digital DNA-DNA hybridisation.
- The draft genome sequence of *S. chiangmaiensis* TA4-1 can benefit comparative genomic research on other *Streptomyces* species that produce bioactive compounds.
- Elucidating the genome sequence of *S. chiangmaiensis* TA4-1 may assist in the investigation of the biosynthesis of bioactive compounds.

2. Background

Streptomyces chiangmaiensis is a Gram-positive bacterium isolated from the Southeast Asian stingless bee, *Tetragonilla collina*, collected in Chiang Mai Province, Thailand [1]. *Streptomyces* is well known for its diverse secondary metabolites [2]. Although the specific bioactive compounds of *S. chiangmaiensis* are not well documented. *Streptomyces* species are rich sources of antibiotics, antifungals, antivirals, antitumors, antihypertensives, and immunosuppressants [2,3]. *Streptomyces* is a promising candidate for the biocontrol of phytopathogenic microorganisms because of its broad biotechnological potential [4]. Further research is required to investigate the bioactive compounds produced by *S. chiangmaiensis* and their potential applications [1].

3. Data Description

Here we present the draft genome sequence data for *S. chiangmaiensis* TA4-1 (Fig. 1). Gene clusters containing genes that produce bioactive compounds and their predicted phenotypes are also included.

The genome consisted of 776 contigs with a total size of 9,707,984 bp, an N50 value of 32,937 bp, and a GC content of 69.73 % (Table 1).

The draft genome of *S. chiangmaiensis* TA4-1 was found to be 99.82 % complete, with an estimated contamination of 1.35 %. The digital DNA-DNA hybridisation (dDDH) and average nucleotide identity (ANI) values show that *S. yaanensis* CGMCC 4.7035 has the highest dDDH value (32.7 %) and ANIm value (88.50 %) compared with TA4-1 (Table 2). Fig. 2 presents the phylogenomic tree of strain TA4-1 and its closely related type strains.

Analysis of gene clusters involved in secondary metabolite biosynthesis revealed that *S. chiangmaiensis* TA4-1 can synthesise antimicrobial agents, including albaflavenone [6] and ε-poly-L-lysine [7] (Table 3). The study also demonstrated that TA4-1 can produce flaviolin 1,3,6,8-tetrahydroxynaphthalene, which has UV protection properties [8]. Additionally, *S. chiangmaiensis* TA4-1 can synthesise the iron chelator desferrioxamine [9].

The draft genome sequence data can be used as a reference genome for bacterial taxonomy and is expected to aid in identifying the genes responsible for the synthesis of bioactive compounds and secondary metabolites in *S. chiangmaiensis* TA4-1.
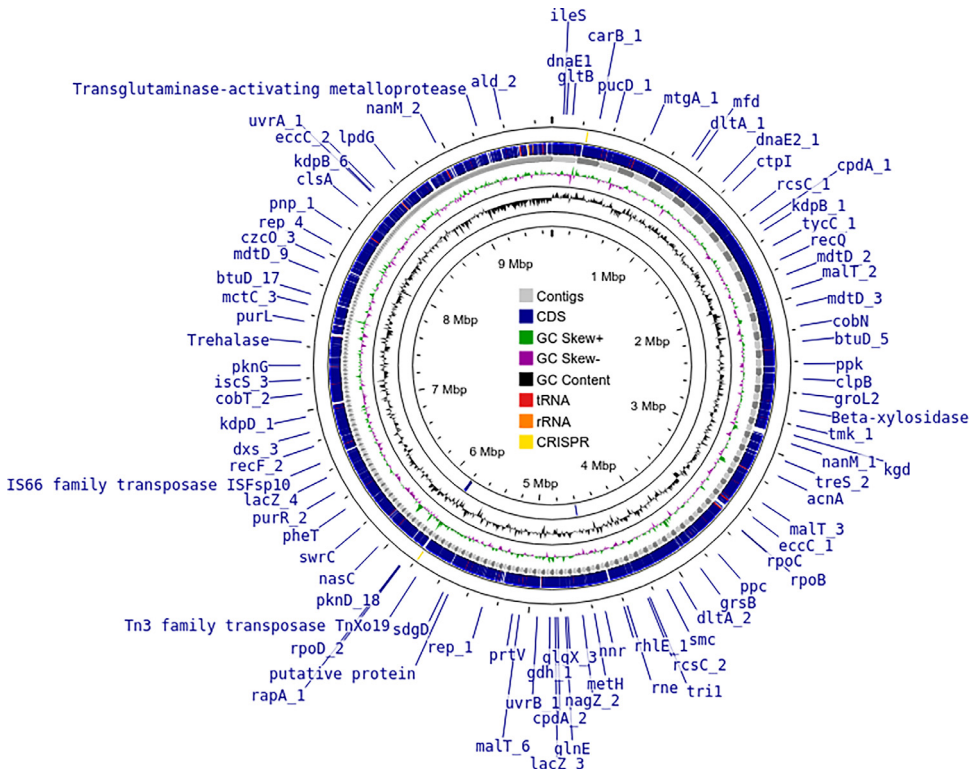

Fig. 1. The genome map of *S. chiangmaiensis* TA4-1 was created using Proksee [5]. Blue arrows indicate coding DNA sequences (CDSs), and grey arrows represent contigs. The green and purple peaks represent GC skew+ and GC skew-, respectively, and the black peaks indicate GC content.

Table 1

Genomic features and assembly statistics of *S. chiangmaiensis* TA4-1.

Attribute	*S. chiangmaiensis* TA4-1
Genome size (bp)	9,707,984
Number of contigs	776
Genome coverage	195×
GC content (%)	69.73
Largest contig (bp)	187,615
N50	32,937
N75	15,872
L50	80
L75	182
Total gene	9,376
Total CDS	9,330
tRNA	40
rRNA	3
ncRNA	3
CRISPR repeat	1

Table 2

Digital DNA-DNA hybridisation (dDDH) and average nucleotide identity (ANI) values were calculated for strain TA4-1 and closely related *Streptomyces* species.

No.	Species	Strain	Accession number	dDDH (%)	ANIb (%)	ANIm (%)
1	*Streptomyces yaanensis*	CGMCC 4.7035	CP134053	32.7	84.36	88.50
2	*Streptomyces colonosanans*	MUSC 93	MLYP00000000	31.2	82.87	87.97
3	*Streptomyces glomeratus*	DSM 41457	JAKEEB000000000	30.9	83.82	87.85
4	*Streptomyces cynarae*	HUAS 13-4	CP106793	30.5	83.23	87.71
5	*Streptomyces nodosus*	ATCC 14899	CP009313	30.3	82.84	87.60
6	*Streptomyces nodosus*	DSM 40109	JACHMR000000000	30.2	82.82	87.58
7	*Streptomyces macrolidinus*	RY43-2	JAMWMR000000000	28.9	81.91	87.25
8	*Streptomyces avermitilis*	NBRC 14893	BA000030	27.0	81.49	86.38
9	*Streptomyces avermitilis*	DSM 46492	LMWT01000000	27.0	81.54	86.39
10	*Streptomyces avermitilis*	MA-4680	BAVY00000000	27.0	81.46	86.38
11	*Streptomyces speibonae*	NRRL B-24240	JNXM00000000	25.9	79.43	86.13
12	*Streptomyces leeuwenhoekii*	C34	AZSD00000000	25.8	79.72	86.21
13	*Streptomyces harenosi*	PRKS01-65	WYCT01000000	25.8	79.63	86.16
14	*Streptomyces leeuwenhoekii*	DSM 42122	LN831790	25.6	79.65	86.18
15	*Streptomyces chromofuscus*	DSM 40273	CP063374	25.5	79.82	86.08
16	*Streptomyces indiaensis*	DSM 43803	JAKEIO000000000	25.2	79.76	85.87
17	*Streptomyces thermoviolaceus* subsp. *apingens*	JCM 4312	BMSR01000000	24.9	78.82	85.93
18	*Streptomyces fuscus*	GXMU-J15	JASJUS000000000	24.8	79.40	85.84
19	*Streptomyces thermodiastaticus*	JCM 4840	BNBV01000000	24.6	78.88	85.81
20	*Streptomyces pluripotens*	MUSC 135	CP021080	24.0	78.64	85.59

dDDH, digital DNA-DNA hybridisation; ANIb, average nucleotide identity algorithm using BLAST; ANIm, average nucleotide identity algorithm using MUMmer.
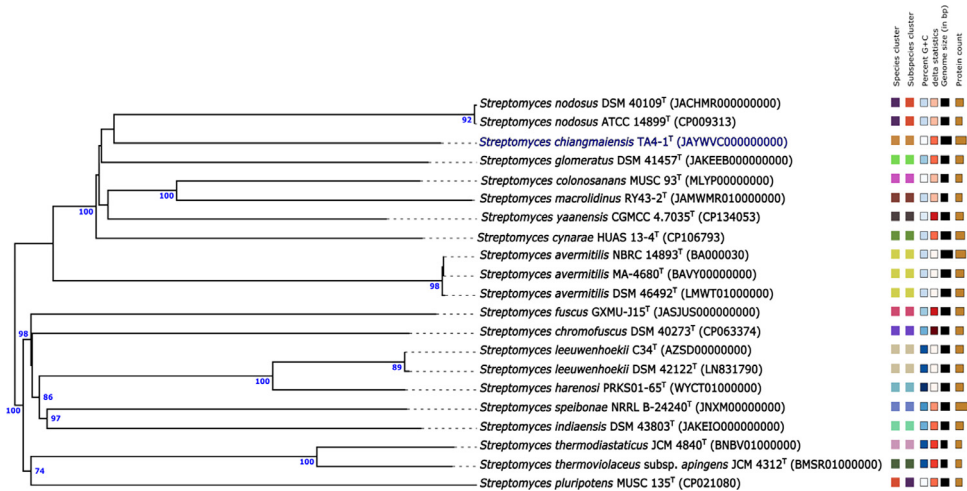
Fig. 2. The phylogenomic tree was reconstructed using whole-genome sequence data from *S. chiangmaiensis* TA4-1 and its closely related type strain on the TYGS platform. Branch numbers were determined based on pseudo-bootstrap support values greater than 70 % from 100 replicates using Genome Blast Distance Phylogeny (GBDP), with an average branch support of 75.0 %.

Table 3

Secondary metabolite biosynthesis gene clusters identified by antiSMASH.

Contig	Type	Position (bp)	Closest known biosynthetic class	Predicted secondary metabolites	Similarity
5	Terpene	65,852-86,355	Terpene	Geosmin	100 %
22	Ectoine	20,948-31,346	Other	Ectoine	100 %
28	T2PKS	24,922-63,772	Polyketide	Spore pigment	75 %
29	Siderophore	9,805-39,577	Other	Desferrioxamine B/Desferrioxamine E	100 %
41	T3PKS	3,937-45,001	Polyketide	Flaviolin/1,3,6,8-tetrahydroxynaphthalene	100 %
53	Terpene	6,045-27,091	Terpene	Albaflavenone	100 %
57	NAPAA	21,704-41,027	NRP	ε-Poly-L-lysine	100 %
268	Terpene	1-7,255	Terpene	Hopene	53 %

NAPAA, non-alpha poly-amino acids like ε-Polylysin; NRP, non-ribosomal peptide; T2PKS, type II polyketide synthases; and T3PKS, type III polyketide synthases.

4. Experimental Design, Materials, and Methods

4.1. Bacterial Strain and Genomic DNA Preparation

The type strain of *S. chiangmaiensis* (TA4-1 = TISTR 1982) was obtained from the Thailand Institute of Scientific and Technological Research (TISTR), Pathum Thani, Thailand. The genomic DNA (gDNA) of strain TA4-1 was extracted from cells grown in International *Streptomyces* Project-2 (ISP-2) broth at 37°C for 48 h. Bacterial cells were collected by centrifugation at 10,000 rpm for 5 min and transferred to microcentrifuge tubes. The cell pellet was suspended in 500 μL of TNE buffer (0.1 mM Tris HCl, 0.01 mM EDTA, 1M NaCl). Next, 50 μL of lysozyme (40 mg/mL) and 25 μL of RNAse A (25 mg/mL) were added, and the mixture was incubated at 37°C for 60 min. Subsequently, 60 μL of 10 % SDS and 40 μL of 1 mg/mL proteinase K were mixed and incubated at 50°C for 20 min. Next, a mixture of phenol, chloroform, and isoamyl alcohol (25:24:1) was added, and the tube was inverted to facilitate mixing before being centrifuged at 5,000 rpm for 15 min. The upper layer was transferred to a new tube and mixed with 50 μL of 3M

Na-acetate and one volume of phenol, chloroform, and isoamyl alcohol (25:24:1). The mixture was then centrifuged at 5,000 rpm for 15 min. Subsequently, the top layer was collected, and DNA precipitation was performed using 1.25 mL of ice-cold absolute ethanol. The mixture was then incubated at -20°C for 5 min, followed by centrifugation at 7,000 rpm for 15 min to collect the DNA pellet. The gDNA pellet was washed three times with 1 mL of 70 % ice-cold ethanol and centrifuged at 10,000 rpm for 5 min. The gDNA was then dried in a dry bath at 50°C and dissolved in 50 μL of nuclease-free water. The quality of gDNA was assessed using agarose gel electrophoresis and NanoDrop spectrophotometry (Thermo Scientific, USA).

4.2. Whole Genome Sequencing and Assembly

The Nextera XT DNA library preparation kit (Illumina, CA, USA) was used to generate sequencing libraries from 1 ng of DNA. Raw sequencing reads were obtained using a NextSeq 550 sequencer (Illumina, CA, USA) and a NextSeq 500/550 high output kit v2.5 (300 cycles, 2 × 150-bp reads). Quality assessments, adapter trimming, and quality filtering were performed using fastp v0.23.4 with default settings [10]. Unicycler v0.5.0 was used for *de novo* genome assembly of the raw reads with default settings [11]. Genome assembly metrics were evaluated using QUAST v5.0.2 with default parameters [12].

4.3. Taxonomic Identification of the Strain

The assessment of genome quality was conducted using CheckM v1.1.2 with default settings [13]. Digital DNA-DNA hybridisation (dDDH) analysis and a phylogenomic tree derived from the whole genome sequences of TA4-1 and its related type strains were conducted using the Type (Strain) Genome Server (TYGS) [14]. The ANI values for strain TA4-1 and closely related *Streptomyces* species were calculated using JSpeciesWS [15].

4.4. Genome Annotation and Analysis of Gene Clusters Involved in Secondary Metabolite Biosynthesis

Proksee [5] generated a genomic map of *S. chiangmaiensis* TA4-1, which was then annotated using the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) with default settings [16]. Furthermore, we conducted annotation and analysis of gene clusters responsible for secondary metabolite biosynthesis using antiSMASH v7.0 with default settings [17].

Limitations

Next-generation sequencing techniques produce large amounts of data. However, the resulting *de novo* genome assemblies are often incomplete, which makes them susceptible to annotation errors. This is particularly true for the imprecise estimation of genes that may exist in the TA4-1 draft genome.

Ethics Statement

This study did not involve the use of human or animal subjects. The authors declare that this manuscript is original and has not been published elsewhere.

Data Availability

Streptomyces chiangmaiensis strain TA4-1, whole genome shotgun sequencing project (Original data) (The National Center for Biotechnology Information (NCBI) Genbank database).

CRediT Author Statement

Montri Yasawong: Methodology, Data curation, Writing – original draft, Writing – review & editing; **A'liyatur Rosyidah:** Methodology, Data curation, Writing – original draft; **Thunwarat Songngamsuk:** Methodology, Data curation; **Manassanan Phatcharaharikarn:** Methodology, Data curation; **Phongsakorn Ganta:** Methodology, Data curation; **Panjamaphon Chanthasena:** Methodology, Data curation; **Nuannoi Chudapongse:** Methodology, Data curation; **Napatsorn Santapan:** Methodology, Data curation; **Wissarut Srisakvarangkool:** Methodology, Data curation; **Supavadee Kerdtoob:** Methodology, Data curation; **Nawarat Nantapong:** Conceptualization, Data curation, Supervision, Writing – original draft, Writing – review & editing.

Acknowledgements

Declaration of Competing Interest

The authors declare that they have no conflicts of interest that could have influenced the work reported in this paper.

References

[1] Y. Promnuan, T. Kudo, M. Ohkuma, P. Chantawannakul, *Streptomyces chiangmaiensis* sp. nov. and *Streptomyces lannensis* sp. nov., isolated from the South-East Asian stingless bee (*Tetragonilla collina*), Int. J. Syst. Evol. Microbiol. 63 (2013) 1896–1901, doi:10.1099/ijs.0.045930-0.
[2] V. Sharma, R. Kaur, R. Salwan, *Streptomyces*: host for refactoring of diverse bioactive secondary metabolites, 3. Biotech. 11 (2021) 340, doi:10.1007/s13205-021-02872-y.
[3] J. Bérdy, Bioactive microbial metabolites, J. Antibiot. (Tokyo) 58 (2005) 1–26, doi:10.1038/ja.2005.1.
[4] S. Khan, S. Srivastava, A. Karnwal, T. Malik, *Streptomyces* as a promising biological control agents for plant pathogens, Front. Microbiol. 14 (2023) 1285543, doi:10.3389/fmicb.2023.1285543.
[5] J.R. Grant, E. Enns, E. Marinier, A. Mandal, E.K. Herman, C.Y. Chen, M. Graham, G. Van Domselaar, P. Stothard, Proksee: in-depth characterization and visualization of bacterial genomes, Nucleic. Acids. Res. 51 (2023) W484–W492, doi:10.1093/nar/gkad326.
[6] T. Kobayashi, Y. Kon, H. Abe, H. Ito, Concise total synthesis of albaflavenone utilizing sequential intramolecular aldol condensation: determination of absolute configuration, Org. Lett. 24 (2014) 6397–6399, doi:10.1021/ol503202d.
[7] L. Wang, C. Zhang, J. Zhang, Z. Rao, X. Xu, Z. Mao, X. Chen, Epsilon-poly-L-lysine: Recent advances in biomanufacturing and applications, Front. Bioeng. Biotechnol. (2021) 748976, doi:10.3389/fbioe.2021.748976.
[8] R. Mishra, D. Dhakal, J.M. Han, H.N. Lim, H.J. Jung, T. Yamaguchi, J.K. Sohng, Production of a novel tetrahydroxynaphthalene (THN) derivative from *Nocardia* sp. CS682 by metabolic engineering and its bioactivities, Molecules. 2 (2019) 244, doi:10.3390/molecules24020244.
[9] D. Bellotti, M. Remelli, B Deferoxamine, A natural, excellent and versatile metal chelator, Molecules. 26 (2021) 3255, doi:10.3390/molecules26113255.
[10] S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics. 34 (2018) i884–i890, doi:10.1093/bioinformatics/bty560.
[11] 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.
[12] A. Mikheenko, A. Prjibelski, V. Saveliev, D. Antipov, A. Gurevich, Versatile genome assembly evaluation with QUAST-LG, Bioinformatics. 34 (2018) i142–i150, doi:10.1093/bioinformatics/bty266.
[13] D.H. Parks, M. Imelfort, C.T. Skennerton, P. Hugenholtz, G.W. Tyson, CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes, Genome Res. 25 (2015) 1043–1055, doi:10.1101/gr.186072.114.

[14] J.P. Meier-Kolthoff, M. Göker, TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy, Nat. Commun. 10 (2019) 2182, doi:10.1038/s41467-019-10210-3.

[15] M. Richter, R. Rosselló-Móra, F. Oliver Glöckner, J. Peplies, JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison, Bioinformatics. 32 (2016) 929–931, doi:10.1093/bioinformatics/btv681.

[16] T. Tatusova, M. DiCuccio, A. Badretdin, V. Chetvernin, E.P. Nawrocki, L. Zaslavsky, A. Lomsadze, K.D. Pruitt, M. Borodovsky, J. Ostell, NCBI prokaryotic genome annotation pipeline, Nucleic. Acids. Res. 44 (2016) 6614–6624, doi:10.1093/nar/gkw569.

[17] K. Blin, S. Shaw, H.E. Augustijn, Z.L. Reitz, F. Biermann, M. Alanjary, T. Weber, antiSMASH 7.0: New and improved predictions for detection, regulation, chemical structures and visualisation, Nucl. Acids Res. 51 (2023) W46–W50, doi:10.1093/nar/gkad344.