The Draft Whole-Genome Sequence of the Antibiotic Producer *Empedobacter haloabium* ATCC 31962 Provides Indications for Its Taxonomic Reclassification

Henrike Miess,[a] Patricia Arlt,[a] Alexander Kristian Apel,[a] Tilmann Weber,[b] Kay Nieselt,[c] Friederike Hanssen,[d] Stefan Czemmel,[d] Sven Nahnsen,[d] (ID) Harald Gross[a,e]

[a]Department of Pharmaceutical Biology, Pharmaceutical Institute, University of Tübingen, Tübingen, Germany
[b]The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark
[c]Integrative Transcriptomics, Institute for Bioinformatics and Medical Informatics, University of Tübingen, Tübingen, Germany
[d]Quantitative Biology Center (QBiC), University of Tübingen, Tübingen, Germany
[e]German Centre for Infection Research (DZIF), Partner Site Tübingen, Tübingen, Germany

ABSTRACT Strain ATCC 31962 was formerly taxonomically classified as *Empedobacter haloabium* and reported to be the producer of the lipopeptide antibiotic empedopeptin. Here, we report the draft genome sequence of ATCC 31962, which encodes regions that suggest a distinct biosynthetic capacity and suggests its taxonomic reclassification.

*E*mpedobacter haloabium ATCC 31962 (syn. designation G393-B445) was isolated in the early 1980s from a soil sample collected in Yamate-Dori, Tokyo, Japan, and was suggested to be a new species of the genus *Empedobacter* (1). It exhibited antibacterial activity in a screening panel and was found to produce the highly potent antibiotic empedopeptin (syn. BMY-28117 and Bu-2517) (2, 3), which acts as a cell wall biosynthesis inhibitor by complex formation with peptidoglycan precursors (4). In order to investigate the complete biosynthetic capacity for secondary metabolism, to locate the biosynthetic gene cluster of empedopeptin, and to clarify the taxonomic position of the strain for safety reasons, the genome of ATCC 31962 was sequenced.

Strain ATCC 31962 was grown in 20 ml meat medium (2% soluble starch, 1% glucose, 0.2% meat extract, 0.2% Bacto yeast extract, 0.5% N-Z-Case, 0.2% $CaCO_3$, and double-distilled water [ddH_2O] [pH 7.0]) for 2 to 3 days at 27°C on a rotary shaker (140 rpm). For genomic DNA isolation, the Qiagen genomic DNA purification kit was used in combination with 100/G Genomic-tips according to the manufacturer's protocol, except that for the bacterial lysis, the handled volumes were doubled, and the incubation time at 50°C was prolonged until a clear lysate was obtained.

The whole-genome sequence of ATCC 31962 was obtained using a combined strategy of paired-end sequencing (NEBNext sample preparation kit, 2 × 76 bp) with an Illumina GA IIx instrument and mate pair sequencing (Nextera mate pair sample preparation kit v2, 151 bp) with an Illumina MiSeq instrument. Subsequently, FastQC v0.11.2 (5) was used to check both libraries for adapter content and base quality. From the paired-end library, 47,704,862 reads of 75 bp were obtained. FastQC identified adapter contamination, and therefore reads were adapter clipped, and the remaining 41,254,498 reads were subject to *de novo* assembly. Within the frame of the mate pair approach, 3,395,400 paired-end reads of 150 bp were sequenced. FastQC did not reveal any adapter content, and base quality was high, so all original reads were subjected to the second assembly. In both cases, the reads were assembled using SOAPdenovo v.1.05 (6, 7). The contigs of both libraries were combined and then subjected to the string-based assembly of MADAM v.beta-version:06-2013 (8). Default parameters were

used for all software mentioned above, except that k = 63 was used for the library with a small insert size and k = 109 for the library with a large insert size. Overall, the final assembly yielded a 6,678,028-nucleotide draft genome at 102-fold coverage, consisting of 106 contigs (N_{50}, 132,565 bp; L_{50}, 18) in total. The *in silico*-determined G+C content of 66.5% is in full agreement with the experimentally determined value of 66.5 ± 1.5% (1). The assembled contigs were annotated with the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) v4.8 (9), yielding a total of 5,690 predicted protein-coding sequences.

A BLAST analysis of the partial 16S rRNA of strain ATCC 31962 (GenBank accession number NR_125708) identified *Massilia* sp. strain YMA4 (KX444135) and *Massilia armeniaca* ZMN-3 (CP028324) as the closest cultured representatives, both sharing 99.5% 16S rRNA gene sequence identity with ATCC 31962. Using autoMLST (10), the overall genome analysis was shown not to support the initial taxonomic classification of the strain as an *Empedobacter* species. According to an average nucleotide identity (ANI) analysis, the closest related strains are *Massilia* sp. strain NR4-1 (ANI, 83.6%) and *Rugamonas rubra* ATCC 43154^T (ANI, 83.5%).

Data availability. This whole-genome sequencing (WGS) project has been deposited at DDBJ/ENA/GenBank under the accession number VPFC00000000. The sequencing reads have been deposited under the accession numbers SRX6060211 and SRX6060212. All reads have been deposited in the SRA and are associated with BioProject number PRJNA532449.

REFERENCES

1. Konishi M, Sugawara K, Hanada M, Tomita K, Tomatsu K, Miyaki T, Kawaguchi H, Buck RE, More C, Rossomano VZ. 1984. Empedopeptin (BMY-28117), a new depsipeptide antibiotic. I. Production, isolation and properties. J Antibiot (Tokyo) 37:949–957. https://doi.org/10.7164/antibiotics.37.949.
2. Sugawara K, Numata K, Konishi M, Kawaguchi H. 1984. Empedopeptin (BMY-28117), a new depsipeptide antibiotic. II. Structure determination. J Antibiot (Tokyo) 37:958–964. https://doi.org/10.7164/antibiotics.37.958.
3. Kawaguchi H, Konishi M, Sugawara K, Tomita K. October 1983. Antibiotic compound. US patent 4,409,210. https://patents.google.com/patent/US4409210.
4. Müller A, Münch D, Schmidt Y, Reder-Christ K, Schiffer G, Bendas G, Gross H, Sahl HG, Schneider T, Brötz-Oesterhelt H. 2012. Lipodepsipeptide empedopeptin inhibits cell wall biosynthesis through Ca^{2+}-dependent complex formation with peptidoglycan precursors. J Biol Chem 287:20270–20280. https://doi.org/10.1074/jbc.M112.369561.
5. Andrews S. 2011. FastQC. Babraham Institute, Cambridge, United Kingdom. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
6. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J. 2010. *De novo* assembly of human genomes with massively parallel short read sequencing. Genome Res 20:265–272. https://doi.org/10.1101/gr.097261.109.
7. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW, Wang J. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. GigaScience 1:18. https://doi.org/10.1186/2047-217X-1-18.
8. Seitz A, Nieselt K. 2017. Improving ancient DNA genome assembly. PeerJ 5:e3126. https://doi.org/10.7717/peerj.3126.
9. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. 2016. NCBI Prokaryotic Genome Annotation Pipeline. Nucleic Acids Res 44:6614–6624. https://doi.org/10.1093/nar/gkw569.
10. Alanjary M, Steinke K, Ziemert N. 2019. AutoMLST: an automated Web server for generating multi-locus species trees highlighting natural product potential. Nucleic Acids Res 47:W276–W282. https://doi.org/10.1093/nar/gkz282.