Genome Sequence of Streptomyces sp. Strain RTd22, an Endophyte of the Mexican Sunflower

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters.

Citation

Published Version
doi:10.1128/genomeA.00693-16

Citable link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:27822280

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA
We report here the complete genome sequence of *Streptomyces* sp. strain RTd22, an endophytic actinobacterium that was isolated from the roots of the Mexican sunflower *Tithonia diversifolia*. The bacterium’s 11.1-Mb linear chromosome is predicted to encode a large number of unknown natural products.

**Endophytic bacteria**, in particular those in the genus *Streptomyces*, are potential reservoirs of new antibiotic and chemotherapeutic compounds. As part of a larger research program aimed at discovering biologically active small molecules from bacterial endophytes, we isolated *Streptomyces* sp. strain RTd22 from the Mexican sunflower, *Tithonia diversifolia* (*Asteraceae*). The plants were grown and harvested in Ribeirão Preto, São Paulo, Brazil, and the isolation and identification of endophytic bacteria were performed as previously described (1).

A draft genome sequence of *Streptomyces* sp. RTd22 was generated from paired-end libraries constructed with the Nextera library preparation kit. The sequencing (2× 250-bp [300 cycle kit]) was performed using V2 Illumina sequencing chemistry run on a MiSeq instrument and produced 7.8 million reads. After filtering and adapter and quality trimming, the reads were assembled using the Lasergene SeqMan NGen version 3.1 (DNAStar) assembler program, converted to Consed-compatible ACE format, and checked using Consed (2, 3). The draft assembly consisted of 199 contigs (made into 97 scaffolds), with an average coverage of 70×. The 199 contigs comprise 11,180,448 bp, with a G+C content of 71.3%. Annotation was carried out using Artemis (4). Genome annotation using RAST (5) predicted 9,577 coding sequences, including 93 RNA genes.

To overcome the assembly issues that are commonly associated with short-read data from high-G+C organisms, we resequenced RTd22 using PacBio single-molecule real-time (SMRT) sequencing technology (6). Specifically, an insert library of an 11.5 kb was prepared, and sequence data were generated from 4 SMRT cells run on a Pacific Biosciences RSII instrument using P6-C4 chemistry. *De novo* assembly was performed using the Hierarchical Genome Assembly Process (HGAP) (7), and after manual cur-ration to remove low-quality bases, a single contiguous sequence of 11,142,275 bp with an average G+C content of 71.3% was produced. The RAST annotation server allowed the identification of 9,667 predicted protein-coding genes, including 88 RNA genes. Analysis of the 11.1-Mb RTd22 chromosome using antiSMASH (8) predicted ~40 biosynthetic gene clusters (BGCs) for secondary metabolite production. Among the 40 predicted BGCs, very few can be confidently assigned to a known natural product based on gene and nucleotide conservation. One exception is a BGC that is predicted to encode a himastatin-like molecule (9). More generally, the chromosome appears to be enriched in terpene-related genes with other abundant classes of BGCs, including nonribosomal peptides and several polyketide clusters that encode polyene macrolides.

Altogether, genome sequencing has revealed that the endophyte *Streptomyces* sp. RTd22 encodes a large chemical reservoir that merits further studies.

**Nucleotide sequence accession numbers.** The whole-genome sequence has been deposited at DDBJ/ENA/GenBank under the accession numbers LXIC00000000 (Illumina) and CP015726 (PacBio). The version described in this paper is the first version.

**ACKNOWLEDGMENTS**

Illumina sequencing was performed at the Department of Biochemistry, University of Cambridge, Cambridge, England, and PacBio RSII sequencing was performed by the Duke Center for Genomic and Computational Biology Sequencing and Genomic Technologies Shared Resource core facility, Durham, NC.

We thank the Harvard Medical School Information Technology Department for access to and maintenance of the Orchestra High Performance Computing cluster.
**FUNDING INFORMATION**

This work, including the efforts of Monica T Pupo, was funded by São Paulo Research Foundation (FAPESP) (2013/07600-3, 2013/50954-0, and 2008/09540-0). This work, including the efforts of Luciana Gonzaga de Oliveira, was funded by São Paulo Research Foundation (FAPESP) (2011/06209-3). This work, including the efforts of Fernanda O Chagas, was funded by São Paulo Research Foundation (FAPESP) (2009/17695-6). This work, including the efforts of Larissa V Bacha, was funded by São Paulo Research Foundation (FAPESP) (2011/11703-7). This work, including the efforts of Raphael Conti, was funded by São Paulo Research Foundation (FAPESP) (2008/00812-7). This work, including the efforts of Rita C Pessotti, was funded by São Paulo Research Foundation (FAPESP) (2011/12910-6). This work, including the efforts of Jon Clardy, was funded by HHS | National Institutes of Health (NIH) (U19TW009872). This work, including the efforts of Fernanda O Chagas, was funded by MCTI | Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (150572/2015-8). This work, including the efforts of Antonio Ruzzini, was funded by Gouvernement du Canada | Canadian Institutes of Health Research (CIHR) (201511MFE).

**REFERENCES**


