genomeA
Journals.ASM.org

De Novo Assembly of the *Streptomyces* sp. Strain Mg1 Genome Using PacBio Single-Molecule Sequencing

B. Christopher Hoefler,[a] Kranti Konganti,[b] Paul D. Straight[a]

Department of Biochemistry and Biophysics, Texas A&M University, College Station, Texas, USA[a]; Whole Systems Genomics Initiative, Texas A&M University, College Station, Texas, USA[b]

We report a draft genome assembly of *Streptomyces* sp. strain Mg1, a competitive soil isolate with multiple secondary metabolite gene clusters.

Received 17 June 2013 **Accepted** 24 June 2013 **Published** 1 August 2013

Citation Hoefler BC, Konganti K, Straight PD. 2013. *De novo* assembly of the *Streptomyces* sp. strain Mg1 genome using PacBio single-molecule sequencing. Genome Announc. 1(4):e00535-13. doi:10.1128/genomeA.00535-13.

Address correspondence to Paul D. Straight, paul_straight@tamu.edu.

Streptomyces sp. strain Mg1 is a competitive soil bacterium with a complex secondary metabolism. An intriguing characteristic of Mg1 is its ability to cause lysis and degradation of *Bacillus subtilis* cells and colonies (1). The genome was sequenced to enable prediction of secondary metabolites and determination of their relative contributions to the competitive functions of Mg1.

Whole-genome shotgun sequencing of the Mg1 strain was carried out using PacBio SMRT sequencing technology (2). For assembly of the Mg1 genome, we applied the recently described hierarchical genome assembly process (HGAP) to eight SMRT cells of sequencing data generated from an 8- to 10-kb insert library (3). The final assembly is 8.7 Mb in seven contigs. Greater than 90% of the predicted genome size is contained within one large 7.8-Mb contig. The remaining sequence is divided into smaller contigs 50 to 500 kb in length.

The microbial genomes of *Streptomyces* sp. are challenging to sequence. These genomes are characteristically high in GC content and possess large (>8-kb) rRNA gene clusters (4, 5). Many contain biosynthetic gene clusters encoding polyketide synthases (PKSs) and nonribosomal peptide synthetases (NRPSs), large multimodular enzymes with repetitive domain structures (6). These sequence features are difficult to assemble using second-generation approaches, primarily because short read lengths limit the ability of assembly algorithms to resolve low-complexity and repetitive regions (7). Due to their extraordinary length, single-molecule sequencing reads are capable of spanning long repeats which, along with a distinct lack of coverage bias, greatly simplify the process of genome assembly. Previously, the Mg1 genome was assembled into 466 contigs (GenBank, ABJF00000000). Comparison to the current assembly shows many collapsed repeats that could not be resolved and nearly 1 Mb of sequence missing due to large coverage gaps. High coverage of short reads from 454 GS-FLX titanium and Illumina HiSeq 100-bp paired-end data did not resolve the problems with the previous assembly.

With this updated assembly, we were able to predict multiple secondary metabolite gene clusters using antiSMASH (8). All of the predicted PKS gene clusters from the previous assembly were broken at contig boundaries, including the one that produces chalcomycin A (9). In the current assembly, the chalcomycin gene cluster is fully contiguous along with seven other predicted PKS and NRPS gene clusters. Additional PKS and NRPS gene clusters remain fragmented, but the ability to mine the Mg1 genome for secondary metabolites and predict their structures has been greatly improved.

The maturation of single-molecule sequencing provides an unprecedented way for researchers to assemble and finish microbial genomes. Assembly of a nearly finished Mg1 genome using only PacBio technology illustrates the accessibility of single-molecule sequencing for studying the medically and industrially significant *Actinobacteria*. Improvements in genome contiguity and accuracy aid analyses that require large-scale ordering of the genome sequence, such as the mining of secondary metabolite gene clusters.

Nucleotide sequence accession numbers. This whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession no. ATCJ00000000. The version described in this paper is the first version, ATCJ01000000.

ACKNOWLEDGMENTS

This work was supported by a Catalyst grant from the Whole Systems Genomics Initiative (WSGI) of Texas A&M University, a Welch Foundation grant (A-1796), and an NSF-CAREER grant (MCB-1253215).

We appreciate the kind assistance of Olivier Fedrigo with our genome assembly and acknowledge the Duke Institute for Genome Sciences and Policy (IGSP) for PacBio sequencing.

REFERENCES

1. **Barger SR, Hoefler BC, Cubillos-Ruiz A, Russell WK, Russell DH, Straight PD.** 2012. Imaging secondary metabolism of streptomyces sp. Mg1 during cellular lysis and colony degradation of competing Bacillus subtilis. Antonie Van Leeuwenhoek **102:**435–445.
2. **Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, deWinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A,**

Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S. 2009. Real-time DNA sequencing from single polymerase molecules. Science **323:**133–138.
3. **Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J.** 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods **5:**563–569.
4. **Bentley SD, Chater KF, Cerdeño-Tárraga A-M, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D, Bateman A, Brown S, Chandra G, Chen CW, Collins M, Cronin A, Fraser A, Goble A, Hidalgo J, Hornsby T, Howarth S, Huang C-H, Kieser T, Larke L, Murphy L, Oliver K, O'Neil S, Rabbinowitsch E, Rajandream M-A, Rutherford K, Rutter S, Seeger K, Saunders D, Sharp S, Squares R, Squares S, Taylor K, Warren T, Wietzorrek A, Woodward J, Barrell BG, Parkhill J, Hopwood DA.** 2002. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature **417:**141–147.
5. **Ohnishi Y, Ishikawa J, Hara H, Suzuki H, Ikenoya M, Ikeda H, Yamashita A, Hattori M, Horinouchi S.** 2008. Genome sequence of the streptomycin-producing microorganism *Streptomyces griseus* IFO 13350. J. Bacteriol. **190:**4050–4060.
6. **Fischbach MA, Walsh CT.** 2006. Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. Chem. Rev. **106:**3468–3496.
7. **Koren S, Harhay GP, Smith TP, Bono JL, Harhay DM, Mcvey DS, Radune D, Bergman NH, Phillippy AM.** 2013. Reducing assembly complexity of microbial genomes with single-molecule sequencing. arXiv preprint, Cornell University Library, Ithaca, NY. http://arxiv.org/ftp/arxiv/papers/1304/1304.3752.pdf.
8. **Blin K, Medema MH, Kazempour D, Fischbach MA, Breitling R, Takano E, Weber T.** 2013. antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Res. **41:** W204–W212. doi:10.1093/nar/gkt449.
9. **Asolkar RN, Maskey RP, Helmke E, Laatsch H.** 2002. Chalcomycin B, a new macrolide antibiotic from the marine isolate *Streptomyces* sp. B7064. J. Antibiot. (Tokyo) **55:**893–898.