Draft Genome Sequence of *Oleiagrimonas soli* 3.5X^T, a Type Species in a Newly Identified Genus, Isolated from an Oil Field in China

Yong Huang, Tingting Fang, Hui Wang, Haiyan Zhou

State Joint Laboratory on Environmental Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing, People's Republic of China

Y.H. and T.F. contributed equally to this work.

Oleiagrimonas gudaosoli 3.5X^T was isolated from an oil field and identified as a new member of a novel genus. The draft genome sequence of this strain, which comprises 3,379,958 bp encoding 3,010 open reading frames (ORFs), can provide insight into the life style of this newly identified genus in petroleum-contaminated soil.

*O*leiagrimonas soli 3.5X^T (=NBRC 110685^T =KCTC 42351^T =CPCC 100614^T) (also known as *Xanthomonadaceae* bacterium 3.5X) (1) was originally isolated from an oil field in Shandong Province (China) during a study on the characterization of halophilic bacteria that show optimal polycyclic aromatic hydrocarbon (PAH)-degrading activity in petroleum-contaminated soil. The strain was isolated using PAHs as the sole carbon source on an agar plate; however, it showed no ability to degrade PAHs in liquid medium. The phylogenetic position of the strain was determined to be within the family *Xanthomonadaceae* and designated a new genus and species (1). To illustrate its potential role in degrading PAHs as well as other hydrocarbons, the genome of this strain was sequenced.

Genomic DNA from the strain was extracted and precipitated following the cetyltrimethylammonium bromide (CTAB) protocol (2). The whole-genome sequence of strain 3.5X was performed using Illumina MiSeq technology (Major bio, China) after the construction of two sequencing libraries, a 400-bp paired-end library and a 3,000-bp mate-pair library. In total, 787.1 Mbp (paired-end) and 380.0 Mbp (mate-pair) filtered data, which correspond to an average coverage of 232.8× and 112.4×, respectively, were generated. The raw reads from the two libraries were assembled together using SOAPdenovo version 1.05 (3) and Gap-Closer (3), yielding 22 contigs with the largest length of 883,955 bp and N_{50} of 389,485 bp. These contigs were further arranged into 6 scaffolds, with the largest length of 3,072,697 bp. The total assembly length of the genome is 3,379,958 bp, and the G+C content is 63.8%.

Glimmer 3.0 (4) was used to predict open reading frames (ORFs). The potential function of these predicted ORFs was annotated by using BLASTp (5) against several databases including NCBI-NR, the KEGG database (6), and the GO database (7). The rRNA and tRNA gene predictions were separately conducted by RNAmmer (8) and the tRNAscan-SE program (9). The determined draft genome contains 3,010 ORFs, within which only 1,647 genes have predicted functions in the KEGG database. Three rRNA genes (16S, 23S, and 5S rRNA) and 46 tRNA genes were predicted.

From the metabolic pathway predicted by KEGG, several genes related to hydrocarbon degradation were predicted, e.g., genes encoding methyltransferase-like protein in the PAH degradation pathway and phenol 2-monooxygenase for toluene or resorcinol degradation. However, these genes just play peripheral roles in hydrocarbon degradation, while key genes, like aromatic dioxygenase and alkane mono-oxygenase, are not present in the genome. It seems that this strain cannot utilize hydrocarbons in oil as substrates directly. Conversely, gene annotation revealed the existence of 25 genes related to a bacterial secretion system and 13 genes associated with protein export, corresponding to the appearance of sticky colloids around the strain colony on the agar plate. These results indicate that this strain has the ability to secrete extracellular polymeric substances, which may facilitate hydrocarbon degradation by other microbes.

Nucleotide sequence accession numbers. This whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under accession no. JROI00000000. The version described in this paper is version JROI00000000.1.

REFERENCES

1. **Fang T, Wang H, Huang Y, Zhou H, Dong P**. 27 February 2015. *Oleiagrimonas soli* gen. nov., sp. nov., a genome-sequenced gammaproteobacterium isolated from an oil field in China. Int J Syst Evol Microbiol. http://dx.doi.org/10.1099/ijs.0.000158.
2. **Wilson K**. 2001. Preparation of genomic DNA from bacteria. Current Protoc Mol Biol Chapter 2:Unit 2.4. http://dx.doi.org/10.1002/0471142727.mb0204s56.
3. **Li Y, Hu Y, Bolund L, Wang J**. 2010. State of the art *de novo* assembly of human genomes from massively parallel sequencing data. Hum Genomics 4:271–277. http://dx.doi.org/10.1186/1479-7364-4-4-271.

4. **Delcher AL, Bratke KA, Powers EC, Salzberg SL**. 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics **23:**673–679. http://dx.doi.org/10.1093/bioinformatics/btm009.

5. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ**. 1990. Basic local alignment search tool. J Mol Biol **215:**403–410. http://dx.doi.org/10.1016/S0022-2836(05)80360-2.

6. **Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M**. 2014. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res **42:**D199–D205. http://dx.doi.org/10.1093/nar/gkt1076.

7. **Jones CE, Baumann U, Brown AL**. 2005. Automated methods of predicting the function of biological sequences using GO and blast. BMC Bioinformatics **6:**272. http://dx.doi.org/10.1186/1471-2105-6-272.

8. **Lagesen K, Hallin P, Rødland EA, Stærfeldt HH, Rognes T, Ussery DW**. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res **35:**3100–3108. http://dx.doi.org/10.1093/nar/gkm160.

9. **Schattner P, Brooks AN, Lowe TM**. 2005. The tRNAscan-SE, snoscan and snoGPS Web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res **33:**W686–W689. http://dx.doi.org/10.1093/nar/gki366.