Complete Genome Sequence of *Sphingopyxis macrogoltabida* Strain 203N (NBRC 111659), a Polyethylene Glycol Degrader

Yoshiyuki Ohtsubo,[a] Shouta Nonoyama, Yuji Nagata,[a] Mitsuru Numata,[b] Keiko Tsuchikane,[b] Akira Hosoyama,[b] Atsushi Yamazoe,[b] Masataka Tsuda,[a] Nobuyuki Fujita,[b] Fusako Kawai[c]

Department of Environmental Life Sciences, Graduate School of Life Sciences, Tohoku University, Sendai, Japan[a]; Biological Resource Center, National Institute of Technology and Evaluation, Tokyo, Japan[b]; Center for Fiber and Textile Science, Kyoto Institute of Technology, Kyoto, Japan[c]

We determined the complete genome sequence of *Sphingopyxis macrogoltabida* strain 203N, a polyethylene glycol degrader. Because the PacBio assembly (285× coverage) seemed to be full of nucleotide-level mismatches, the Newbler assembly of MiSeq mate-pair and paired-end data was used for finishing and the PacBio assembly was used as a reference. The PacBio assembly carried 414 nucleotide mismatches over 5,953,153 bases of the 203N genome.

*S*phingopyxis macrogoltabida strain 203 was isolated from soil as the polyethylene glycol (PEG)-utilizing *Flavobacterium* sp. strain 203 (1). Later, the strain was designated the type strain of *Sphingomonas macrogoltabidus* (2) and reidentified as *Sphingopyxis macrogoltabida* (3), based on the taxonomical standards proposed by Yabuuchi et al. (4). The strain was deposited to the National Institute of Technology and Evaluation (Tokyo, Japan) and stocked under the number NBRC 15033. The complete genome of NBRC 15033 was determined, but the genes for PEG utilization were missing, and repeated cultivation was assumed to be the reason for the loss (5). From a laboratory stock, we recovered a strain, designated 203N, harboring the *pegA* gene (6, 7) and capable of growing on PEG.

Here, we report the complete genome sequence of *S. macrogoltabida* 203N. To determine the complete sequence, we obtained PacBio data from Macrogen Japan. The total number of reads obtained was 237,846 with an N_{50} length of 9,733 bp and a total length of 1.7 Gb. The reads were assembled by HGAP3, and three circular contigs corresponding to the main chromosome and two plasmids were obtained. However, we found that the sequences differ considerably to those of NBRC 15033 (5). Besides one genomic rearrangement, which was predicted to cause the loss, and 10 differences related to insertion sequences, a huge number (approximately 400) of nucleotide-level mismatches were counted. We also obtained Illumina MiSeq reads from the very DNA solution used for PacBio sequencing, and the assembled contig sequences suggested that PacBio assembly was erroneous. Replacing each part of the PacBio assembly by a corresponding MiSeq contig seemed inappropriate for correcting the errors, because the nucleotide-level mismatches were located throughout the genome, and contigs deriving from repeats in the genome might carry variation bases. Therefore, we decided to start the finishing from the Newbler assembly of the MiSeq reads obtained from mate-pair and PCR-free paired-end libraries. The finishing

was facilitated by using ShortReadManager, GenoFinisher, and AceFileViewer (AFV) (8), which have been used to determine complete genome sequences, often enabling the complete *in silico* finishing, especially when the PCR-free kit of the Illumina sequence library preparation was used. In the finishing, the PacBio assembly was used as a reference to search the correct paths of contigs that fill each gap in scaffolds. The correct sequences of the paths were determined by AFV. The finished sequence was confirmed by FinishChecker, wherein genomic *k*-mers not found in the MiSeq reads were searched and corrected as necessary. Thus, the complete genome sequence of 203N was determined. We found 414 nucleotide-level mismatches to the PacBio assembly, most of which were found at homopolymeric stretches, which is not a characteristic error pattern of Illumina sequencing reads but may be one for PacBio.

The sequences were annotated by the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) and curated using GenomeMatcher (9). While referring to the annotation data obtained from the Microbial Genome Annotation Pipeline (http://www.migap.org) (10), we corrected start codon positions and added genes that were missing in the PGAP annotation.

Nucleotide sequence accession numbers. The genome sequence of *Sphingopyxis macrogoltabida* strain 203N has been deposited in NCBI/GenBank under the accession numbers CP013344 to CP013346. *Sphingopyxis macrogoltabida* strain 203N is available from the Biological Resource Center, National Institute of Technology and Evaluation (Tokyo, Japan). Its deposit number is NBRC 111659.

ACKNOWLEDGMENTS

REFERENCES

1. **Kawai F, Kimura T, Tani Y, Yamada H**. 1984. Involvement of a polyethylene glycol (PEG)-oxidizing enzyme in the bacterial metabolism of PEG. Agric Biol Chem **48**:1349–1351.
2. **Takeuchi M, Kawai F, Shimada Y, Yokota A**. 1993. Taxonomic study of polyethylene glycol-utilizing bacteria: emended description of the genus *Sphingomonas* and new descriptions of *Sphingomonas macrogoltabidus* sp. nov., *Sphingomonas sanguis* sp. nov. and *Sphingomonas terrae* sp. nov. Syst Appl Microbiol **16**:227–238. http://dx.doi.org/10.1016/S0723 -2020(11)80473-X.
3. **Takeuchi M, Hamana K, Hiraishi A**. 2001. Proposal of the genus *Sphingomonas* sensu stricto and three new genera, *Sphingobium*, *Novosphingobium* and *Sphingopyxis*, on the basis of phylogenetic and chemotaxonomic analyses. Int J Syst Evol Microbiol **51**:1405–1417. http://dx.doi.org/ 10.1099/00207713-51-4-1405.
4. **Yabuuchi E, Yano I, Oyaizu H, Hashimoto Y, Ezaki T, Yamamoto H**. 1990. Proposals of *Sphingomonas paucimobilis* gen. nov. and comb. nov., *Sphingomonas parapaucimobilis* sp. nov., *Sphingomonas yanoikuyae* sp. nov., *Sphingomonas adhaesiva* sp. nov., *Sphingomonas capsulata* comb. nov., and two genospecies of the genus *Sphingomonas*. Microbiol Immunol **34**:99–119. http://dx.doi.org/10.1111/j.1348 -0421.1990.tb00996.x.
5. **Ohtsubo Y, Nagata Y, Numata M, Tsuchikane K, Hosoyama A, Yamazoe A, Tsuda M, Fujita N, Kawai F**. 2015. Complete genome sequence of *Sphingopyxis macrogoltabida* type strain NBRC 15033, originally isolated as a polyethylene glycol degrader. Genome Announc **3**(6):e01401-15. http://dx.doi.org/10.1128/genomeA.01401-15.
6. **Sugimoto M, Tanabe M, Hataya M, Enokibara S, Duine JA, Kawai F**. 2001. The first step in polyethylene glycol degradation by sphingomonads proceeds via a flavoprotein alcohol dehydrogenase containing flavin adenine dinucleotide. J Bacteriol **183**:6694–6698. http://dx.doi.org/10.1128/ JB.183.22.6694-6698.2001.
7. **Tani A, Charoenpanich J, Mori T, Takeichi M, Kimbara K, Kawai F**. 2007. Structure and conservation of a polyethylene glycol-degradative operon in sphingomonads. Microbiology **153**:338–346. http://dx.doi.org/ 10.1099/mic.0.2006/000992-0.
8. **Ohtsubo Y, Maruyama F, Mitsui H, Nagata Y, Tsuda M**. 2012. Complete genome sequence of *Acidovorax* sp. strain KKS102, a polychlorinated-biphenyl degrader. J Bacteriol **194**:6970–6971. http:// dx.doi.org/10.1128/JB.01848-12.
9. **Ohtsubo Y, Ikeda-Ohtsubo W, Nagata Y, Tsuda M**. 2008. GenomeMatcher: a graphical user interface for DNA sequence comparison. BMC Bioinformatics **9**:376. http://dx.doi.org/10.1186/1471-2105-9-376.
10. **Sugawara H, Ohyama A, Mori H, Kurokawa K**. 2009. Microbial genome annotation pipeline (MiGAP) for diverse users, abstr S-001, p 1–2. Abstr. 20th Int. Conf. Genome Informatics, Kanagawa, Japan.