The HKU Scholars Hub   The University of Hong Kong   香港大學學術庫

Title	**Draft Genome Sequence of the Haloacid-Degrading Burkholderia caribensis Strain MBA4**
Author(s)	**Pan, Y; Kong, KF; Tsang, JSH**
Citation	**Genome Announcements, 2014, v. 2 n. 1, p. e00047-14**
Issued Date	**2014**
URL	**http://hdl.handle.net/10722/196403**
Rights	**Genome Announcements. Copyright © American Society for Microbiology.**

Draft Genome Sequence of the Haloacid-Degrading *Burkholderia caribensis* Strain MBA4

Yanling Pan, Ka Fai Kong, Jimmy S. H. Tsang

Molecular Microbiology Laboratory, School of Biological Sciences, The University of Hong Kong, Hong Kong SAR, China

***Burkholderia caribensis* MBA4 was isolated from soil for its ability to utilize 2-haloacid. An inducible haloacid operon, encoding a dehalogenase and a permease, is mainly responsible for the biotransformation. Here, we report the draft genome sequence of this strain.**

Address correspondence to Jimmy S. H. Tsang, jshtsang@hku.hk.

Haloacetates such as monochloroacetate (MCA) are toxic and mutagenic and can be produced incidentally during disinfection of water. *Burkholderia caribensis* MBA4 is a Gram-negative bacterium that can utilize 2-haloacid as a growth substrate. This bacterium was characterized for its production of a dimeric hydrolytic dehalogenase (Deh4a) (1, 2) that removes the halogen from the carbon backbone. Here we describe the draft genome sequence of *Burkholderia caribensis* MBA4.

Analysis of *B. caribensis* MBA4 with pulsed-field gel electrophoresis showed that it has a genome size of more than 9 Mb with at least three replicons (data not shown). Whole-genomic sequencing was obtained with 454 GS FLX Titanium and Illumina HiSeq 2000. With low-quality short reads discarded, the 454 sequencing has 929,485 reads and 380,525,001 bp after trimming. Four sets of Illumina paired-end libraries with insert sizes of 100, 300, 500, and 2,000 bp were constructed and sequenced. After trimming and filtering, the four libraries have 37,483,321, 36,788,695, 23,594,431, and 12,689,821 high-quality paired-end reads, respectively. The average read lengths were 61, 61, 69, and 39 bp, respectively. The overall coverage is about 750-fold. Illumina paired-end and 454 reads were *de novo* assembled using CLC Genomic Workbench 6.0.1 (CLC bio, Aarhus, Denmark) with default settings. SSPACE basic 2.0 (3) was used to join contigs into scaffolds with information derived from paired-end reads. Moreover, 47,627 *de novo*-assembled transcripts from nine sets of RNA-seq data were mapped to the scaffolds to (i) remove some of the internal gaps, (ii) remove ambiguous base pairs, and (iii) join the scaffolds together. Standard PCR and Sanger-sequencing technology were applied to fill gaps inside the scaffolds. Multiplex PCR was used to amplify unknown regions between scaffolds, and some scaffolds were linked after subsequent cloning and sequencing. As a result, 14 scaffolds were obtained with 79 component contigs of >200 bp. Contig relationships were maintained in the GenBank submission by the inclusion of an AGP (A Golden Path) file. The total size of the contigs is 9,418,480 bp. The N_{50} of the contigs is 217,392 bp and the longest contig is 1,305,062 bp. The GC content was determined to be 62.48%, which is consistent with a result obtained from high-performance liquid chromatography (HPLC) analysis.

The draft genome was annotated automatically with the Rapid Annotations using Subsystems Technology (RAST) server (4) and the Prokaryotic Genomes Automatic Annotation Pipeline (PGAAP) from NCBI (5). The draft genome contains 9,082 genes, including 8 rRNA and 52 tRNA genes. Furthermore, there were 624 tandem repeats identified by Tandem Repeats Finder (6). Among the 9,022 protein-coding sequences, 76% were annotated as encoding known proteins while the remaining 24% encode hypothetical products. Among these RAST-annotated genes, 3,666 coding DNA sequences (CDS) were assigned to 27 subsystems. Analysis of the CDS with the KEGG Automatic Annotation Server (version 1.6a) (7) has specified 34 groups with 191 pathways.

Nucleotide sequence accession numbers. This whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession number AXDD00000000. The version described in this paper is version AXDD01000000.

ACKNOWLEDGMENTS

REFERENCES

1. **Tsang JSH, Pang BCM.** 2000. Identification of the dimerization domain of dehalogenase IVa of *Burkholderia cepacia* MBA4. Appl. Environ. Microbiol. **66:**3180–3186. http://dx.doi.org/10.1128/AEM.66.8.3180-3186.2000.
2. **Tsang JSH, Sallis PJ, Bull AT, Hardman DJ.** 1988. A monobromoacetate dehalogenase from *Pseudomonas cepacia* MBA4. Arch. Microbiol. **150:**441–446. http://dx.doi.org/10.1007/BF00422284.
3. **Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W.** 2011. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics **27:**578–579. http://dx.doi.org/10.1093/bioinformatics/btq683.
4. **Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T,**

Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. 2008. The RAST server: rapid annotations using subsystems technology. BMC Genomics **9**:75. http://dx.doi.org/10.1186/1471-2164-9-75.

5. **Angiuoli SV, Gussman A, Klimke W, Cochrane G, Field D, Garrity G, Kodira CD, Kyrpides N, Madupu R, Markowitz V, Tatusova T, Thomson N, White O.** 2008. Toward an online repository of standard operating procedures (SOPs) for (meta)genomic annotation. Omics **12**:137–141. http://dx.doi.org/10.1089/omi.2008.0017.

6. **Benson G.** 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. **27**:573–580. http://dx.doi.org/10.1093/nar/27.2.573.

7. **Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M.** 2007. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. **35**:W182–W185. http://dx.doi.org/10.1093/nar/gkm321.