Data in Brief

Isolation and complete genome sequencing of *Mimivirus bombay*, a Giant Virus in sewage of Mumbai, India

Anirvan Chatterjee, Farhan Ali, Disha Bange, Kiran Kondabagil *

Indian Institute of Technology Bombay, Powai, Mumbai, Maharashtra 400076, India

**Abstract**

We report the isolation and complete genome sequencing of a new *Mimiviridae* family member, infecting *Acanthamoeba castellanii*, from sewage in Mumbai, India. The isolated virus has a particle size of about 435 nm and a 1,182,200-bp genome. A phylogeny based on the DNA polymerase sequence placed the isolate as a new member of the *Mimiviridae* family lineage A and was named as *Mimivirus bombay*. Extensive presence of *Mimiviridae* family members in different environmental niches, with remarkably similar genome size and genetic makeup, point towards an evolutionary advantage that needs to be further investigated. The complete genome sequence of *Mimivirus bombay* was deposited at GenBank/EMBL/DDBJ under the accession number KU761889.

© 2016 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

**Keywords:** Mimivirus, NCLDV, Giant virus, CRISPR

**Organism/cell line/tissue**

*Mimivirus bombay*

**Sex**

NA

**Sequencer or array type**

Illumina MiSeq v2 150 x 2 PE

**Data format**

analyzed, complete genome FASTA sequence

**Experimental factors**

virus grown in *Acanthamoeba castellanii*

**Experimental features**

de novo genome assembly and annotation

**Consent**

not applicable

**Sample source location**

Mumbai, India, City, 19.180158 N, 72.848614 E

**Direct link to deposited data**

http://www.ebi.ac.uk/ena/data/view/KU761889

http://www.ncbi.nlm.nih.gov/nuccore/KU761889

**NCBI Sequence graphics**

https://www.ncbi.nlm.nih.gov/nuccore/1020259557/report=graph

*Corresponding author: Department of Biosciences and Bioengineering, Indian Institute of Technology Bombay, Powai, Mumbai 400076, India.

E-mail addresses: kirankondabagil@iitb.ac.in, kirankondabagil@gmail.com (K. Kondabagil).

http://dx.doi.org/10.1016/j.gdata.2016.05.013

2213-5960/© 2016 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Whole genome shotgun sequencing

Library preparation was performed at the Genotypic Technology's (Bengaluru, India) Genomics facility according to the SureSelect\textsuperscript{XT} Library Prep protocol outlined in the Sure Select\textsuperscript{XT} whole genome library prep for Illumina multiplexed sequencing protocol (Cat #5500–0121). Twenty five nanogram of genomic DNA was fragmented and the adapter-tag was added using Sure Select\textsuperscript{XT}. Amplitied adapter-tagged libraries were purified using high prep beads clean up kit (MACGIO, USA). The libraries were quantified using Qubit fluorometer and quality validated by running an aliquot on D1000 Tape (Cat# 5067–5582) using D1000 Tape Station Kit (Agilent, Cat# 5067–5583). After quality check, the library was sequenced using IlluminaMiSeq v2 2 × 150 bp paired-end sequencing.

Genome assembly and annotation

Adapter trimming and read filtering for QV > 30 was performed using Agilent SureCall suite. De novo assembly was performed using multiple assemblers including SOAPdenovo\textsuperscript{2} (15), A5-miseq (5), Velvet (18) and SPAdes (3), and were evaluated using QUAST (10). MAUVE (6) was used to reorder the contigs and generate consensus FASTA. Open reading frames (ORFs) were predicted with GeneMarkS (4), individually annotated using Blastp (2) and the results were retrieved using custom Python scripts. Phylogenetic analysis was performed using MEGA-CC Linux distribution (11). A5 miseq provided the best assembly parameters with a median coverage of 714 × and N50 of 906,835. All

![Fig. 1. Transmission electron micrograph of Mimivirus bombay (MVB).](image)

![Fig. 2. Amino acid sequence of MVB ORF#318, annotated as DNA polymerase, was used as input sequence for blastp query against non-redundant protein sequence database.](image)
contigs were aligned to BLAST NR database using MEGABLAST [16] and the consensus FASTA was generated by reordering the 7 contigs using MAUVE [6]. MVB has a genome size of 1,182,200 bp with 898 predicted ORFs. The annotated genome was uploaded to NCBI using BankIt web based submission tool.

Data description
Transmission electron microscopy revealed virus particles of about 435 nm in size (Fig. 1), similar to some recently reported giant viruses known as Nucleo-Cytoplasmic Large DNA Viruses (NCLDV) [17]. Illumina Basespace web tool (Kraken metagenomics) taxonomically classified 98% of the total 3,017,739 reads (the trimmed and QC filtered) as Mimiviridae. Hence the isolate was named as Mimivirus bombay (MVB). Further, a Maximum Likelihood (ML) based phylogeny of DNA mimiviruses [1]. The GC content of MVB (28%) is also comparable to other mimiviruses [10].

The assembled complete genome was deposited to NCBI under accession number KU761889.1.

Nucleotide accession number
The assembled complete genome was deposited to NCBI under accession number KU761889.1.

Acknowledgements
This work is supported by IIT Bombay Seed grant (11IRCCG004) to KK. AC is supported by IIT Bombay Post-Doctoral Fellowship. FA and DB were supported by Department of Biotechnology (DBT) Masters Program.

References
[1] S. Aher, P. Colson, B. La Scola, D. Raoult, Giant viruses of amoebas: an update. Front. Microbiol. 7 (2016) 349.
[2] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool. J. Mol. Biol. 215 (1990) 403–410.
[3] A. Bankievich, S. Nurm, D. Antipov, A.A. Gurevich, M. Dvorkin, A.S. Kulikov, V.M. Lesin, S.I. Nikolenko, S. Pham, A.D. Pfyfillsky, A.V. Pfyhill, A.V. Sirotkin, N. Vyahhi, G. Tesler, M.A. Alekseyev, P.A. Pevzner, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19 (2012) 455–477.
[4] J. Besemer, A. Lomsadze, M. Borodovsky, GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res. 29 (2001) 2607–2618.
[5] D. Coil, G. Jospin, A.E. Darling, A5-miseq: an updated pipeline to assemble microbial genomes from illumina MiSeq data. Bioinformatics 31 (2015) 587–589.
[6] A.C. Darling, B. Mau, F.R. Blattner, N.T. Perna, Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 14 (2004) 1394–1403.
[7] I. Grissa, C. Vergnaud, C. Pourcel, CRISPRFinder: a website to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res. 36 (2008) W145–W148.
[8] I. Grissa, G. Vergnaud, C. Pourcel, The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinf. 8 (2007) 172.
[9] I. Grissa, C. Vergnaud, C. Pourcel, CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res. 35 (2007) W52–W57.
[10] A. Gurevich, V. Saveliev, N. Vyahhi, G. Tesler, QUAST: quality assessment tool for genome assemblies. Bioinformatics 29 (2013) 1072–1075.
[11] S. Kumar, G. Stecher, D. Peterson, K. Tamura, MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. Bioinformatics 28 (2012) 2685–2686.
[12] M. Legendre, A. Lartigue, L. Bertaux, S. Jeudy, J. Bartoli, M. Lescot, J.M. Alempic, C. Bruley, K. Labadie, S. Shimakova, E. Rivkina, Y. Coute, C. Abergel, J.M. Claverie, In-depth study of Mollivirus sibericum, a new 30,000-y-old giant virus infecting Acanthamoeba. Proc. Natl. Acad. Sci. U. S. A. 112 (2015) E5327–E5335.
[13] A. Levasseur, M. Bekliz, E. Chabrier, P. Pantoratti, B. La Scola, D. Raoult, MIMIVIRE is a defence system in mimivirus that confers resistance to virophage. Nature 531 (2016) 249–252.
[14] T.M. Lowe, S.R. Eddy, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25 (1997) 955–964.
[15] R. Luo, B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, C. Yu, B. Wang, Y. Li, C. Han, D.W. Cheung, S.M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T.W. Lam, SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1 (2012) 18.
[16] A. Moegils, C. Coulouris, Y. Raytselis, T.L. Madden, R. Agarwala, A.A. Schaffer, Database indexing for production MEGABLAST searches. Bioinformatics 24 (2008) 1757–1764.
[17] D. Raoult, S. Audic, C. Robert, C. Abergel, P. Renesto, H. Ogata, B. La Scola, M. Suzan, J.M. Claverie, The 1.2-megabase genome sequence of Mollivirus sibericum, a new 30,000-y-old giant virus infecting Acanthamoeba. Proc. Natl. Acad. Sci. U. S. A. 112 (2015) E5327–E5335.
[18] D.R. Zerbino, E. Birney, Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18 (2008) 821–829.