



A New Putative *Caulimoviridae* Genus Discovered through Air Metagenomics

Alberto Rastrojo,^a Andrés Núñez,^b  Diego A. Moreno,^b Antonio Alcamí^a

^aCentro de Biología Molecular Severo Ochoa, Consejo Superior de Investigaciones Científicas and Universidad Autónoma de Madrid (CSIC-UAM), Madrid, Spain

^bEscuela Técnica Superior de Ingenieros Industriales, Universidad Politécnica de Madrid (ETSII-UPM), Madrid, Spain

ABSTRACT Members of the *Caulimoviridae* family are important plant pathogens. These circular double-stranded DNA viruses may integrate into the host genome, although this integration is not required for the viral replication cycle. Here, we describe three complete genomes belonging to a new putative *Caulimoviridae* genus discovered through air metagenomics.

Caulimoviridae is the only known double-stranded DNA (dsDNA) virus family infecting plants that replicates by reverse transcription (1). Members of this family have circular genomes of 7,000 to 8,200 bp with discontinuities in both strands coding for 1 to 8 open reading frames (ORFs) (2). The replication cycle is episomal and does not require integration into the host genome. However, integration can occur during the nonhomologous end-joining repair of dsDNA breaks in host genomes, leaving a fingerprint of past infections (3). The study of these endogenous viral elements suggests that this family would have emerged approximately 320 million years ago (4).

Here, we describe three complete viral genomes belonging to the family *Caulimoviridae*. These genomes were obtained from air samples collected in Madrid, Spain (40.439881°N, 3.689409°W) using different devices, namely, a Hirst spore trap, a Surface Air System DUO 360 instrument, and a Burkard multivial cyclone sampler (our unpublished data). A PowerSoil DNA isolation kit was used to extract total DNA. Samples were then sequenced using Illumina technology, obtaining 40 million paired-end reads for each sample (2×125 nucleotides). Raw reads were quality filtered using PRINSEQ (mean quality score of 25 and length of >75) (5), assembled with IDBA_UD with default parameters (6), and classified using a BLAST search against the NCBI nonredundant database (e-value, $<1e-3$; score, >50) (7). Three contigs of ~ 7 kb were assigned to the family *Caulimoviridae* and were circularized using Minimus2 (8). These viral genomes are 98.8 to 99.3% identical to each other and therefore belong to the same species. All three genomes have a minus-strand primer binding site, a polypurine tract, and polyadenylation signals, and they have only one large ORF coding a 2,129-amino acid polyprotein with several domains, movement protein (amino acids [aa] 57 to 222), coat protein (aa 602 to 838), aspartic protease (aa 991 to 1124), reverse transcriptase (aa 1144 to 1619), and RNase H (aa 1627 to 1740) (9). All of these features are characteristic of members of the genus *Petuvirus* (2). However, the reverse transcriptase (RT)-RNase H domain shares only 40% identity with *Petunia vein clearing virus* (PVCV), the unique member of the genus *Petuvirus*.

By searching against the complete or near-complete endogenous viral genomes described by Diop et al. (4), we were able to find the closest relative, *Pinus taeda gymnodovirus 2* (PtGy2), which shares a 77% identity at the nucleotide level with the three new genomes. PtGy2 contains five ORFs, but the rearrangement the ORFs generated a single ORF coding a 2,106-aa polyprotein that is 79% identical to the

Received 1 August 2018 Accepted 13 September 2018 Published 11 October 2018

Citation Rastrojo A, Núñez A, Moreno DA, Alcamí A. 2018. A new putative *Caulimoviridae* genus discovered through air metagenomics. *Microbiol Resour Announc* 7:e00955-18. <https://doi.org/10.1128/MRA.00955-18>.

Editor David A. Baltrus, University of Arizona

Copyright © 2018 Rastrojo et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Alberto Rastrojo, arastrojo@cbm.csic.es, or Antonio Alcamí, aalcamí@cbm.csic.es.

proteins of the newly discovered viruses. This identity increased to 84% when the RT-RNase H domain was examined. Therefore, the new genomes could represent a replication competent version of the PtGy2 endogenous element, which could have been fragmented because of the accumulation of several mutations or indels due to its integration in the genome of *Pinus taeda*. Interestingly, ~80% of the shotgun metagenomic reads were assigned to *Pinus taeda*. Additionally, we were able to detect by PCR these new viruses in a *Pinus nigra* sample from the vicinity of where the air samples were collected, suggesting that these new viruses are likely *Pinus* pathogens. The phylogenetic analysis (10, 11) of the RT-RNase H protein of representative *Caulimoviridae* members showed a strong relation of the new viruses with PtGy2 and with *Picea glauca gymnevovirus 2*, another endogenous element, all of them forming an ancient branch in clade B, to which PVCV belongs (4). In conclusion, we propose that these new genomes may represent a new genus of *Caulimoviridae*-infecting gymnosperms, in contrast to *Petuvirus*-infecting angiosperms, which could represent the replicative counterpart of the endogenous *Gymnevovirus 2* genus recently described.

Data availability. The three complete genome sequences reported here have been deposited in GenBank under the accession numbers [MH551471](#), [MH551472](#), and [MH551473](#). Shotgun raw reads have also been deposited in the European Nucleotide Archive (ENA) under the accession numbers [ERX2313857](#), [ERX2313858](#), and [ERX2313863](#).

ACKNOWLEDGMENTS

We thank Francisco Murilo Zerbini, Pierre-Yves Teycheney, Andrew Geering, and Elliot J. Lefkowitz for their advice. We also thank the Genomics and Next Generation Sequencing Service at the Centro de Biología Molecular Severo Ochoa and Centro de Computación Científica from the Universidad Autónoma de Madrid for their support.

This study was supported by the Community of Madrid (Spain), under the AIRBIOTA-CM Program (S2013/MAE-2874).

REFERENCES

- Harper G, Hull R, Lockhart B, Olszewski N. 2002. Viral sequences integrated into plant genomes. *Annu Rev Phytopathol* 40:119–136. <https://doi.org/10.1146/annurev.phyto.40.120301.105642>.
- Lefkowitz EJ, Dempsey DM, Hendrickson RC, Orton RJ, Siddell SG, Smith DB. 2018. Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res* 46:D708–D717. <https://doi.org/10.1093/nar/gkx932>.
- Geering ADW, Maumus F, Copetti D, Choisine N, Zwickl DJ, Zytnicki M, McTaggart AR, Scalabrin S, Vezzulli S, Wing RA, Quesneville H, Teycheney P-Y. 2014. Endogenous florendoviruses are major components of plant genomes and hallmarks of virus evolution. *Nat Commun* 5:5269. <https://doi.org/10.1038/ncomms6269>.
- Diop SI, Geering ADW, Alfama-Depauw F, Loaec M, Teycheney P-Y, Maumus F. 2018. Tracheophyte genomes keep track of the deep evolution of the *Caulimoviridae*. *Sci Rep* 8:572. <https://doi.org/10.1038/s41598-017-16399-x>.
- Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27:863–864. <https://doi.org/10.1093/bioinformatics/btr026>.
- Peng Y, Leung HCM, Yiu SM, Chin FYL. 2012. IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28:1420–1428. <https://doi.org/10.1093/bioinformatics/bts174>.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. <https://doi.org/10.1186/1471-2105-10-421>.
- Treangen TJ, Sommer DD, Angly FE, Koren S, Pop M. 2011. Next generation sequence assembly with AMOS. *Curr Protoc Bioinformatics Chapter 11:Unit 11.8*. <https://doi.org/10.1002/0471250953.bi1108s33>.
- Alva V, Nam S-Z, Söding J, Lupas AN. 2016. The MPI Bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis. *Nucleic Acids Res* 44:W410–W415. <https://doi.org/10.1093/nar/gkw348>.
- Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59:307–321. <https://doi.org/10.1093/sysbio/syq010>.
- Lefort V, Longueville J-E, Gascuel O. 2017. SMS: smart model selection in PhyML. *Mol Biol Evol* 34:2422–2424. <https://doi.org/10.1093/molbev/msx149>.