Arch Virol (2017) 162:2041–2045
DOI 10.1007/s00705-017-3298-0

BRIEF REPORT

Genome sequences of a capulavirus infecting *Plantago lanceolata* in the Åland archipelago of Finland

Hanna Susi[1] · Anna-Liisa Laine[1] · Denis Filloux[2] · Simona Kraberger[3] · Kata Farkas[3] · Pauline Bernardo[2] · Mikko J. Frilander[4] · Darren P. Martin[5] · Arvind Varsani[3,6,7] · Philippe Roumagnac[2]

Received: 23 December 2016 / Accepted: 2 February 2017 / Published online: 10 March 2017

Abstract The discovery and full-genome sequences of two isolates of a fourth capulavirus species are reported. The viruses were discovered during a viral metagenomics survey of uncultivated *Plantago lanceolata* plants in the Åland archipelago of south western Finland. The newly discovered viruses apparently produce no symptoms in *P. lanceolata*. They have a genome organization that is very similar to that of the three known capulavirus species and additionally share between 62.9 and 67.1% genome-wide sequence identity with the isolates of these species. It is therefore proposed that these viruses be assigned to a new capulavirus species named "*Plantago lanceolata latent virus*".

Metagenomics approaches [16] coupled with the rolling circle amplification (RCA) [9, 11, 18] technique have had a particularly profound impact on our appreciation and knowledge of the prevalence, pervasiveness and extreme diversity of single-stranded DNA (ssDNA) viruses in the environment [19]. Among the best studied of these ssDNA viruses have been the plant-infecting viruses in the family *Geminiviridae*. Recently a new genus, called *Capulavirus* [4, 5, 17] has been proposed to accommodate some of the highly divergent viruses that have been discovered in this family. Three distinct species of capulaviruses were discovered between 2010 and 2011 infecting a wild spurge (*Euphorbia caput-medusae*) in South Africa (*Euphorbia caput-medusae latent virus*; [4]), alfalfa (*Medicago sativa*) plants in France (*Alfalfa leaf curl virus*; [17]), and French bean (*Phaseolus vulgaris*) plants in India (*French bean severe leaf curl virus*, accession number JX094280). Besides being extremely divergent, these viruses all have genome organizations that are unique amongst the geminiviruses [4, 5] and for at least one of them, alfalfa leaf curl virus (ALCV), *Aphis craccivora*, an invasive aphid species with an almost global distribution, is a vector [17].

GenBank Accession #: KT214389, KT214390.

Electronic supplementary material The online version of this article (doi:10.1007/s00705-017-3298-0) contains supplementary material, which is available to authorized users.

✉ Philippe Roumagnac
philippe.roumagnac@cirad.fr

[1] Department of Biosciences, Metapopulation Research Centre, University of Helsinki, P.O. Box 65, 00014 Helsinki, Finland

[2] CIRAD-INRA-SupAgro, UMR BGPI, Campus International de Montferrier-Baillarguet, 34398 Montpellier Cedex-5, France

[3] School of Biological Sciences and Biomolecular Interaction Centre, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand

[4] Institute of Biotechnology, University of Helsinki, Helsinki, Finland

[5] Computational Biology Group, Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Cape Town 7925, South Africa

[6] Structural Biology Research Unit, Department of Clinical Laboratory Sciences, University of Cape Town, Cape Town 7701, South Africa

[7] The Biodesign Center for Fundamental and Applied Microbiomics, Center for Evolution and Medicine and School of Life sciences, Arizona State University, Tempe, AZ, USA

In 2013, wild *Plantago lanceolata* L. (*Plantaginaceae*) samples were collected in the Åland archipelago of south western Finland as part of a project aimed at characterizing virus communities within these plants. Twelve pooled samples, each containing leaf material from 12-14 individual *P. lanceolata* plants, were collected. The nucleic acid extraction and sequencing approach of Kreuze et al. [12] was used with slight modifications for the *P. lanceolata* plants. Total RNA was extracted from 100 mg of each pool of 12-14 *P. lanceolata* fresh leaf material using Trizol (Invitrogen, USA) following the manufacturer's instructions. Small RNA libraries were directly generated from total RNAs. Small RNAs ligated with 3' and 5' adapters were reverse transcribed and PCR amplified (98°C for 30 sec; 13 cycles of 98°C for 10 sec, 60°C for 30 sec, 72°C for 15 sec and a final extension of 72°C for 10 min) to create cDNA libraries selectively enriched for fragments having adapter molecules at both ends. The last step was an acrylamide gel purification of the 140-150 nt amplified cDNA constructs (corresponding to cDNA inserts from siRNAs + 120 nt from the adapters). Small RNA libraries were checked for quality and quantified using a 2100 Bioanalyzer (Agilent, USA). The library was then sequenced on one lane of a HiSeq system (Illumina, USA) as single-end 50 base reads. Raw reads were then cleaned to eliminate Illumina adapters and low quality regions (cut-off Phred quality score of 25) using cutadapt [13]. *De novo* assemblies of cleaned reads were performed using Velvet [22] and CAP3 [10] with a minimal contig size set at 45 bp.

Between 10 and 21 million high quality sequence reads in the size range of siRNAs were obtained after filtering raw reads for the twelve *P. lanceolata* sample pools. Pool 7, which contained cDNAs from twelve *P. lanceolata* plants, yielded 27,803 contigs by *de novo* assembly, seven of which showed significant degrees of similarity to the capulavirus, Euphorbia caput-medusae latent virus (EcmLV), based on BLASTX analysis [2]. Two of the seven contigs apparently corresponded to the capsid protein gene (*cp*) and the other five to the replication-associated protein gene (*rep*).

DNA was separately extracted from the twelve *P. lanceolata* plants of pool 7 using the DNeasy Plant Mini Kit (Qiagen, Germany). DNA was used as a template for PCR amplification of the complete genome using a pair of abutting primers with a *Pst*I overlapping site. These primers were designed based on the contigs from the *de novo* assembly (Pla_pstIF: 5'-CTG CAG ATC ATT GTA TAA ATA CTG TCC CAA ATA CG-3'; Pla_pstIR: 5'-CTG CAG TAT CTG TGA TAT TTG TAT ACA AAT TTC TGA C-3'). The amplification was carried out using Kapa Hifi Hotstart DNA polymerase (KAPA Biosystems, USA) with the following thermal cycling conditions: 96°C for 3 min, 25 cycles of 98°C (20 s), 60°C (30 s), 72°C (3 min), and a final extension of 72°C for 3 min. The amplicons were resolved on a 0.7% agarose gel and from the 12 plants in pool 7, two (plants 7-5 and 7-11) had ~2.8 kb amplicons. These ~2.8 kb amplicons were excised, gel purified and cloned in to the plasmid pJET1.2. The resulting clones were Sanger sequenced by primer walking at Macrogen Inc. (South Korea). The contigs produced were assembled using DNA Baser V4 (Heracle BioSoft S.R.L. Romania) and pairwise identities between the assembled genomes and those of other capulaviruses available in GenBank were determined using SDT v1.2 [15].

The two *P. lanceolata* derived genomes share 97.8% genome-wide pairwise identity with one another, and 62.9-67.1% identity with the known capulaviruses, ALCV (n = 27), EcmLV (n = 16) and French bean severe leaf curl virus (FbSLCV, n = 2; Supplementary Figure 1). Circular DNA molecules were amplified by rolling circle amplification (RCA) from the two *P. lanceolata* plants from which these isolates were obtained (plants 7-5 and 7-11) using the φ*29* DNA polymerase (TempliPhi^{TM}, GE Healthcare, USA) as previously described [18]. RCA products were digested with *Afl*II for 3 h at 37°C. Only one band, of ~2.8 kbp, was resolved by electrophoresis following *Afl*II digestion of the RCA products. No fragments were detected within the size range of known geminivirus-associated sequences. Plants 7-5 and 7-11 did not exhibit any conspicuous symptoms (such as chlorotic mosaic,

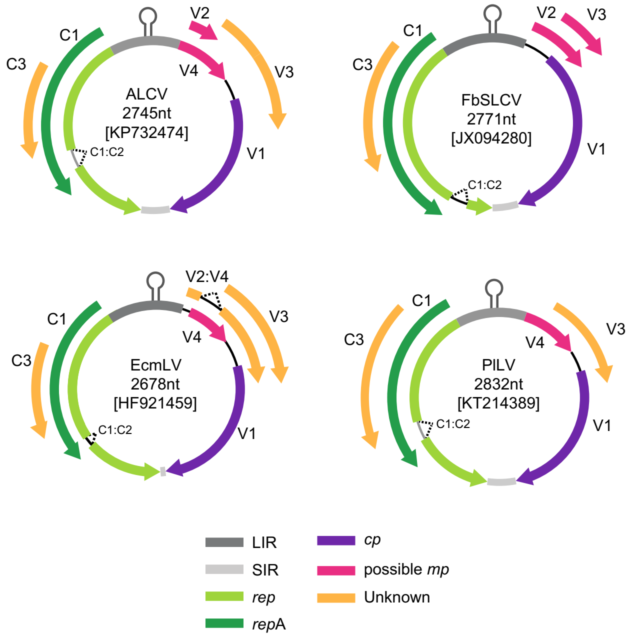

Fig. 1 Genome organizations of the known capulaviruses: alfalfa leaf curl virus (ALCV) from France, Euphorbia caput-medusae latent virus (EcmLV) from South Africa, French bean severe leaf curl virus (FbSLCV) from India and Plantago lanceolata latent virus from Finland

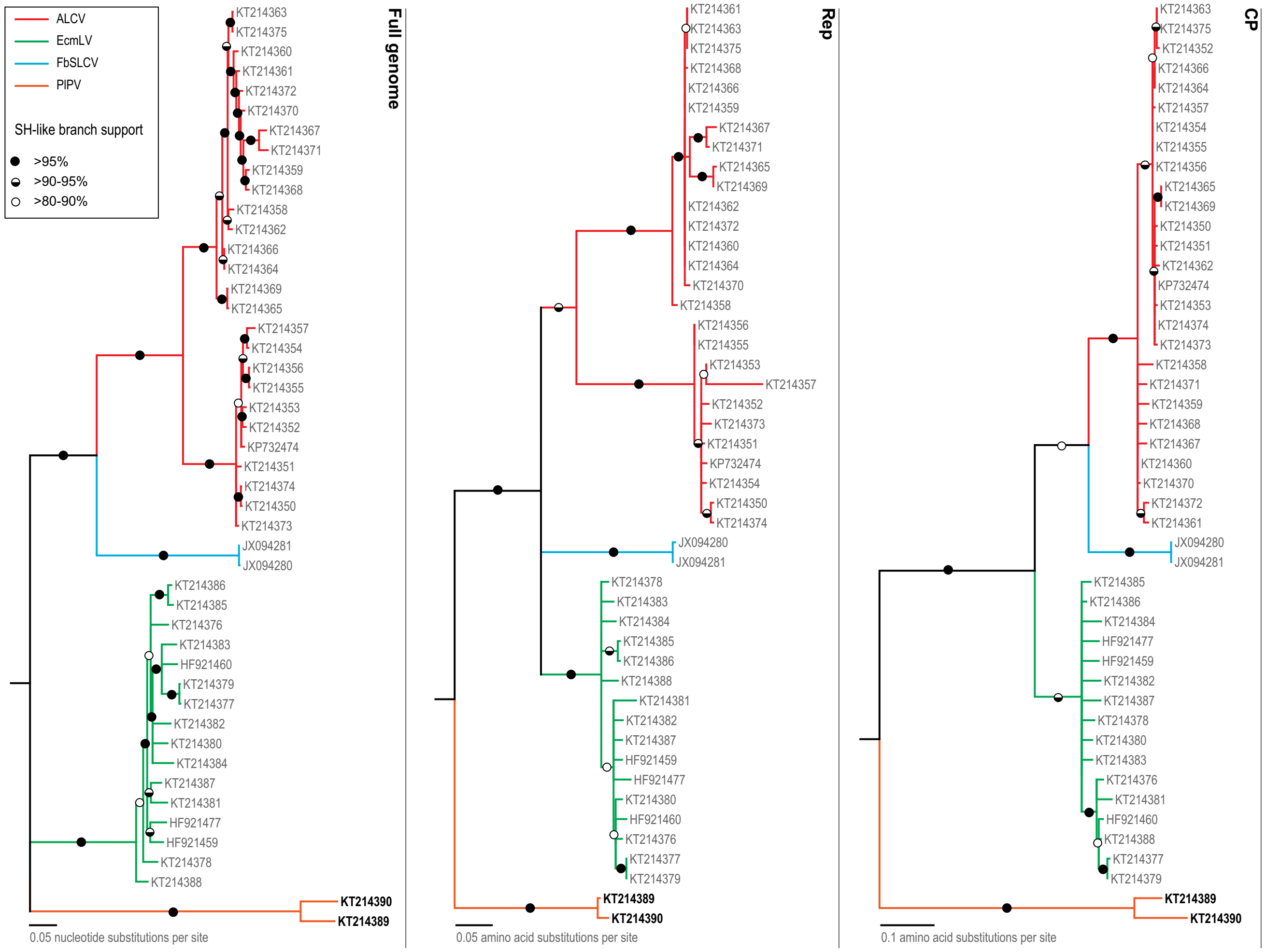

Fig. 2 Maximum likelihood phylogenetic trees of complete capulavirus genome nucleotide sequences, capulavirus replication-associated protein (Rep) amino acid sequences and capulavirus capsid protein (CP) amino acid sequences. The trees are rooted with representative sequences of each of the seven other known geminivirus genera. Branches with less than 80% approximate likelihood ratio test support have been collapsed

chlorotic streak or leaf deformation) that differentiated them from plants that contained no siRNA reads with any significant degrees of similarity to known capulaviruses. This suggests that, like EcmLV in *E. caput-medusae*, this virus might infect *P. lanceolata* without inducing symptoms (i.e. it too might be defined as a latent virus).

The arrangement of open reading frames (ORFs; Figure 1) within the 2832 bp circular DNA of clones obtained from both plants 7-5 and 7-11 is most similar to those reported for the three known capulavirus species [5], with three overlapping complementary-sense ORFs (C1, C2 and C3), two intergenic regions and three virion-sense ORFs (V1, V3 and V4; by analogy with ALCV; Figure 1).

Although the genomes obtained from the *P. lanceolata* plants are most closely related to EcmLV, ALCV and FbSLCV, they only share 62.9-67.1% genome-wide pairwise identity with these three capulavirus species: a degree of similarity which is below the lowest of the species demarcation thresholds recommended by the ICTV for any of the established geminivirus genera (78% for mastreviruses, 77% for curtoviruses, 75% for turncurtoviruses, 91% for begomoviruses and 75% for eragroviruses; [14, 20, 21]). We therefore propose that the apparently novel capulaviruses obtained from the two *P. lanceolata* plants should be assigned to a new capulavirus species named "*Plantago lanceolata latent virus*" with individual isolates being named equivalently and abbreviated to PlLV.

We assembled a nucleotide sequence dataset consisting of the 47 known capulavirus genomes and two amino acid sequence datasets containing capulavirus Rep and CP amino acid sequences together with amino acid sequences from one member of each of the seven other established geminivirus genera (as outgroups). The sequences in each dataset were aligned using MUSCLE [7] and the resulting alignment was used to infer a maximum likelihood phylogenetic tree using PhyML3.0 [8]. Whereas the full-genome phylogenetic tree was inferred using the GTR + I+G nucleotide substitution model (selected as the best fitting model by jModelTest; [6]), the Rep and CP phylogenetic trees were inferred using the LG + G+I and WAG + G+I amino acid substitution models, respectively (inferred to be the best fitting models using ProtTest; [1]). Approximate likelihood ratio tests (aLRT) [3] were used to infer relative supports for branches (with branches having <80% support being collapsed). In all three of the generated phylogenetic trees, the capulavirus derived sequences cluster together relative to the other geminiviruses (which were used to root these trees but are not shown in the figures) with sequences from each of the four capulavirus species consistently clustering within distinct clades with greater than 90% aLRT support (Figure 2).

The Rep of PlLV shares 64-74% amino acid identity with those of other capulaviruses (Supplementary Figure 2) and 29-45% with those of other geminiviruses. The CP of PlLV shares 46-53% amino acid identity (Supplementary Figure 3) with those of other capulaviruses and 20-24% with those of other geminiviruses.

We therefore conclude that these PlLV should be considered as isolates of a novel capulavirus species. Furthermore we suggest both that, as with ALCV, it may also be aphid transmitted, and that, as with EcmLV, its genomes may be encapsidated in twinned icosahedral virions [17]. The discovery of PlLV extends the known geographical and host-ranges of capulaviruses and reinforces the possibility raised in other plant virus metagenomics studies [4, 17] that, in the Eastern hemisphere at least, members of this new geminivirus genus may be as prevalent in the environment as begomoviruses or mastreviruses: viruses belonging to the two most thoroughly sampled geminivirus genera. The reason they may have evaded detection for so long could simply be that, as appears to be the case with EcmLV and PlLV, they might predominantly cause asymptomatic infections in their natural hosts.

Acknowledgements This work was supported by INRA, Département Santé des Plantes et Environnement (grant AAP SPE 2015) awarded to PR. DPM and AV are supported by the National Research Foundation of South Africa. PR is supported by EU grant FP7-PEOPLE-2013-IOF (N° PIOF-GA-2013-622571).

Compliance with ethical standards

This study was funded by INRA, Département Santé des Plantes et Environnement (grant AAP SPE 2015) awarded to PR.

Conflict of interest Author DPM and AV have received research grants from the National Research Foundation of South Africa. Author PR has received an EU grant FP7-PEOPLE-2013-IOF (N° PIOF-GA-2013-622571). Author DPM, AV and PR declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Abascal F, Zardoya R, Posada D (2005) ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104–2105
2. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410
3. Anisimova M, Gascuel O (2006) Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst Biol 55:539–552
4. Bernardo P, Golden M, Akram M, Naimuddin Nadarajan N, Fernandez E, Granier M, Rebelo AG, Peterschmitt M, Martin DP, Roumagnac P (2013) Identification and characterisation of a highly divergent geminivirus: evolutionary and taxonomic implications. Virus Res 177:35–45
5. Bernardo P, Muhire B, Francois S, Deshoux M, Hartnady P, Farkas K, Kraberger S, Filloux D, Fernandez E, Galzi S, Ferdinand R, Granier M, Marais A, Monge Blasco P, Candresse T, Escriu F, Varsani A, Harkins GW, Martin DP, Roumagnac P (2016) Molecular characterization and prevalence of two capulaviruses: Alfalfa leaf curl virus from France and *Euphorbia caput-medusae* latent virus from South Africa. Virology 493:142–153
6. Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 9:772
7. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797
8. Guindon S, Dufayard JF, 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
9. Haible D, Kober S, Jeske H (2006) Rolling circle amplification revolutionizes diagnosis and genomics of geminiviruses. J Virol Methods 135:9–16
10. Huang XQ, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9:868–877
11. Inoue-Nagata AK, Albuquerque LC, Rocha WB, Nagata T (2004) A simple method for cloning the complete begomovirus genome using the bacteriophage phi 29 DNA polymerase. J Virol Methods 116:209–211
12. Kreuze JF, Perez A, Untiveros M, Quispe D, Fuentes S, Barker I, Simon R (2009) Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. Virology 388:1–7
13. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet 17(1):10
14. Muhire B, Martin DP, Brown JK, Navas-Castillo J, Moriones E, Zerbini FM, Rivera-Bustamante R, Malathi VG, Briddon RW, Varsani A (2013) A genome-wide pairwise-identity-based proposal for the classification of viruses in the genus Mastrevirus (family Geminiviridae). Arch Virol 158:1411–1424
15. Muhire BM, Varsani A, Martin DP (2014) SDT: a virus classification tool based on pairwise sequence alignment and identity calculation. PLoS One 9:e108277
16. Roossinck MJ, Martin DP, Roumagnac P (2015) Plant virus metagenomics: advances in virus discovery. Phytopathology 105:716–727

17. Roumagnac P, Granier M, Bernardo P, Deshoux M, Ferdinand R, Galzi S, Fernandez E, Julian C, Abt I, Filloux D, Mesleard F, Varsani A, Blanc S, Martin DP, Peterschmitt M (2015) Alfalfa leaf curl virus: an aphid-transmitted geminivirus. J Virol 89:9683–9688

18. Shepherd DN, Martin DP, Lefeuvre P, Monjane AL, Owor BE, Rybicki EP, Varsani A (2008) A protocol for the rapid isolation of full geminivirus genomes from dried plant tissue. J Virol Methods 149:97–102

19. Simmonds P, Adams MJ, Benko M, Breitbart M, Brister JR, Carstens EB, Davison AJ, Delwart E, Gorbalenya AE, Harrach B, Hull R, King AMQ, Koonin EV, Krupovic M, Kuhn JH, Lefkowitz EJ, Nibert ML, Orton R, Roossinck MJ, Sabanadzovic S, Sullivan MB, Suttle CA, Tesh RB, van der Vlugt RA, Varsani A, Zerbini FM (2017) Consensus statement: virus taxonomy in the age of metagenomics. Nat Rev Microbiol 15(3):161–168

20. Varsani A, Martin DP, Navas-Castillo J, Moriones E, Hernandez-Zepeda C, Idris A, Murilo Zerbini F, Brown JK (2014) Revisiting the classification of curtoviruses based on genome-wide pairwise identity. Arch Virol 159:1873–1882

21. Varsani A, Navas-Castillo J, Moriones E, Hernandez-Zepeda C, Idris A, Brown JK, Murilo Zerbini F, Martin DP (2014) Establishment of three new genera in the family Geminiviridae: Becurtovirus, Eragrovirus and Turncurtovirus. Arch Virol 159:2193–2203

22. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829