Arch Virol (2009) 154:1189–1193
DOI 10.1007/s00705-009-0412-y

VIROLOGY DIVISION NEWS

# Suggestions for a nomenclature of endogenous pararetroviral sequences in plants

**Christina Staginnus · M. L. Iskra-Caruana · B. Lockhart · T. Hohn · K. R. Richert-Pöggeler**

Received: 19 December 2008 / Accepted: 25 May 2009 / Published online: 12 June 2009


## Introduction

The term "pararetrovirus" was introduced by Temin [1] for animal (*Hepadnaviridae*) and plant viruses (*Caulimoviridae*) that, in contrast to retroviruses, have a DNA genome and do not integrate into the host genome for replication. Like retroviruses, pararetroviruses use a reverse transcriptase for their replication.

Endogenous pararetroviruses (EPRVs) in plants represent counterparts of members of the virus family *Caulimoviridae* integrated in their host's genome. Despite the non-integrative replication cycle of members of the *Caulimoviridae*, a growing number of integrated viral sequences have been reported and are still being identified in various plant genomes [2–5].

Most of the integrants are silent, repetitive genome components. However, some of these sequences may still be able to replicate and initiate viral infection under certain conditions, according to their structural and sequence integrity and their genomic and/or epigenetic context.

## Suggestions for a uniform nomenclature of endogenous virus sequences

Facing the rapidly growing diversity of EPRVs discovered in plant genomes, the need for a uniform nomenclature is obvious. According to the multi-copy nature of EPRVs, important differences in sequence composition and structure of integrants have been observed.

It would be highly desirable for a nomenclature system (1) to distinguish endogenous from episomal caulimovirid sequences, (2) to discriminate potentially functional integrants from passive and pseudogene host genome components and (3) to describe the element's viral activity in a specific genomic context.

In some genomes, a wide variety of EPRVs has been identified, comprising viral sequences with or without exogenous virus counterparts [6, 7] as well as rearranged and functional forms of a specific virus genome [3].

Like their exogenous counterparts, EPRVs can be classified as petuvirus-like elements, badnavirus-like elements or as members of the genera *Caulimovirus*, *Cavemovirus* or *Tungrovirus* according to the number and arrangement of open reading frames (ORFs) and nucleotide sequence homologies with episomal viruses (see Table 1).

C. Staginnus (✉)
Plant Molecular Biology, Biocentre, University of Frankfurt am Main, Marie-Curie-Str. 9, 60439 Frankfurt am Main, Germany
e-mail: staginnus@em.uni-frankfurt.de

K. R. Richert-Pöggeler (✉)
Julius Kühn-Institut (JKI), Federal Research Centre for Cultivated Plants, Institute for Epidemiology and Pathogen Diagnostics, Messeweg 11-12, 38104 Braunschweig, Germany
e-mail: katja.richert-poeggeler@jki.bund.de

M. L. Iskra-Caruana
CIRAD, UMR BGPI-TA A54/K, 34398 Montpellier Cedex 5, France

B. Lockhart
Department of Plant Pathology, University of Minnesota, St. Paul, MN 55108, USA

T. Hohn
Botanical Institute, University of Basel, Schoenbeinstr. 6, 4056 Basel, Switzerland

**Table 1** Selection of EPRVs identified in plant genomes, their current nomenclature and their homologous exogenous viruses

| Host plant | Structural classification | Activatable EPRVs | Non-activatable EPRVs | Corresponding episomal virus | References |
|---|---|---|---|---|---|
| *Musa balbisiana* | Badnavirus | BSOLV EPRV, BSGFV EPRV-EPRV 7 | BSMysV, BSImV, BSOLV, BSGFV-EPRV 9 | BSMysV, BSImV, BSOLV, BSGFV | [3, 6, 7, 16–19] |
| *Oryza sativa (japonica, indica)* also: *Oryza* sp. | Tungrovirus | – | ERTBV-A; ERTBV-B; ERTBV-C | RTBV | [8] |
| *Petunia hybrida* also: *Petunia* sp. | Petuvirus | ePVCV | ePVCV | PVCV | [9, 20] |
| *Nicotiana edwardsonii, Nicotiana tabacum* also: *Nicotiana* sp. | Cavemovirus | TVCV | TVCV | TVCV[a] | [21] |
| *Nicotiana sylvestris, N. tabacum* also: *Nicotiana* sp. | Cavemovirus | – | *Ns*EPRV (TEPRV, TPVL) | TVCV[a] | [22] |
| *Nicotiana tomentosiformis, N. tabacum* also: *Nicotiana* sp. | Cavemovirus | – | *Nto*EPRV | TVCV[a] | [14] |
| *Solanum tuberosum* | Cavemovirus | – | *Sotu*I, *Sotu*III | TVCV[a] | [10] |
| *Solanum* subsection *Lycopersicon* | Cavemovirus | – | *Lyc*EPRV | TVCV[a] | [11] |

*BSOLV* Banana streak virus Obino I'Ewai species, *BSMysV* Banana streak virus Mysore species, *BSImV* Banana streak virus Imove species, *TVCV* Tobacco vein clearing virus

[a] However, existence of an independent external virus is uncertain

So far, authors use the prefix "E-"or "e-" for "endogenous" (e.g. ERTBV[1] [8], ePVCV[2] [9]) or the suffix "-EPRV" (BSGFV[3] EPRV; [3]) in connection with the virus name to distinguish integrated viral sequences from the homologous episomal virus. In other examples, they are named after the host plant from which they have been isolated, analogous to the nomenclature of transposons [e.g. *Sotu* (in *Solanum tuberosum*) or *LycEPRV* (in several *Solanum* subsection Lycopersicon species); [10–12]].

One major point for the nomenclature of plant endogenous virus sequences is to identify a significant relationship between viruses and integrated sequences. Usually, the highest matches of sequence identity are considered relevant. Based on existing sequence comparisons (Table 2), we suggest a threshold level of at least 80% nucleotide identity over 80% of the sequence within the polymerase (POL) reading frame ("ORF3" in Table 2) to confirm the affiliation of an endogenous sequence to a virus. This value is based on the suggestions of Wicker et al. [12] for the distinction of transposable elements and the ICTV rules for distinguishing species in the family *Caulimoviridae* [13]. It remains to be seen if this threshold value is appropriate when more EPRV sequences become available. Additionally, the comparison of integrated virus sequences can identify distinct EPRVs in the same host genome (e.g. *NsEPRV* and *NtoEPRV* in *Nicotiana tabacum;* [14]).

Another important feature for the classification of an integrant is whether it is functional and can trigger a virus infection. Thus, sequences in question have to be isolated and their infectivity has to be proven experimentally by infection of the respective host plants.

These considerations led us to the following suggestions for a uniform distinction between integrated pararetrovirus sequences (EPRVs) and their homologous exogenous viruses (see also Table 3):

1. **If the endogenous sequence can be affiliated to an exogenous virus,** viral sequences integrated into the host genome should be marked by the **prefix "e"** (**e**ndogenous) in front of the virus abbreviation in cases where no information about the status of the integrant is yet available (eBSOLV, ePVCV). If further information is available, the prefix **"ea"** (**e**ndogenous and **a**ctivatable) in connection with the virus abbreviation should be applied. Episomal viruses should be referred to following the ICTV nomenclature.
   - 1.1 **Functional endogenous copies** are able to release a replication-competent viral genome with high similarity to an exogenous virus and should be marked by the prefix **"ea"** followed by the virus name (eaPVCV, eaBSGFV-7).
   - 1.2 An integrant related to an exogenous counterpart, but **not known to be functional as a virus per se** should be named with the prefix **"e"** and the virus name (e.g. eTVCV, eBSVGFV-9). The integrant itself is incapable of making a functional virus: e.g. no transition from the endogenous to the episomal form is known, or the sequence lacks functional ORFs due to mutations. However, activation of "eEPRV" by recombination with

[1] RTBV: *Rice tungro bacilliform virus.*

[2] PVCV: *Petunia vein clearing virus.*

[3] BSGFV: *Banana streak virus Goldfinger species.*

**Table 2** Observed sequence similarities between selected members of the family Caulimoviridae and EPRVs for identification and classification of the endogenous forms

| ORF | Sequence similarity between virus and EPRV | | | | |
|---|---|---|---|---|---|
| | RTBV/ERTBV [8] (nt) % | CsVMV/TPVL = TEPRV = *Ns*EPRV [14, 22, 23] (aa) % | TVCV [21]/TPVL = TEPRV = *Ns*EPRV [14, 22, 23] (aa) % | TVCV/*Nto*EPRV [14] (aa) % | *Ns*EPRV/*Nto*EPRV [14] (aa) % |
| 1 | 49 | 21 | 72 [14, 22, 23] | 88 | 69 |
| 2 | n.a. | 30 | 78 [14, 22, 23] | 91 | 77 |
| 3 | 51 | 43 | 87 [14, 22, 23] | 94 | 86 |
| 4 | 47 | 23 | 61 [14, 22, 23] | 90 | 61 |

*Ns*EPRV (formerly TPVL or TERPV) was affiliated with TVCV rather than with CsVMV due to higher amino acid similarities. *Ns*EPRV and *Nto*EPRV represent different families of integrated viral sequences in the genome of *N. tabacum* according to sequence divergence. The affiliation with ERTBV to RTBV is currently being discussed (Geering pers. comm.)

aa sequence comparison at the amino acid level, nt sequence comparison at the nucleotide level

*CsVMV* Cassava vein mosaic virus, *n.a.* not available

**Table 3** Proposal for a uniform nomenclature of integrated viral sequences

| Pararetroviruses | Endogenous pararetrovirus (EPRV) sequences in a plant genome | | |
|---|---|---|---|
| | Episomal form known and **functional EPRV** | Episomal form known and **no functional EPRV** | **No** episomal form known & **no functional EPRV** |
| Virus name (ICTV nomenclature) | **ea** (virus name) | **e** (virus name) | (Host initials)**-EPRS** |
| e.g. BSGFV, PVCV, RTBV, TVCV | eaBSGFV-7, eaPVCV | eBSGFV-9 | SotuEPRS |
| Episomal replicating unit. Integration is not obligatory for replication. | EPRV sequences integrated in the genomes of *Musa balbisiana* and *Petunia hybrida* cause infection [3, 9] | Integrated sequences of this locus are not responsible for infection [3] | Only partial sequences of the POL, TAV and intergenic region isolated so far from potato (*Solanum tuberosum;* [10]). |

Examples for integrants found in banana, petunia and potato are presented

*POL* polymerase comprising reverse transcriptase and RNAseH, *TAV* transactivator/viroplasmin

exogenous counterparts cannot be ruled out. Moreover, "eEPRVs" may fulfill other purposes in the host, such as providing virus resistance [15].

1.3 In cases where it is necessary to distinguish different integrated copies from each other, a numerical or alphabetical index (such as eaBSGFV-7 and eBSGFV-9) should be introduced.

2. When **no exogenous virus is currently known** or when only small fragments of a viral genome have been identified, the **host plant initials** plus the **suffix "EPRS"** for endogenous pararetroviral sequence" should be chosen (e.g. SotuEPRS for *Solanum tuberosum* endogenous pararetroviral sequence, Table 4).

For some integrated viral sequences, homologous viruses may be unknown, e.g. because they are extinct or have not been discovered yet. Therefore, we suggest using **e-** or **ea-(virus name)** only in cases where *the exogenous form is known.* If there is any doubt about the existence of a corresponding pararetrovirus, **(plant initials)-EPRS** should be chosen; e.g. integrants in the rice genome that reveal weak homology to RTBV ([8], Table 2) were suggested to be classified as OsEPRS (*Oryza sativa* endogenous pararetroviral sequence) according to recent sequence alignments (see Table 3, Geering pers. comm., Table 4). As soon as more sequence information becomes available, existing EPRV names have to be changed accordingly.

We hope that these suggestions will provide some guidance for developing a uniform scheme for the nomenclature of integrated viral sequences. We encourage a discussion leading to future improvements of the nomenclature. Anybody who wants to comment on this proposal is welcome to do so using the following website: http://talk.ictvonline.org/[4].

[4] To view and participate in the discussion, users will have to create an account by clicking on "Join" in the upper right-hand corner of the web page (http://talk.ictvonline.org/). This is open to anyone who chooses to participate. Once registered, anyone can create a new discussion thread under the "General ICTV Discussions" link that is found by clicking on the "Discussions" top menu line.

**Table 4** Current nomenclature and suggested names according to the proposed guidelines of selected EPRVs

| Host plant | Activatable EPRVs | New names | Non-activatable EPRVs | New names | Corresponding episomal virus |
|---|---|---|---|---|---|
| *Musa balbisiana* | BSOLV<br>BSImV<br>BSGFV | eaBSOLV<br>eaBSImV<br>eaBSGFV-7<br>(only a specific locus is shown to generate the episomal virus!) | BSMysV, BSImV, BSOLV, BSGFV-9 | eBSMysV (no functionality known)<br>eBSImV<br>eBSOLV, eBSGFV-9 (other loci than the functional one) | BSMysV, BSImV, BSOLV, BSGFV |
| *Oryza sativa (japonica, indica)* also: *Oryza* sp. | – | – | ERTBV-A; ERTBV-B; ERTBV-C | OsEPRS (POL nucleotide identity <80%) | Unknown |
| *Petunia hybrida* also: *Petunia* sp. | ePVCV | eaPVCV<br>(only a specific locus is shown to generate the episomal virus!) | ePVCV | ePVCV (other loci than the functional one) | PVCV |
| *Nicotiana edwardsonii, Nicotiana tabacum* also: *Nicotiana* sp. | TVCV[a] | eTVCV<br>(existence of an independent external virus is uncertain) | TVCV[a] | eTVCV<br>(existence of an independent external virus is uncertain) | TVCV[a] |
| *Nicotiana sylvestris, N. tabacum* also: *Nicotiana* sp. | – | – | NsEPRV (TEPRV, TPVL) | NsEPRS[b] (data regarding nucleotide identity missing) | Unknown |
| *Nicotiana tomentosiformis, N. tabacum* also: *Nicotiana* sp. | – | – | NtoEPRV | NtoEPRS[b] (data regarding nucleotide identity missing) | Unknown |
| *Solanum tuberosum* | – | – | *Sotu*I, *Sotu*III | SotuEPRS<br>(fragments only) | Unknown |

*BSOLV* Banana streak virus Obino I'Ewai species

*BSMysV* Banana streak virus Mysore species

*BSImV* Banana streak Imove species

*TVCV* Tobacco vein clearing virus

[a] However, existence of an independent external virus is uncertain

[b] Should be renamed if nucleotide identities are confirmed and exceed 80%

**Acknowledgments** The authors thank M.A. Grandbastien for valuable comments on the manuscript and R. Hull for helpful advice in defining a threshold level for sequence comparisons.

## References

1. Temin HM (1985) Reverse transcription in the eukaryotic genome: retroviruses, pararetroviruses, retrotransposons and retrotranscripts. Mol Biol Evol 2:455–468
2. Staginnus C, Richert-Pöggeler KR (2006) Endogenous pararetroviruses: two-faced travelers in the plant genome. Trends Plant Sci 11(10):485–491
3. Gayral P et al (2008) A single banana streak virus integration event in the banana genome as the origin of infectious endogenous pararetrovirus. J Virol 82:6697–6710
4. Pahalawatta V et al (2008) A new and distinct species in the genus Caulimovirus exists as an endogenous plant pararetroviral sequence in its host, Dahlia variabilis. Virology 376:253–257. doi:10.1016/j.virol.2008.03.003
5. Hohn T et al (2008) Evolution of integrated plant viruses. In: Roossinck MJ (ed) Plant virus evolution. Springer, Berlin, pp 53–81. doi:10.1007/978-3-540-75763-4
6. Harper G et al (2005) The diversity of banana streak virus isolates in Uganda. Arch Virol 150:2407–2420
7. Geering ADW et al (2005) Banana contains a diverse array of endogenous badnaviruses. J Gen Virol 86:511–520
8. Kunii M et al (2004) Reconstruction of putative DNA virus from endogenous rice tungro bacilliform virus-like sequences in the rice genome: implications for integration and evolution. BMC Genomics 5:80 doi:10.1186/1471-2164-5-80 (http://www.biomedcentral.com/1471-2164/5/80)
9. Richert-Pöggeler KR et al (2003) Induction of infectious Petunia vein clearing (pararetro) virus from endogenous provirus in petunia. EMBO J 22:4836–4845
10. Hansen CN et al (2005) Characterization of pararetrovirus-like sequences in the genome of potato (*Solanum tuberosum*). Cytogenet Genome Res 110:559–565
11. Staginnus C et al (2007) Endogenous pararetroviral sequences in tomato (*Solanum lycopersicum*) and related species. BMC Plant Biol 7:24. doi:10.1186/1471-2229-7-24
12. Wicker T et al (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973–982. doi:10.1038/nrg2165
13. Fauquet C et al (2005) Virus taxonomy: classification and nomenclature of viruses: eighth report of the International

Committee on the Taxonomy of Viruses. Elsevier/Academic Press, San Diego
14. Gregor W et al (2004) A distinct endogenous pararetrovirus family in Nicotiana tomentosiformis, a diploid progenitor of polyploid tobacco. Plant Physiol 134:1191–1199
15. Mette MF et al (2002) Endogenous viral sequences and their potential contribution to heritable virus resistance in plants. EMBO J 21:461–469
16. Geering ADW et al (2001) Analysis of the distribution and structure of integrated banana streak virus DNA in a range of Musa cultivars. Mol Plant Pathol 2:207–213
17. Geering ADW et al (2005) Characterisation of banana streak Mysore virus and evidence that its DNA is integrated in the B genome of cultivated Musa. Arch Virol 150:787–796
18. Harper G et al (1999) Detection of episomal banana streak badnavirus by IC-PCR. J Virol Methods 79:1–8
19. Ndowora T et al (1999) Evidence that badnavirus infection in *Musa* can originate from integrated pararetroviral sequences. Virology 255:214–220
20. Harper G et al (2002) Viral sequences integrated into plant genomes. Annu Rev Phytopathol 40:119–136
21. Lockhart BE et al (2000) Characterization and genomic analysis of tobacco vein clearing virus, a plant pararetrovirus that is transmitted vertically and related to sequences integrated in the host genome. J Gen Virol 81:1579–1585
22. Jakowitsch J et al (1999) Integrated pararetroviral sequences define a unique class of dispersed repetitive DNA in plants. Proc Natl Acad Sci USA 96:13241–13264
23. Matzke M et al (2004) Endogenous pararetroviruses of allotetraploid *Nicotiana tabacum* and its diploid progenitors. *N. sylvestris* and *N. tomentosiformis*. Biol J Linn Soc 82:627–638