Tracheophyte genomes keep track of the deep evolution of the *Caulimoviridae*

Authors

Seydina Diop[1], Andrew D.W. Geering[2], Françoise Alfama-Depauw[1], Mikaël Loaec[1], Pierre-Yves Teycheney[3] and Florian Maumus[1@]

Affiliations

[1] URGI, INRA, Université Paris-Saclay, 78026 Versailles, France;
[2] Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, GPO Box 267, Brisbane, Queensland 4001, Australia
[3] UMR AGAP, CIRAD, INRA, SupAgro, 97130 Capesterre Belle-Eau, France

[@]**Correspondance**

florian.maumus@inra.fr

Abstract

Endogenous viral elements (EVEs) are viral sequences that are integrated in the nuclear genomes of their hosts and are signatures of viral infections that may have occurred millions of years ago. The study of EVEs, coined paleovirology, provides important insights into virus evolution. The *Caulimoviridae* is the most common group of EVEs in plants, although their presence has often been overlooked in plant genome studies due to misidentification by automatic annotation programs. We have refined methods for the identification of caulimovirid EVEs and interrogated the genomes of a broad diversity of plant taxa, from algae to advanced flowering plants. Evidence is provided that almost every vascular plant (tracheophyte), including the most primitive taxa (clubmosses, ferns and gymnosperms) contains caulimovirid EVEs, many of which represent previously unrecognized evolutionary branches. In angiosperms, EVEs from at least two and as many as five different caulimovirid genera were frequently detected and florendoviruses were the most widely distributed, followed by petuviruses. For reasons that are unknown, citrus and castor bean contained particularly high densities of caulimovirid EVEs, about 10 times higher than the average across all seed plants. From the analysis of the distribution of different caulimovirid genera within different plant species, we propose a working evolutionary scenario in which this family of viruses has emerged during the Silurian era (approx. 420 million years ago) when land plants first emerged.

Introduction

Although the field of viral metagenomics is exponentially expanding the repertoire of viral genome sequences available for evolutionary studies (Roossinck, 2016), it only provides a picture of viral diversity over a very short geological time scale. However, viruses can leave molecular records in the genomes of their hosts in the form of endogenous viral elements (EVEs). EVEs are viral sequences that have transferred into the nuclear genomes of their hosts by either active or passive integration mechanisms and retained over extended periods of time, sometimes millions of years. The study of EVEs does allow the evolution of viruses to be traced, much like a fossil record (Aiewsakun and Katzourakis, 2015). For example, the study of endogenous retroviruses has led to the conclusion that retroviruses have a marine origin and that they developed in parallel with their vertebrate hosts more than 450 million years ago (MYA; (Aiewsakun and Katzourakis, 2017)).

Plant EVEs were first discovered a little more than 20 years ago (Bejarano et al., 1996) and have only received a fraction of the research attention directed towards endogenous retroviruses in humans and animals. Most characterized plant EVEs are derivatives of viruses in the family *Caulimoviridae* (Teycheney and Geering, 2011). The *Caulimoviridae* is one of the four families of reverse-transcribing viruses or virus-like retrotransposons that occur in eukaryotes, and is the only family of viruses with a double-stranded DNA genome that infects plants (https://talk.ictvonline.org/). Eight different genera of the *Caulimoviridae* are currently recognized by the International Committee for the Taxonomy of Viruses (ICTV), of which five (petu-, badna-, caulimo-, solendo- and soymovirus) have EVE counterparts in at least one plant genome (Teycheney and Geering, 2011) (Mushegian and Elena, 2015). Recently, (Geering et al., 2014) showed that EVEs from an additional tentative genus of the *Caulimoviridae*, called 'Florendovirus', are widespread in the genomes of cultivated and wild angiosperms and provided evidence for the oldest EVE integration event yet reported in plants, at 1.8 MYA (Geering et al., 2014). Although no extant florendoviruses has been identified, their endogenous fossils account for almost half of all recognized *Caulimoviridae*, illustrating that fossil viral species characterized from EVEs may outnumber extant ones and that they have the potential to help refine viral taxonomy. The discovery of endogenous florendoviruses in ANITA grade plant species also showed that beyond mesangiosperms, *Caulimoviridae* host range spans or once spanned in more basal angiosperms. Furthermore, contrary to extant viral retroelements, endogenous florendovirus genomes revealed a bipartite genome organization (Geering et al., 2014). These findings demonstrate that analyzing the genetic footprints left by viruses in plant genomes can contribute to a better understanding of the long-term processes driving the evolution of the *Caulimoviridae*.

In this study, we report on the discovery of EVEs from several novel viruses in a significant number of plant genomes and we propose to create nine new virus genera in the *Caulimoviridae* to accommodate these new viral species. We show that the *Caulimoviridae* host range extends to the Euphyllopyte and Lycopodiophytes clades, surpassing that of any plant virus family. By analyzing the distribution of different genera of the *Caulimoviridae* within different plant species, we unveil a complex pattern of associations and propose a scenario in which *Caulimoviridae* would have emerged together with early land plants approximately 420 million years ago and evolved principally through vertical transmission.

Results

Augmenting the diversity of endogenous caulimovirids

The reverse transcriptase (RT) domain is the most conserved domain in the genome of viral retroelements and is used for classification (Xiong and Eickbush, 1990) (Hansen and Heslop-Harrison, 2004). The strong sequence conservation of this domain allows high quality alignments to be generated, even for distantly related taxa. We have thus used a collection of RT domains from known exogenous and endogenous caulimovirids to search for related sequences across the breadth of the Viridiplantae (four green algae, two basal land plants, four gymnosperms, and 62 angiosperms; Supplementary Table 1) using BLAST.

Initially, over 8,400 protein-coding sequences were retrieved, all containing an RT domain with a best reciprocal hit against members of the *Caulimoviridae*, as opposed to the closely related *Metaviridae* (Ty3-*gypsy* group LTR retrotransposons). To provide a preliminary classification, sequences with at least 55% amino acid identity to each other were clustered and then iteratively added to our reference set of RT domains to build a phylogenetic network. The successive networks were examined manually and representative sequences from each cluster were kept only when creating substantially divergent branches so as to cover an extended diversity of caulimovirid RT with a core sequence assortment. While this network-based approach cannot be taken as phylogenetic reconstruction, it provided a practical method to explore diversity.

In the final phylogenetic network (Figure 1), 17 groups were identified, hereafter referred to as operational taxonomic units (OTUs). Remarkably, nine of these OTUs lacked recognized members of the *Caulimovidae*. Four of these novel OTUs were exclusively composed of sequences from gymnosperms, thereby representing a new and significant host range extension for the *Caulimoviridae*. These OTUs were named Gymnendovirus 1 to 4. Two other novel OTUs were composed of RTs from various angiosperms and were named Xendovirus and Yendovirus. The last three novel OTUs were small in size, comprising sequences from one or two plant species (*Petunia inflata* and *Petunia axillaris*; *Vitis vinifera*; *Glycine max*; named species-wise: Petunia-, Vitis-, and Glycine-endovirus). This initial search therefore enabled uncovering a significantly augmented diversity of caulimovirid RTs.

Endogenous caulimovirid RT (ECRT) density across the Viridiplantae

Using the sequences from the final phylogenetic network (Figure 1) to perform a second, more comprehensive search for ECRTs in our collection of plant genomes, we detected 14,895 genomic loci representing high-confidence ECRT candidates. Remarkably, ECRTs were found in nearly all seed plants, ranging from gymnosperms (ginkgo and conifers) to angiosperms. However, none were detected in green algae and non-seed basal land plants for which complete genomes are publicly available (Supplementary Table 1). Quantitatively, over one-thousand ECRTs were detected in the genome assemblies of the gymnosperms *Picea glauca* (white spruce) and *Pinus taeda* (loblolly pine), as well as from the solanaceous species *Capsicum annuum* (bell pepper) (Figure 2A). In general, we observed a positive correlation between plant genome size and the number of ECRTs, although there were notable exceptions such as the monocot *Zea mays* (maize), which has a relatively large genome at 2.1 Gb but no detectable ECRT. Five other seed plants from our sample also lacked ECRTs, including two other monocots (*Zostera marina* and *Oryza brachyantha*) and three dicots in the order *Brassicales* (*Arabidopsis thaliana*, *Schrenkiella parvula* and *Carica papaya*). When the number of

ECRTs was normalized against genome size, *Citrus sinensis* (sweet orange) and *Ricinus communis* (castor bean) had the highest densities at 2.3 and 2 ECRTs per Mb, respectively (Figure 2B). The ANITA grade angiosperm *Amborella trichopoda* also had a relatively high density of ECRTs (1 ECRT per Mb) compared to an average density of 0.2 ECRT per Mb across the 62 seed plant species that were examined.

Caulimovirid sequences are also detected in ferns and a clubmoss

As ECRTs were detected in gymnosperm genomes, we extended our search to the genomes of ferns (class Polypodiopsida, basal seed plants), which represent an intermediate bifurcation in the evolution of the Viridiplantae but for which no complete genome is publicly available. We retrieved genomic contigs from six fern genomes that have recently been sequenced at low coverage (approximately 0.4 to 2 x genome size equivalent (Wolf et al., 2015)) and screened this data set for the presence of ECRTs. A total of twenty-one protein-coding ECRTs were detected in genomic contigs from five out of the six fern species examined (Supplementary Table 1). Phylogenetic network reconstruction using representative fern ECRTs revealed that they form two novel OTUs that were named Fernendovirus 1 and 2 (Supplementary Figure 2).

To further explore the association of *Caulimoviridae* with species from basal lineages of the Viridiplantae, 1,000 plant transcriptomes generated by the 1KP initiative (Matasci et al., 2014) (Wickett et al., 2014) were interrogated. From this sample, we found two transcript contigs (2.4 and 2.8 kilobases long, respectively) in *Botrypus virginianus* (identifier BEGM-2004510) and *Lindsaea linearis* (identifier NOKI-2097008), which contained ECRTs. Remarkably, we identified one more transcript contig (identified as ENQF-2084799, 2kb) that contained an ECRT in the clubmoss *Lycopodium annotinum*, which belongs to *Lycopoda*, the most basal radiation of vascular plants (Tracheophyta).

Phylogenetic reconstruction

Complete or near complete viral genomes were reconstructed from each novel OTU except Fernendovirus (Supplementary file 1). From the fern genomic data sets, we were able to reconstruct fragments of Fernendovirus 1 & 2 genomes that contain genome coverage for phylogenetic analysis. We also used the complete genomes of type species of the eight currently recognized genera in the family *Caulimoviridae* (*Badnavirus*, *Caulimovirus*, *Cavemovirus*, *Petuvirus*, *Rosadnavirus*, *Solendovirus*, *Soymovirus*, *Tungrovirus*), those of two unassigned *Caulimoviridae*, Blueberry fruit drop associated virus (BFDaV, (Diaz-Lara and Martin, 2016)) and Rudbeckia flower distortion virus (RuFDV, (Lockhart et al., 2017)), and those of endogenous caulimovirids from the tentative genera Orendovirus (Geering et al., 2010) and Florendovirus (Geering et al., 2014). From this library of caulimovirid genomes, we aligned the genomic regions containing the conserved protease, reverse transcriptase and ribonuclease H1 domains to build a maximum likelihood phylogenetic tree (Figure 3).

This tree confirmed that the reconstructed genomes clade within the *Caulimoviridae*. In agreement with previous studies (Geering et al., 2014), the tree revealed two sister clades (hereafter referred to as clade A and B). Clade A comprised sequences from representatives of Xendovirus and Yendovirus OTUs and from members of genera *Caulimovirus*, *Soymovirus*, *Rosadnavirus*, *Solendovirus*, *Cavemovirus*, *Badnavirus*, *Tungrovirus* and Orendovirus, as well as from unassigned RuFDV and BFDaV. Noteworthy, the two sequences reconstructed from the Xendovirus OTU appear to clade at distant positions in the phylogenetic tree: the one from *Gossypium raimondii* (cotton) is most closely related to the cluster encompassing sequences from genera *Cavemovirus* and *Solendovirus* whereas the one reconstructed from *Fragaria vesca* (wild strawberry) is closer to BFDaV. This sequence from *Fragaria vesca* was hence reclassified in an additional OTU named Zendovirus. The representative sequence for Yendovirus, reconstructed from *Capsicum annuum* (bell pepper), is most closely related to the clade comprising representatives of the genera *Badnavirus*, *Tungrovirus* and Orendovirus.

Clade B comprised sequences from representatives of the four gymnendovirus OTUs and those from ferns and clubmoss, as well as sequences from genera Florendovirus and *Petuvirus*. Sequences from Fernendovirus 1 & 2 clade together, the clubmoss sequence being collapsed in Fernendovirus 1 (Figure 3). Petuvirus clade is sister to Fernendovirus 2 though with a significant level of uncertainty (bootsrap value=49%). Gymnendovirus 1 is sister to the Fernendo/Petuvirus clade whereas Gymnendovirus 3 and 4 form a cluster that is sister to Florendovirus. Gymnendovirus 2 sequences group in a clade that is sister to all other clade B viruses. Noteworthy, reconstructed genomes from novel OTUs found in single dicotyledon species (Petunia-, Vitis-, and Glycine-endovirus) were discarded from the phylogenetic reconstruction as they appeared to significantly weaken the robustness of the tree. They could however be placed unambiguously in clade A (data not shown).

ECRT distribution across seed plant genomes

To address the distribution of caulimovirid EVEs in our collection of plant genomes, we determined the most likely phylogenetic position within the reference *Caulimoviridae* phylogenetic tree proposed above (Figure 3) for the 14,895 ECRTs that we collected from seed plant genomes using the pplacer program (Matsen et al., 2010). For this, we extracted ECRT loci extending upstream and downstream so as to retrieve potential sequences containing the contiguous fragment corresponding to the protease, RT and ribonuclease H1 domains. Using more relaxed length criteria, we extracted a total of 134 ECRT loci from the fern genomic data set that we also attempted to place on our reference tree.

Applying this strategy, we were able to assign unambiguous phylogenetic position on specific OTUs to a total of 13,834 ECRTs (Figure 4), the remaining ECRT loci being placed on inner nodes of the reference tree. Overall, we observed striking differences between Caulimoviridae genera for both the number of ECRT loci and the number of plant species in which they were found. For instance, Florendovirus ECRT loci were the most abundant, amounting to an overall total of 5k copies and they were also found in the highest number of host species (46 of the 62 seed plant species that were screened). *Petuvirus* ECRT loci were also well represented, with an overall total of 1.9k copies found in a total of 27/62 seed plant species, especially in dicots. Among the novel OTUs, ECRTs classified as Yendovirus were found in the largest number of species, including monocots and dicots (Figure 4).

Most importantly, the detailed distribution of *Caulimoviridae* in plant genomes reveals striking differences between lycopods, ferns, gymnosperms and angiosperms (Figure 4). No single OTU spans several plant divisions on Figure 4 (which describes plant genomic contents) but Fernendovirus 1 sequences are found in both fern genomes and lycopod transcriptome. Fern genomes contain exclusively ECRT loci that are classified as Fernendovirus 1 & 2. In corollary, Fernendovirus ECRTs are found only in club moss and ferns. Gymnosperm genomes enclose exclusively ECRT loci that are assigned to one of the four Gymnendovirus OTUs, all of which being undetected outside of gymnosperms. Among gymnosperms, the three conifer genomes analyzed contain a mixture ECRTs from the four Gymnendovirus genera. By contrast, only ECRT loci classified as Gymnendovirus 2 were detected in *Ginkgo biloba* (*Ginkgoales*). Within angiosperms, we also observed a dichotomy for the distribution of ECRTs between monocots and dicots. On one hand, Yendovirus, Badnavirus, Orendovirus and Florendovirus ECRTs are common in monocots. On the other hand, Petuvirus, Florendovirus, Xendovirus, Cavemovirus/Solendovirus and Yendovirus ECRTs are the most widely distributed in dicots, Florendovirus and Yendovirus hence being remarkably well represented in both dicots and monocots.

Discussion

Endogenous viral elements are considered relics of past infections, and an extrapolation of the results from this study is that nearly every tracheophyte plant species in the world has at some point in its evolutionary history been subject to infection by at least one, and sometimes five distinct viral species/genera from the family *Caulimoviridae*. This finding attests to the tremendous adaptability of the *Caulimoviridae*, surpassing any other groups of plant viruses. Members of the *Caulimoviridae* have likely also had a large influence on plant evolution, either as pathogens or donors of novel genetic material to the plant genome.

Our findings provide insights into the evolutionary history of the *Caulimoviridae*. Nine new OTUs were detected, each of which likely representing several virus genera. In addition, we detected sequences defining three additional OTUs that were initially detected in only one plant species (or genus in the case of Petunia) (Figure 1). Considering that we have discovered nine putative novel genera in the family *Caulimoviride* by screening over 60 genomes from a variety of land plants, compared to the eight current genera in this family, one can reasonably assume that the systematic search for caulimovirid EVEs in the incoming flow of plant genomic resources will further increase the amount of viral species in this family, and probably the number of OTUs in the *Caulimoviridae*.

A defining moment in the evolution of the *Caulimoviridae* appears to be the development of vasculature in plants. The presence of a 30K movement protein is an important feature of the *Caulimoviridae* that distinguishes it from the *Metaviridae*, and this protein is crucial for the formation of systemic infection by allowing intercellular trafficking of macromolecules through increasing the size exclusion limit of plasmodesmata (Link and Sonnewald, 2016). Although algae contain plasmodesmata, which superficially resemble those of higher plants, they are not homologous and their molecular structure is different from that of higher plants (Brunkard and Zambryski, 2017). While the acquisition of a 30K movement protein would have provided a selective advantage for ancestral caulimovirids to colonize the tracheophytes, it would not have facilitated infection of more primitive plant forms.

In light of the results, one of the most surprising findings was the absence of ECRTs in a select number of plant species such as *Arabidopsis thaliana*. Given that close relatives of *A. thaliana*, such as *A. lyrata*, contain ECRTs, it is unlikely that *A. thaliana* has avoided infection either through chance or because of disease resistance. One of the special traits of *A. thaliana* is its very small genome size, a reason it has developed as a model plant species. Comparisons of *A. thaliana* with large-genome plant species such as *Nicotiana tabacum* and *Hordeum vulgare* suggest that there are marked differences in double-stranded break repair (DSB) mechanisms (Orel and Puchta, 2003) (Vu et al., 2017). The frequency and size of DNA insertions after DSB is much higher for *H. vulgare* compared to *A. thaliana*, and conversely, the size of deletions much higher for *A. thaliana*. These intrinsic differences in DSB between plant species have likely acted against the accumulation of EVEs in *A. thaliana*.

The correlation between plant genome size and the number of ECRTs (Figure 2A) may also suggest that extensive heterochromatic regions as found in large genomes are relatively permissive to the retention of *Caulimoviridae* insertions compared to gene-rich regions. In this regard, Florendovirus EVEs are so widespread in dicolyledon species that their absence from the medium sized genomes of *Medicago truncatula* (412 Mbp) and *Cannabis sativa* (585 Mbp) is remarkable and could reflect acquired resistance in these species. In contrast, peculiar dynamics of *Caulimoviridae* integration could explain the relatively high density of ECRTs observed in the sweet orange and castor bean genomes. Finally, the absence of ECRT in the genome of the monocotyledon plant *Zostera marina* is not surprising considering its marine lifestyle (Olsen et al., 2016) and it stresses that the return to the sea undergone by some flowering plants also provided an escape from terrestrial viruses.

We can assume that the extinction of viral genera is common over long evolutionary scales, due to the extinction of vectors or the development of plant resistance. Therefore, one can expect that only a fraction of ancestral *Caulimoviridae* is represented in their modern descendants, whether endogenous or exogenous. Therefore our results need to be interpreted with the consideration that, most probably, many unknown *Caulimoviridae* genera are lacking from our phylogeny. Moreover, incomplete lineage sorting can cause two sequences separated by a whole world of extinct/unavailable sequences to appear most closely related in phylogenetic reconstructions. We also consider that the most parsimonious path to explain the spread of viruses within host taxonomy is vertical transmission. Vertical transmission is well supported by a co-evolutionary study of Florendovirus EVEs and their host species (Geering et al., 2014) and it is probably also the leading route followed by retroviruses along with their host species (Hayward et al., 2015). Following a co-evolutionary scenario by pure vertical transmission, a virtually complete phylogeny of *Caulimoviridae* genera should mirror the one of host species. If the last common host associated with two different genera, then the vertical transmission of each ancestral genus would lead to two paralog clusters, each mirroring host evolution, just as ancestrally duplicated genes would show in modern species.

Indeed, we can note a relatively deep cluster in clade B in which *Caulimoviridae* evolution to a large extent mirrors host evolution over major divisions of land plants (Figure 5). This cluster is hereafter referred to as modern mirror (mm) cluster. Within the mm cluster, the Caulimoviridae sequence isolated from clubmoss is most closely related to ECRT from ferns and is included in Fernendovirus 1 OTU. *Petuvirus* is sister to both Fernendovirus OTUs but with significant uncertainty. A cluster containing sequences from angiosperms and gymnosperms branches at the base of this group. The finding that the phylogeny of some *Caulimoviridae* mirrors the evolution of plant species prompted us to further elaborate on the evolutionary history of the *Caulimoviridae* based principally on vertical transmission in order to interpret the observed *Caulimoviridae*-plant associations.

Reconciling our data with a vertical transmission-based co-evolutionary scenario, the mm cluster would be unique in that its last common ancestor would be represented by modern descendants that associate with species from the four plant divisions (angiosperms, gymnosperms, ferns, and clubmoss). In this cluster, the phylogenetic relationships between viruses and between plants can be interpreted by coevolution through vertical transmission while introducing very few viral speciation events and gaps filling due to extinct or non-sampled sequences. As vertical transmission goes forward, ancient *Caulimoviridae* would have associated with the ancestor of modern clubmoss species. The position of the genus that associates with modern clubmoss in the mm cluster, *i.e.* Fernendovirus 1, supports this hypothesis.

Deeper in the evolutionary history of the *Caulimoviridae*, our working scenario would involve the existence of a last common ancestor (LCA) of Caulimoviridae at latest during the emergence of clubmoss, *i.e.* towards the end of the Silurian era 420 MYA (Hedges et al., 2006). Before the emergence of ferns about 380 MYA, the speciation of Caulimoviridae LCA would have evolved into the ancestors of clade A and B and both would have continued vertical transmission with their host. The absence of ECRTs from modern descendants in the (non-mm) clusters could then be explained mainly by their relatively ancient death and/or by limited sampling (Figure 5).
We consider that an alternative scenario that would involve a significant number of host swaps between plant divisions is counter intuitive considering the current data, because several of these swaps would actually overlap with plant evolutionary history, as exemplified in the mm cluster. Therefore, we currently favor at the earliest Silurian origin of the *Caulimoviridae* followed by viral speciation and then principally by vertical transmission from the emergence of ferns to later bifurcations in plant evolution.

Supplementary material

Supplementary file 1: reconstructed ancestral sequences of members of the *Caulimoviridae* from novel OTUs (fasta format).

Figure legends

Figure 1: Augmented diversity of the Caulimoviridae. Core of a phylogenetic network constructed using an alignment of amino acid reverse transcriptase (RT) sequences from reference genera, representative endogenous caulimovirid RTs (ECRTs) and Ty3/Gypsy LTR retrotransposons. The full network is available in Supplementary Figure 1. This representation allows determining 17 Caulimoviridae OTUs. OTU names have dashed lime green outline when they include no known reference genera (referred to as novel OTUs). Each fill color corresponds to a different OTU except for OTUs comprising only a representative ECRT sequence that are colored with dark grey and named after the only host plant genome they were detected in at this stage (Petunia-, Vitis-, and Glycine-virus). * RT clustering at 55% identity groups Cavemovirus and Solendovirus into a single OTU (OTU 8). ** Sequences grouped in the Xendovirus OTU appeared to be paraphyletic after phylogenetic reconstruction (see Figure 3).

Figure 2: Highly variable ECRT numbers and density across plants. (A) Number of ECRTs found in each plant genome as function of Log10 genome size expressed in megabases (assembly gaps excluded). Logarhitmic trendline indicates moderate correlation between the number of ECRT and genome size (R^2=0.544). (B) Density of ECRTs per megabase in each plant genome as function of Log10 genome

size expressed in megabases (assembly gaps excluded). In (A) and (B), arrows indicate a sample of outlier dots and the corresponding plant species name.

Figure 3: Phylogeny of the *Caulimoviridae*. Phylogenetic tree obtained by maximum likelihood search from a multiple sequence alignment of the genomic regions containing protease, reverse transcriptase and ribonuclease H1 domains from known (black) and novel (red) Caulimoviridae genera. The sequences from Gypsy and Ty3 LTR retrotransposons are used as outgroups. Bootstrap support values below 50% are not shown. Sequences from members of the novel genera are available in supplementary data. Closely related sequences were collapsed into branches. The sequences contained in each branch are as follows. Orendovirus: Aegilops tauschii virus (AtV), Brachypodium distachyon virus (BdV); *Tungrovirus*: *Rice tungro bacilliform virus* (RTBV), Rice tungro bacilliform virus isolate west Bengal (RTBV); *Badnavirus*: *Commelina yellow mottle virus* (ComYMV), *Banana streak OL virus* (BSOLV); Yendovirus: Capiscum annuum virus; Zendovirus: Fragaria vesca virus; Blueberry: Blueberry fruit drop associated virus (BFDaV); Caulimovirus: Cauliflower mosaic virus (CaMV), Figwort mosaic virus (FMV); Rudbeckia: Rudbeckia flower distortion virus (RuFDV); *Soymovirus*: Soybean chlorotic mottle virus (SoyCaulimoviridae), Peanut chlorotic streak virus (PCSV); *Solendovirus*: Sweet potato vein clearing virus (SPVCV), Tobacco vein clearing virus (TVCV); *Cavemovirus*: Cassava vein mosaic virus (CsVMV), Sweet potato collusive virus (SPCV); Petuvirus: Petunia vein clearing virus (PVCV); *Rosadnavirus*: *Rose yellow vein virus* (RYVV); Florendovirus: Fragaria vesca virus (FvesV), Mimulus guttatus virus (MgutV); Gymnendovirus 1: Pinus taeda Gymnendovirus 1, Picea glauca Gymnendovirus 1; Gymnendovirus 2: Pinus taeda Gymnendovirus 2, Picea glauca Gymnendovirus 2, Ginkgo biloba Gymnendovirus 2; Gymnendovirus 3: Pinus taeda Gymnendovirus 3; Gymnendovirus 4: Pinus taeda Gymnendovirus 4, Picea glauca Gymnendovirus 4; Fernendovirus 1: Cystopteris protrusa Fernendovirus 1 contig 1, and the transcript scaffolds BEGM-2004510 from Botrypus virginianus, NOKI-2097008 from Lindsaea linearis, and ENQF-2084799 from Lycopodium annotinum; Fernendovirus 2: Dipteris conjugata Fernendovirus 2 Contigs 2, 4 and 1319.

Figure 4: Distribution of endogenous Caulimoviridae in Euphyllophyte. The left tree represents a cladogram of Euphyllophyte species investigated in this study. The name of major branches and nodes is indicated. The top tree represents the topology of the phylogenetic tree obtained in Figure 3. At the intersection of these two trees, color code indicates the number of ECRT loci classified into each Caulimoviridae genus for each plant species. Abbreviations of Caulimoviridae genera are as follows: Pe (Petuvirus), Gy1 (Gymnendovirus 1), Gy2 (Gymnendovirus 2), Gy3 (Gymnendovirus 3), Gy4 (Gymnendovirus 4), Fe1 (Fernendovirus 1), Fe2 (Fernendovirus 2), Flo (Florendovirus), Soy (Soymovirus), Rud (Rudbeckia flower distortion virus), Cau (Caulimovirus), Blu (Blueberry fruit drop associated virus), Zen (Zendovirus), Xen (Xendovirus), Yen (Yendovirus), CaS (Cavemovirus +Solendovirus), Ros (Rosadnavirus), Bad (Badnavirus), Tun (Tungrovirus), Ore (Orendovirus).

Figure 5: Working scenario of Caulimoviridae deep evolution. The left tree is the same as in Figure 3 where the deepest Caulimoviridae node was annotated as LCA (last common ancestor) and its two daughter nodes were annotated Clade A and Clade B. The modern mirror (mm) cluster is delimited by a green rectangle. Well supported branches within the mm cluster have been casted with dashed blue lines onto a cladogram, on the right, that recapitulates the relationships between major trachepohyte divisions (branches are annotated following the date of emergence of the different plant divisions in million years ago (MYA).

Supplementary Figure 1: Overview of the phylogenetic network used to build Figure 1.

Supplementary Figure 2: ECRT ORFs collected from ferns cluster as two novel OTUs. Representative sequences identified in fern genomes were appended to the collection of sequences represented in Figure 1. The resulting library has been re-aligned with MUSCLE and phylogenetic network was built using SplitsTree. The branches containing fern sequences have been empirically grouped into two novel OTUs.

Methods

Discovery and clustering of novel Caulimoviridae OTUs

We built a library containing an assortment of amino acid (aa) sequences from 54 RT domains including four from *Retroviridae*, 6 from Ty3/Gypsy LTR retrotransposons, 41 from 8 different known Caulimoviridae genera (Florendovirus, Caulimovirus, Tungrovirus, Cavemovirus, Solendovirus, Badnavirus, Soymovirus, and Petuvirus), 2 from Picea glauca identified ahead as belonging to new Caulimoviridae genera, and the one from the DIRS-1 element. We compared this library to a collection of 72 genome assemblies from Viridiplantae species (listed in Supplementary Table 1) using tBLASTn with default parameters (except –e=1e-5). The hit genomic loci were merged when overlapping and their coordinates were extended 120 bases upstream and downstream. Extended hit loci were translated and the protein sequences of length >=200aa were compared to the initial RT library using BLASTp with default parameters (except –e=1e-5). Queries with best alignment score against Caulimoviridae over at least 170 residues were selected for further analysis. For each plant species, the selected set of RT aa sequences have been clustered following sequence similarity using the UCLUST program (Edgar, 2010) with identity threshold set at 80%. The longest sequence from each resulting cluster was considered as the representative sequence and it was appended to the initial RT library. To detect potential false positives, each set of sequences (each consisting of the initial RT library and cluster representatives from one species) was aligned using MUSCLE followed by filtering of lower fit sequences using two rounds of trimAl v1.2 (Capella-Gutierrez et al., 2009) to remove poorly aligned sequences (-resoverlap 0.75 -seqoverlap 50) separated by one round to remove gaps from the alignment (-gt 0.5). The representative sequences from each plant species that passed this selection have then been combined into a single file and appended to the initial RT library to be clustered with UCLUST using an identity threshold of 55%. At this level of similarity, aa RT sequences from every Caulimoviridae genera fall into distinct clusters except those from Cavemovirus and Solendovirus that cluster together. Starting with the first cluster, one or more sequences presenting high quality alignment and containing several conserved residues as determined contextually for each cluster were then manually selected to be representative of the diversity observed within each cluster. The following clusters were processed similarly while keeping the representative sequences selected from previously processed clusters. Clusters containing ECRT sequences from only one plant species were analyzed only when they contained at least 3 sequences. After processing each cluster individually, a total of 56 ECRT sequences detected here and 20 RT from known genera have been selected for their remarkable divergence. Together with 4 RT sequences from Ty3/Gypsy LTR retrotransposons, these combined sequences (hereafter referred to as “diverse library”) were aligned with the GUIDANCE2 (Sela et al., 2015) program using MAFFT (Katoh and Standley, 2013) to generate bootstrap supported MSA and to remove columns (--

colCutoff) with confidence score below 0.95 (16/244 columns removed in the RT sequence from Caulimovirus CaMV). The resulting MSA was then used to build the phylogenetic network shown in Figure 1 and Supplementary Figure 1 with SplitsTree4 (Huson and Bryant, 2006) applying the NeighborNet method with uncorrected P distance model and 1,000 bootstrap tests. Manual analysis of this network enabled the empirical discrimination of 17 distinct OTUs among Caulimoviridae sequences.

Because initial search allowed discovering several novel OTUs, we repeated ECRT mining in plant genomes using the diverse library as query. This second search is also designed to be more sensitive as it takes into account DNA sequences instead of uninterrupted ORFs. The workflow is identical to the one employed for the initial search until obtaining the set of extended hit loci. These were directly compared to the diverse library using BLASTx with default parameters (except –e=1e-5). Queries with best alignment score against any Caulimoviridae with an alignment length above 80% of subject length (set generically to 576 bp considering an average size of RT domains of 240 aa) were selected for phylogenetic placement.

Phylogenetic analysis

Fragments of virus sequence were assembled using CodonCode aligner 6.0.2 using default settings or using VECTOR NTI Advance 10.3.1 (Invitrogen) operated using default settings, except that the values for maximum clearance for error rate and maximum gap length were increased to 500 and 200, respectively.

Phylogenetic reconstruction was performed using the contiguous nucleotide sequences corresponding to the protease, reverse transcriptase and ribonuclease H domains. Whole sequences from Caulimoviridae genera representatives and Ty3 and Gypsy LTR retrotransposons were first aligned with global method using MAFFT v7.3 (Katoh and Standley, 2013). The core genomes was extracted and re-aligned by local method using MAFFT. The resulting alignment was tested for different evolutionary models with pmodeltest v1.4 (from ETE 3 package (Huerta-Cepas et al., 2016)) which inferred the GTRGAMMA model. Phylogenetic inference with maximum likelihood was then performed using RaxML v8.2 (Stamatakis, 2014) under the predicted model with 500 ML bootstrap replicates.

The resulting tree was then used as a reference to classify the ECRT loci mined from plant genomes. We first added query sequences from each plant species separately to the reference alignment and aligned each library using Mafft v7.3 (with options --addfragment, --keeplength and by reordering). We then tested the most likely placement of each ECRT sequence on to the reference tree using pplacer v1.1 alpha19 (Matsen et al., 2010) with the option (--keep-at-most 1) with allows to keep one placement for each query sequence. The python package Taxit was used to construct a reference package which we used to run pplacer.

References

AIEWSAKUN, P. & KATZOURAKIS, A. (2015) Endogenous viruses: Connecting recent and ancient viral evolution. *Virology,* 479-480**,** 26-37.

AIEWSAKUN, P. & KATZOURAKIS, A. (2017) Marine origin of retroviruses in the early Palaeozoic Era. *Nat Commun,* 8**,** 13954.

BEJARANO, E. R., KHASHOGGI, A., WITTY, M. & LICHTENSTEIN, C. (1996) Integration of multiple repeats of geminiviral DNA into the nuclear genome of tobacco during evolution. *Proc Natl Acad Sci U S A,* 93**,** 759-64.

BRUNKARD, J. O. & ZAMBRYSKI, P. C. (2017) Plasmodesmata enable multicellularity: new insights into their evolution, biogenesis, and functions in development and immunity. *Curr Opin Plant Biol,* 35**,** 76-83.

CAPELLA-GUTIERREZ, S., SILLA-MARTINEZ, J. M. & GABALDON, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics,* 25**,** 1972-3.

DIAZ-LARA, A. & MARTIN, R. R. (2016) Blueberry fruit drop-associated virus: A New Member of the Family Caulimoviridae Isolated From Blueberry Exhibiting Fruit-Drop Symptoms. *The American Phytopathological Society,* 100**,** 2211-2214.

EDGAR, R. C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics,* 26**,** 2460-1.

GEERING, A. D., MAUMUS, F., COPETTI, D., CHOISNE, N., ZWICKL, D. J., ZYTNICKI, M., MCTAGGART, A. R., SCALABRIN, S., VEZZULLI, S., WING, R. A., QUESNEVILLE, H. & TEYCHENEY, P. Y. (2014) Endogenous florendoviruses are major components of plant genomes and hallmarks of virus evolution. *Nat Commun,* 5**,** 5269.

GEERING, A. D., SCHARASCHKIN, T. & TEYCHENEY, P. Y. (2010) The classification and nomenclature of endogenous viruses of the family Caulimoviridae. *Arch Virol,* 155**,** 123-31.

HANSEN, C. & HESLOP-HARRISON, J. S. (2004) Sequences and phylogenies of plant pararetroviruses, viruses, and transposable elements. *Advances in Botanical Research,* 41**,** 165-193.

HAYWARD, A., CORNWALLIS, C. K. & JERN, P. (2015) Pan-vertebrate comparative genomics unmasks retrovirus macroevolution. *Proc Natl Acad Sci U S A,* 112**,** 464-9.

HEDGES, S. B., DUDLEY, J. & KUMAR, S. (2006) TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics,* 22**,** 2971-2.

HUERTA-CEPAS, J., SERRA, F. & BORK, P. (2016) ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol Biol Evol,* 33**,** 1635-8.

HUSON, D. H. & BRYANT, D. (2006) Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol,* 23**,** 254-67.

KATOH, K. & STANDLEY, D. M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol,* 30**,** 772-80.

LINK, K. & SONNEWALD, U. (2016) Interaction of Movement Proteins with Host Factors, Mechanism of Viral Host Cell Manipulation and Influence of MPs on Plant Growth and Development. IN KLEINOW, T. (Ed.) *Plant-Virus Interactions: Molecular Biology, Intra- and Intercellular Transport.* Cham, Springer International Publishing.

LOCKHART, B., MOLLOV, D., OLSZEWSKI, N. & GOLDSMITH, N. (2017) Identification, transmission and genomic characterization of a new member of the family Caulimoviridae causing a flower distortion disease of Rudbeckia hirta. *Virus Res,* In Press, Corrected Proof.

MATASCI, N., HUNG, L. H., YAN, Z., CARPENTER, E. J., WICKETT, N. J., MIRARAB, S., NGUYEN, N., WARNOW, T., AYYAMPALAYAM, S., BARKER, M., BURLEIGH, J. G., GITZENDANNER, M. A., WAFULA, E., DER, J. P., DEPAMPHILIS, C. W., ROURE, B., PHILIPPE, H., RUHFEL, B. R., MILES, N. W., GRAHAM, S. W., MATHEWS, S., SUREK, B., MELKONIAN, M., SOLTIS, D. E., SOLTIS, P. S., ROTHFELS, C., POKORNY, L., SHAW, J. A., DEGIRONIMO, L., STEVENSON, D. W., VILLARREAL, J. C., CHEN, T., KUTCHAN, T. M., ROLF, M., BAUCOM, R. S., DEYHOLOS, M. K., SAMUDRALA, R., TIAN, Z., WU, X., SUN, X., ZHANG, Y., WANG, J., LEEBENS-MACK, J. & WONG, G. K. (2014) Data access for the 1,000 Plants (1KP) project. *Gigascience,* 3**,** 17.

MATSEN, F. A., KODNER, R. B. & ARMBRUST, E. V. (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics,* 11**,** 538.

MUSHEGIAN, A. R. & ELENA, S. F. (2015) Evolution of plant virus movement proteins from the 30K superfamily and of their homologs integrated in plant genomes. *Virology,* 476**,** 304-15.

OLSEN, J. L., ROUZE, P., VERHELST, B., LIN, Y. C., BAYER, T., COLLEN, J., DATTOLO, E., DE PAOLI, E., DITTAMI, S., MAUMUS, F., MICHEL, G., KERSTING, A., LAURITANO, C., LOHAUS, R., TOPEL, M., TONON, T., VANNESTE, K., AMIREBRAHIMI, M., BRAKEL, J., BOSTROM, C., CHOVATIA, M., GRIMWOOD, J., JENKINS, J. W., JUETERBOCK, A., MRAZ, A., STAM, W. T., TICE, H., BORNBERG-BAUER, E., GREEN, P. J., PEARSON, G. A., PROCACCINI, G., DUARTE, C. M., SCHMUTZ, J., REUSCH, T. B. & VAN DE PEER, Y. (2016) The genome of the seagrass Zostera marina reveals angiosperm adaptation to the sea. *Nature,* 530**,** 331-5.

OREL, N. & PUCHTA, H. (2003) Differences in the processing of DNA ends in Arabidopsis thaliana and tobacco: possible implications for genome evolution. *Plant Mol Biol,* 51**,** 523-31.

ROOSSINCK, M. J. (2016) Deep sequencing for discovery and evolutionary analysis of plant viruses. *Virus Res*.

SELA, I., ASHKENAZY, H., KATOH, K. & PUPKO, T. (2015) GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res,* 43**,** W7-14.

STAMATAKIS, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics,* 30**,** 1312-1313.

TEYCHENEY, P. Y. & GEERING, A. D. (2011) Endogenous viral sequences in plant genomes. *Recent advances in plant virology.* Norfolk, Caister Academic Press.

VU, G. T. H., CAO, H. X., REISS, B. & SCHUBERT, I. (2017) Deletion-bias in DNA double-strand break repair differentially contributes to plant genome shrinkage. *New Phytol,* 214**,** 1712-1721.

WICKETT, N. J., MIRARAB, S., NGUYEN, N., WARNOW, T., CARPENTER, E., MATASCI, N., AYYAMPALAYAM, S., BARKER, M. S., BURLEIGH, J. G., GITZENDANNER, M. A., RUHFEL, B. R., WAFULA, E., DER, J. P., GRAHAM, S. W., MATHEWS, S., MELKONIAN, M., SOLTIS, D. E., SOLTIS, P. S., MILES, N. W., ROTHFELS, C. J., POKORNY, L., SHAW, A. J., DEGIRONIMO, L., STEVENSON, D. W., SUREK, B., VILLARREAL, J. C., ROURE, B., PHILIPPE, H., DEPAMPHILIS, C. W., CHEN, T., DEYHOLOS, M. K., BAUCOM, R. S., KUTCHAN, T. M., AUGUSTIN, M. M., WANG, J., ZHANG, Y., TIAN, Z., YAN, Z., WU, X., SUN, X., WONG, G. K. & LEEBENS-MACK, J. (2014) Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc Natl Acad Sci U S A,* 111**,** E4859-68.

WOLF, P. G., SESSA, E. B., MARCHANT, D. B., LI, F. W., ROTHFELS, C. J., SIGEL, E. M., GITZENDANNER, M. A., VISGER, C. J., BANKS, J. A., SOLTIS, D. E., SOLTIS, P. S., PRYER, K. M. & DER, J. P. (2015) An Exploration into Fern Genome Space. *Genome Biol Evol,* 7**,** 2533-44.

XIONG, Y. & EICKBUSH, T. H. (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J,* 9**,** 3353-62.

Figures

Figure 1

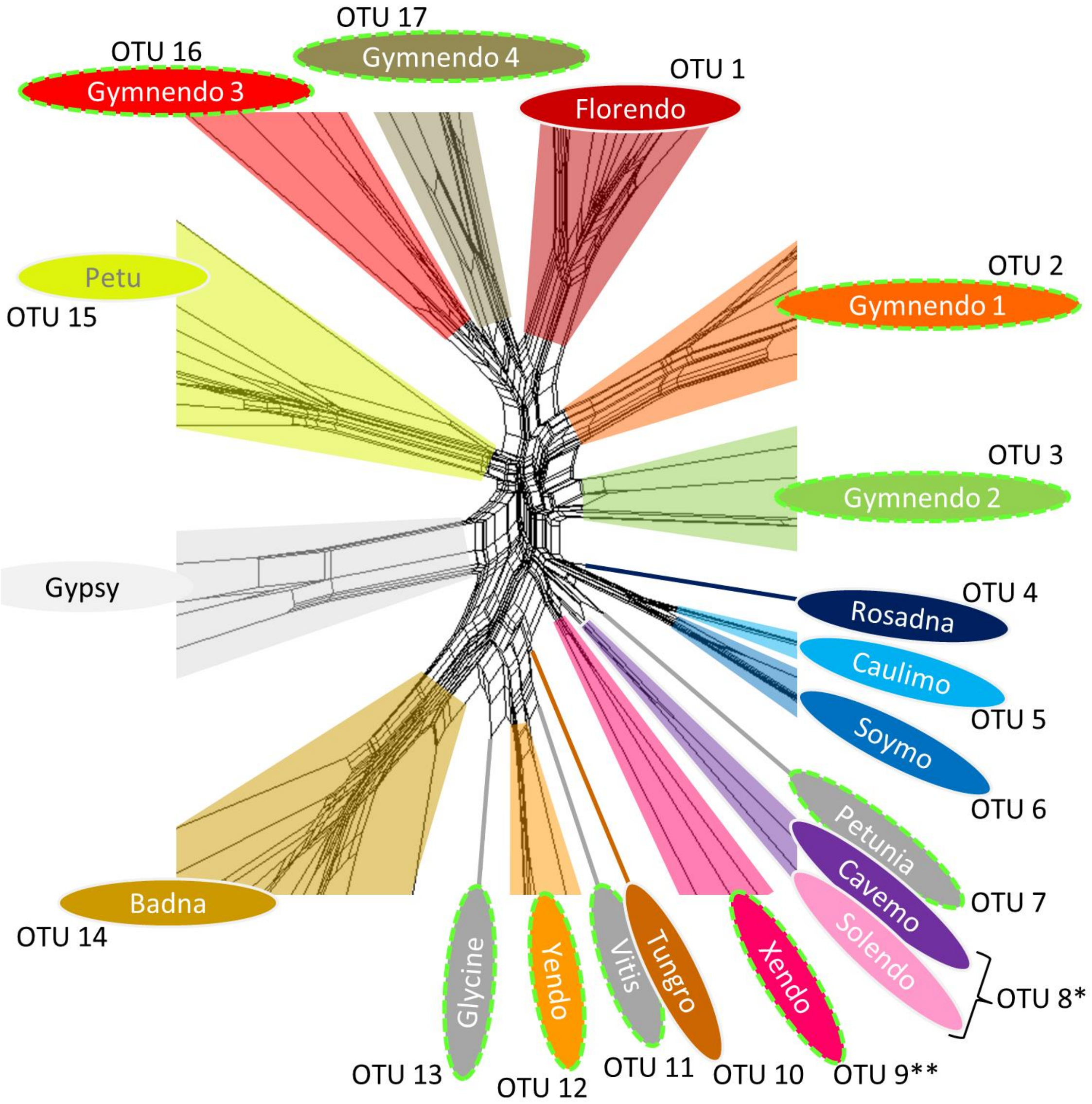


Figure 2

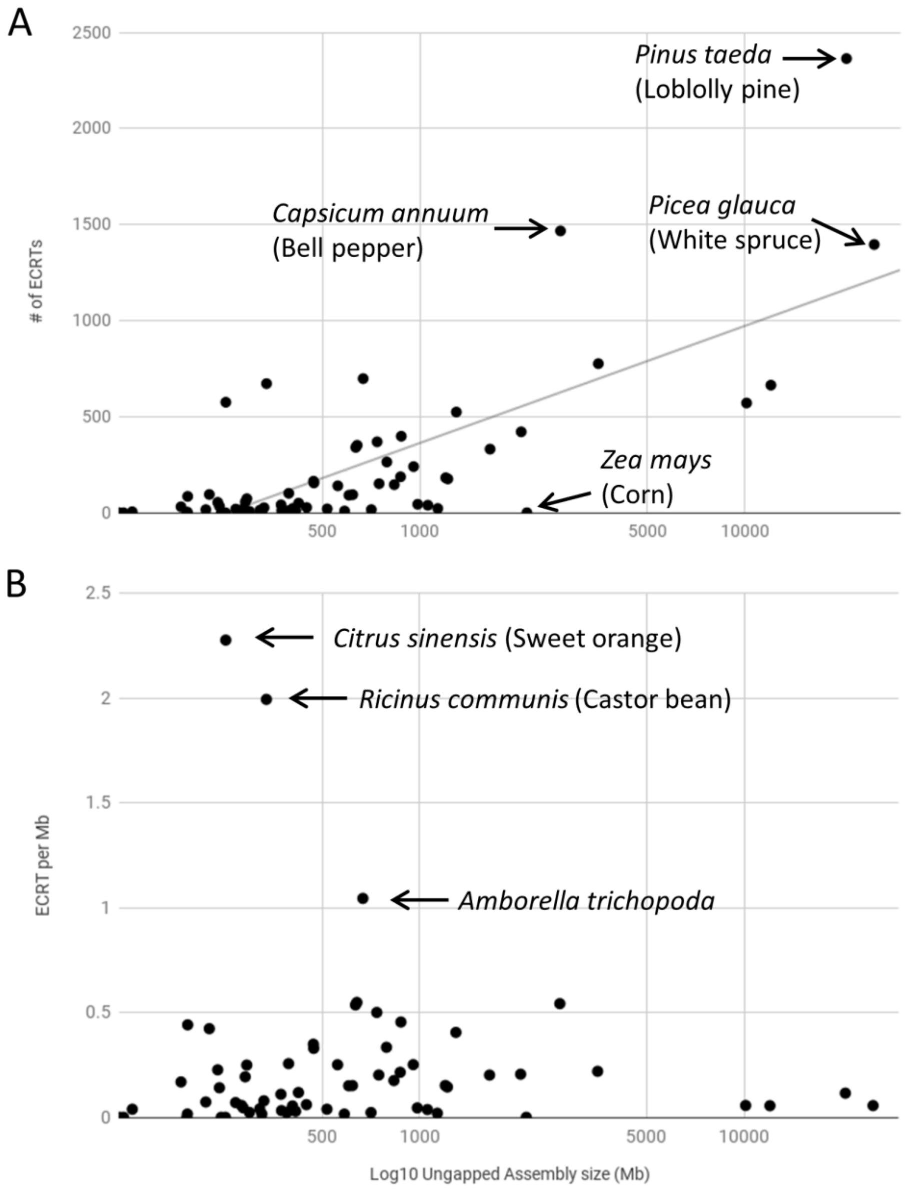

A
2500
2000
1500
1000
500
0
# of ECRTs
500
1000
5000
10000
Pinus taeda
(Loblolly pine)
Capsicum annuum
(Bell pepper)
Picea glauca
(White spruce)
Zea mays
(Corn)
B
2.5
2
1.5
1
0.5
0
ECRT per Mb
Citrus sinensis (Sweet orange)
Ricinus communis (Castor bean)
Amborella trichopoda
500
1000
5000
10000
Log10 Ungapped Assembly size (Mb)

Figure 3

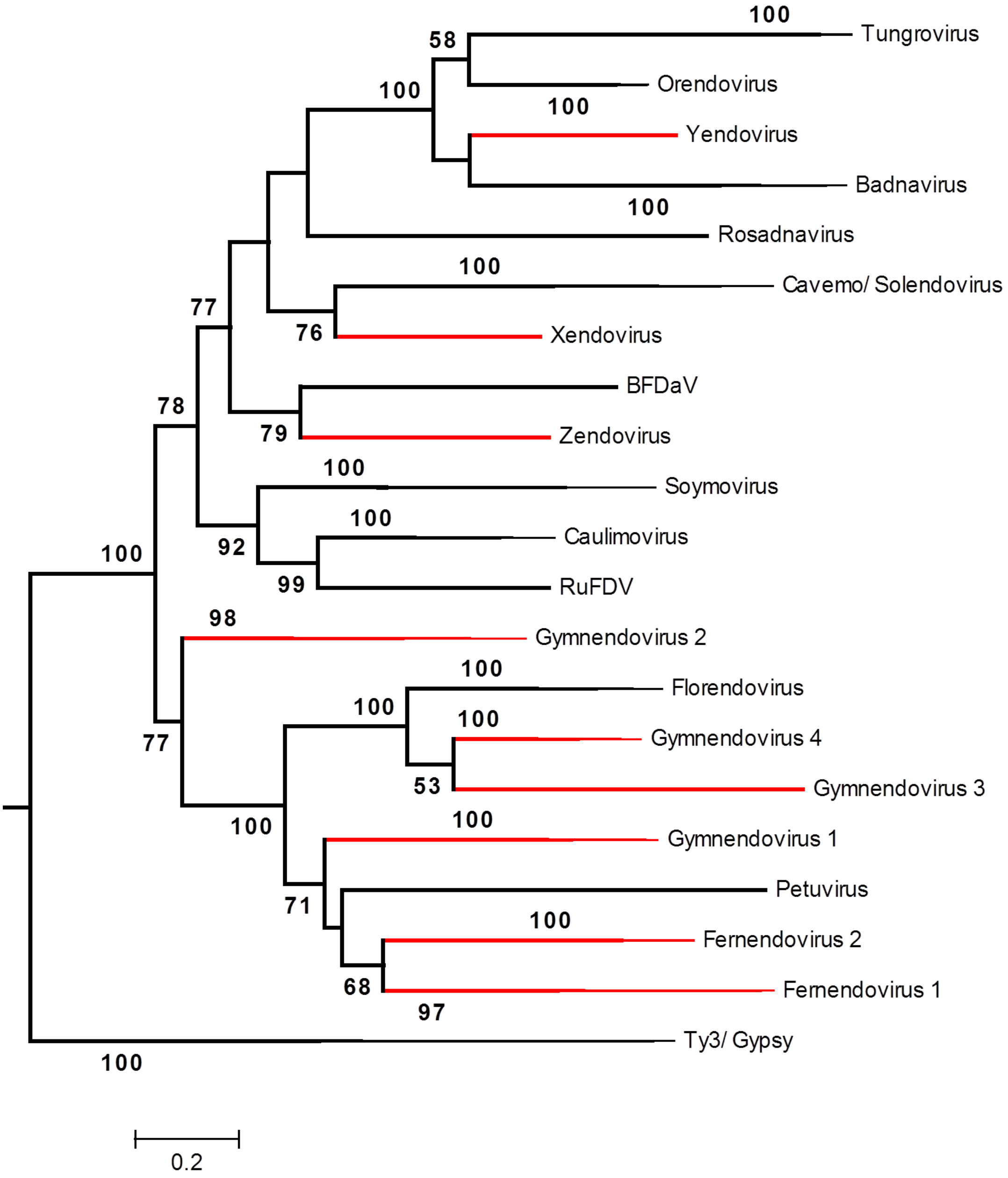

100
58
Tungrovirus
100
Orendovirus
100
Yendovirus
Badnavirus
100
Rosadnavirus
100
Cavemo/ Solendovirus
77
76
Xendovirus
BFDaV
78
79
Zendovirus
100
Soymovirus
100
Caulimovirus
100
92
99
RuFDV
98
Gymnendovirus 2
100
Florendovirus
100
100
Gymnendovirus 4
77
53
Gymnendovirus 3
100
100
Gymnendovirus 1
71
Petuvirus
100
Fernendovirus 2
68
Fernendovirus 1
97
Ty3/ Gypsy
100
0.2

Figure 4

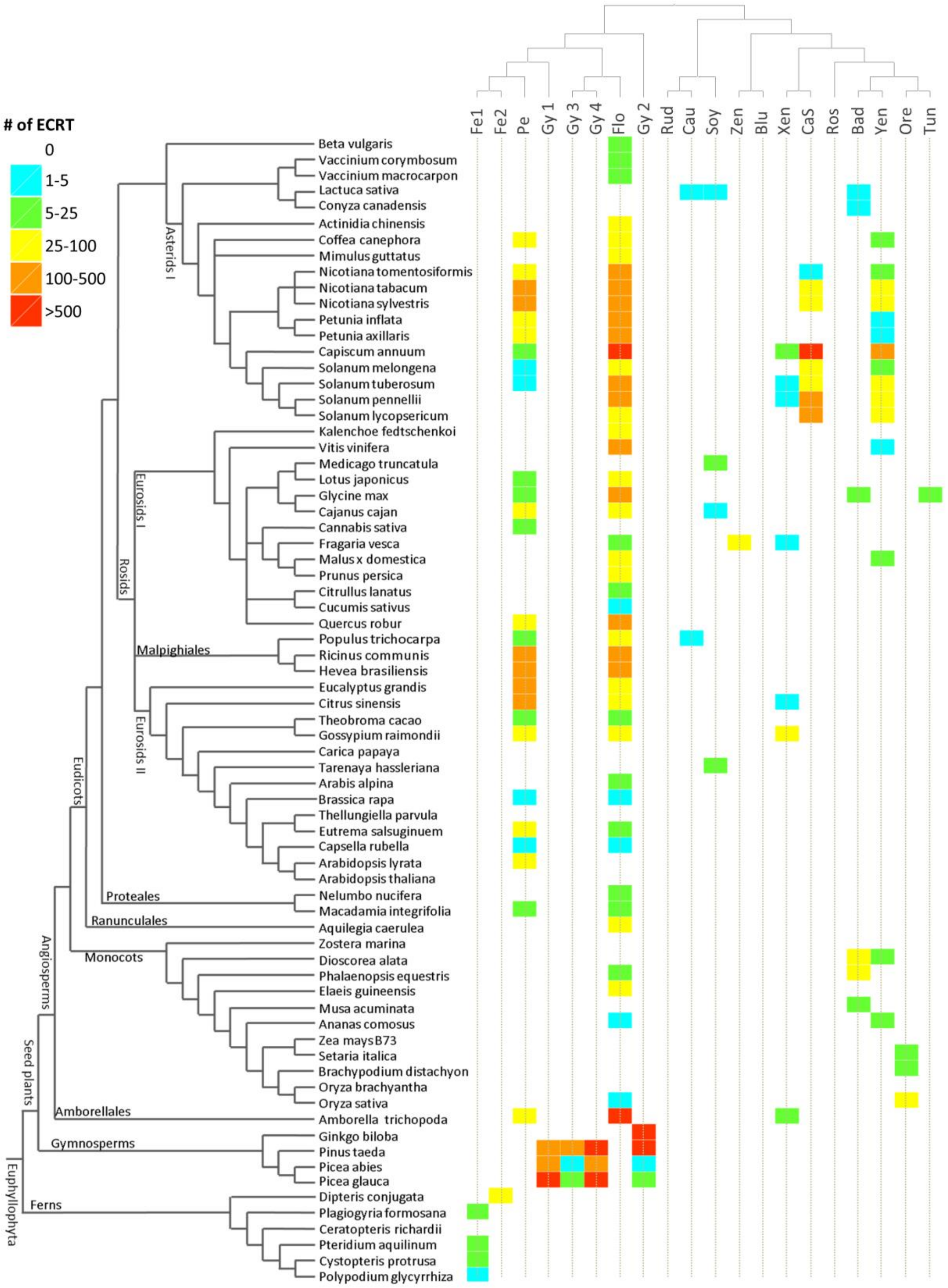
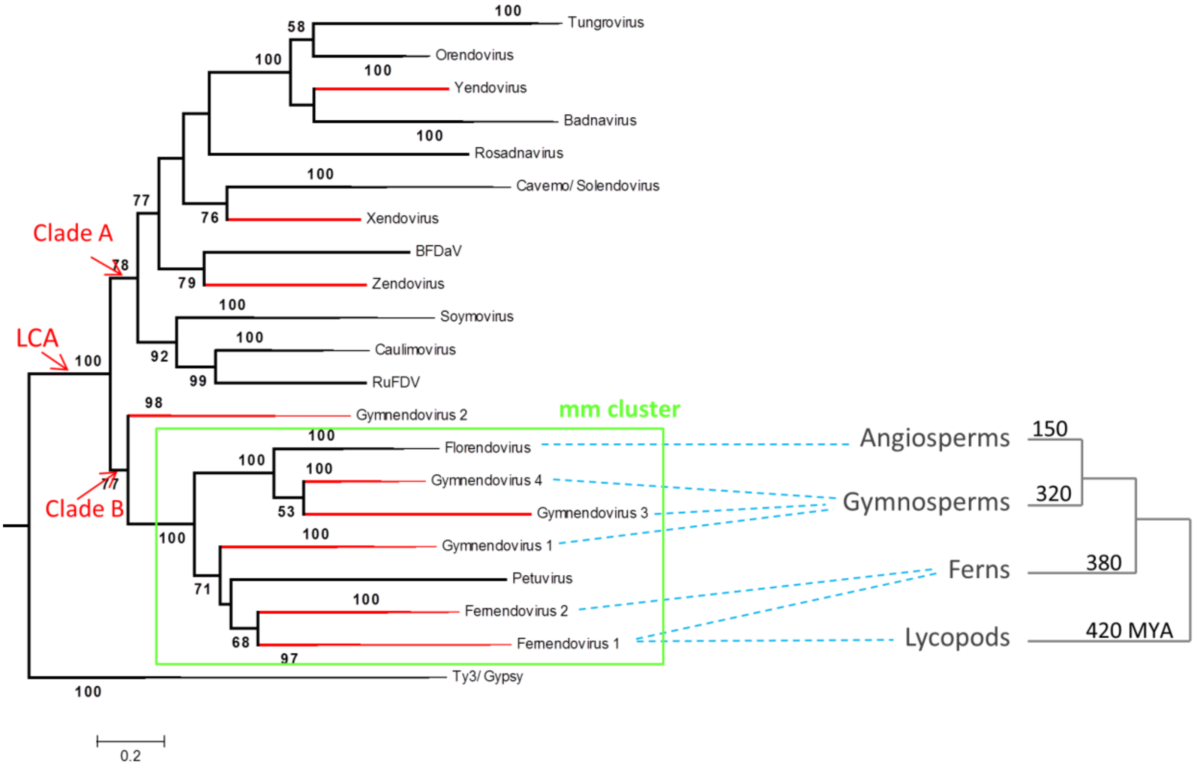
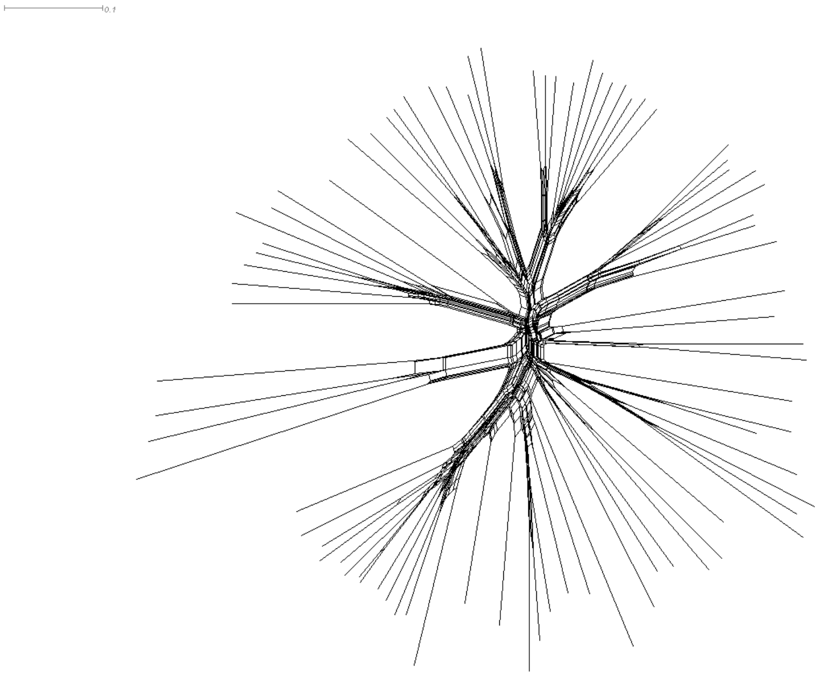

# of ECRT
0
🟦 1-5
🟩 5-25
🟨 25-100
🟧 100-500
🟥 >500
Fe1
Fe2
Pe
Gy 1
Gy 3
Gy 4
Flo
Gy 2
Rud
Cau
Soy
Zen
Blu
Xen
CaS
Ros
Bad
Yen
Ore
Tun
Beta vulgaris
Vaccinium corymbosum
Vaccinium macrocarpon
Lactuca sativa
Conyza canadensis
Actinidia chinensis
Coffea canephora
Mimulus guttatus
Nicotiana tomentosiformis
Nicotiana tabacum
Nicotiana sylvestris
Petunia inflata
Petunia axillaris
Capiscum annuum
Solanum melongena
Solanum tuberosum
Solanum pennellii
Solanum lycopsericum
Kalenchoe fedtschenkoi
Vitis vinifera
Medicago truncatula
Lotus japonicus
Glycine max
Cajanus cajan
Cannabis sativa
Fragaria vesca
Malus x domestica
Prunus persica
Citrullus lanatus
Cucumis sativus
Quercus robur
Populus trichocarpa
Ricinus communis
Hevea brasiliensis
Eucalyptus grandis
Citrus sinensis
Theobroma cacao
Gossypium raimondii
Carica papaya
Tarenaya hassleriana
Arabis alpina
Brassica rapa
Thellungiella parvula
Eutrema salsuginuem
Capsella rubella
Arabidopsis lyrata
Arabidopsis thaliana
Nelumbo nucifera
Macadamia integrifolia
Aquilegia caerulea
Zostera marina
Dioscorea alata
Phalaenopsis equestris
Elaeis guineensis
Musa acuminata
Ananas comosus
Zea mays B73
Setaria italica
Brachypodium distachyon
Oryza brachyantha
Oryza sativa
Amborella trichopoda
Ginkgo biloba
Pinus taeda
Picea abies
Picea glauca
Dipteris conjugata
Plagiogyria formosana
Ceratopteris richardii
Pteridium aquilinum
Cystopteris protrusa
Polypodium glycyrrhiza
Asterids I
Eurosids I
Rosids
Malpighiales
Eurosids II
Eudicots
Proteales
Ranunculales
Monocots
Angiosperms
Amborellales
Seed plants
Gymnosperms
Ferns
Euphyllophyta

Figure 5

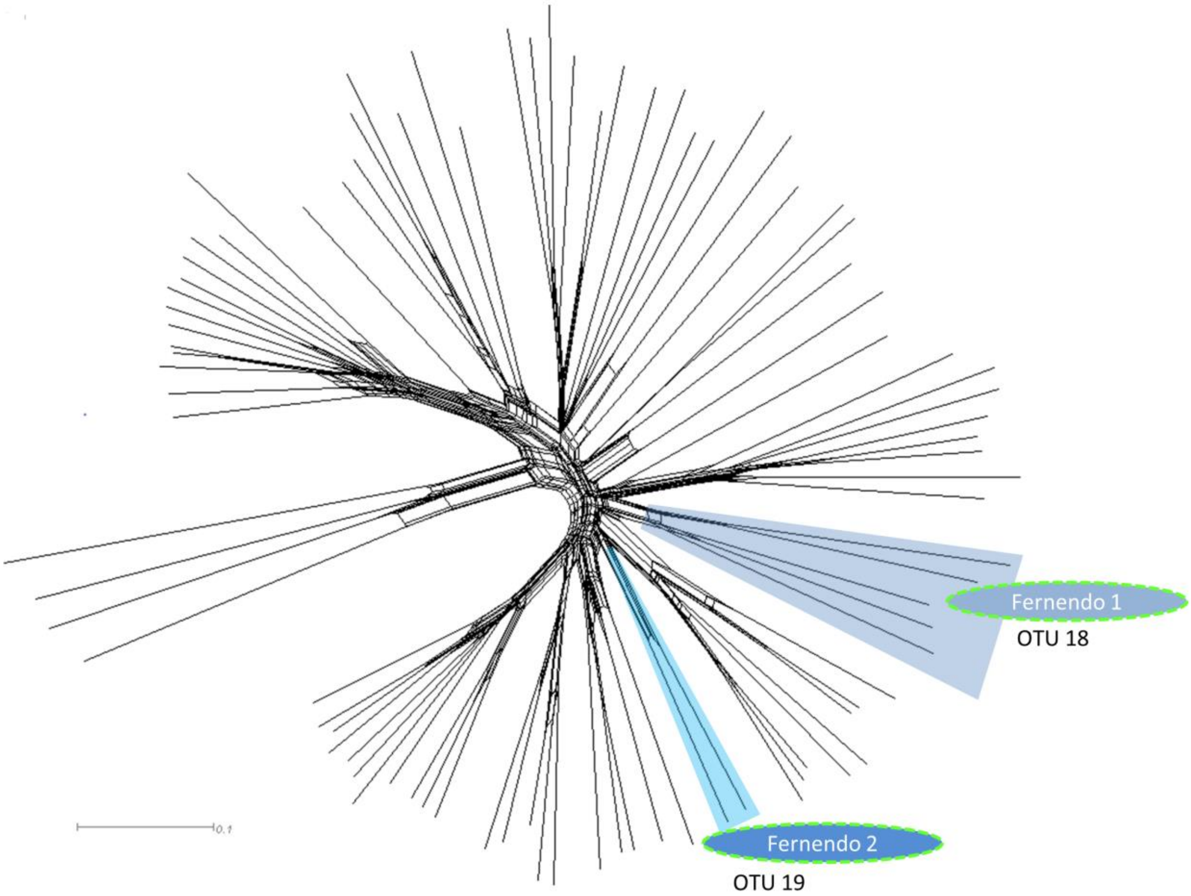

100
58
Tungrovirus
100
Orendovirus
100
Yendovirus
Badnavirus
100
Rosadnavirus
100
Caverno/ Solendovirus
77
76
Xendovirus
Clade A
78
BFDaV
79
Zendovirus
100
Soymovirus
LCA
100
Caulimovirus
100
92
99
RuFDV
98
Gymnendovirus 2
mm cluster
100
Florendovirus
100
100
Gymnendovirus 4
77
Clade B
53
Gymnendovirus 3
100
100
Gymnendovirus 1
71
Petuvirus
100
Fernendovirus 2
68
Fernendovirus 1
97
Ty3/ Gypsy
100
0.2
Angiosperms
150
Gymnosperms
320
Ferns
380
Lycopods
420 MYA

Supplementary figures

Supplementary figure 1



Supplementary figure 2


Fernendo 1
OTU 18
Fernendo 2
OTU 19
0.1