The diversity and host range of endogenous caulimovirids illuminate the ancient origin and early evolution of the *Caulimoviridae*

Héléna Vassilieff[1], Saad Serfraz[2], Nathalie Choisne[1], Andrew D.W. Geering[3], Pierre Lefeuvre[4,5], Pierre-Yves Teycheney[4], Florian Maumus[1]

[1] Université Paris-Saclay, INRAE, BioinfOmics, URGI, 78026, Versailles, France
[2] CABB, University of Agriculture, Faisalabad, 38000, Pakistan
[3] QAAFI, The University of Queensland, Brisbane, QLD 4072, Australia
[4] CIRAD, UMR PVBMT, F-97410 Saint-Pierre de La Réunion, France
[5] CIRAD, UMR PVBMT, Department of Plant Protection, College of Agriculture, Can Tho University, Can Tho City, Vietnam

Present address:
NC and FM: Université Paris-Saclay, INRAE, AgroParisTech, Institute Jean-Pierre Bourgin for Plant Sciences (IJPB), 78000, Versailles, France
HV: Molecular Evolution, Department of Cell and Molecular Biology, Uppsala University, 752 36 Uppsala, Sweden

Authors' ORCID:
FM: 0000-0001-7325-0527
HV: 0000-0002-8578-8153
SS: 0000-0001-8187-8226
NC: 0000-0001-6024-1664
ADWG: 0000-0002-5743-6804
PL: 0000-0003-2645-8098
PYT: 0000-0002-9754-0745

Corresponding authors:
HV: helena.vassilieff@icm.uu.se
PYT: pierre-yves.teycheney@cirad.fr
FM: florian.maumus@inrae.fr

Abstract

Endogenous viral elements (EVEs) result from the integration of viral sequences into the genome of their hosts. EVEs can be considered molecular fossils of ancient or unknown viruses. Through paleovirological approaches, EVEs can help unravel the evolutionary trajectories of related viruses across extended temporal scales. In plants, most characterized EVEs belong to the family *Caulimoviridae*, the only family of retrotranscribed plant viruses, and are termed endogenous caulimovirids (ECVs). Previous studies of ECVs have proposed either cospeciation or host-switching as the primary mode of *Caulimoviridae* transmission through plant divisions, both scenarios assuming an emergence predating the euphyllophytes. However, robust assessments of ECV diversity and hosts were hindered by the underrepresentation of basal plant genome assemblies.

Taking advantage of newly available genomic resources, we analyzed a dataset of 93 plant genome sequences representing all subdivisions of embryophytes, including basal groups such as ferns and lycophytes. We identified 47,135 ECVs in 75 genomes from lycophytes, ferns, gymnosperms, and angiosperms. Based on their diversity, these ECVs were classified into 71 operational taxonomic units (OTUs), 35 of which were previously undescribed. Our work revealed an unexpected diversity of *Caulimoviridae* in tracheophytes and identified a new clade restricted to gymnosperms. By comparing the phylogenetic relationships of *Caulimoviridae* with host plant taxonomy, we propose a new macroevolutionary scenario in which cospeciation with the tracheophytes is the main driver of *Caulimoviridae* diversification.

Introduction

Endogenous viral elements (EVEs) are viral sequences that are integrated into the genome of their host organism and are transmitted vertically from one generation to the next, like a normal cellular gene. In plants, the most common types of EVE originate from viruses in the family *Caulimoviridae* and are referred to as endogenous caulimovirids (ECVs; Vassilieff *et al.* 2023). Viruses in the *Caulimoviridae* possess circular, non-covalently closed double-stranded DNA genomes that mandatorily encode a movement protein (MP), a capsid protein (CP), and a polymerase polyprotein (Pol) containing aspartic protease (AP), reverse transcriptase (RT), and ribonuclease H1 (RH) domains (Teycheney et al. 2020). They are currently classified into 11 genera based on virion morphology, genome organization, and transmission mode. Within each genus, viral species are distinguished based on phylogenetic relationships inferred from the RT-RH domain, using a species demarcation threshold of 80% nucleotide sequence identity. Some genera possess auxiliary proteins unique to their lineage, such as the aphid transmission factor found in the genera *Caulimovirus* and *Soymovirus* (Teycheney et al. 2020).

The *Caulimoviridae* is one of five families in the order *Ortervirales*, along with the *Retroviridae*, *Belpaoviridae* (Bel/Pao elements), *Metaviridae* (Ty3/Gypsy elements), and *Pseudoviridae* (Ty1/Copia elements) (Krupovic et al. 2018). All viruses in the *Ortervirales* share a replication strategy whereby the genome alternates between dsDNA and ssRNA through cycles of transcription and reverse transcription. Viruses in the *Caulimoviridae* differ from all other *Ortervirales* by lacking a proviral stage in the replication cycle involving integration into the chromosomes through the action of an integrase enzyme (Krupovic et al. 2018; Teycheney et al. 2020). Nevertheless, caulimovirid sequences have been widely and repeatedly integrated into plant genomes, probably as filler sequences during repair of dsDNA breaks in chromosomes (Vassilieff et al. 2023).

Since the first ECVs were reported at the end of the 1990s (Harper et al. 1999; Jakowitsch et al. 1999; Ndowora et al. 1999), the number of available plant genome sequences has grown exponentially, enabling ECVs to be identified relatively easily using a bioinformatics approach. Of all the ECVs that have been characterized, only a small number are known to be replication-competent and capable of initiating new infections upon activation by biotic or abiotic stresses (Lockhart et al. 2000; Dallot et al. 2001; Richert-Pöggeler et al. 2003). Still, several dozen complete viral genomes have been assembled from ECVs for which corresponding episomal forms are not known. These assembled genomes share characteristics of the *Caulimoviridae*, including sizes ranging from 7.1 to 9.8 kbp and a core suite of genes encoding replication, structural, and movement proteins. Phylogenetic analyses place these ECVs within the *Caulimoviridae* family, although many do not group with the 108 viral species and eleven genera currently recognized by the ICTV (Geering et al. 2014; Diop et al. 2018; Gong and Han 2018; Teycheney et al. 2020; de Tomás and Vicient 2022). Most complete caulimovirid genomes assembled from ECVs were reconstructed by aligning multiple ECV sequences belonging to the same viral species, all sharing at least 90% nucleotide sequence identity in the region encoding the RT-RH domain.

Previous studies have shown that the distribution of ECVs extends across all major lineages of euphyllophytes (ferns, gymnosperms, and angiosperms), far exceeding the host range of known extant viruses in the *Caulimoviridae* (Vassilieff et al. 2023). Initial work by Mushegian and Elena (2015) identified homologues of the caulimovirid MP domain in the genomes of

several gymnosperm and fern species. Subsequent studies by Gong and Han (2018) and Diop *et al.* (2018) confirmed and expanded these findings to a broader diversity of euphyllophytes. Notably, Diop et al. (2018) identified a transcript from the lycophyte *Lycopodium annotinum* containing a caulimovirid RT sequence, suggesting that the host range of the *Caulimoviridae* may extend to this taxonomic group as well. However, the scope of these studies was constrained by the limited availability of genomic and transcriptomic data at the time they were conducted. The recent surge in high-throughput sequencing data, especially the addition of plant genomes from previously underrepresented or undocumented taxa such as lycophytes, now allows for a more thorough exploration of ECVs in tracheophytes. This expanded dataset also enables a broader evaluation of the host range of the *Caulimoviridae* family and the refinement of existing evolutionary hypotheses.

In this study, we have investigated the presence and diversity of ECVs in a large dataset of 73 plant genome assemblies covering basal and key tracheophyte species, supplemented by transcriptomic data from additional botanical families for which publicly available genome assemblies are still lacking (Figure 1). Thousands of novel ECVs were discovered, extending the host range of the *Caulimoviridae* to all divisions of the tracheophytes. By combining our results with those of previous studies (Geering et al. 2010; Llorens et al. 2011; Geering et al. 2014; Diop et al. 2018; Gong and Han 2018; Debat and Bejerman 2022; de Tomás and Vicient 2022), we propose a new model for the macroevolution of the *Caulimoviridae* that suggests long-term host-virus cospeciation disrupted by host extinction events.

Results

Novel ECVs further extend the ancestral host range of the *Caulimoviridae*

To investigate the presence and diversity of ECVs in understudied plant taxa, we screened 73 plant genomes (Figure 1; Supp. Table 1) for endogenous caulimovirid RT domains (ECRTs) using the “marker miner” pipeline of Caulifinder branch B described by Vassilieff *et al.* (2022). The plant genome dataset comprised seven bryophytes and 66 tracheophytes, including nine lycophytes, 17 ferns, 27 gymnosperms, three basal angiosperms (ANA-grade), and ten mesangiosperms (one *Chloranthaceae* and nine Magnoliids). This screening yielded 20,267 amino acid ECRT sequences (aa-ECRTs) across 55 plant species (Figure 2a; Figure 3). Based on the default value of 80% minimum amino acid (aa) identity used by Caulifinder branch B, these sequences were grouped by host species into 1,684 clusters (Figure 2a). The longest sequence from each cluster was selected as the representative aa-ECRT (aa-ECRT rep) for further analysis.

No aa-ECRTs were detected in any of the seven bryophyte genomes analysed. In contrast, aa-ECRTs were detected in the genomes of three lycophytes and twelve ferns, as well as in all gymnosperm, basal angiosperm, *Chloranthaceae,* and *Magnoliidae* species that were examined (Figure 3). The number of aa-ECRTs per genome varied greatly, even within the same plant division. This variability was particularly pronounced in gymnosperms, where the number of aa-ECRTs ranged from three in *Metasequoia glyptostroboides* to 3,352 in *Pinus tabuliformis*. A positive correlation was observed between the number of aa-ECRTs and host

genome size (p-value = $9e^{-12}$). However, genome size alone only accounted for part of this variation (R^2 = 0.5; Supp. Figure 1).

The 1,684 aa-ECRT rep were filtered to select the most divergent sequences. For this, we performed iterations of sequence clustering, alignment with 98 aa-RT reference caulimovirid sequences representing both virus genera recognized by ICTV (Teycheney et al. 2020) and putative virus genera assembled from ECVs in previous studies (Geering et al. 2010; Geering et al. 2014; Diop et al. 2018; Gong and Han 2018; Rastrojo et al. 2018; Debat and Bejerman 2022; de Tomás and Vicient 2022) (Supp. Table 2) and manual curation, resulting in the selection of a subset of 261 aa-ECRT rep (Figure 2a).

For plant taxa with limited or no available genomic resources, we searched transcriptomic datasets for aa-ECRTs whose sequences differed by more than 35% from homologous sequences collected in plant genomes or reference sequences. To achieve this, we analysed publicly available datasets from NCBI GenBank (Sayers et al. 2025), the 1KP project (Leebens-Mack et al. 2019), and two transcriptomic studies dedicated to ferns and lycophytes, respectively (Xia et al. 2022; Ali et al. 2024). This analysis led to the selection of two additional aa-ECRTs from the transcriptomes of the ferns *Brainea insignis* (GFUE01038349.1) and *Pteris vittata* (GGXV01034871.1), which were added to the aa-ECRT rep subset, bringing the total number to 263.

A phylogenetic network was constructed after combining these sequences with the 98 aa-RT reference caulimovirid sequences, along with eight outgroup sequences from *Metaviridae* and *Retroviridae* (Supp. Figure 2). Visualisation of the resulting network guided the selection of a 62% amino acid identity threshold, which allowed a clear separation of viral genera recognized by the ICTV into distinct operational taxonomic units (OTUs) (Blaxter et al. 2005). Using this threshold, the 263 aa-ECRT rep sequences and the 98 aa-RT reference sequences clustered into 71 OTUs (Figure 2a; Supp. Figure 2; Supp Table 3). Thirty-six of these OTUs included at least one of the 98 aa-RT reference caulimovirid sequences and are hereafter referred to as “reference OTUs”. These were named after their corresponding ICTV-recognized genus (e.g. OTU Soymovirus) or after previously defined OTUs (e.g. OTU Florendovirus).

As expected, some of the 71 OTUs resulting from this work corresponded to subdivisions of OTUs defined in previous works, which used lower sequence identity thresholds (Diop et al. 2018; de Tomás and Vicient 2022) (Supp. Table 3). For instance, the OTU Gymnendovirus 2 defined by Diop *et al.* (2018) was split into three distinct OTUs in our work, two of which also included β-type gymnosperm endogenous caulimovirus-like viruses (β-GECV) described by Gong and Han (2018). The remaining 35 OTUs lacked any reference sequences and were considered novel. These OTUs were composed mostly of sequences retrieved from plant genomes or transcriptomes that had not previously been analysed for the presence of aa-ECRTs. Gymnosperms emerged as a particularly rich source of novel OTUs. Notably, *Wollemia nobilis*, considered a living fossil in the southern conifer family *Araucariaceae*, harbored five novel OTUs.

Phylogenetic analysis

To determine the phylogenetic placement of the 35 novel OTUs, we conducted a phylogenetic analysis based on nucleotide sequences encoding the RT-RH domain, the most conserved region within caulimovirid genomes (Figure 2b) (Teycheney et al., 2020). Nucleotide sequences were extracted from the 66 tracheophyte genomes of the dataset (Supp. Table 4) using the “sequence retriever” pipeline of Caulifinder branch A (Vassilieff et al. 2022). This pipeline reconstructs caulimovirid consensus sequences by aligning a minimum of five ECV copies with at least 85% identity from the same host genome. The 7 bryophyte genomes of the original dataset were not included in the analysis, since they do not contain ECVs (see above). After filtering out LTR retrotransposons, 9,031 ECV consensus sequences were obtained from 49 plant genomes (Supp. Table 4), whereas none were identified in six of the nine lycophyte and in eleven of the seventeen fern genomes analyzed.

After assigning each ECV consensus sequence containing an RT-RH domain (n=6,503) to one of the 71 previously defined OTUs, 47 ECV consensus sequences were selected to represent 22 of the 35 novel OTUs (see Materials and Methods). Among these, 17 OTUs were represented by two to four consensus sequences (Supp. Figure 3). For 17 of the 18 remaining novel OTUs, where either no consensus sequence (12 OTUs) or only a single consensus (5 OTUs)was available, RT-RH coding sequences were selected from 21 individual genomic ECV copies identified using Caulifinder branch B (Vassilieff et al. 2022), and from one sequence from the transcriptome of *Brainea insignis* (GFUE01038349.1) (Supp. Figure 3). In total, 69 RT-RH nucleotide sequences representing 34 of the 35 novel OTUs were obtained. The only exception was OTU 34, for which only highly mutated copies were recovered, precluding the selection of a representative sequence (Supp. Table 6).

Furthermore, 40 RT-RH nucleotide sequences were retrieved from representative members of the reference OTUs (Figure 2b). To enhance sequence and host diversity within reference OTUs that were previously represented by only a few sequences, Caulifinder branches A and B were applied to seven angiosperm genomes (Figure 2b; Supp. Table 5). This yielded an additional 33 RT-RH sequences (20 consensus and 13 genomic copies). For three of the 36 reference OTUs, no representative sequences could be recovered. Altogether, 142 RT-RH sequences representing 67 OTUs were compiled for phylogenetic analysis (Figure 2b; Supplementary Table 6).

Two phylogenetic trees were constructed using Maximum Likelihood (ML) and Bayesian approaches (Supp. Figure 4). Both methods produced trees with highly congruent topologies. The Bayesian tree was selected for further analysis, as it exhibited stronger support for basal nodes (posterior probability ≥ 0.95). This tree was rooted using *Saccharomyces cerevisiae* Ty3 virus, the exemplar species for the *Metaviridae*. Clades containing a single OTU were collapsed to produce a simplified version (Figure 4). The 142 caulimovirid RT-RH sequences clustered according to their respective OTUs, except for the sequences from OTU ‘Wendovirus + Xendovirus gp S. tuberosum’, which are polyphyletic, and the single sequence from OTU 30, which is nested within those of OTU 1 (Supp. Figure 4; Figure 4).

According to our phylogenetic hypothesis, OTU 19 is sister to Clades A and B, as defined by Diop *et al.* (2018), and was therefore tentatively assigned to a novel clade, Clade C. Seven of

the 34 novel OTUs (Table 1) were grouped within Clade A. Of these, OTU 21 was sister to all the other Clade A members, whereas the remaining six were more closely related to either previously described ECVs or to ICTV-recognized *Caulimoviridae* genera. In Clade B, 26 novel OTUs (Table 1) were distributed across the topology, with most showing close relationships to previously described ECVs. The sole exception was OTU 23, which clustered within the group that includes *Petuvirus*, the only Clade B OTU currently recognised as a genus by ICTV.

New insights were obtained into evolutionary relationships among previously identified caulimovirid groups. For example, Gymnendovirus 2, as originally proposed by Diop *et al.* (2018), was found to be polyphyletic. One subset of Gymnendovirus 2 (including gp P. glauca and PtaeV_2) grouped with Welwitschia mirabilis viruses 1&2 (WMV1 and WMV2) (Debat and Bejerman 2022) and three novel OTUs in a distinct subclade (B.1). Other members of Gymnendovirus 2 (e.g., gp G. biloba and gp G. biloba 2 + GbilV) formed separate branches at the base of subclade B.2 (B.2.Gymnendovirus 2 gp G. biloba 1 and B.2. Gymnendovirus 2 gp G. biloba 2). Compared to the previous study of Diop *et al.* (2018), the phylogenetic tree also revealed that Fernendovirus 1 and 2 are more distantly related, Fernendovirus 2 occupies a more basal position, and Fernendovirus 1 is more closely related to Gymnendovirus 1.

Characterization of caulimovirid Clade C

In both the ML and Bayesian phylogenetic trees (Supp. Figure 4), OTU 19 formed a polytomy with the outgroup (Ty3) and *Caulimoviridae* Clades A and B, leaving its taxonomic placement within the *Caulimoviridae* unresolved. To clarify the phylogenetic position of OTU 19 within the order *Ortervirales*, we included amino acid sequences of the RT domain (aa-RT) from sequences of OTU 19 and representatives of Clades A and B of the *Caulimoviridae* alongside aa-RT representatives of the four other *Ortervirales* families (Supp. Table 7) in an ML phylogenetic analysis. This phylogenetic analysis followed the RT-based framework previously applied to classify virus families within the order *Ortervirales* (Krupovic et al. 2018).

The resulting phylogenetic tree resolved the initial polytomy (Figure 5a), positioning OTU 19 as a sister clade to Clades A and B. OTU 19 includes sequences from *Wollemia nobilis* and *Araucaria angustifolia*, two critically endangered species in the *Araucariaceae* (Stull et al. 2021). To confirm the phylogenetic placement of OTU 19, we examined the phylogenetic relationships of the putative MP domains encoded by OTU 19 sequences from the genomes of *W. nobilis* and *A. angustifolia*, alongside MP sequences from other caulimovirid clades and from 14 plant virus families (Butkovic et al. 2023). While MPs from many viral families formed polyphyletic groups, *Caulimoviridae* MP sequences clustered into a distinct monophyletic group, within which OTU 19 sequences were deeply nested (Figure 5b).
A representative genomic assembly sequence for OTU 19 was constructed from *W. nobilis* ECRTs and tentatively named Wollendovirus 1 (WolV1; Figure 5c). The WolV1 genome is 9,631 bp in length and includes a 16 bp sequence (TGGTATCAGAGCCAGG) that is complementary to the 3' end of the host cytosolic tRNAmet. This sequence is identical to the tRNA primer binding sites observed in other caulimovirid genomes (Teycheney et al., 2020), although its genomic position is unique compared to previously characterized caulimovirids. The WolV1 genome contains two putative ORFs. The first encodes a large polyprotein (2,107 aa, 243 kDa) featuring a movement protein (MP) motif, a zinc finger (ZnF) domain typically at the C-terminus of caulimovirid capsid proteins, and the canonical aspartic protease (AP),

reverse transcriptase (RT), and ribonuclease H (RH1) domains characteristic of the *Caulimoviridae*. The second ORF encodes a 568 aa (67 kDa) protein of unknown function presenting no homolog in GenBank.
The strong similarity in genomic composition and organisation between the WolV1 genome and those of known caulimovirids suggests a synapomorphic relationship. Combined with the close phylogenetic relationship between OTU 19 MP and RT domains with those of other caulimovirids, this evidence supports the classification of Clade C as a distinct new lineage within the *Caulimoviridae* family.

Distribution of ECVs across plant genome assemblies

To assess the global distribution of ECVs across plant taxa, we searched for nucleotide sequences of endogenous caulimovirid RTs (nt-ECRTs) within the original dataset of 73 land plant genomes previously used to identify novel ECVs. To expand the taxonomic breadth and improve the representation of angiosperms in our dataset, 20 additional angiosperm genomes were included (Supp. Table 5, Supp. Table 8), resulting in a final dataset of 93 land plant genomes. Importantly, the use of nucleotide sequences rather than amino acid sequences enabled the inclusion of mutated sequences with interrupted protein translational reading frames. This analysis revealed 47,135 nt-ECRTs across 75 tracheophytes genomes. In nearly all the genomes where aa-ECRTs had previously been detected, the number of nt-ECRTs was significantly higher, on average threefold greater, with the most extreme increase (18.7-fold) observed in *Metasequoia glyptostroboides* (Supp. Table 9). No nt-ECRT was found in genomes in which no aa-ECRT had been found previously.

Similar to aa-ECRTs, the abundance of nt-ECRTs was positively correlated with the host genome size (r^2 = 0.45, p=$9e^{-13}$). This trend held true within specific plant lineages after corrections of p-value for multiple testing using the Benjamini-Hocheberg method (Benjamini and Hochberg 1995): lycophytes (r^2 = 0.85, p=$4e^{-04}$), gymnosperms (r^2 = 0.46, p=$2.8e^{-04}$), and angiosperms (r^2 = 0.43, p=0.008). However, the correlation was weaker in ferns (r^2 = 0.24, p=0.03) (Supp. Figure 5).

All nt-ECRTs were assigned to OTUs based on their best BLASTx hit against the representative aa-RT sequences of the 71 previously defined OTUs, using a minimum identity threshold of 65% to ensure reliable and unambiguous assignment to a specific OTU (Figure 2b, Figure 6, Supp. Table 10). 47,135 nt-ECRTs identified were assigned to 67 OTUs: 32,303 were assigned to the 33 reference OTUs, while the remaining 14,832 were distributed among the 34 novel OTUs. The distribution of nt-ECRTs varied substantially between OTUs in both sequence abundance and host plant diversity. For example, *Capsicum annuum* contained the highest number of nt-ECRTs associated with a single OTU, with 1,945 nt-ECRTs assigned to the OTU *Solendovirus* (Figure 6, Supp. Table 10). Florendoviruses exhibited the broadest host range, with 6,982 nt-ECRTS identified across 28 angiosperm species. At the other end of the spectrum, OTU 21, one of the novel OTUs described in this study, was represented by only two nt-ECRTS, each found in a separate genome of the *Nymphaeaceae* species *Nymphaea colorata* and *Nymphaea thermarum*.

Strong lineage-specific patterns were associated with the three caulimovirid clades. Within Clade A, nt-ECRTs were predominantly detected in angiosperm genomes (Figure 6) with four notable exceptions: (i) OTU 5 nt-ECRTs were found exclusively in the two *Gnetaceae*

(gymnosperms) analysed (*Gnetum gnemon* and *Gnetum montanum*); (ii) OTU 15 nt-ECRTs were detected only in five of the seven *Cupressaceae* (gymnosperms) genomes analysed (*Sequoia sempervirens*, *Sequoiadendron giganteum*, *Cryptomeria japonica*, *Juniperus communis* and *Cupressus sempervirens*), (iii) Wendovirus 4-related nt-ECRTs were identified in both angiosperms (*Lindenbergia philippensis*) and ferns (*Azolla pinnata* and *Rumohra adiantiformis*) and (iv) OTU 10 nt-ECRTs were restricted to the fern *Rumohra adiantiformis*, with additional evidence from transcriptomic data suggesting their presence in the ferns *Brainea insignis* and *Pteris vittata*. Furthermore, OTU 21, which occupied a basal position within Clade A, was found exclusively in *Nymphaeaceae*, a basal lineage of angiosperms. In clade B, nt-ECRTs were primarily confined to lycophyte, fern, and gymnosperm genomes, with two key exceptions. The putative genus Florendovirus and the ICTV-recognized genus *Petuvirus* were both restricted to angiosperms and often present in high copy numbers. Additionally, OTU 1 (subclade B.1) displayed a broad host range encompassing both gymnosperms and angiosperms, including early-diverging lineages such as the ANA-grade families *Nymphaeaceae* and *Amborellaceae*. In contrast, its sister lineage, OTU 30, was found exclusively in magnoliids. Finally, the host range of Clade C nt-ECRTs confirmed its specificity to the *Araucariaceae* family. These sequences were detected in *Wollemia nobilis* (genus *Wollemia*) and *Araucaria angustifolia* (genus *Araucaria*), with additional evidence suggesting their presence in raw genomic reads from *Agathis dammara*. This restricted distribution further supports the monophyly and host specificity of Clade C within this gymnosperm family.

Identification of caulimovirid sequences in plant transcriptomes

Although transcriptomic data do not permit quantitative estimates of ECV content, they can nonetheless reveal the presence or absence of specific OTUs in host species, whether the transcripts originate from ECVs or episomal caulimovirids. To further investigate caulimovirid-host interactions in basal plant lineages, we searched for caulimovirid RTs using Caulifinder branch B in the same transcriptomic dataset used above, which included sequences from NCBI GenBank (Sayers et al., 2025), the 1KP project (Leebens-Mack et al., 2019), and two studies dedicated to ferns and lycophytes, respectively (Ali et al., 2024; Xia et al., 2022).

This search identified 529 transcripts encoding caulimovirid RT domains, of which 491 originated from angiosperm transcriptomes (Supp. Table 11). To facilitate confident taxonomic assignment, we selected a subset of 194 transcripts that span at least 80% of the RT domain. Notably, this subset included 13 transcripts from taxonomic groups not represented in our genomic datasets: eight from ferns, two from gymnosperms, and three from ANA grade angiosperms (Table 2; Supp. Figure 6). Among the fern-derived sequences, originating from the orders *Cyatheales*, *Polypodiales*, *Marattiales*, and *Ophioglossales*, the latter two representing basal fern lineages (Nitta et al. 2022), three were assigned to OTUs within subclade ‘B.2.Fernendovirus 1'. The remaining five did not meet the 62% identity threshold in tBLASTn searches, yet two of them, originating from *Botrypus virginianus* (*Ophioglossales*, *Ophioglossaceae*) and *Lindsaea linearis* (*Polypodiales*, *Lindsaeaceae*), grouped within subclade ‘B.2.Fernendovirus 1’ in a phylogenetic analysis (Supp. Figure 7). The two gymnosperm-derived transcripts from species in the families *Zamiaceae* (order *Cycadales*) and *Podocarpaceae* (order *Araucariales*) were assigned to OTUs 1 and 3, respectively. Regarding the ANA-grade angiosperms, all three transcripts from *Austrobaileya scandens* (order *Austrobaileyales*) were assigned to the Florendovirus OTU, two of them clustering

within the “B.2.Florendovirus” subclade near OsatBV in a phylogenetic analysis (Supp. Figure 8) (Geering et al. 2014). Additionally, five transcripts from the family *Lycopodiaceae*, a group underrepresented in our genomic dataset, were examined. Three were assigned to subclade ‘B.2.Fernendovirus 1', while the remaining two unexpectedly clustered with Badnavirus sequences from angiosperms. This latter result may reflect host switching events or contamination during library preparation.

Evidence for patterns of cospeciation

To investigate the macroevolutionary history of the *Caulimoviridae*, we integrated evidence of host-virus associations derived from host genomic and transcriptomic data with information from known episomal caulimovirids. Figure 7 presents a condensed interactome that maps these associations at the level of viral clades or subclades and plant families or orders. Together with the distribution data shown in Figure 6, this synthesis enabled us to explore potential cases of cospeciation and host-switching.

We hypothesize that the evolutionary history of the *Caulimoviridae* has been largely shaped by cospeciation with their plant hosts, under the assumption that caulimovirid species exhibit a high degree of host specificity, typically restricted to a single plant family. In such a model, host speciation would drive viral diversification, resulting in congruent host and virus phylogenies. While the overall host-*Caulimoviridae* distribution does not fully support strict phylogenetic congruence, several localized patterns are consistent with cospeciation. These patterns suggest the co-existence of multiple *Caulimoviridae* lineages with partially overlapping host ranges. In this scenario, viral speciation may have occurred independently of host divergence, followed by parallel cospeciation within individual lineages. This would result in some plant species hosting distinct caulimovirid lineages that have co-evolved separately with their respective hosts (Figure 8). In points i) through v), we detail several distribution subpatterns that support the presence of ancient, co-existing viral lineages within the *Caulimoviridae*.

(i) **Clade C sequences**, all grouped in a single OTU, were exclusively identified in three extant species of the *Araucariaceae* family - *Wollemia nobilis*, *Araucaria angustifolia*, and *Agathis dammara* - each representing one of the family’s three extant genera. These species are geographically distributed across Australia, South America, and South East Asia, respectively, regions that became isolated following the breakup of Gondwana, a process that began around 167 MYA in the Late Jurassic and concluded approximately 100 MYA during the Cretaceous (Chatterjee et al. 2013; Reguero et al. 2014). Molecular dating suggests that these three species share a last common ancestor (LCA) that existed at least 150 MYA (Stull et al. 2021).

The restricted host range of Clade C and its correspondence with the biogeographic history of the *Araucariaceae* suggest that the last common ancestor of this plant family hosted episomal Clade C viruses, which subsequently co-diverged with each lineage. This supports the hypothesis that Clade C ECVs represent genomic integrations of modern descendants from an ancient *Caulimoviridae* lineage that originated no later than the mid-Jurassic and evolved in parallel with their slow-evolving hosts. The aa-ECRTs of Clade C exhibit high pairwise identities, up to 81.3%, 82.9% and 84% identity between *W. nobilis* and *A. angustifolia*, *A.*

angustifolia and *A. dammara*, and *W. nobilis* and *A. dammara*, respectively. This remarkable sequence conservation over such an extended evolutionary timeframe may reflect slow evolutionary rates in both the virus and its host, potentially due to strong selective constraints. Consequently, Clade C may represent a fossil caulimovirid lineage. Alternatively, the presence of closely related Clade C sequences in geographically distant hosts could result from more recent intercontinental transfers, possibly mediated by migratory animals such as birds.

(ii) **Clade A** sequences were predominantly found in angiosperms, yet six paraphyletic OTUs within this clade (*Solendovirus*, Xendovirus gp G. raimondii, *Badnavirus*, *Caulimovirus*, Unclassified gp V. Vinifera, and OTU 21) also occurred in ANA-grade lineages (Figure 6). OTU 21, which is restricted to *Nymphaeaceae*, occupies a basal position in Clade A, suggesting that diversification of this clade predated or coincided with early angiosperm evolution. Some Clade A OTUs extended beyond angiosperms. OTU 15 appeared in five *Cupressaceae* species, including *Cupressus sempervirens* and *Sequoia sempervirens*, whose LCA dates at least 170 MYA (Stull et al. 2021). Its presence here may reflect either an ancient host jump from a pre-angiosperm lineage into early *Cupressaceae* or the long-term persistence of a viral lineage now represented solely by these sequences.

Two closely related OTUs (OTU 10 and Wendovirus 4) were detected in ferns of the orders *Polypodiales* (*Brainea insignis*, *Pteris vittata*, *Rumohra adiantiformis*) and *Salviniales* (*Azolla pinnata*). These fern lineages diverged over 280 MYA, well before angiosperms arose (Lehtonen et al. 2017). Their sporadic occurrence in ferns and gymnosperms could indicate (i) the retention of ancient Clade A lineages that infected early vascular plants, now only detectable in these hosts, (ii) multiple independent host-jump events, or (iii) lineage extinction or sampling gaps in other taxa, limiting detection elsewhere. Collectively, the presence of Clade A ECRTs in both ferns and gymnosperms points to the coexistence of distinct ancient viral lineages, shaped by a combination of cospeciation, host-switching, and uneven genomic integration.

(iii) **Subclade B.1** sequences were identified across a wide phylogenetic spectrum of spermatophyte hosts, encompassing several *Magnoliidae* species, members from two ANA-grade angiosperm clades, and multiple distantly related gymnosperm families. These include *Gnetaceae*, *Welwitschiaceae*, *Pinaceae*, *Araucariaceae*, and the basal lineages *Cycadaceae* (*Cycas panzhihuaensis*) and *Zamiaceae* (*Dioon edule,* based on transcriptomic evidence, Table 2). These taxa trace back to an LCA of seed plants (spermatophytes) estimated to have lived approximately 360 MYA (Stull et al. 2021). The broad and deep distribution of subclade B.1 sequences across both early diverging gymnosperms and basal angiosperms suggests the long-term cospeciation of this viral lineage since the early evolution of spermatophytes, without significant onward transmission into more recent angiosperms.

Remarkably, sequences from Welwitschia mirabilis virus 1 and 2 (WMV1 and WMV2), derived from RNAseq data, clustered with subclade B.1. Previous work from (Debat and Bejerman 2022) suggested that these sequences represent episomal viral genomes, providing indirect evidence that subclade B.1 caulimovirid infections are ongoing in gymnosperm hosts and that subclade B.1 includes both endogenous and episomal representatives persisting across deep evolutionary time in seed plants.

(iv) **Subclade B.2** comprises three basally branching, paraphyletic OTU groups associated exclusively with gymnosperm hosts, which may represent ancient viral lineages within the

Caulimoviridae: (i) the 'B.2.Gymnendovirus 2 gp G. biloba 2' group, found only in *Ginkgo biloba*; (ii) the 'B.2.Gymnendovirus 2 gp G. biloba 1' group, which exhibits a broader host range including the extant gymnosperm families *Araucariaceae*, *Cycadaceae*, *Ginkgoaceae*, *Cupressaceae*, *Taxaceae* and *Podocarpaceae* (*Podocarpus rubens,* identified through transcriptomic data, Table 2); (iii) the B.2.1 group composed of four novel OTUs (9, 27, 28 and 29) identified in *Ginkgo biloba* and several *Cupressaceae* species. The distribution of these three B.2 groups across diverse and deeply divergent gymnosperm lineages, including so-called "living fossils" like *Ginkgo biloba*, suggests that they may be remnants of ancient viral clades that emerged early in gymnosperm evolution, potentially dating back to the origin of seed plants approximately 360 MYA. Their persistence in distinct families today may reflect lineage-specific retention following widespread extinction of related viral forms, or highly constrained co-evolution with long-lived host lineages.

(v) In contrast to the groups described in point (iv), the remaining members of subclade B.2 exhibit a broader distribution across diverse tracheophyte lineages, pointing to an even deeper origin for some of these viral groups and a complex pattern of diversification. The 'B.2.Fernendovirus 2' group represents the most basal lineage within subclade B.2. It is detected in the genomes of two distantly related ferns — *Alsophila spinulosa* (order *Cyatheales*, family *Cyatheaceae*) and *Dipteris conjugata* (order *Gleicheniales*, family *Dipteridaceae*) — whose LCA dates back at least 350 MYA (Lehtonen et al. 2017). Given that Fernendovirus 2 has not been identified outside ferns, we find little support for a recent cross-division host switch into ferns. However, host switches within the fern clade remain a plausible mechanism explaining its current distribution. Importantly, this basal group is sister to the rest of subclade B.2, which further divides into two monophyletic groups.

The first monophyletic group includes the 'B.2.Gymnendovirus 1' group, found in *Cupressaceae* genomes, and the 'B.2.Fernendovirus 1' group, found across several orders of ferns, including the basal order *Ophioglossales*, and lycophytes (e.g., *Lycopodiaceae*). The distribution of these viruses across both ferns and lycophytes — two ancient vascular plant lineages — suggests that this viral clade may have originated near the base of vascular plants (*ca* 444 MYA; (Morris et al. 2018)).

The second monophyletic group contains two major branches. The first branch encompasses 'B.2.Petuvirus', found in ANA-grade angiosperms, magnoliids, and eudicots, and 'B.2.2', spanning highly divergent gymnosperm families, including *Ginkgoaceae*, *Gnetaceae*, *Cupressaceae*, and *Araucariaceae*. This broad host range suggests either a long evolutionary history with multiple ancient host-jump events or a deeply rooted caulimovirid lineage that diversified alongside both gymnosperms and angiosperms. The second branch encompasses 'B.2.3', exclusively detected in the *Araucariaceae*, indicating a narrow and potentially ancient family-specific association, 'B.2.Florendovirus', widespread in both basal angiosperms and mesangiosperm, consistent with extensive diversification within flowering plants, and 'B.2.Gymnendovirus 3-4', confined to the *Pinaceae*, supporting a family-specific viral lineage within conifers.

Together, the distribution of subclade B.2 viruses across ferns, lycophytes, gymnosperms, and angiosperms illustrates a mosaic of ancient cospeciation events, lineage-specific persistence, and probable host-switching, reflecting the deep and complex macroevolutionary history of this subclade within the *Caulimoviridae*.

Discussion

Our study greatly advances the understanding of *Caulimoviridae* diversity by identifying 35 new operational taxonomic units (OTUs), with 34 enriching the known diversity within Clades A and B, and one defining a novel Clade C. These findings confirm that viruses in the *Caulimoviridae* have colonized all major divisions of tracheophytes over evolutionary time and considerably broaden the known host range of this viral family. Clade A OTUs are predominantly associated with angiosperms, although some are also found in gymnosperms and ferns. In contrast, Clade B OTUs span the full spectrum of tracheophytes, including lycophytes, a division previously represented by only a single reverse transcriptase sequence.

The detection of ECRTs in multiple lycophyte genomes and transcripts provides definitive evidence that this plant division is indeed part of the *Caulimoviridae* host range. Clade C, by contrast, appears narrowly confined to the *Araucariaceae* family of gymnosperms. The plant genome dataset available at the time of this study remains incomplete and unevenly distributed across tracheophyte diversity. Reference genomes were unavailable for seven of the eleven fern orders, four of the twelve gymnosperm families, and the basal angiosperm order *Austrobaileyales*. Broader sampling may uncover additional *Caulimoviridae*-host associations and further viral diversity that could revise our understanding of host-virus coevolution. Likewise, expanded transcriptomic and metatranscriptomic data will likely reveal novel footprints of infection or endogenous integration events.

By expanding both the known sequence diversity and host range of the *Caulimoviridae*, our study opens new avenues for investigating the origins and macroevolution of this virus family. The observed host distribution of *Caulimoviridae* across plant lineages does not support a model of strict cospeciation from a single ancestral lineage (Figures 6 and 7) but instead points to parallel co-evolution of distinct caulimovirid lineages with their respective hosts (Figure 8). For instance, we observed Clade A and B ECRTs across ferns, gymnosperms, and angiosperms, as well as several “orphan” lineages, particularly among gymnosperms at the base of subclade B.2.

Several subpatterns consistent with cospeciation support the early emergence of distinct *Caulimoviridae* lineages. Nonetheless, the overall host-virus distribution is complex, with some lineages detected only in a few host species within their expected host range, assuming continuous cospeciation. These gaps may stem partly from limitations in detection: the absence of evidence for host-virus interaction is not evidence of absence. Most episomal caulimovirids have been discovered in symptomatic plants, and ECVs can only be detected if integration occurred in the germline and was retained. Additionally, selective pressure imposed by viral infections may have led to the recurrent evolution of antiviral defenses in host plants, independently across lineages. This stochastic emergence of resistance could have driven many *Caulimoviridae* lineages to extinction, shrinking their modern host range. Some of the ECVs detected in our study may represent extinct viral genera. Importantly, the evolutionary history of tracheophytes includes several mass extinction events, which could have directly impacted their associated viral lineages. Lycophytes, ferns, and gymnosperms underwent a major extinction between the Late Permian (~250 MYA) and Late Triassic (~200 MYA) (Silvestro et al. 2015; Lehtonen et al. 2017; McLoughlin 2021; Spencer et al. 2021). A second mass extinction at the Cretaceous-Paleogene boundary affected all tracheophyte groups (~66 MYA; Silvestro et al., 2015; McLoughlin, 2021). These events likely caused the

disappearance of entire plant lineages and their associated viral lineages, further eroding signals of cospeciation (Figure 7).

By integrating host distributions with viral phylogenetic relationships, we propose that four major caulimovirid lineages can delineated: Clade A, Clade C, and subclades B.1 and B.2 within Clade B. Of these, subclade B.2 stands out for its broadest host range, spanning all four major tracheophyte divisions: angiosperms, gymnosperms, ferns, and lycophytes. Notably, the “B.2 Fernendovirus 1” lineage includes viral elements from both ferns and lycophytes, suggesting that subclade B.2 may have originated in the last common ancestor of tracheophytes (Figure 8).

The detection of elements from B.2 subclade in the genomes and transcriptomes of multiple fern orders, including early-diverging *Ophioglossales*, supports a scenario in which the last common ancestor of euphyllophytes was already infected by B.2-like caulimovirids. In lycophytes, ‘B.2 Fernendovirus 1' elements were found only in *Lycopodiaceae* species, and not in representatives of *Isoetaceae* or *Selaginellaceae*. Given that *Lycopodiaceae* diverged from these families approximately 361-403 MYA (Testo et al. 2018; Xia et al. 2022), we cannot determine whether these viruses were inherited vertically or reflect an ancient host jump from a fern to a *Lycopodiaceae* ancestor. The more parsimonious explanation is that subclade B.2 originated with the euphyllophyte LCA during the Devonian period, at least 402 MYA (Morris et al. 2018). Further sampling of *Isoetaceae* and *Selaginellaceae* — represented here by six genomes and 18 transcriptomes — could provide critical evidence to trace *Caulimoviridae* evolution even deeper, possibly to the LCA of all vascular plants.

Based on our phylogenetic hypothesis, this model implies that the divergence between Clades A and B predates the emergence of subclade B.2, and that the split between Clades A, B, and C occurred even earlier, likely within the LCA of euphyllophytes. This suggests an ancient diversification of the *Caulimoviridae* into three distinct clades, which co-evolved with early euphyllophytes and later gave rise to their modern sub-lineages (Figure 8). The rapid expansion of land plants during the Siluro-Devonian Revolution likely created ecological niches that facilitated the early diversification and establishment of the *Caulimoviridae* across tracheophytes.

Materials and Methods

Plant genome datasets

To identify and characterize endogenous caulimovirid sequences (ECVs), we assembled three complementary plant genomic datasets. The first dataset targeted understudied plant groups to explore novel ECV diversity. It comprised 73 publicly available embryophyte genome assemblies sourced from the China National Genebank, Phytozome, NCBI Genomes, Project Living Fossils, and Fern GCl Assembly (Supp. Table 1). This dataset included seven bryophyte genomes and 66 genomes from basal tracheophytes: nine lycophytes (three *Isoetaceae*, three *Selaginellaceae*, and three *Lycopodiaceae*), 17 ferns (one *Gleicheniales*, two *Cyatheales*, five *Salviniales*, and nine *Polypodiales*), 27 gymnosperms (seven

Cupressaceae, two *Taxaceae*, two *Araucariaceae*, eleven *Pinaceae*, two *Gnetaceae*, one *Cycadaceae*, one *Welwitschiaceae*, and one *Ginkgoaceae*), three ANA-grade angiosperms (two *Nymphaeaceae* and one *Amborellaceae*), one *Chloranthales*, and nine *Magnoliidae* (one *Aristolochiaceae*, two *Magnoliaceae*, two *Calycanthaceae*, and four *Lauraceae*). The second dataset was designed to complement the first by including seven angiosperm genomes where ECVs had already been documented: *Lindenbergia philippensis*, *Citrus sinensis, Rosa chinensis*, *Ricinus communis*, *Nicotiana sylvestris*, *Vitis vinifera*, and *Glycine max* (Supp. Table 4). These additional genomes were supplemented to the 73 from the first plant genome dataset, resulting in a set of 81 genomes from which we collected RT-RH1 sequences for phylogenetic analysis. To ensure comprehensive taxonomic coverage of the ECV distribution analysis across tracheophytes, a third dataset was compiled, including 14 additional angiosperm genomes representing major lineages: *Dioscorea alata*, *Musa balbisiana*, *Brachypodium distachyon*, *Oryza sativa*, *Papaver somniferum, Nelumbo nucifera*, *Beta vulgaris*, *Vaccinium darrowii*, *Rudbeckia hirta*, *Solanum lycopersicum*, *Capsicum annuum, Arabidopsis lyrata*, *Pyrus communis*, and *Fragaria vesca*, while excluding the *Nicotiana sylvestris* genome (Supp. Table 8).

Collection of reference *Caulimoviridae* amino acid RT sequences

A reference library of 106 amino acid reverse transcriptase (aa-RT) sequences homologous to the RT domain of cauliflower mosaic virus (CaMV; GenBank accession number V00141, positions 4,449-5,648) was assembled (Supp. Table 2). It includes 98 aa-RT sequences from members of the family *Caulimoviridae* (Supp. Table 2), comprising:

(i) 16 sequences from episomal viruses, covering all eleven ICTV-recognized genera (Teycheney et al., 2020): *Badnavirus* (n=4), *Caulimovirus* (n=1), *Cavemovirus* (n=1), *Dioscovirus* (n=1), *Petuvirus* (n=1), *Rosadnavirus* (n=1), *Ruflodivirus* (n=1), *Solendovirus* (n=1), *Soymovirus* (n=3), *Tungrovirus* (n=1) and *Vaccinivirus* (n=1);

(ii) 3 sequences from putative caulimovirids: Welwitschia mirabilis virus 1 and 2 (WMV1 and WMV2) (Debat & Bejerman, 2022) and Pinus nigra virus (Rastrojo et al. 2018);

(iii) 79 sequences from ECVs, including: one sequence from the proposed genus Orendovirus, represented by Aegilops tauschii virus (Geering et al. 2010); 11 sequences from the proposed genus Florendovirus (Geering et al. 2014); 56 additional sequences from the proposed genera Gymnendovirus 1–4 (n=22), Xendovirus (n=2), Yendovirus (n=4), Zendovirus (n=1), Petuvirus-like (n=1), and Fernendovirus 1-2 (n=4); 19 sequences from characterized members of genera *Badnavirus* (n=10), *Caulimovirus* (n=1), *Petuvirus* (n=7), and *Soymovirus* (n=1) (Diop et al. 2018); five Beta Endogenous Caulimovirus-like Viruses (BECVs) described by Gong & Han (2018); six sequences from the proposed genus Wendovirus (de Tomás and Vicient 2022).

To root the phylogenetic trees, eight outgroup aa-RT sequences were added: seven from *Metaviridae* and *Retroviridae* curated in the GyDB database (Llorens et al. 2011), and one additional *Metaviridae* RT sequence in the *Anthoceros agrestis* genome, using Caulifinder branch B.

Caulifinder libraries

Caulifinder branch A is distributed with two core libraries. The “Caulimoviridae_ref_genomes” library comprises full-length reference caulimovirid genomic sequences and is used for the

initial detection of ECVs. The “baits” library contains ORFs and conserved protein domains from various members of the order *Ortevirales* and is used to filter out false positives (Vassilieff et al. 2022; Vassilieff et al. 2023). Caulifinder branch B is delivered with two distinct libraries. “Caulimo_RT_probes” is composed of caulimovirid aa-RT sequences and is used for the initial detection of ECRTs. “Tree_RT_set” contains a collection of aa-RT sequences from *Caulimoviridae* and other *Ortervirales*, and serves to filter out false positives and to generate phylogenetic trees of the detected ECRTs.
To improve detection sensitivity and taxonomic resolution, the complete genome sequences of Gymnendovirus 5, 6, 7, and 8, as well as Gnetovirus and Sequoiavirus (Serfraz, 2021; Serfraz, personal communication), were integrated into Caulifinder branch A reference genome library. All translated protein sequences from these genomes that were ≥ 100 aa in length were incorporated into the “baits” library. The corresponding aa-RT sequences were also added to both the “Caulimo_RT_probes” and “Tree_RT_set_library” of Branch B. To further enhance the discrimination of bona fide viral sequences from unrelated elements, we supplemented the “baits” library with 545 RT and pol domain sequences from *Metaviridae* and *Retroviridae*, retrieved from the GyDB database (Llorens et al., 2011). These augmented resources are hereafter referred to as customized libraries.

Transcriptomic data

To identify ECVs and *Caulimoviridae* transcripts, we analyzed four publicly available transcriptomic datasets: (1) NCBI Transcriptome Assemblies from the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/), (2) the “1,000 Plants (1KP)” project (Leebens-Mack *et al*., 2019; https://sites.google.com/a/ualberta.ca/onekp/), (3) a curated set of lycophyte transcriptomes kindly provided by Xia *et al.*, (2022), and (4) fern transcriptomes compiled by Ali *et al.* (2024; https://conekt.sbs.ntu.edu.sg/species/).

Detection and Clustering of endogenous caulimovirid aa-RT domains (aa-ECRTs)

To identify amino acid sequences corresponding to endogenous caulimovirid RT domains (aa-ECRTs), Caulifinder branch B (Vassilieff *et al.*, 2022) was executed on the dataset composed of 73 plant genome assemblies using customized RT libraries. This pipeline was also applied to transcriptomic datasets from the 1 KP project, as well as curated lycophyte and fern transcriptomes. In parallel, a tBLASTn search was performed against the NCBI TSA database using the same custom RT library sequences as queries. Due to the large size of the TSA database, this search was conducted online targeting transcriptomic data from five major plant groups: Bryophytes (taxid:3208), Lycophytes (taxid:1521260), Ferns (taxid:241806), Gymnosperms (taxid:1437180), and Angiosperms (taxid:3398). Representative aa-ECRT sequences retrieved from these sources were subsequently combined with the previously assembled reference library of 106 RT amino acid sequences.

All sequences were aligned using Guidance v2.02 (Sela et al. 2015) with MAFFT v7.511 (Katoh and Standley 2013), applying a column confidence cutoff of 0.98 to ensure alignment robustness. A phylogenetic network was then constructed using SplitsTree v4.14.2 (Huson and Bryant 2006) under default parameters (neighbor-net method) to visualize sequence diversity, detect potential reticulation, and guide clustering thresholds.

Pairwise similarity comparisons were performed with BLASTp (Camacho et al. 2009), and the results were used to cluster sequences with the Markov Cluster Algorithm (MCL) (Van Dongen 2008) with an identity cutoff at 62% and an inflation parameter I = 2. Representative sequences for each cluster were selected based on the following criteria: (i) a minimum length of 150 amino acids, (ii) presence of the conserved 'DD' catalytic motif (aligned to positions 145–146 of the CaMV RT domain), (iii) acceptable alignment quality (Supp. Figure 9). This clustering process yielded a set of candidate sequences, which were manually curated at every step, i.e. at the sequence level first (i) and (ii) and OTU level second (iii) (Supp. Figure 9). Ambiguous sequences were retained only when they represented potentially novel or phylogenetically unique diversity, as confirmed by OTU clustering and phylogenetic placement in the Caulifinder branch B tree (Supp. Figure 9).

Phylogeny of the *Caulimoviridae*

The phylogenetic reconstruction of the *Caulimoviridae* family was based on nucleotide sequences encoding the RT-RH1 domain homologous to that of the cauliflower mosaic virus (CaMV, GenBank accession number V00141, positions 4,449-5,648), the type virus of genus *Caulimovirus* (Teycheney et al. 2020). To investigate the evolutionary relationships of novel *Caulimoviridae* OTUs identified in this study, homologous nucleotide sequences were recovered from plant genome and transcriptome datasets. Representative sequences were compiled from three primary sources:

(i) Plant genome analysis using Caulifinder branch A: the "sequence retriever" module of Caulifinder (Vassilieff et al., 2022) was run with default parameters on a set of 80 embryophyte genomes, comprising the original 73 genomes plus seven additional angiosperm genomes (*Lindenbergia philippensis*, *Citrus sinensis*, *Rosa chinensis*, *Ricinus communis*, *Nicotiana sylvestris*, *Vitis vinifera*, and *Glycine max*). For each genome, consensus sequences were generated when at least five ECV copies with ≥85% nucleotide identity were identified. Only consensus sequences spanning the RT-RH domain were retained. Sequences exhibiting <41% BLASTp identity to any of the 98 reference aa-ECRT sequences were excluded, as this threshold marks the highest observed similarity between a caulimovirid RT (2_Wono) and that of the *Metaviridae* Ty3, ensuring conservative filtering. Sequences were manually selected by retaining consensus with at least three copies covering the RT-RH domain.

(ii) Recovery from genomic copies: for the OTUs for which no or a single consensus could be obtained, unique ECRT loci were extracted using positional information from Caulifinder branch B via the "marker miner" module. For each identified locus, a 2 kbp flanking region was extracted on both 5′ and 3′ ends to ensure complete domain recovery.
(iii) Transcriptomic data: for *Brainea insignis*, for which a complete genome sequence was not available, an RT-RH domain-containing sequence was retrieved from transcriptomic data (NCBI, accession: GFUE01038349.1).

To assign sequences to OTUs, a BLASTx search (e-value ≤ $1e^{-5}$) was performed between them and the 361 aa-RT sequences used to define the OTU. Each of the retrieved sequences was assigned to its top hit OTU if it shared ≥62% amino acid identity over a minimum of 150 residues. Nucleotide sequences less than 1,000 bp in length or those generating long branches in preliminary phylogenetic trees were discarded. Up to four representative sequences were selected per OTU. Sequences were aligned using MAFFT v7.511 (Katoh &

Standley, 2013) with the G-INS-i strategy and 1,000 iterative refinements. Phylogenetic trees were inferred using two complementary approaches:
(i) Maximum Likelihood (ML): Trees were constructed using IQ-TREE v2.1.2 (Minh et al. 2020) with 1,000 ultrafast bootstrap replicates. The best-fit substitution model (GTR+F+I+G4) was selected using ModelFinder (Kalyaanamoorthy et al. 2017) based on the Bayesian Information Criterion (BIC).

(ii) Bayesian Inference: Bayesian phylogenies were generated with MrBayes v3.2.7 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003; Ronquist et al. 2012), under the GTR+I+Γ model. Analyses were run for 2,000,000 generations across four Markov chains, sampling every 500 generations. Convergence was assessed using Tracer v1.7.2 (Rambaut et al. 2018), and a 10% burn-in was applied before computing the consensus tree and posterior probabilities.

Quantification and classification of nt-ECRTs across plant genomes

To quantify the prevalence of nt-ECRTs across our plant genome dataset, each nt-ECRTs identified by Caulifinder branch B (Vassilieff et al., 2022) were queried against the representative aa-ECRT library used to define OTUs, using BLASTx (e-value $\leq 1\times10^{-5}$). Assignments to OTUs were made according to each nt-ECRT's top BLASTx hit, requiring ≥ 65% amino acid identity over at least 150 residues (the length of the smallest reference aa-RT domain).

We adopted a 65% identity cutoff, more stringent than the 62% threshold used during the clustering of aa-RTs into OTUs, to minimize misclassification and enhance specificity in our quantification of nt-ECRT abundance across diverse plant genomes.

Classification of RT sequences detected in transcriptomic data

The classification of nt-ECRTs recovered from transcriptomic datasets was conducted using BLASTx against the set of aa-RTs sequences used to define OTUs..
Each transcript-derived nt-ECRT was assigned to an OTU based on its best BLASTx hit, provided the alignment met a minimum identity threshold of 62% across at least 150 amino acids, matching the criteria used during OTU clustering. This consistent framework allowed for the reliable classification of transcript-derived sequences within the established *Caulimoviridae* diversity landscape.

Viral genome assembly from ECVs

The WolV1 genome was assembled from ECVs found in the *W. nobilis* genome as defined previously (Geering et al. 2014). Briefly, the 5' and 3' genomic positions of nt-ECRTs assigned to OTU 19 were extended by 10 kb, and the sequences of the corresponding loci were extracted. These sequences were compared between them using an all-against-all BLASTn search to identify high-scoring sequences of different loci containing identical or near-identical sequences. Fragments of virus sequence were assembled 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.

Phylogeny of *Ortervirales*

To explore the evolutionary relationships within the order *Ortervirales*, a representative set of 52 amino acid sequences corresponding to the RT domains of viruses from its five recognized families (*Belpaoviridae*, *Pseudoviridae*, *Retroviridae*, *Metaviridae*, and *Caulimoviridae*) was compiled (Llorens *et al.*, 2011) (Krupovic *et al.*, 2018) (Supp. Table 7).
Sequences were aligned using MAFFT v7.511 (Katoh & Standley, 2013) with the G-INS-i global alignment strategy and up to 1,000 iterations. An ML phylogeny was reconstructed using IQ-TREE v1.6.12 (Minh *et al.*, 2020), based on 1,000 bootstrap replicates for branch support. The rtREV+G4 substitution model was selected automatically using the Bayesian Information Criterion (BIC) via ModelFinder (Kalyaanamoorthy *et al.*, 2017).

Phylogeny of the 30K movement protein (MP) family

The phylogenetic analysis of the 30K movement protein (MP) family was conducted using amino acid sequences derived from *Caulimoviridae* and homologs from other plant virus families. MP domains from caulimovirid nucleotide sequences were initially identified by annotating protein domains in ECV loci and consensus sequences, and reference genomes using rpsblastn (Camacho et al. 2009) with the CDD database (e-value ≤ 1) (Wang, Chitsaz, Myra K Derbyshire, et al. 2023). Ambiguous matches were further validated against the Pfam database using Hmmer3 (Sonnhammer et al. 1997; Mistry et al. 2021). A total of 46 sequences representing the diversity of clades A, B, and C OTUs were selected and translated, of which ten belonged to clade C. MP domains retrieved from plant genomic loci were classified following the OTU assignment of their most proximal ECRT within a range of 5 kbp upstream and downstream, a rationale supported by the distances separating the RT and MP domains in ICTV-characterized genomes (Teycheney et al. 2020).

To contextualize caulimovirid MP diversity within broader viral evolution, an additional 286 MP sequences were sourced from Butkovic et al. (Butkovic et al. 2023), spanning 14 plant virus families: *Alphaflexiviridae*, *Aspiviridae*, *Betaflexiviridae*, *Bromoviridae*, *Botourmiaviridae*, *Fimoviridae*, *Geminiviridae*, *Kitaviridae*, *Mayoviridae*, *Phenuiviridae*, *Rhabdoviridae*, *Secoviridae*, *Tospoviridae,* and *Virgaviridae*.

In total, 332 MP sequences were aligned using MAFFT 7.511 (Katoh & Standley 2013) with the G-INS-I global alignment strategy and up to 1000 iterations. The alignment was trimmed with trimAl 1.4 using the -gt 0.5 option to remove poorly aligned columns. An ML phylogeny was inferred using IQ-TREE 2.1.4 (Minh et al. 2020), with 1000 Sh-aLRT and 1000 ultrafast bootstrap replicates. The LG+F+G4 substitution model was selected based on BIC via ModelFinder (Kalyaanamoorthy et al. 2017). The resulting tree was midpoint-rooted to facilitate the interpretation of clade relationships.

Search for Clade C RTs in *Agathis* dammara raw sequencing data

To investigate the potential presence of OTU 19 (Clade C) RT sequences within the *Agathis* genus, we analyzed Illumina short-reads whole-genome sequencing data from *Agathis dammara*, available through the NCBI Sequence Read Archive (SRA) (SRR15616211). This dataset, originally generated for a mitogenome sequencing project (NCBI bioproject 757934), comprises 3.4 Gbp of sequence data, representing approximately 0.12% of the estimated 27 Gbp *A. dammara* genome size (Zonneveld 2012). We used the OTU 19 rep-aa-ECRT rep sequence from *W. nobilis* as a query in tblastn searches conducted via the NCBI web interface

(Camacho et al. 2009; Sayers et al. 2025). To improve the likelihood of detecting homologous hits across the entire RT domain, the probe sequence was segmented into consecutive 50 aa fragments. Each fragment was used to query the short reads dataset independently. For every chunk, the ten highest-scoring reads were retrieved and pooled for assembly using the CAP3 program (Huang and Madan 1999). This approach yielded a 529 nt contig, which was translated into a 175 aa RT sequence. We designated this sequence as the OTU 19 representative aa-ECRT from *A. dammara* following validation through BLASTx against the curated set of 361 reference RT sequences used in OTU classification.

Acknowledgements

The authors thank Christophe Plomion (INRAE) for providing access to the *Cupressus sempervirens* genome sequence ahead of publication. The authors thank Anamarija Butkovic for providing access to plant viral 30K MP sequences and Thiery Candresse for critical reading of the manuscript. HV was funded by a joint INRAe-CIRAD PhD fellowship. PYT is funded by the European Regional Development Fund (ERDF, contract REU005756) and the Conseil Régional de La Réunion through the project “Dispositif de partenariat en santé et biodiversité”. URGI is supported by Saclay Plant Sciences-SPS (ANR-17-EUR-0007) and the PlantBioinfoPF platform.

References

Ali ZM, Tan QW, Lim PK, Chen H, Pfeifer L, Julca I, Lee JM, Classen B, Vries S de, Vries J de, et al. 2024. Comparative transcriptomics in ferns reveals key innovations and divergent evolution of the secondary cell walls. *bioRxiv*:2024.08.27.609851.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* 57:289–300.

Benton MJ, Wilf P, Sauquet H. 2022. The Angiosperm Terrestrial Revolution and the origins of modern biodiversity. *New Phytol.* 233:2017–2035.

Blaxter M, Mann J, Chapman T, Thomas F, Whitton C, Floyd R, Abebe E. 2005. Defining operational taxonomic units using DNA barcode data. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 360:1935–1943.

Butkovic A, Dolja VV, Koonin EV, Krupovic M. 2023. Plant virus movement proteins originated from jelly-roll capsid proteins. *PLoS Biol.* 21:e3002157.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

Chatterjee S, Goswami A, Scotese CR. 2013. The longest voyage: Tectonic, magmatic, and paleoclimatic evolution of the Indian plate during its northward flight from Gondwana to Asia. *Gondwana Res.* 23:238–267.

Dallot S, Acuña P, Rivera C, Ramírez P, Côte F, Lockhart BE, Caruana ML. 2001. Evidence

that the proliferation stage of micropropagation procedure is determinant in the expression of banana streak virus integrated into the genome of the FHIA 21 hybrid (Musa AAAB). *Arch. Virol.* 146:2179–2190.

Debat H, Bejerman N. 2022. A glimpse into the DNA virome of the unique “living fossil” Welwitschia mirabilis. *Gene* 843:146806.

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.

Donoghue PCJ, Harrison CJ, Paps J, Schneider H. 2021. The evolutionary emergence of land plants. *Curr. Biol.* 31:R1281–R1298.

Geering ADW, Maumus F, Copetti D, Choisne N, Zwickl DJ, Zytnicki M, McTaggart AR, Scalabrin S, Vezzulli S, Wing RA, et al. 2014. Endogenous florendoviruses are major components of plant genomes and hallmarks of virus evolution. *Nat. Commun.* 5:5269.

Geering ADW, Scharaschkin T, Teycheney P-Y. 2010. The classification and nomenclature of endogenous viruses of the family Caulimoviridae. *Arch. Virol.* 155:123–131.

Gong Z, Han G-Z. 2018. Euphyllophyte Paleoviruses Illuminate Hidden Diversity and Macroevolutionary Mode of Caulimoviridae. *J. Virol.* 92:e02043-17.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinforma. Oxf. Engl.* 29:1072–1075.

Harper G, Osuji JO, Heslop-Harrison JS, Hull R. 1999. Integration of banana streak badnavirus into the Musa genome: molecular and cytogenetic evidence. *Virology* 255:207–213.

Herendeen PS, Friis EM, Pedersen KR, Crane PR. 2017. Palaeobotanical redux: revisiting the age of the angiosperms. *Nat. Plants* 3:1–8.

Huang X, Madan A. 1999. CAP3: A DNA sequence assembly program. *Genome Res.* 9:868–877.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754–755.

Huson DH, Bryant D. 2006. Application of Phylogenetic Networks in Evolutionary Studies. *Mol. Biol. Evol.* 23:254–267.

Jakowitsch J, Mette MF, van der Winden J, Matzke MA, Matzke AJM. 1999. Integrated pararetroviral sequences define a unique class of dispersed repetitive DNA in plants. *Proc. Natl. Acad. Sci.* 96:13241–13246.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14:587–589.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* 30:772–780.

Krupovic M, Blomberg J, Coffin JM, Dasgupta I, Fan H, Geering AD, Gifford R, Harrach B,

Hull R, Johnson W, et al. 2018. Ortervirales: New Virus Order Unifying Five Families of Reverse-Transcribing Viruses. *J. Virol.* 92:e00515-18.

Kumar S, Suleski M, Craig JM, Kasprowicz AE, Sanderford M, Li M, Stecher G, Hedges SB. 2022. TimeTree 5: An Expanded Resource for Species Divergence Times. *Mol. Biol. Evol.* 39:msac174.

Leebens-Mack JH, Barker MS, Carpenter EJ, Deyholos MK, Gitzendanner MA, Graham SW, Grosse I, Li Z, Melkonian M, Mirarab S, et al. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574:679–685.

Lehtonen S, Silvestro D, Karger DN, Scotese C, Tuomisto H, Kessler M, Peña C, Wahlberg N, Antonelli A. 2017. Environmentally driven extinction and opportunistic origination explain fern diversification patterns. *Sci. Rep.* 7:4831.

Llorens C, Futami R, Covelli L, Domínguez-Escribá L, Viu JM, Tamarit D, Aguilar-Rodríguez J, Vicente-Ripolles M, Fuster G, Bernet GP, et al. 2011. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Res.* 39:D70-74.

Lockhart BE, Menke J, Dahal G, Olszewski NE. 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.

McCourt RM, Lewis LA, Strother PK, Delwiche CF, Wickett NJ, de Vries J, Bowman JL. 2023. Green land: Multiple perspectives on green algal evolution and the earliest land plants. *Am. J. Bot.* 110:e16175.

McLoughlin S. 2021. Gymnosperms. In: Alderton D, Elias SA, editors. Encyclopedia of Geology (Second Edition). Oxford: Academic Press. p. 476–500. Available from: https://www.sciencedirect.com/science/article/pii/B9780081029084000680

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* 37:1530–1534.

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, et al. 2021. Pfam: The protein families database in 2021. *Nucleic Acids Res.* 49:D412–D419.

Morris JL, Puttick MN, Clark JW, Edwards D, Kenrick P, Pressel S, Wellman CH, Yang Z, Schneider H, Donoghue PCJ. 2018. The timescale of early land plant evolution. *Proc. Natl. Acad. Sci. U. S. A.* 115:E2274–E2283.

Mushegian AR, Elena SF. 2015. Evolution of plant virus movement proteins from the 30K superfamily and of their homologs integrated in plant genomes. *Virology* 476:304–315.

Ndowora T, Dahal G, LaFleur D, Harper G, Hull R, Olszewski NE, Lockhart B. 1999. Evidence that badnavirus infection in Musa can originate from integrated pararetroviral sequences. *Virology* 255:214–220.

Nitta JH, Schuettpelz E, Ramírez-Barahona S, Iwasaki W. 2022. An open and continuously updated fern tree of life. *Front. Plant Sci.* 13:909768.

Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. 2018. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst. Biol.* 67:901–904.

Rastrojo A, Núñez A, Moreno DA, Alcamí A. 2018. A New Putative Caulimoviridae Genus Discovered through Air Metagenomics. *Microbiol. Resour. Announc.* 7:e00955-18.

Reguero MA, Gelfo JN, López GM, Bond M, Abello A, Santillana SN, Marenssi SA. 2014. Final Gondwana breakup: The Paleogene South American native ungulates and the demise of the South America–Antarctica land connection. *Glob. Planet. Change* 123:400–413.

Richert-Pöggeler KR, Noreen F, Schwarzacher T, Harper G, Hohn T. 2003. Induction of infectious petunia vein clearing (pararetro) virus from endogenous provirus in petunia. *EMBO J.* 22:4836.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61:539–542.

Sayers EW, Beck J, Bolton EE, Brister JR, Chan J, Connor R, Feldgarden M, Fine AM, Funk K, Hoffman J, et al. 2025. Database resources of the National Center for Biotechnology Information in 2025. *Nucleic Acids Res.* 53:D20–D29.

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–W14.

Silvestro D, Cascales-Miñana B, Bacon CD, Antonelli A. 2015. Revisiting the origin and diversification of vascular plants through a comprehensive Bayesian analysis of the fossil record. *New Phytol.* 207:425.

Sonnhammer EL, Eddy SR, Durbin R. 1997. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* 28:405–420.

Spencer V, Nemec Venza Z, Harrison CJ. 2021. What can lycophytes teach us about plant evolution and development? Modern perspectives on an ancient lineage. *Evol. Dev.* 23:174–196.

Stull GW, Qu X-J, Parins-Fukuchi C, Yang Y-Y, Yang J-B, Yang Z-Y, Hu Y, Ma H, Soltis PS, Soltis DE, et al. 2021. Gene duplications and phylogenomic conflict underlie major pulses of phenotypic evolution in gymnosperms. *Nat. Plants* 7:1015–1025.

Testo W, Field A, Barrington D. 2018. Overcoming among-lineage rate heterogeneity to infer the divergence times and biogeography of the clubmoss family Lycopodiaceae. *J. Biogeogr.* 45:1929–1941.

Teycheney P-Y, Geering ADW, Dasgupta I, Hull R, Kreuze JF, Lockhart B, Muller E, Olszewski N, Pappu H, Pooggin MM, et al. 2020. ICTV Virus Taxonomy Profile: Caulimoviridae. *J. Gen. Virol.* 101:1025–1026.

The Angiosperm Phylogeny Group. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot. J. Linn.*

Soc. 181:1–20.

de Tomás C, Vicient CM. 2022. Genome-wide identification of Reverse Transcriptase domains of recently inserted endogenous plant pararetrovirus (Caulimoviridae). *Front. Plant Sci.* 13:1011565.

Van Dongen S. 2008. Graph Clustering Via a Discrete Uncoupling Process. *SIAM J. Matrix Anal. Appl.* 30:121–141.

Vassilieff H, Geering ADW, Choisne N, Teycheney P-Y, Maumus F. 2023. Endogenous Caulimovirids: Fossils, Zombies, and Living in Plant Genomes. *Biomolecules* 13:1069.

Vassilieff H, Haddad S, Jamilloux V, Choisne N, Sharma V, Giraud D, Wan M, Serfraz S, Geering ADW, Teycheney P-Y, et al. 2022. CAULIFINDER: a pipeline for the automated detection and annotation of caulimovirid endogenous viral elements in plant genomes. *Mob. DNA* 13:31.

Vassilieff H, Maumus F, Haddad S, Jamilloux V, Teycheney P-Y. 2023. Caulifinder banks. Available from: https://entrepot.recherche.data.gouv.fr/dataset.xhtml?persistentId=doi:10.57745/ADFNMB

Wang J, Chitsaz F, Derbyshire Myra K, Gonzales NR, Gwadz M, Lu S, Marchler GH, Song JS, Thanki N, Yamashita RA, et al. 2023. The conserved domain database in 2023. *Nucleic Acids Res.* 51:D384–D388.

Wang J, Chitsaz F, Derbyshire Myra K., Gonzales NR, Gwadz M, Lu S, Marchler GH, Song JS, Thanki N, Yamashita RA, et al. 2023. The conserved domain database in 2023. *Nucleic Acids Res.* 51:D384–D388.

Xia Z-Q, Wei Z-Y, Shen H, Shu J-P, Wang T, Gu Y-F, Jaisi A, Yan Y-H. 2022. Lycophyte transcriptomes reveal two whole-genome duplications in Lycopodiaceae: Insights into the polyploidization of Phlegmariurus. *Plant Divers.* 44:262–270.

Zonneveld BJM. 2012. Genome sizes of all 19 Araucaria species are correlated with their geographical distribution. *Plant Syst. Evol.* 298:1249–1255.

Figures

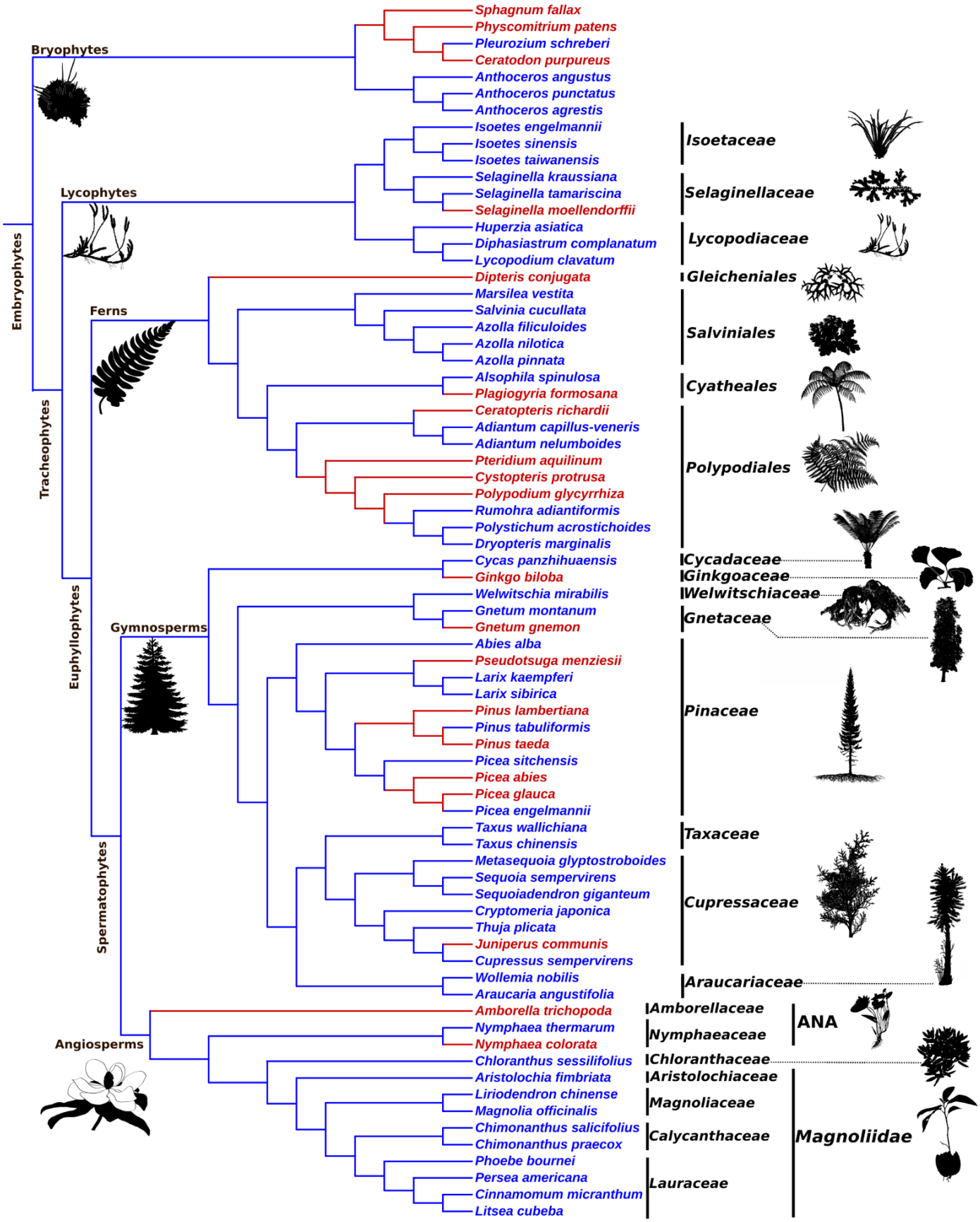

Embryophytes
Bryophytes
Tracheophytes
Lycophytes
Euphyllophytes
Ferns
Spermatophytes
Gymnosperms
Angiosperms
Sphagnum fallax
Physcomitrium patens
Pleurozium schreberi
Ceratodon purpureus
Anthoceros angustus
Anthoceros punctatus
Anthoceros agrestis
Isoetes engelmannii
Isoetes sinensis
Isoetes taiwanensis
Selaginella kraussiana
Selaginella tamariscina
Selaginella moellendorffii
Huperzia asiatica
Diphasiastrum complanatum
Lycopodium clavatum
Dipteris conjugata
Marsilea vestita
Salvinia cucullata
Azolla filiculoides
Azolla nilotica
Azolla pinnata
Alsophila spinulosa
Plagiogyria formosana
Ceratopteris richardii
Adiantum capillus-veneris
Adiantum nelumboides
Pteridium aquilinum
Cystopteris protrusa
Polypodium glycyrrhiza
Rumohra adiantiformis
Polystichum acrostichoides
Dryopteris marginalis
Cycas panzhihuaensis
Ginkgo biloba
Welwitschia mirabilis
Gnetum montanum
Gnetum gnemon
Abies alba
Pseudotsuga menziesii
Larix kaempferi
Larix sibirica
Pinus lambertiana
Pinus tabuliformis
Pinus taeda
Picea sitchensis
Picea abies
Picea glauca
Picea engelmannii
Taxus wallichiana
Taxus chinensis
Metasequoia glyptostroboides
Sequoia sempervirens
Sequoiadendron giganteum
Cryptomeria japonica
Thuja plicata
Juniperus communis
Cupressus sempervirens
Wollemia nobilis
Araucaria angustifolia
Amborella trichopoda
Nymphaea thermarum
Nymphaea colorata
Chloranthus sessilifolius
Aristolochia fimbriata
Liriodendron chinense
Magnolia officinalis
Chimonanthus salicifolius
Chimonanthus praecox
Phoebe bournei
Persea americana
Cinnamomum micranthum
Litsea cubeba
Isoetaceae
Selaginellaceae
Lycopodiaceae
Gleicheniales
Salviniales
Cyatheales
Polypodiales
Cycadaceae
Ginkgoaceae
Welwitschiaceae
Gnetaceae
Pinaceae
Taxaceae
Cupressaceae
Araucariaceae
Amborellaceae
Nymphaeaceae
ANA
Chloranthaceae
Aristolochiaceae
Magnoliaceae
Calycanthaceae
Lauraceae
Magnoliidae

Figure 1: Datasets assembly for searching ECVs

The cladogram displays the species names of the 73 plant genome sequences used to detect new ECVs. Species previously examined for ECVs detection are highlighted in red (Geering et al. 2010; Geering et al. 2014; Mushegian and Elena 2015; Diop et al. 2018; Gong and Han 2018; de Tomás and Vicient 2022), while species investigated for the first time are highlighted in blue. The cladogram reflects major divisions of land plants (Embryophytes), following established phylogenetic frameworks (The Angiosperm Phylogeny Group 2016; Lehtonen et al. 2017; Morris et al. 2018; Donoghue et al. 2021; Stull et al. 2021; Benton et al. 2022; Nitta et al. 2022; McCourt et al. 2023). These plant divisions are labeled at the corresponding nodes. Taxonomic groupings are shown on the right, with granularity adapted to lineage diversity: family level for lycophytes, gymnosperms, and angiosperms; order level for ferns; ANA group for the paraphyletic clades comprising *Amborellales*, *Nymphaeales,* and *Austrobaileyales*; and subclass level for *Magnoliidae.* Illustrative plant silhouettes were obtained from the PhyloPic repository (https://www.phylopic.org/).

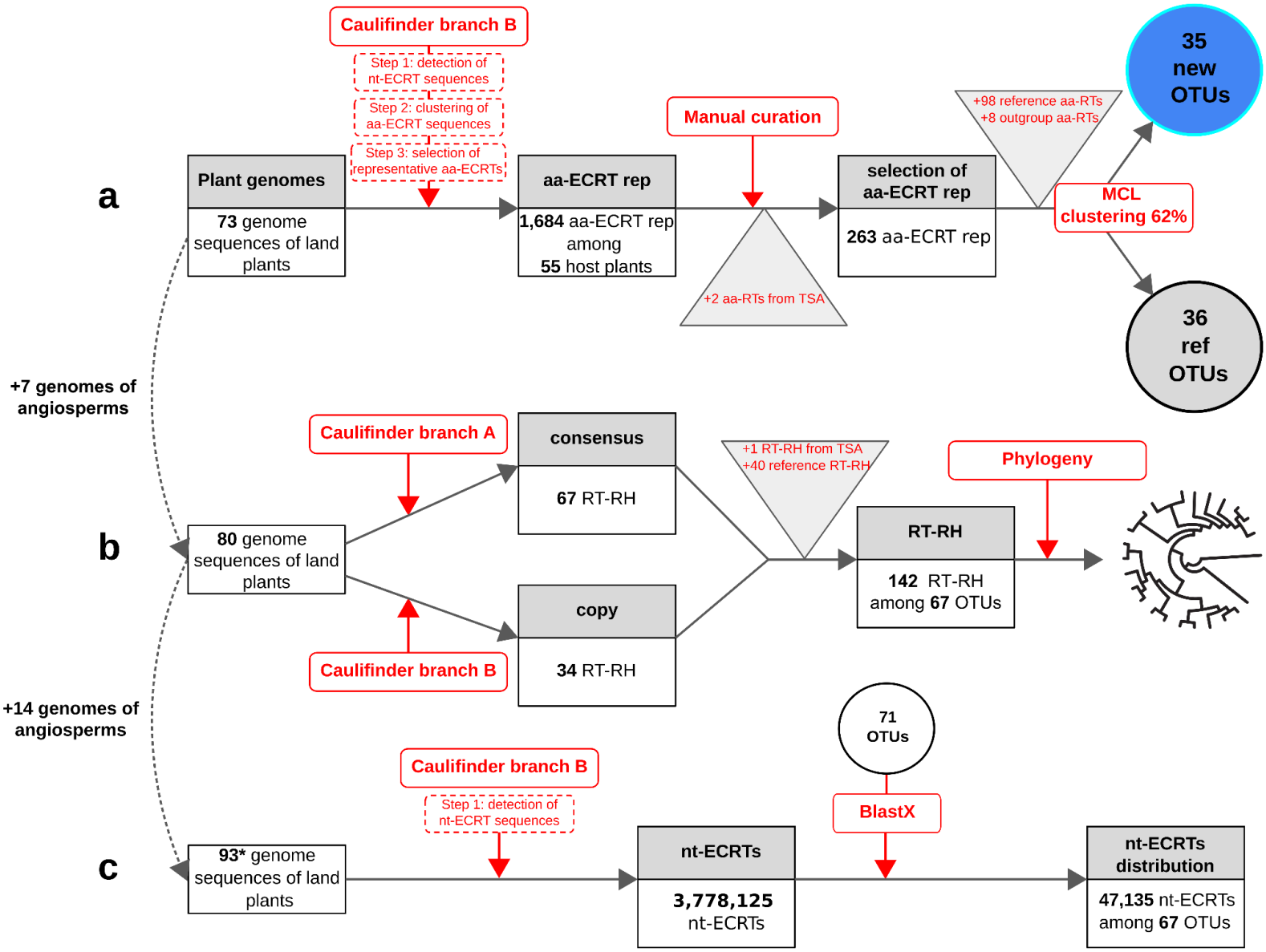
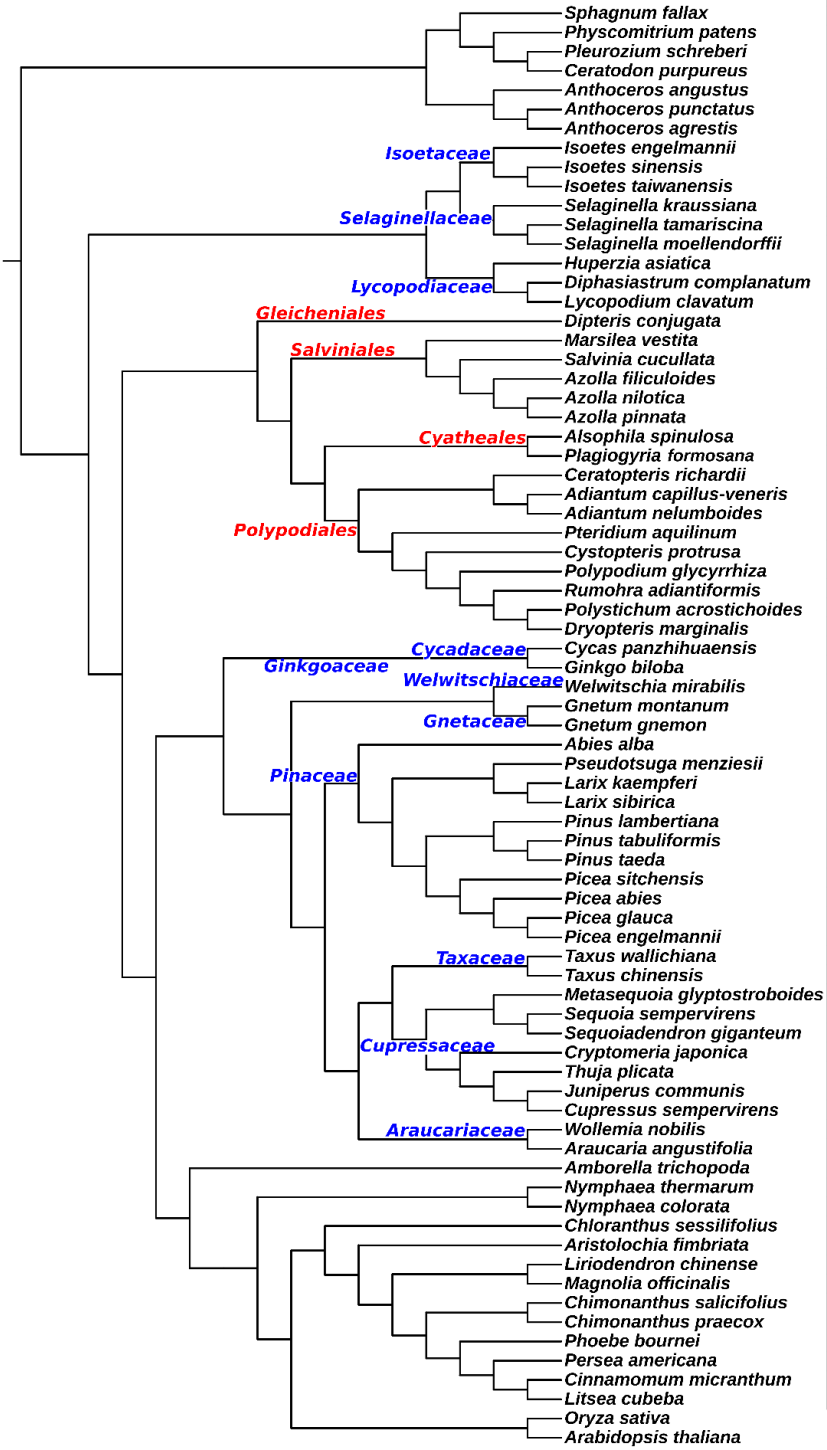


Figure 2: Detection, selection and distribution of novel ECVs using Caulifinder pipeline

This figure outlines the workflows used to identify and classify novel endogenous caulimovirid elements (ECVs) using the Caulifinder pipeline (Vassilieff *et al.*, 2022). a) Selection of representative amino acid reverse transcriptase domain sequences (aa-ECRTs) from Caulifinder branch B output for OTU definition. A total of 263 representative aa-ECRTs (aa-ECRT rep) were selected. Host plant species identified solely through transcriptomic data (*Brainea insignis* and *Pteris vittata*) are marked with an asterisk (*). b) Reconstruction of nucleotide consensus sequences and retrieval of RT-RH domain copies using output from Caulifinder branches A and Caulifinder branch B (Vassilieff et al. 2022). This resulted in 102 RT-RH sequences used for phylogenetic analysis. c) Detection of nucleotide RT domain sequences (nt-ECRTs) via Caulifinder branch B, followed by their classification in predefined OTUs. The asterisk (*) refers to the removal of the *Nicotiana sylvestris* genome from the last dataset.



Tag	aa-ECRT	aa-ECRT rep	Ref OTU	New OTU	Size (Mbp)	N50 (Kbp)
Sfalla	0	0	0	0	395	21,139
Phype	0	0	0	0	472	17,436
Pleusch	0	0	0	0	220	0,154
Cepur	0	0	0	0	363	25,16
Anthan	0	0	0	0	119	1,092
Anpunc	0	0	0	0	132	1,716
Anthag	0	0	0	0	117	17,269
Ieng	0	0	0	0	641	4
Isine	0	0	0	0	2,132	86,664
Itaiw	0	0	0	0	1,658	14,867
Selkra	0	0	0	0	115	3
Selta	0	0	0	0	301	408
Selmo	0	0	0	0	213	1,750
Hupas	73	8	1	0	7,94	57,017
Dicom	3	2	1	0	1,745	59,474
Lycla	34	13	1	0	2,305	1,013
Dico	13	2	1	0	232	0,637
Mvestita	0	0	0	0	1,033	2,973
Sacu	0	0	0	0	232	72
Azfi	0	0	0	0	623	27
Aznil	0	0	0	0	5	965
Azpi	1	1	1	0	167	30
Alspin	779	22	0	2	6,234	92,484
Plafo	1	1	1	0	46	0,922
Cerri	0	0	0	0	7,462	181,963
Adica	64	28	0	1	4,823	159,32
Adine	53	6	0	1	6,272	1,374
Ptaq	1	1	1	0	620	0,708
Cypo	4	4	1	0	43	0,842
Pogly	1	1	1	0	53	0,771
Rhadi	28	14	2	1	2,307	1
Polac	14	8	2	0	451	0,941
Dryma	37	17	1	2	1,113	2
Cypan	86	32	1	2	10,483	952,06
Gbilo	203	28	3	2	9,555	724,989
Welmi	34	21	2	0	6,867	359,756
Gnmont	401	82	2	1	4,072	475
Gnegne	346	82	3	1	3,257	1,862
aalba5	562	62	4	1	18,167	26
Psemen	963	50	5	0	14,673	387
Lakae	1314	51	4	1	13,188	436
Lasib	363	59	5	1	12,342	10
Pilam	1564	186	6	0	21,300	2,917
Pitab	3352	75	6	0	25,423	2,107,675
Ptaeda	2171	75	6	0	22,104	111
Pisit	969	36	6	0	18,225	10
Pabies	511	65	4	0	12,301	3,183
Pglau	1111	56	6	0	27,14	50
Picen	1130	51	6	0	24,944	404
Tawal	305	67	0	2	11,119	987,156
Tachi	462	49	0	1	10,238	903,737
Megly	3	3	0	2	8,686	10,628
Sesem	874	46	1	3	26,55	61,328
Segig	44	20	1	3	8,126	690,55
Cryja	31	15	1	4	9,047	754,841
Thupli	72	11	1	3	9,096	2,591
Juco	31	27	2	3	11,041	836
Cusem	76	12	0	3	9,986	919,939
Wono	141	37	0	5	11,596	8,829
Aran	12	12	0	3	18,463	4,935
AmTr	589	96	3	1	706	4,927
Nyther	62	14	3	2	368	275
Nycol	47	14	3	2	408	25,517
Chlose	139	22	3	0	2,169	145,775
Arifi	75	15	1	0	258	33,25
Lichin	84	8	2	0	1,742	2,204
Magof	62	13	2	1	1,685	76,619
Chisa	316	15	2	1	853	75,743
Chipr	476	17	2	1	737	63,914
Phobo	51	8	1	3	989	2,053
Persam	39	8	1	2	1,057	47,368
Cinmic	70	9	2	2	73	50,351
Licub	20	6	2	2	1,326	113,312

Bryophytes

Lycophytes

Ferns

Gymnosperms

ANA

Chloranthales

Magnoliidae

Figure 3: Distribution of aa-ECRTs, aa-ECRTs rep, and OTUs across plant genomes
The cladogram on the left illustrates the phylogenetic relationships among the 73 plant species analyzed. Major plant lineages are annotated, including bryophytes, lycophytes, ferns, gymnosperms, and angiosperms, with the latter subdivided into ANA grade, *Chloranthales*, and *Magnoliidae*. Taxonomic groups are labeled at the family level (blue) and order level (red). For each plant species, the following metrics are presented: species label, total number of detected aa-ECRTs, number of representative aa-ECRTs (aa-ECRTs rep), number of reference OTUs detected (OTU ref), number of newly identified OTUs (New OTU), genome size (in Mbp), and N50 contig value (in kbp), as computed by QUAST (Gurevich et al. 2013).

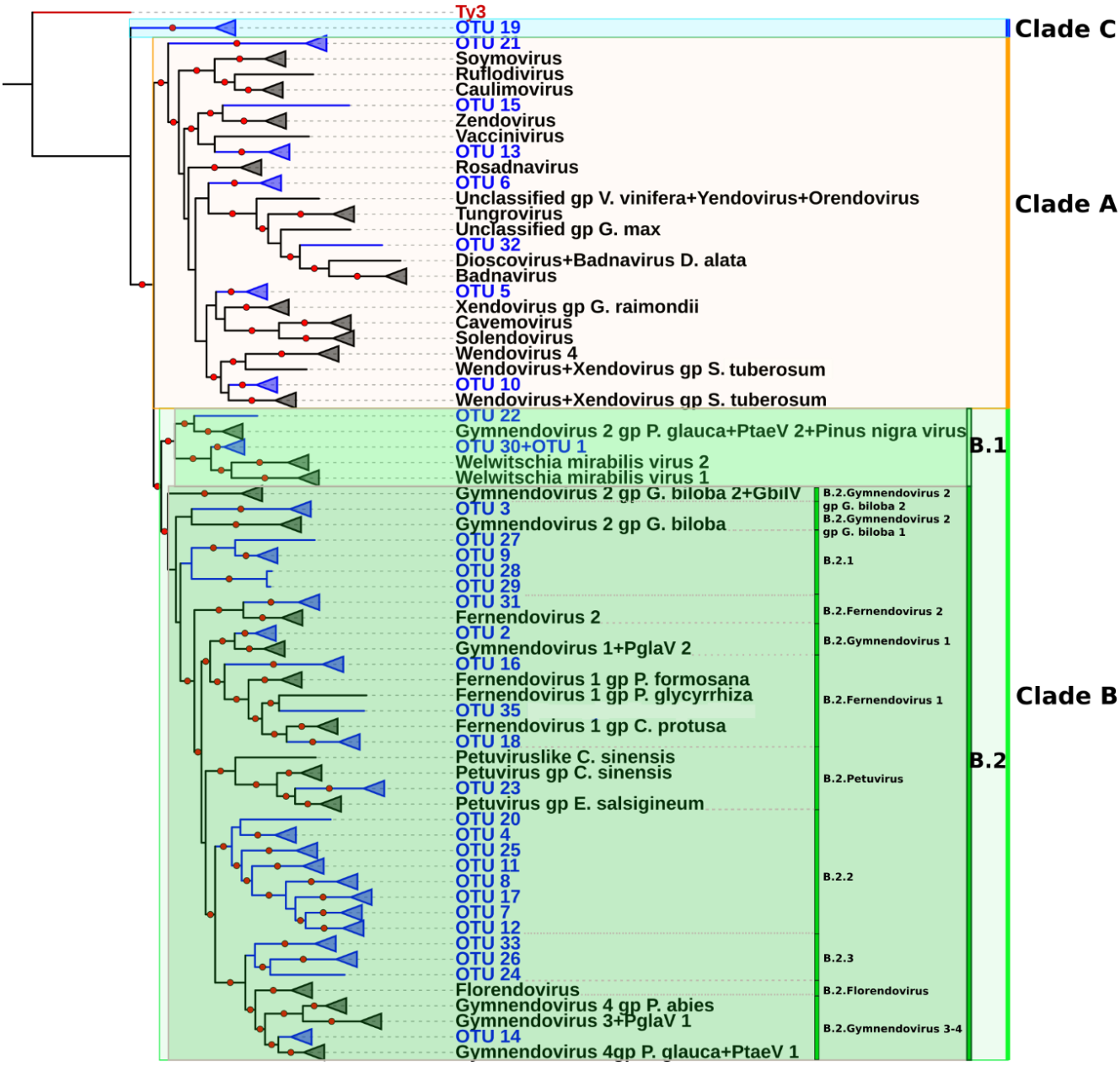


Figure 4: Bayesian phylogeny of *Caulimoviridae*, with collapsed OTU branches
This phylogenetic tree comprises 142 nucleotide sequences corresponding to the RT-RH domain of *Caulimoviridae*. Bayesian inference was conducted using MrBayes, and the tree is rooted with a *Metaviridae* Ty3 sequence (highlighted in red). Red circles indicate nodes with posterior probability supports ≥ 0.95. Triangles represent collapsed monophyletic groups corresponding to individual OTUs. Polyphyletic OTUs are shown in multiple locations; "Wendovirus+Xendovirus gp S. tuberosum" appears twice. OTUs newly identified in this study are highlighted in blue, while previously defined (reference) OTUs are shown in black. The tree is divided into three major clades, color-coded as follows: Clade A (orange), Clade B (green), and Clade C (blue).

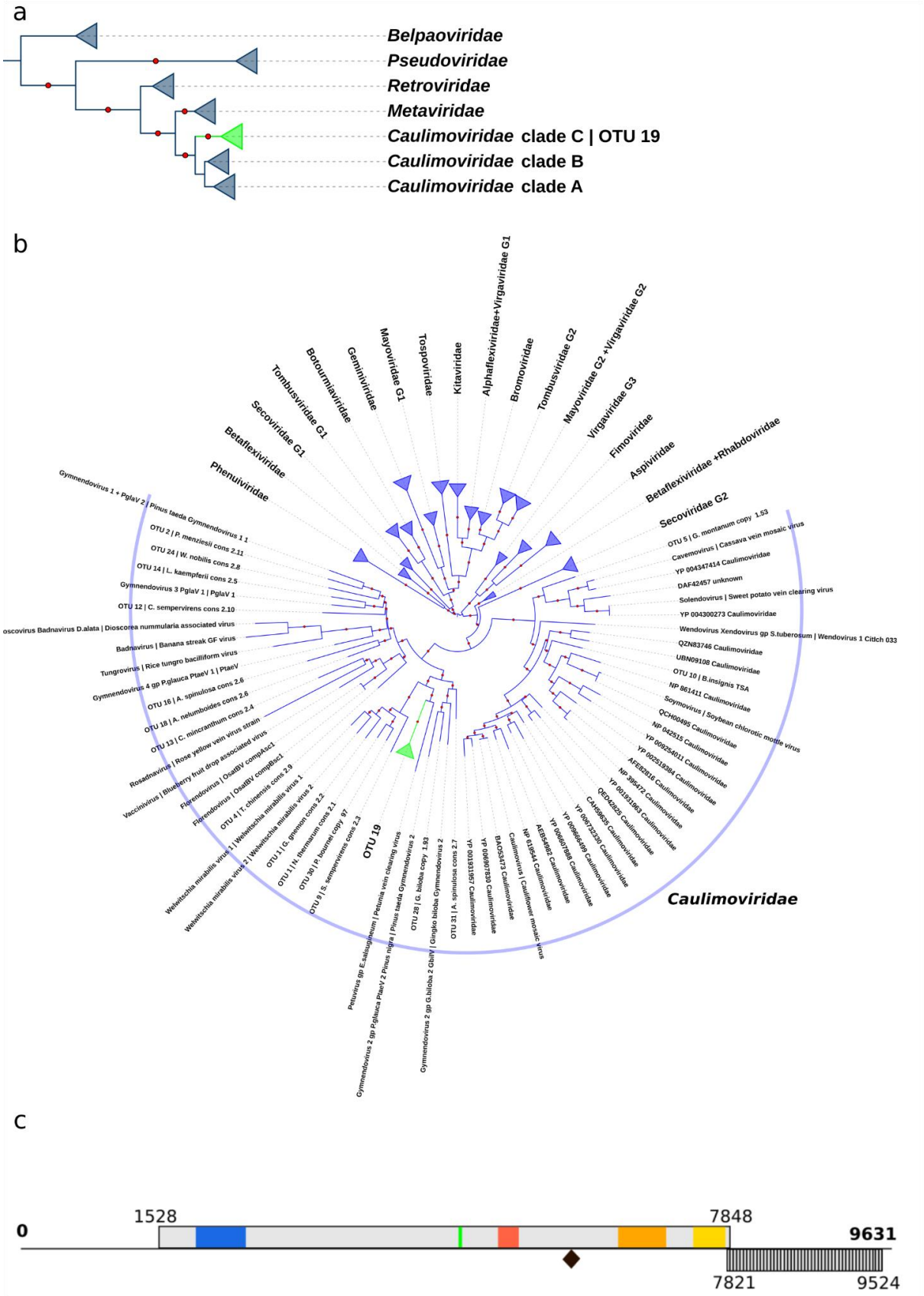

a
Belpaoviridae
Pseudoviridae
Retroviridae
Metaviridae
Caulimoviridae clade C | OTU 19
Caulimoviridae clade B
Caulimoviridae clade A
b
Alphaflexiviridae+Virgaviridae G1
Bromoviridae
Tombusviridae G2
Mayoviridae G2 +Virgaviridae G2
Virgaviridae G3
Fimoviridae
Aspiviridae
Betaflexiviridae +Rhabdoviridae
Secoviridae G2
Kitaviridae
Tospoviridae
Mayoviridae G1
Geminiviridae
Botourmiaviridae
Tombusviridae G1
Secoviridae G1
Betaflexiviridae
Phenuiviridae
Caulimoviridae
OTU 19
c
0
1528
7848
9631
7821
9524

Figure 5: Taxonomic placement of Clade C and genome structure of Wollendoviruses
(a) ML phylogeny of reverse transcriptase domains from 52 *Ortervirales* sequences. Red dots mark branches with bootstrap support ≥ 50%. Clades are collapsed for *Caulimoviridae* and by family for other *Ortervirales* members.
(b) ML phylogeny of the 30K movement protein (MP) family, including 46 *Caulimoviridae* sequences identified using Caulifinder and 286 sequences from Butkovic *et al.* (2024), representing the following plant viral families: *Alphaflexiviridae*, *Aspiviridae*, *Betaflexiviridae*, *Bromoviridae*, *Botourmiaviridae*, *Caulimoviridae*, *Fimoviridae*, *Geminiviridae*, *Kitaviridae*, *Mayoviridae*, *Phenuiviridae*, *Rhabdoviridae*, *Secoviridae*, *Tospoviridae,* and *Virgaviridae*. The tree is midpoint-rooted. For *Caulimoviridae* sequences, OTU IDs are listed, followed by sequence names; NCBI accession numbers are provided for reference sequences. OTU 19 is highlighted in green. Red dots indicate branches supported by bootstrap values >= 80 %. For non-caulimovirid families, branches are collapsed by family and represented as triangles.
(c) Genome organisation of the proposed Wollendovirus type member, WolV1, reconstructed from sequence assembly. Open reading frames (ORFs), predicted using ORF Finder (https://www.ncbi.nlm.nih.gov/orffinder/), are shown as light grey rectangles. ORF of unknown function is shaded. The black diamond indicates the tRNAMet primer binding site, corresponding to genome position +1. Conserved domains identified via CD-search (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) against the CDD database (Wang, Chitsaz, Myra K. Derbyshire, et al. 2023) are color-coded: viral movement protein (MP; PF01107) in blue; retropepsin (AP; CD00303) in red; reverse transcriptase (RT; CD01647) in orange; RNase H1 (RH1; CD06222) in yellow; zinc finger in green.

nt-ECRT
1-5
5-25
25-100
100-500
>500
B.2
B.1
B.2.Gymnendovirus 3-4
B.2.Florendovirus
B.2.3
B.2.2
B.2.Petuvirus
B.2.Fernendovirus 1
B.2.Gymnendovirus 1
B.2.Fernendovirus 2
B.2. Gymnendovirus 2
B.2.1
Gymnendovirus 4 gp P. glauca
Gymnendovirus 3
Gymnendovirus 4 gp P. abies
Florendovirus
OTU 14
OTU 24
OTU 26
OTU 33
OTU 32
OTU 7
OTU 17
OTU 8
OTU 11
OTU 4
OTU 25
OTU 20
OTU 6
Gymnendovirus E. salsugineum
OTU 23
Petuvirus C. sinensis
Petuvirusike C. sinensis
OTU 18
OTU 35
Gymnendovirus 1 C. protusa
Fernendovirus 1 P. glycyrrhiza
Fernendovirus 1 P. formosana
OTU 1
OTU 16
Fernendovirus 1
Fernendovirus 2
OTU 2
OTU 31
OTU 29
OTU 28
OTU 9
OTU 27
OTU 3
Gymnendovirus 2 gp G. biloba
Gymnendovirus 2 gp G. biloba 2
Welwitschia mirabilis virus 1
Welwitschia mirabilis virus 2
OTU 12
OTU 30
OTU 22
Gymnendovirus 2 gp P. glauca
Mosto-Xendovirus stub
OTU 10
Wendovirus 4
Solendovirus
Cavemovirus
Kenaovirus Grai
OTU 5
Badnavirus
Dioscovirus
OTU 32
Unclassified Gmax
Tungrovirus
Unclassified Vvin
OTU 6
Cocadnavirus
OTU 13
Vaccinivirus
Zendovirus
OTU 15
Rubulovirus
Rufiovirus
Soymovirus
OTU 21
OTU 19
Isoetes engelmannii
Isoetes sinensis
Isoetes taiwanensis
Selaginella kraussiana
Selaginella tamariscina
Selaginella moellendorffii
Huperzia asiatica
Diphasiastrum complanatum
Lycopodium clavatum
lycophytes
Dipteris conjugata
Marsilea vestita
Salvinia cucullata
Azolla filiculoides
Azolla nilotica
Azolla pinnata
Alsophila spinulosa
Plagiogyria formosana
Ceratopteris richardii
Adiantum capillus-veneris
Adiantum nelumboides
Pteridium aquilinum
Cystopteris protrusa
Polypodium glycyrrhiza
Rumohra adiantiformis
Polystichum acrostichoides
Dryopteris marginalis
ferns
Cycas panzhihuaensis
Ginkgo biloba
Welwitschia mirabilis
Gnetum gnemon
Gnetum montanum
Abies alba
Pseudotsuga menziesii
Larix kaempferi
Larix sibirica
Pinus lambertiana
Pinus tabuliformis
Pinus taeda
Picea sitchensis
Picea abies
Picea glauca
Picea engelmannii
Taxus wallichiana
Taxus chinensis
Metasequoia glyptostroboides
Sequoia sempervirens
Sequoiadendron giganteum
Cryptomeria japonica
Thuja plicata
Juniperus communis
Cupressus sempervirens
Wollemia nobilis
Araucaria angustifolia
gymnosperms
Amborella trichopoda
Nymphaea thermarum
Nymphaea colorata
Chloranthus sessilifolius
Aristolochia fimbriata
Liriodendron chinense
Magnolia officinalis
Chimonanthus salicifolius
Chimonanthus praecox
Phoebe bournei
Persea americana
Cinnamomum micranthum
Litsea cubeba
Dioscorea alata
Musa balbisiana
Brachypodium distachyon
Oryza sativa
Papaver somniferum
Nelumbo nucifera
Beta vulgaris
Vaccinium darrowii
Rudbeckia hirta
Lindenbergia philippensis
Solanum lycopersicum
Capsicum annuum
Vitis vinifera
Citrus sinensis
Arabidopsis lyrata
Ricinus communis
Glycine max
Pyrus communis
Fragaria vesca
Rosa chinensis
angiosperms
429 400 300 200 100

Figure 6: Heatmap of nt-ECRT abundance by OTU across plant hosts

The heatmap displays the distribution of nt-ECRTs across pant genomes, grouped by host species (left) and viral OTUs (top right).

Left: the host plant phylogeny is adapted from the TimeTree (Kumar et al. 2022) database (accessed on 20/09/24). Geological periods are shown along the top, with red bars indicating the standard deviation of estimated emergence times for major plant lineages, based on the following sources: Lycophytes (432.5–392.8 MYA; Morris *et al.*, 2018), Ferns (411.5- 384 MYA Morris *et al.*, ; Lehtonen *et al.*, 2017), Gymnosperms (380–360 MYA Rothwell *et al.*, 1989 ; Stull *et al.*, 2021), and Angiosperms (246.5–130 MYA; Herendeen *et al.*, 2017).

Top: the *Caulimoviridae* cladogram is derived from the Bayesian tree of 142 RT-RH nucleotide sequences (see Figure 4), with major clades color-coded as follows: Clade A (orange), Clade B (green), and Clade C (blue). For visualization purposes, the polyphyletic OTU 'Wendovirus+Xendovirus gp S. tuberosum' (Figure 4) is represented here as a single, monophyletic clade.

For clarity some OTU names have been shortened: Xendovirus gp G. raimondii to Xendovirus Grai; Unclassified gp V. vinifera+Yendovirus+Orendovirus to Unclassified Vvin; Dioscovirus+Badnavirus D. alata to Dioscovirus; Wendovirus+Xendovirus gp S. tuberosum to Wendo-Xendovirus stub; Gymnendovirus 2 gp P. glauca+PtaeV 2+Pinus nigra virus to Gymnendovirus 2 gp P. glauca; Gymnendovirus 2 gp G. biloba 2 + GbilV to Gymnendovirus 2 gp G. biloba 2; Gymnendovirus 1 + PglaV2 to Gymnendovirus1.

Color intensity in the heatmap corresponds to the number of nt-ECRT copies assigned to each OTU in a given host genome.

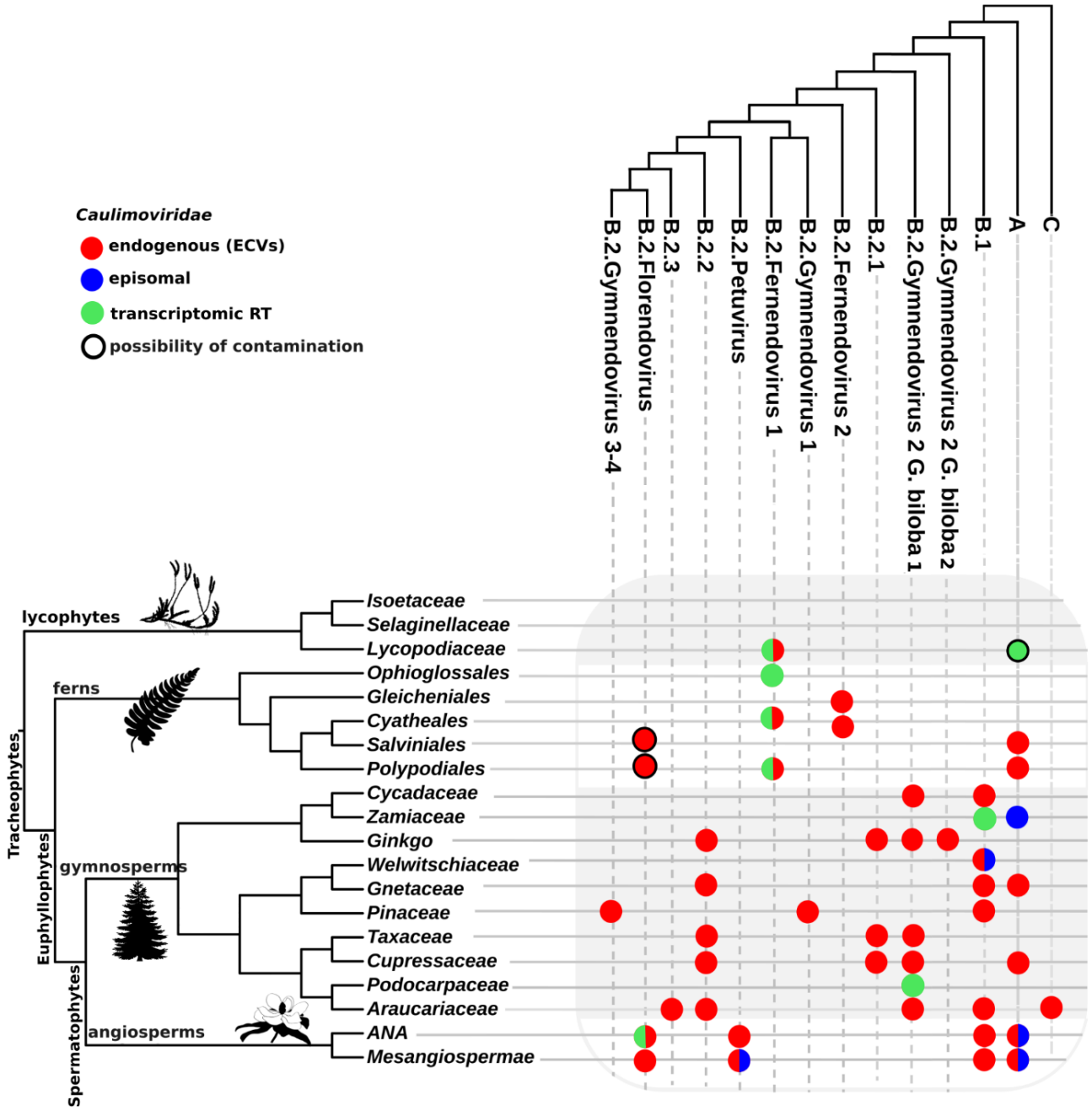


Figure 7: Distribution of *Caulimoviridae* clades and subclades across plant taxa
Left: a plant cladogram depicts the major plant divisions of embryophytes for which either genomic or transcriptomic data were analyzed. Lycophytes and gymnosperms are resolved to the family level due to broader data availability. Ferns, with more limited genomic representation, are shown at the order level. Angiosperms are split into basal ANA grade and *Mesangiospermae*, which includes *Magnoliidae*, *Chloranthales*, monocots, and dicots.
Top: A viral cladogram summarizes the major *Caulimoviridae* clades and subclades as defined in Figure 3. Intersection circles indicate the detection of *Caulimoviridae* sequences within each host taxon, with the following color code. Blue: detection of episomal viral forms, including

Petunia vein clearing viruses (B.2 Petuvirus; Teycheney et al., 2020a), Welwitschia mirabilis viruses 1 and 2 (Debat and Bejerman, 2022), cycad leaf necrosis virus (clade A, genus *Badnavirus*; Alvarez-Quinto et al., 2020), and additional ICTV-recognized *Caulimoviridae* (Teycheney et al., 2020a). Red: detection of ECVs in host genomes. Green: detection of caulimovirid transcripts in transcriptomic datasets. Circles combining multiple colors indicate multiple forms of evidence (E.g., genomic + transcriptomic). Silhouettes are sourced from the PhyloPic image repository (https://www.phylopic.org/).

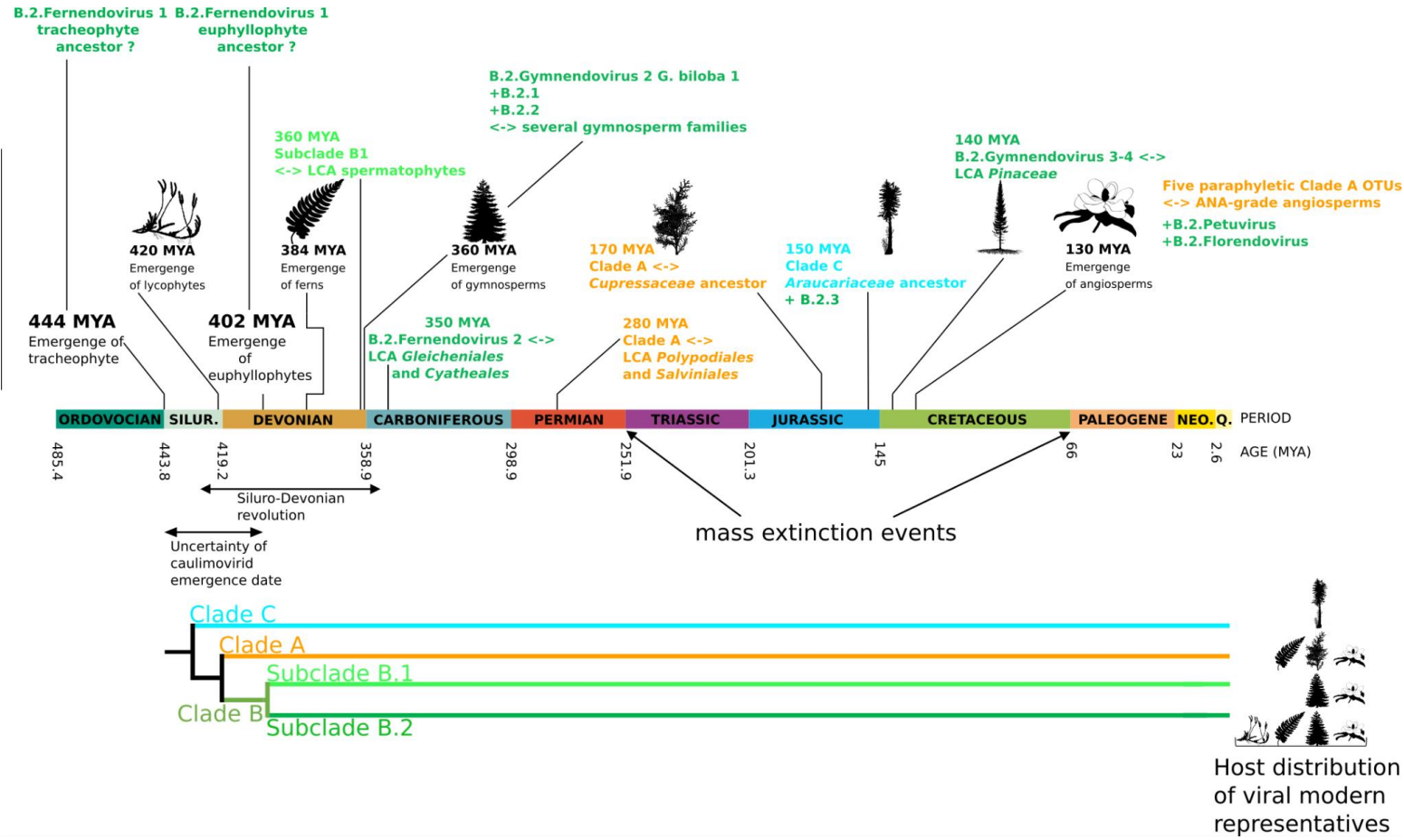


Figure 8: Evolution of *Caulimoviridae* diversity over time in relation to host range

The figure integrates viral diversity with plant evolutionary history along a geological timescale adapted from (McLoughlin 2021).

Timeline: key evolutionary milestones are indicated, including (i) the emergence of tracheophytes and euphyllophytes (Morris et al. 2018); (ii) the divergence of major vascular plant lineages: lycophytes, ferns, gymnosperms and angiosperms (Herendeen et al. 2017; Lehtonen et al. 2017; Morris et al. 2018; Stull et al. 2021; Benton et al. 2022); (iii): the last common ancestor (LCA) of *Gleicheniales-Cyatheales* and *Polypodiales-Salviniales* (Lehtonen et al. 2017); and (iv) the LCA *Cupressaceae* and *Araucariaceae* (Stull et al. 2021). The known host range of each clade is summarized and matched to key plant divergence points on the geological timeline, illustrating the potential cospeciation and diversification of *Caulimoviridae* alongside their plant hosts. Bottom: A simplified *Caulimoviridae* cladogram (adapted from Figure 4) highlights the three main viral clades: A (orange), B (green), and C (blue). Silhouettes representing plant taxa are sourced from the PhyloPic image repository (https://www.phylopic.org/).

Tables

Table 1: Distribution of new and reference OTUs across *Caulimoviridae* clades
The table presents the distribution of newly identified and previously defined reference OTUs across the three major clades of the *Caulimoviridae* family: Clade A, Clade B, and Clade C. The classification is based on the phylogenetic reconstruction shown in Figure 4.

	Clade A	**Clade B**	**Clade C**
new OTU	5, 6, 10, 13, 15, 21, 32	1, 2, 3, 4, 7, 8, 9, 11, 12, 14, 16, 17, 18, 20, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 33, 35	19
ref OTU	Badnavirus, Caulimovirus, Cavemovirus, Dioscovirus+Badnavirus_D.alata, Rosadnavirus, Ruflodivirus, Solendovirus, Soymovirus, Tungrovirus, Unclassified_gp_G.max, Unclassified_gp_V.vinifera+Yendovirus+Orendovirus, Vaccinivirus, Wendovirus+Xendovirus_gp_S.tuberosum, Wendovirus_4, Xendovirus_gp_G.raimondii, Zendovirus	Fernendovirus_1_gp_C.protusa, Fernendovirus_1_gp_P.formosana, Fernendovirus_1_gp_P.glycyrrhiza, Fernendovirus_2, Gymnendovirus_1+PglaV_2, Gymnendovirus_2_gp_G.biloba, Gymnendovirus_2_gp_G.biloba_2+GbilV, Gymnendovirus_2_gp_P.glauca+PtaeV_2+Pinus_nigra, Gymnendovirus_3+PglaV_1, Gymnendovirus_4_gp_P.abies, Gymnendovirus_4_gp_P.glauca+PtaeV_1, Petuvirus_gp_C.sinensis, Petuvirus_gp_E.salsugineum, Petuviruslike_C.sinensis, Welwitschia_mirabilis_virus_1, Welwitschia_mirabilis_virus_2	NA

Table 2: nt-RTs detected in the three transcriptome sources

This table presents details of the 18 transcripts retained from the search for caulimovirid RTs in plant transcriptomes. Each entry includes the following information :

(i) Taxonomic information of the plant: division (column 1), species (column 2), genus (column 3), family (column 4);

(ii) Blast analysis results against the 361 aa-RT sequences used to define OTUs: OTU with the best hit (column 5), identity percentage of the best hit (column 6), transcript length in bp (column 7);

(iii) Reference information: source publication (column 8), transcript accession number (column 9).

Plant division	Species	Genus	Family	OTU best hit	identity	size (nt)	Origin	Accession
Lycophytes	*Lycopodium annotinum*	*Lycopodiales*	*Lycopodiaceae*	Fernendovirus 1 P. formosana	76.2 %	2028	1kp	scaffold-ENQF-2084799-Lycopodium
	Pseudolycopodiella caroliniana	*Lycopodiales*	*Lycopodiaceae*	Badnavirus	72.6 %	540	1kp	scaffold-UPMJ-2015667-Pseudolycopodiella
	Palhinhaea hainanensis	*Lycopodiales*	*Lycopodiaceae*	Badnavirus	67.2 %	1362	https://doi.org/10.1016/j.pld.2021.08.004	Palha_Gene.73649_Cluster-24281.0
	Palhinhaea hainanensis	*Lycopodiales*	*Lycopodiaceae*	35	62.9 %	804	https://doi.org/10.1016/j.pld.2021.08.004	Palha_Gene.103642_Cluster-38429.0
	Lycopodium zonatum	*Lycopodiales*	*Lycopodiaceae*	Fernendovirus 1 P. formosana	65.7 %	1128	https://doi.org/10.1016/j.pld.2021.08.004	Lyzo_Gene.43739_c45295_g1_i1
Ferns	*Botrypus virginianus*	*Ophioglossales*	*Ophioglossaceae*	Fernendovirus 1 P. formosana	47.8 %	2791	1kp	scaffold-BEGM-2004510-Botrypus
	Unclassified Marattia	*Marattiales*	*Marattiaceae*	31	56.1 %	166	1kp	scaffold-UXCS-2117714-Marattia
	Sceptridium dissectum	*Ophioglossales*	*Ophioglossaceae*	25	48.2 %	3425	1kp	scaffold-EEAQ-2089744-Sceptridium
	Lindsaea linearis	*Polypodiales*	*Lindsaeaceae*	Fernendovirus 1 P. formosana	50.2 %	2371	1kp	scaffold-NOKI-2097008-Lindsaea
	Lonchitis hirsuta	*Polypodiales*	*Lonchitidaceae*	Fernendovirus 1 C. protusa	79 %	6523	1kp	scaffold-VVRN-2010621-Lonchitis
	Amblovenatum opulentum	*Polypodiales*	*Thelypteridaceae*	Fernendovirus 1 P. glycyrrhiza	55.2 %	465	https://doi.org/10.1101/2024.08.27.609851	Aop_g40581
	Cibotium barometz	*Cyatheales*	*Cibotiaceae*	Fernendovirus 1 C. protusa	77.9 %	483	https://doi.org/10.1101/2024.08.27.609851	Cba_g18761
	Nephrolepis biserrata	*Polypodiales*	*Nephrolepidaceae*	Fernendovirus 1 P. formosana	67.9 %	1362	https://doi.org/10.1101/2024.08.27.609851	Nbi_g42169

Gymnosperms	*Dioon edule*	*Cycadales*	*Zamiaceae*	1	71 %	509	1kp	scaffold-WLIC-2000696-Dioon
	Podocarpus rubens	*Pinales*	*Podocarpaceae*	3	62.3 %	6552	1kp	scaffold-XLGK-2060667-Podocarpus
ANA	*Austrobaileya scandens*	*Austrobaileyales*	*Austrobaileyaceae*	Florendovirus	79.3 %	1747	1kp	scaffold-FZJL-2010292-Austrobaileya
	Austrobaileya scandens	*Austrobaileyales*	*Austrobaileyaceae*	Florendovirus	66 %	4766	1kp	scaffold-FZJL-2022942-Austrobaileya
	Austrobaileya scandens	*Austrobaileyales*	*Austrobaileyaceae*	Florendovirus	78.5 %	583	1kp	scaffold-FZJL-2162197-Austrobaileya