Endogenous florendoviruses are major components of plant genomes and hallmarks of virus evolution

Andrew D.W. Geering1,*, Florian Maumus2,*, Dario Copetti3,4, Nathalie Choisne2, Derrick J. Zwick5, Matthias Zytnicki2, Alistair R. McTaggart1, Simone Scalabrin6, Silvia Vezzulli7, Rod A. Wing3,4, Hadi Quesneville2 & Pierre-Yves Teycheney8

The extent and importance of endogenous viral elements have been extensively described in animals but are much less well understood in plants. Here we describe a new genus of Caulimoviridae called ‘Florendovirus’, members of which have colonized the genomes of a large diversity of flowering plants, sometimes at very high copy numbers (>0.5% total genome content). The genome invasion of Oryza is dated to over 1.8 million years ago (MYA) but phylogeographic evidence points to an even older age of 20–34 MYA for this virus group. Some appear to have had a bipartite genome organization, a unique characteristic among viral retroelements. In Vitis vinifera, 9% of the endogenous florendovirus loci are located within introns and therefore may influence host gene expression. The frequent colocation of endogenous florendovirus loci with TA simple sequence repeats, which are associated with chromosome fragility, suggests sequence capture during repair of double-stranded DNA breaks.
Horizontal gene transfer constitutes a significant lateral force in species evolution. For multicellular eukaryotes, only DNA that is transferred into germline nuclei is able to be transmitted to the progeny. Most documented examples of horizontal gene transfer involve the transfer of DNA of either prokaryotic or viral origin. A range of endogenous viral elements (EVEs) originating from ancestral viruses with all combinations of genome (single-stranded (ss) and double-stranded (ds) DNA, ss and dsRNA, and positive and negative sense ssRNA) have been found in eukaryotic genomes. Of these, only the retroviruses (family Retroviridae) have an active integration mechanism. For all other viruses, endogenization is thought to be accidental and may involve incorporation of viral DNA during non-homologous end-joining repair of dsDNA breaks in the chromosome or, alternatively, some hybrid mechanism involving the enzymatic machinery of retrotransposons.

Although the integration of viral sequences into host genomes can induce deleterious mutations, EVEs can also have beneficial outcomes for the host. For example, retrovirus long terminal repeat (LTR) promoters in the human genome contribute to the gene regulatory network by acting as alternative promoters for mRNA transcription, as binding sites for transcription factors such as p53, Oct4 and Nanog, or as promoters of antisense or non-coding RNA. The insertion of endogenous retroviral sequences near to or in non-coding parts of a gene can also modify gene expression as a result of local epigenetic remodelling, the insertion of polyadenylation signals and the generation of new splicing sites.

Some EVEs have also contributed proteins that have assumed new and sometimes vital roles in normal host physiology. For example, syncitin-A, which derives from an endogenous retrovirus Env gene, is essential to early development of the placenta in mammals. Several endogenous retroviruses also contribute to virus-derived immunity against closely related exogenous viruses by producing proteins that block cell entry, disrupt virus replication or movement, or ameliorate disease symptoms. In plants, the gem gene of the grass Festuca pratensis, which is linked to delayed leaf senescence (the ‘stay-green’ phenotype), has a partiviral origin.

EVEs are also of interest to the scientific community because they are essentially fossils of viruses that existed in eons gone by and therefore offer unique insights into virus evolution and biogeography. For example, by screening for orthologous EVE loci in different selections of the wild banana Musa balbisiana, it has been possible to date the endogenization events (and therefore the minimum ages) of two extant badnavirus species, Banana streak GF virus and Banana streak IM virus, to c. 640,000 years ago. In another example from the animal kingdom, the discovery of an endogenous lentivirus in a Madagascan prosimian, the greenalga Chlamydomonas reinhardtii, found in flowering plants and form a novel lineage within the family Caulimoviridae (see below), they were named ‘Florendovirus’, after Flora, the Roman goddess of flowers.
The reconstructed genomes of the majority of florendoviruses contain two open reading frames (ORFs) in different translational frames (Supplementary Data 3) such as for Lotus japonicus A virus (Fig. 1). ORF1 encodes a putative 205–216 kDa polyprotein with movement protein (MP), coat protein (CP) with zinc finger, aspartic protease (AP), RT and RH1 domains (Supplementary Data 4). ORF2 encodes a putative 45–58 kDa protein that lacks significant homology to reference proteins and protein domains, and appears to be specific to the florendoviruses. Although the domains and their order is conserved, the genome organizations of Amborella trichopoda B virus and Glycine max virus differ from that of other florendoviruses; there is a single ORF in the former and three ORFs in the latter, caused by a division of ORF1 from that of other florendoviruses; there is a single ORF in the genome of V. vinifera B virus (VvinBV) and V. vinifera D virus (VvinDV). Hence, components A and B sequences have the same overall structures as in V. vinifera albeit the ORF2 homologues are more divergent. Hence, components A and B appear to encode complementary (and partially redundant) sets of proteins that together constitute complete florendovirus proteomes. The virus from Oryza was called Oryza sativa B virus (OsatBV) to distinguish it from previously described endogenous caulimovirids in rice.

Although component B sequences lack an RT domain, conspecificity with selected component A sequences could be established on the basis of very high sequence similarities within the intergenic region (>90% nt identity) and MP domain (>87% nt identity; Supplementary Table 1). Together, domain complementarity and sequence similarity suggest that these two sets of sequences are co-evolving entities of a bipartite viral genome. Lending support to this hypothesis, we detected several loci in the V. vinifera and Oryza genomes where components A and B form compound insertions (Supplementary Fig. 1), suggesting coexistence and probably interaction of the two DNA genomes at the time of capture in the chromosome. Evidence that these are not artefacts generated by the sequence assembly process is provided by the large number of entire or nearly entire component sequences that are present in each plant genome (Supplementary Table 2 and Supplementary Data 5). In addition, the flanking regions of the component sequences show very little redundancy, indicating that segmental duplication or transduplication do not explain the multiplicity of these sequences in V. vinifera (Supplementary Table 3), thereby suggesting that the majority of the copies result from independent integration events. Two virus species with putative bipartite genome organizations are present in V. vinifera, namely Vitis vinifera B virus (VvinBV) and Vitis vinifera D virus (VvinDV) and one each in the monocot species. Although OsatBV and Sorghum bicolour virus share a most recent common ancestor, VvinBV and VvinDV represent parallel evolution events (Fig. 2).

Dating the genome invasion of Oryza. We took advantage of the availability of several closely related Oryza genomes
(Supplementary Table 4) to date the OsatBV endogenization events by searching for orthologous OsatBV loci in the different species. Of the nine Oryza AA genomes and the one Oryza BB genome (Oryza punctata) that were examined, OsatBV (>99% nt identity to the OsatBV consensus) was found in all, except in Oryza brachyantha (Table 2). Related sequences were also found in Leersia perrieri albeit these were significantly different to those in Oryza (80–88% nt identity to the OsatBV consensus). Collectively, a total of 54 different OsatBV loci were identified, of which 13 loci contained A component sequences, 35 loci contained B component sequences and the remaining 6 loci contained a mixture of the two (Supplementary Data 6). Out of 16 loci shared by 2 or more assemblies, only 2 (japo_1_23M and japo_7_27M) had a pattern that was inconsistent with the phylogenetic tree of the species (Fig. 4), and can be explained by incomplete lineage sorting at these loci. Interestingly, one OsatBV locus was shared by all AA-genome types, except Oryza meridionalis, which is the basal lineage of this genome type. The seven Oryza species containing the OsatBV orthologues are distributed across a wide geographical area including Asia, western and sub-Saharan Africa, Madagascar, and Central and South America, suggesting that introgression of the shared OsatBV loci by interspecific hybridization is very unlikely given the considerable geographic barriers. Based on the estimated time of divergence of O. meridionalis from all other AA genome taxa (Fig. 4), these OsatBV insertions have occurred between 1.8 and 2.3 MYA.

When looking for evidence of recent OsatBV insertions, we found an interesting instance where a polymorphic locus between Oryza glaberrima (Chr10:8390000..8409999) and Oryza barthii (Chr10:8419000..8422000) shows an insertion flanked by (TA)_n repeats in the former and the presence of an empty stretch of TA repeats in the latter. Unless this insertion was precisely eliminated in O. barthii, it is likely to be that this polymorphism reflects the endogenization of OsatBV after the divergence of the two species about 120,000 years ago (Fig. 4).

Abundance and distribution in plant genomes. The reconstructed florendovirus genome sequences identified here (Supplementary Data 1) were used to mask plant genomes and we observed highly heterogeneous florendovirus sequence abundance. In Arabidopsis thaliana, only putative traces (c. 3.4 kb) of sequence were detected (Table 1). In contrast, florendovirus sequences make up >1% of the R. communis genome and >0.5% of the J. curcas, A. trichopoda, Citrus clementina and V. vinifera genomes. The same method was used to mask the various Oryza genomes, but this time only using the OsatBV consensus sequences (Table 2). Overall, the OsatBV contribution to the Oryza genomes is relatively modest (≤0.04% of total genome content) and also highly variable between the different species, which may reflect true differences in the copy number but also could be significantly influenced by the quality of the genome assemblies. For reasons that are unclear, component B sequences were overall, twofold more abundant than component A sequences.

To determine whether there was an association between florendovirus sequences and any other genome feature, we focused on the reference genomes of V. vinifera and O. sativa, which are assembled into pseudo-chromosomes. For both species, we found that florendovirus sequences are on average located much closer to TE than to genes (Fig. 5). For V. vinifera, which is florendovirus-rich compared with O. sativa, c. 9% of the loci...
overlap with host genes, with 99% (c. 286 kbp) of these located within introns. To assess whether florendoviral promoters would have been selected for the transcriptional regulation of host genes, we also investigated whether they are frequent in the proximity of genes but we could not establish such a correlation.

Manual examination of various plant genomes led to the observation that florendovirus sequences were frequently flanked by TA dinucleotide simple sequence repeats (TA(n)). The existence of the simple sequence repeat before the integration was confirmed by inspecting several orthologous loci of related

Figure 2 | Phylogenetic relationships within the Caulimoviridae. Phylogram obtained from a maximum likelihood search with DNA sequence data from AP-RT-RHI genes. Bootstrap support (≥ 70%) values from 1,000 replicates above nodes. Posterior probabilities (≥ 0.95) summarized from 29,000 trees in a Bayesian search are shown below nodes. Virus species from each of the recognized genera are Cauliflower mosaic virus (CaMV), Figwort mosaic virus (FMV), Soybean chlorotic mottle virus (SoyCMV), Peanut chlorotic streak virus (PCSV), Rice tungro bacilliform virus (RTBV), Commelina yellow mottle virus (ComYMV), Banana streak OL virus (BSOLV), Sweet potato vein clearing virus (SPVCV), Tobacco vein clearing virus (TVCV), Cassava vein mosaic virus (CsVMV), Sweet potato collusive virus (SPCV), PVCV, Rose yellow vein virus (RYVV, unassigned) and Citrange pararetrovirus (CitPRV, unassigned). The outgroup is Saccharomyces cerevisiae Ty3 virus (SceTy3V). New florendovirus species are colour-coded to indicate the plant family in which they are found: dark blue is Poaceae, light grey is Euphorbiaceae, dark grey is Amborellaceae, olive green is Brassicaceae, pink is Cucurbitaceae, purple is Vitaceae, light blue is Fabaceae, red is Rosaceae, yellow is Solanaceae, light green is Malvaceae, dark grey is Myrtaceae and orange is Rutaceae. Scale bar, 0.4 nucleotide substitutions per site in the nucleotide alignment using the GTRGAMMA model of evolution.
Oryza species that present an empty site. This analysis also revealed that as a result of the insertion, short stretches of sequence can be gained or lost (Supplementary Fig. 2). To quantify the sequence associations, we examined a subset of large (≥500 bp) endogenous florendovirus loci in five different plant genomes (Fig. 6) and found that (TA)n-proximal loci are significantly (P≤0.0001) more frequent than expected by chance in each species addressed. The proportion of (TA)n-proximal loci ranged from 14% in V. vinifera to 46% and 51% in G. max and O. sativa, respectively. Interestingly, an integration bias of the rice tungro bacilliform virus-like sequences towards (TA)n repeats within the Oryza genome has already been described by Kunii et al.\(^24\). Here, our results suggest that this repeated motif is present before insertion of the DNA in the Oryza genome and the broader association of florendoviral sequences with TA stretches supposes a similar situation in a variety of plant species.

**Figure 3** Structure of bipartite florendovirus genomes. Comparison of the genome organizations of Vitis vinifera A virus (VvinAV), Vitis vinifera B virus (VvinBV), Vitis vinifera D virus (VvinDV), Oryza sativa B virus (OsatBV) and Sorghum bicolor virus (SbicV). Genomes have been linearized and following convention, the first nucleotide of the tRNAMET (OsatBV) and Sorghum bicolor virus (SbicV). Genomes have been linearized and following convention, the first nucleotide of the tRNAMET binding site designated the beginning of the genome. Light grey boxes mark ORFs and conserved domains within each ORF are coloured as for Fig. 1. Regions of sequence homology are represented by polygons containing ORFs and conserved domains within each ORF are coloured as for Fig. 1. Regions of sequence homology are represented by polygons containing the nucleotide positions in the virus genomes.

**Figure 4** Placement of 54 Oryza sativa B virus (OsatBV) loci on the Oryza phylogenetic tree. OsatBV insertions were searched in orthologous loci across all 12 Oryza species and placed onto the phylogenetic tree according to the most parsimonious hypothesis. The red numbers represent insertions of A, a mixture of A and B, or B components, respectively. Shared insertions are indicated by arrows pointing at the corresponding branch. Because of the method adopted, the split of the outgroup Leersia perrieri could not be dated (dashed line), while the split of O. brachyantha was fixed to 15 million years ago (MYA). Other node ages are: O. punctata (BB genome), 6.712 MYA; all AA genome species, 2.317 MYA; BB genome species, 1.832 MYA; AA genome species, 0.738 MYA; Asian-African AA species, 0.572 MYA; Asian species, 0.391 MYA; O. japonica ssp. indica-O. nivara, 0.202 MYA; O. sativa ssp. japonica-O. rufipogon, 0.187 MYA; O. glaberrima-O. barthii, 0.120 MYA. The scale bar represents time, with increments of one million years, and labels every five million years.

### Table 2 Variation in the contribution of endogenous Oryza sativa B virus to the genomes of a range of Oryza species and Leersia perrieri.

| Genome            | Hit counts | Genome occupancy (bp) | Genome fraction (%) |
|-------------------|------------|-----------------------|---------------------|
|                   | type       | compAsc1 | compBsc1 | Total | compAsc1 | compBsc1 | Total | compAsc1 | compBsc1 | Total |
| O. s. japonica    | AA         | 11       | 39       | 50    | 25,976   | 98,105   | 124,081 | 0.0070   | 0.0263   | 0.0332 |
| O. rufipogon      | AA         | 39       | 61       | 100   | 21,429   | 40,579   | 62,008  | 0.0063   | 0.0120   | 0.0183 |
| O. s. indica      | AA         | 11       | 30       | 41    | 25,754   | 58,361   | 84,115  | 0.0069   | 0.0156   | 0.0225 |
| O. nivara         | AA         | 12       | 13       | 25    | 13,766   | 14,777   | 28,546  | 0.0041   | 0.0044   | 0.0084 |
| O. glaberima      | AA         | 12       | 25       | 37    | 36,296   | 37,524   | 73,820  | 0.0127   | 0.0132   | 0.0259 |
| O. barthii        | AA         | 5        | 5        | 10    | 8,500    | 4,156    | 12,656  | 0.0028   | 0.0013   | 0.0041 |
| O. glumeapatula   | AA         | 3        | 6        | 9     | 2,561    | 5,475    | 8,036   | 0.0007   | 0.0015   | 0.0022 |
| O. longistaminata | AA         | 19       | 22       | 41    | 3,231    | 5,579    | 8,810   | 0.0009   | 0.0016   | 0.0026 |
| O. meridionalis   | AA         | 14       | 13       | 27    | 6,310    | 5,959    | 12,269  | 0.0019   | 0.0018   | 0.0037 |
| O. punctata       | BB         | 6        | 17       | 23    | 19,540   | 39,584   | 59,124  | 0.0050   | 0.0101   | 0.0150 |
| O. brachyantha    | FF         | —        | —        | —     | —        | —        | —      | —        | —        | —     |
| L. perrieri       | —          | 2        | 10       | 12    | 967      | 40,006   | 40,973  | 0.0004   | 0.0150   | 0.0154 |
| Total             |            | 84       | 141      | 225   | 116,928  | 211,421  | 328,349 | 0.0486   | 0.1027   | 0.1512 |
Figure 5 | Distances between endogenous florendovirus elements (EFEs) and other plant genome features. The mean nucleotide distances that separate EFEs from either transposable elements (TEs) or genes in *Vitis vinifera* cv. Pinot Noir and *Oryza sativa* are shown.

Figure 6 | Physical concomitance of endogenous florendovirus elements (EFEs) and TA dinucleotide ((TA)n) repeats. The percentages of EFE and equal numbers of random loci that are located at less than 1 kbp from (TA)n repeats are shown. (TA)n repeats were detected with Tandem Repeat Finder. Loci sample sizes (n) were: *Citrus clementina* (n = 543), *Prunus persica* (n = 136), *Vitis vinifera* (n = 968), *Glycine max* (n = 468) and *Oryza sativa* (n = 39). The statistical significance of differences in the frequency of association of EFE and random loci to (TA)n repeats was determined using a Mann-Whitney U-test.

Evidence for replication competency. In *V. vinifera*, the sequences of endogenous *Vitis vinifera* A virus and *Vitis vinifera* C virus have decayed to a point that no loci containing an entire viral genome could be identified, nor even a fragment of sequence containing an uninterrupted ORF. In contrast, endogenous *VvBV* is much more likely to be replication competent, as many loci contain entire component sequences and several also had uninterrupted ORFs (Supplementary Table 2 and Supplementary Data 5). The mean fragment length of endogenous *VvBV* was also about double that of *VvAV* (Supplementary Table 2). In a recent study addressing genetic diversity among four phenotypically different somaclonal variants of *V. vinifera* cv. Pinot Noir, insertional polymorphisms of a sequence called Cauliv-1 were observed. Cauliv-1 was classified as a class I LTR TE by the authors but sequence comparisons by us show that it is the same as *VvinBV*. The domestication of grapevine probably began during the Neolithic era (6,000 to 5,000 B.C.)

Endogenous florendoviruses as sources of small RNAs. In general, EVEs are sources of small RNAs (sRNAs, 21–24 nt) that could be involved in antiviral defense mechanisms or play a role in shaping the epigenome. We searched for sRNAs with zero mismatches to the reconstructed florendovirus genome sequences and found corresponding molecules in complementary DNA libraries from *A. trichopoda*, *C. clementina*, *C. sinensis*, *P. trichocarpa*, *Solanus lycopersicum*, *Solanus tuberosum*, *S. bicolor* and *V. vinifera*. Estimates of the number of sRNAs are probably greatly underestimated, as the reconstructed florendovirus genome sequences are consensus sequences and therefore do not reflect the full extent of sequence variation at different loci. Although there were some hotspots within the viral genomes from where the sRNAs derived, no consistent patterns could be ascertained (Supplementary Fig. 4).

Discussion

We have reconstructed representative genomes of a new genus of the *Caulimoviridae*, tentatively named ‘Florendovirus’, from fragments of sequence that have been captured and preserved in plant genomes. A premise of this type of analysis is that following endogenization, the rate of evolution of the sequences greatly slows down, and because selective constraints on the viral sequence are removed those mutations that do occur are random and are eliminated on generation of a consensus sequence. A similar analytical approach has been successfully used to reconstruct the ancestral sequences of a range of TEs and is considered to give a good approximation of the ancestral sequence as long as the endogenous sequences are not so old as to be unrecognizable from the ancestral sequence and that they exist in a sufficiently high copy number to allow determination of a ‘modal’ sequence. One very remarkable feature of the florendoviruses is the extraordinary diversity of host plants (ANITA grade, monocots and dicots), and at a discovery rate of > 50% in the plant genomes that were examined, many additional florendovirus species are still likely to be discovered. From this study alone, the diversity of florendoviruses is greater than any other extant genera of the *Caulimoviridae* except the badnaviruses.

Phylogenetic analyses showed that the proposed genus Florendovirus is sister to PVCV, the type and sole member of the genus *Petuvivirus*, with which it shares the plesiomorphic trait of MP, CP, AP, RT and RH1 precursors occurring in one large polyprotein. This polyprotein is presumably processed by the virus into the mature proteins through the action of the virus.
encoded Ap30. The florendoviruses are readily distinguished from 
PVCCV by the presence of a second ORF, which encodes a putative 
protein of unknown function with no homologue in any other 
caulimovirid. For the majority of species, ORF2 was in a different 
translational reading frame to ORF1, suggesting a mechanism of 
translation using occasional leaky ribosome scanning31. 
Interestingly, the atypical genome organization observed for 
Glycine max virus with split ORF1 is similar to the situation of 
ORF3 from the badnavirus Sweet potato pakakuy virus32. This 
additional division of the virus genome may allow more precise 
control of expression of the structural and enzymatic proteins 
during different parts of the replication cycle as compared with 
post-translational processing of a large polyprotein. 

Bipartite florendovirus genomes represent a unique genome 
organization for viral retroelements. Retroviruses encapsidate two 
identical or nearly identical RNA molecules and therefore their 
genomes are diploid rather than bipartite33. One cannot discount 
the possibility that the putative bipartite florendoviruses 
depended on a helper virus for replication, although this would 
appear unnecessary as when viewed together, components A and 
B are complementary and contribute all gene regulatory and 
protein-coding sequences necessary for replication. Importantly, 
each component contains a complete intergenic region, which is 
neither identical between component A and B sequences. 
Interestingly, the florendovirus bipartite genome structure 
evolved on three independent occasions with a remarkably 
similar outcome in terms of gene organization. However, there 
are no examples of divided genomes in extant members of the 
Caulimoviridae, although it is commonplace in plant RNA 
viruses34. Complementation between cauliflower mosaic virus 
(CaMV) and a CaMV-derived virus vector with a foreign marker 
gene has been observed34, providing evidence that complementation 
between two different genome components of a caulimovirid is at least experimentally possible. Bipartite 
florendovirus genomes may therefore represent unsuccessful 
attempts in the evolution process of viral retroelements.

Analyses of the patterns of integration suggest that florendo-
virus sequences are more likely to be found in TE-rich regions of 
the plant genome and there is also a strong bias towards insertion 
in TA dinucleotide simple sequence repeats. The co-location of 
TEs and florendovirus sequences may simply reflect similar 
selection pressures acting to determine where in the genome these 
elements accumulate, as insertions in gene-rich regions are more 
likely to be deleterious to the individual and therefore the 
insertion less likely to persist in the population due to selection 
pressure35. Insertion in stretches of TA dinucleotides may, 
however, point to the mechanism of integration. It is thought 
that TA dinucleotide-rich areas of sequence are more likely to 
form highly stable secondary structures (for example, hairpins) 
that perturb DNA replication, thereby causing chromosome 
fragility36,37. Florendovirus DNA could then be coopted to act 
as filler DNA to repair the double-stranded DNA breaks by either 
non-homologous end joining or microhomology-mediated end 
joining38.

A minimum age of at least 1.8 million years has been provided 
for endogenous OsatBV, which is approximately three times older 
than the other endogenous caulimovirids that have been 
dated, Banana streak GF virus and Banana streak IM virus14. 
The dating technique that was used does have its intrinsic limitations, 
as the turnover of repetitive elements in plants is relatively rapid. 
For example, in Nicotiana spp., there is near-complete genome 
turnover of repetitive elements in as little as five million years39, 
and for O. sativa, the half-life of LTR retrotransposons is less than 
six million years40. The phylogenetic relationships that were 
observed suggest a much older age of the florendoviruses. For 
instance, the florendoviruses in Eucalyptus, an iconic Australian 
plant genus in the Gondwanan family Myrtaceae, were sister to 
those in Theobroma cacao and Gossypium raimondii, both of 
which originate from South America. The discovery of Eucalyptus 
macrofossils in southern Argentina suggests that this plant genus 
was continuously distributed across the Antarctic land bridge 
between Australia and South America41. This floristic connection 
was broken about 34 MYA when the Drake Passage opened and 
permanent ice sheets formed in Antarctica42,43. The geographic 
distribution of closely related endogenous florendoviruses in 
Eucalyptus grandis, T. cacao and G. raimondii can be explained by 
either vicariance or long-distance dispersal of the most recent 
common ancestor of the viruses across the Pacific Ocean, the 
largest stretch of water in the world. We consider the first 
hythesis much more probable, giving a minimum age of 34 
million years for this virus clade. The florendoviruses from 
Nicotiana benthamiana, S. lycopersicum and S. tuberosum also 
formed a monophyletic clade. N. benthamiana is a member of 
Nicotiana section Solanoventes, a section of this genus largely 
endemic to Australia but with South American ancestors, while S. 
lycopersicum and S. tuberosum originate from South America44. 
The most recent common ancestor of the Australian 
representatives of section Solanoventes is thought to have 
colonized Australia at least 20 MYA45, which could be also 
considered a minimum age for this virus clade.

Definitive conclusions about the replication competency of any 
endogenous florendoviruses described in this study cannot be 
made, although the discovery of polymorphic loci in closely 
related Oryza species suggests that there were cycles of infection 
and endogenization of OsatBV as little as 100,000 years ago. 
Given that endogenous florendoviruses occur in some of the most 
intensively studied crops (for example, grape, rice, cotton, 
soybean, maize, peach, strawberry, potato and tomato), one 
would assume that if they still existed in an exogenous form 
today, then they would have been discovered in more than a 
century of plant virology research, even if serendipitously as a 
contaminant in other virus preparations. Given the long 
association of florendoviruses with their plant hosts, it is possible 
that as a consequence of coevolution, the disease symptoms 
caused by the viruses have attenuated to a point that they no 
longer cause harm to the plants and therefore have not received 
the attention of plant pathologists. Alternatively, endogenization 
may have provided plant immunity to infection by the cognate 
exogenous virus through induction of RNA interference path-
ways17,19, 
causng the viruses to become extinct. Finally, perhaps the 
fluendoviruses (and members of the Caulimoviridae in general) flourished in prehistoric times when a particular vector 
group was abundant, but this vector group has now disappeared 
or greatly diminished in abundance due to environmental 
changes or domestication of the host plant species by humans. 
Supporting this hypothesis, PVCCV, the nearest relative of the 
fluendoviruses, has only ever been grafted-transmitted and 
attacks at mechanical or vector transmission have been 
unsuccessful46. Many other extant members of the 
Caulimoviridae also do not have any known insect vectors, 
such as some caulimoviruses, all soymo-, cava- 
and solendoviruses16, 
and Rose yellow vein virus47. Furthermore, 
Rice tungro bacilliform virus is only transmitted by leafhoppers 
when present in a mixed infection with a non-related helper virus, 
and caulimoviruses have only acquired the ability to be 
transmitted by aphids, the most common vector group 
nowadays, through the acquisition of a novel auxiliary gene, the 
aphid transmission factor18. 

The question remains as to what beneficial functions 
endogenous florendoviruses could confer on the plant. Plant 
defence against virus infection is one possible benefit but it is 
questionable whether this is the current function for several 

39. Interestingly, the atypical genome organization observed for 
Glycine max virus with split ORF1 is similar to the situation of 
ORF3 from the badnavirus Sweet potato pakakuy virus32. This 
additional division of the virus genome may allow more precise 
control of expression of the structural and enzymatic proteins 
during different parts of the replication cycle as compared with 
post-translational processing of a large polyprotein.

40. But this vector group has now disappeared 
or greatly diminished in abundance due to environmental 
changes or domestication of the host plant species by humans. 
Supporting this hypothesis, PVCV, the nearest relative of the 
fluendoviruses, has only ever been grafted-transmitted and 
attacks at mechanical or vector transmission have been 
unsuccessful46. Many other extant members of the 
Caulimoviridae also do not have any known insect vectors, 
such as some caulimoviruses, all soymo-, cava- 
and solendoviruses16, 
and Rose yellow vein virus47. Furthermore, 
Rice tungro bacilliform virus is only transmitted by leafhoppers 
when present in a mixed infection with a non-related helper virus, 
and caulimoviruses have only acquired the ability to be 
transmitted by aphids, the most common vector group 
nowadays, through the acquisition of a novel auxiliary gene, the 
aphid transmission factor18. 

The question remains as to what beneficial functions 
endogenous florendoviruses could confer on the plant. Plant 
defence against virus infection is one possible benefit but it is 
questionable whether this is the current function for several
reasons. First, rather than preventing infection, modern experience with related badnaviruses, petuviruses and solenodoviruses suggests that some endogenous forms of these viruses are paradoxically the major source of infection and when an exogenous viral genome is released, it is able to overcome RNA interference resistance, perhaps by the expression of a silencing suppressor protein. Second, the copy number of most endogenous caulimovirid sequences in plant genomes (for example, this study, Jakowitsch et al. and Kunii et al.) is far in excess of that which is needed to provide efficient silencing of a virus: one hairpin transgene containing sense/anti-sense arms that are as short as 98 nts is capable of providing efficient silencing. Finally, endogenous florendoviruses are widespread in the plant kingdom, and if the cognate exogenous viruses are in fact either very rare or have become extinct, a role in plant defence is somewhat redundant.

It would seem more likely that the endogenous florendoviruses are contributing to plant evolution by acting as sources of novel genetic material at either the coding or transcriptional regulatory levels. A feature of plants that distinguishes them from animals is their highly plastic genome structure: angiosperm genome sizes vary nearly 2,000-fold compared with those of mammals and their highly plastic genome structure: angiosperm genome sizes vary nearly 2,000-fold compared with those of mammals and birds, whose genome sizes vary by no more than 5-fold. It is theorized that this genome plasticity has allowed plants to acquire new biochemical processes or growth patterns in relatively short time frames to adapt to new predation or competition pressures or variable climatic conditions in an unstable environment. It is noteworthy that 9% of the endogenous florendovirus loci in V. vinifera are within plant genes, and of these almost all are present within the introns. The presence of florendovirus sequences in introns possibly has biological consequences by affecting both the structure of the gene transcript as well as the sequences in introns possibly has biological consequences by affecting both the structure of the gene transcript as well as the level of its expression. Overall, florendoviruses appear to have significantly contributed to the evolution of angiosperm genomes and perhaps to the emergence of phenotypes that have been domesticated such as in grape somaclonal variants.

Methods

Discovery and assembly of endogenous viral genomes. Uncharacterized EVEs in the V. vinifera genome were initially identified using the CaMV AP-RT-RH1 (GenBank Accession NP_056728) as the query sequence in a tBLASTN search of the non-redundant nucleotide database of GenBank. High-scoring sequences were then extended by pairwise BLASTN comparisons 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. Following convention for the Caulimoviridae, the first nucleotide of the rnaMHT consensus sequence was designated the beginning of the viral genome and, accordingly, the preceding nucleotide the end of the genome. Once the first viral genomes were assembled, these in turn were used to search for similar sequences in other plant genomes including those available on the NCBI Genomes (chromosomes) and Whole-genome Shotgun Reads databases, Phytozome release v7.0 (www.phytozome.net), the peach genome v1 (http://www.rosaceae.org/peachgenome), the strawberry genome v1 (http://www.rosaceae.org/strawberry_genome), the Jatropha Genome DataBase (http://www.kazusa.or.jp/iptrophala/) and the Amborella Genome Database (http://www.amborella.org/). Accession numbers of sequences used in the analyses and further details of the databases are provided in Supplementary Table 1. Florendovirus genomes were screened for the homology with known protein domains using the InterPro and CDD databases.

Phylogenetic analyses. To investigate evolutionary relationships, AP-RT-RH1 gene sequences (sequences homologous to nts 3,732–5,650 of CaMV, NCBI Accession NC_001497.1) were used. DNA sequences of representatives of each of the genera in the Caulimoviridae, as well as the florendovirus consensus sequences that had uninterrupted reading frames, were conceptually translated and aligned using the MUSCLE algorithm in the MEGA v. 5.05 software package, then back-translated into the nucleotide code. Florendovirus consensus sequences with interrupted reading frames were then added to this alignment using the ‘realign selected sequences’ option of CLUSTALX. Ambiguous regions of the alignment were removed using the GBs programme on the Phylogeny.fr server (available at http://www.phylogeny.fr/). The four domains (AP, RT, tether region and RH1) were analysed together as partitioned loci in the phylogenetic analyses. Two phylogenetic assessment criteria were implemented: Bayesian inference using MrBayes (ref. 35) and maximum likelihood using RAxML. Resulting trees were observed with FigTree (http://tree.bio.ed.ac.uk/software/figtree/). The RAxML analyses were run with a rapid bootstrap analysis (command -f a) under GTRGAMMA using a random starting tree and 1,000 maximum likelihood bootstrap replicates. MrBayes 3 was used to conduct a Markov Chain Monte Carlo with Bayesian inference. Four runs, each consisting of four chains, were implemented until the s.d. of split frequencies was below 0.01. The cold chain was heated at a temperature of 0.25. Substitution model parameters were sampled every 100 generations and trees were saved every 5,000 generations. The convergence of the Bayesian analysis was examined using the cumulative and compare analyses in AWTY (available at http://awty.readthedocs.org/en/latest/content/bedtools-suite.html). For each species, the number of endogenous florendovirus loci and the number of random annotations that are located at <1 kbp from a stretch of TA dinucleotides were then counted.

Estimation of Orzya divergence times. Divergence times within Orzya were estimated on a phylogeny inferred using protein-coding genes from assemblies of the short arm of chromosome 3 (Supplementary Table 4). Sequences from 16 Orzya accessions were included, with L. perrieri serving as the outgroup (Zwickl et al. and Supplementary Table 4). Single-cop syntenic orthologue clusters were collected using the BLAST-Overlap-Synteny pipeline detailed in Zwickl et al. Full gene sequences (including introns) of each locus were aligned using PRANK v.141010 (ref. 62) using the ‘-f’ setting. All individual alignments containing all 17 taxa (n = 187) were concatenated into a single supermatrix (2,055,035 bp). A maximum likelihood phylogeny was inferred from the supermatrix with GARLI version 2.01 (ref. 63). A partitioned model was used that allowed each locus an independent substitution rate, while all loci shared a single general time-reversible nucleotide substitution model with gamma-distributed rate heterogeneity.

The maximum likelihood phylogeny inferred by GARLI was rooted using the outgroup L. perrieri, which was subsequently pruned from the tree. The tree of...
16 Oryza taxa, with the maximum-likelihood branch length estimates obtained from GARLI, was input to PATHd8 v1.0 (ref. 64). To time calibrate the phylogeny, the divergence time of the genus Oryza (the root of the tree) was fixed for the PATHd8 analyses, using a divergence time consistent with several recent studies (15 MYA[65–67]). A time-calibrated ultrametric tree was output by PATHd8. The resulting dated phylogeny was pruned down to the taxon set of interest in this study.

**RNA transcript and sRNA analyses.** Strand-specific RNA-Seq reads from three organs (leaf, root and mixed stage panicle) across 14 Oryza species and L. perrieri were assembled with Trinity[68] to obtain a collection of assembled transcripts. These transcripts were aligned to the two OsatBV component genomes using Bowtie 2 (ref. 69).

Predictions of promoter elements in the florendovirus pregenomic RNA were made by submitting the region of sequence spanning the end of ORF2 and the 3′RNA[71]A as a consensus sequence to analysis using the BDGP Neural Network Promoter Prediction Web site (http://www.fruitfly.org/seq_tools/promoter.html). sRNAs (21–24 nt of florendovirus origin were searched in different tissue types (leaves, flowers, fruits, stolons, xylem) using data and tools provided by the Comparative Sequencing of Plant Small RNAs Web site (http://smallrna.udeu.edu). Reads matching florendovirus sequences with no mismatch were mapped on reconstituted viral genomes using Mosaic version 1.1.021 (http://code.google.com/p/mosaik-aligner/). Density plot and cartography of the reads on viral genomes were generated using S-MART version 1.16 (ref. 70).

**Promoter Prediction Web site (http://www.fruitfly.org/seq_tools/promoter.html).** Predictions of promoter elements in the florendovirus pregenomic RNA were made by submitting the region of sequence spanning the end of ORF2 and the 3′RNA[71]A as a consensus sequence to analysis using the BDGP Neural Network Promoter Prediction Web site (http://www.fruitfly.org/seq_tools/promoter.html).

**References**

1. Brown, J. R. Ancient horizontal gene transfer. Nat. Rev. Genet. 4, 121–132 (2003).
2. Keeling, P. & Palmer, J. Horizontal gene transfer in eukaryotic evolution. Nat. Rev. Genet. 9, 605–618 (2008).
3. Feschotte, C. & Gilbert, C. Endogenous viruses: insights into viral evolution and impact on host biology. Nat. Rev. Genet. 13, 283–296 (2012).
4. Chiba, S. et al. Widespread endogenization of genome sequences of non-retroviral RNA viruses into plant genomes. PLoS Pathog. 7, e1002146 (2011).
5. Teycheney, P.-Y. & Geering, A. D. W. in Recent Advances in Plant Virology (eds Caranta, C., Aranda, M. A., Tepfer, M. & López-Moya, J. Y.) 343–362 (Caister Academic Press, 2011).
6. Liu, H. et al. Widespread horizontal gene transfer from double-stranded RNA viruses to eukaryotic nuclear genomes. J. Virol. 84, 11876–11887 (2010).
7. Cohoon, C. J., Lock, W. M. & Major, J. Endogenous viral LTRs as promoters for human genes: a critical assessment. Gene 448, 105–114 (2009).
8. Faulkner, G. J. et al. The regulated retrotransposon transcriptome of mammalian cells. Nat. Genet. 41, 563–571 (2009).
9. Li, J. F. et al. Mouse endogenous retroviruses can trigger premature transcriptional derepression at a distance. Genome Res. 22, 870–884 (2012).
10. Isbel, L. & Whitelegge, J. Endogenous retroviruses in mammals: an emerging picture of how ERVs modify expression of adjacent genes. Bioessays 34, 734–738 (2012).
11. Macfarlan, T. S. et al. Endogenous retroviruses and neighboring genes are coordinately repressed by LSD1/KDM1A. Genes Dev. 25, 594–607 (2011).
12. Dungrass, A. J. et al. Synctin A knock-out mice demonstrate the critical role in placentaation of a fusogenic, endogenous retrovirus-derived, envelope gene. J. Virol. 86, 7151–7159 (2012).
13. Laco, G. S. & Beachy, R. N. Rice tungro bacilliform virus encodes reverse transcriptase, DNA polymerase, and ribonuclease H activities. Proc. Natl Acad. Sci. USA 91, 2654–2658 (1994).
14. Jakowitsch, J., Mette, M. F., van der Winden, J., Matzke, M. A. & Matzke, A. J. M. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res. 19, 1419–1428 (2009).
15. Zlotorynski, E. et al. Molecular basis for expression of common and rare fragile sites. Mol. Cell. Biol. 23, 7557–7565 (2003).
16. Doillon, L. W., Pierce, C. A., Smith, M. G. Y. & Wang, Y.-H. Role of DNA secondary structures in fragile site breakage along human chromosome 10. Hum. Mol. Genet. 22, 1443–1456 (2013).
17. Huertas, P. DNA resection in eukaryotes: deciding how to fix the break. Nat. Struct. Mol. Biol. 17, 11–16 (2010).
18. Lim, K. Y. et al. Sequence of events leading to near-complete genome turnover in aphid-borne Nicotiana within five million years. New Physiol. 175, 756–763 (2007).
19. Ma, I., Devos, K. M. & Bennetzen, J. L. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res. 14, 860–869 (2004).
20. Gandon, M. A. et al. Oldest known Eucalyptus macrofossilis are from South America. PLoS ONE 6, e21084 (2011).
21. Coxall, H. K., Wilson, P. A., Palike, H., Lear, C. H. & Backman, J. Rapid stepwise onset of Antarctic glaciation and deeper calcification in the Pacific Ocean. Nature 433, 53–57 (2005).
22. Livermore, R., Nankivell, A., Eagles, G. & Morris, P. Paleogene opening of Drake Passage. Earth Planet. Sci. Lett. 27, 47–57 (2010).
23. Olmstead, R. G. Phylogeny and biogeography in Solanaceae, Verbenaceae and Bignoniaceae: a comparison of continental and intercontinental diversification patterns. Bot. J. Linn. Soc. 171, 80–102 (2013).
24. Ladiges, P. Y., Marks, C. E. & Nelson, G. Biogeography of Caulimoviridae. J. Virol. 43, 461–472 (2009).
25. Kejnovsky, E., Leitch, I. J. & Leitch, A. R. Contrasting evolutionary dynamics between angiosperm and mammalian genomes. Trends Ecol. Evol. 24, 572–582 (2009).
51. Zdobnov, E. M. & Apweiler, R. InterProScan—an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17, 847–848 (2001).
52. Marchler-Bauer, A. et al. CDDB: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res. 39, D225–D229 (2011).
53. Tamura, K. et al. MEGAS5: Molecular Evolutionary Genetics Analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731–2739 (2011).
54. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56, 564–577 (2007).
55. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574 (2003).
56. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690 (2006).
57. Nylander, J. A. A., Wilgenbusch, J. C., Warren, D. L. & Swoford, D. L. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics 24, 581–583 (2008).
58. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. Bioinformatics 21, i351–i358 (2005).
59. Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. Considering transposable element diversification in de novo annotation approaches. PLoS ONE 6, e16526 (2011).
60. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27, 573–580 (1999).
61. Zwickl, D. J., Stein, I. C., Wing, R. A., Ware, D. & Sanderson, M. J. Disentangling methodological and biological sources of gene tree discordance on Oryza (Poaceae) chromosome 3. Syst. Biol. 63, 645–659 (2014).
62. Löytynoja, A. & Goldman, N. An algorithm for progressive multiple alignment of sequences with insertions. Proc. Natl Acad. Sci. USA 102, 10557–10562 (2005).
63. Zwickl, D. J. Genetic Algorithm Approaches for the Phylogenetic Analyses of Large Biological Sequence Datasets Under The Maximum Likelihood Criterion (PhD Thesis. Univ. Texas Austin, 2006).
64. Britton, T., Anderson, C. L., Jacquet, D., Lundqvist, S. & Bremer, K. Estimating divergence times in large phylogenetic trees. Syst. Biol. 56, 741–752 (2007).
65. Tang, L. et al. Phylogeny and biogeography of the rice tribe (Oryzeae): evidence from combined analysis of 20 chloroplast fragments. Mol. Phylogenet. Evol. 54, 266–277 (2010).
66. Zou, X. H., Yang, Z., Doyle, J. J. & Ge, S. Multilocus estimation of divergence times and ancestral effective population sizes of Oryza species and implications for the rapid diversification of the genus. New Phytol. 198, 1155–1164 (2013).
67. Jacquemin, J. et al. Fifteen million years of evolution in the Oryza genus shows extensive gene family expansion. Mol. Plant. 7, 642–656 (2014).
68. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652 (2011).
69. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).
70. Zytnicki, M. & Quesneville, H. S-MART, A software toolbox to aid RNA-seq data analysis. PLoS ONE 6, e25988 (2011).

Acknowledgements
We thank the International Peach Genome Initiative (IPGI) for sharing sequence data before publication, Thierry Candresse and Mark Tepfer for fruitful discussions, Michael J. Sanderson for assistance with the Oryza phylogenetic analysis, and OEDC Co-operative Research Programme for financial support. P.-Y.T. is supported by the European Regional Development Fund.

Author contributions
A.D.W.G. and F.M. independently discovered the endogenous florendoviral elements, both assisted with assembling the virus genomes and are equal first authors. M.Z. and P.-Y.T. did the sRNA analyses; A.R.Mc.T., D.J.Z. and A.D.W.G., the phylogenetic analyses; F.M., D.C., N.C., S.S. and S.V., the plant genome mapping; A.D.W.G. and D.C., the transcripts analyses and pairwise sequence comparisons; P.-Y.T., R.W. and H.Q., project coordination; and A.D.W.G., F.M., D.C. and P.-Y.T. have written the manuscript.

Additional information
Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Geering, A. D. W. et al. Endogenous florendoviruses are major components of plant genomes and hallmarks of virus evolution. Nat. Commun. 5:5269 doi: 10.1038/ncomms6269 (2014).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/