RECONSTRUCTING DE NOVO SILENCING OF AN ACTIVE RETROTRANSPOSON IN ARABIDOPSIS

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. Sc. ETH Zurich)

presented by

ARTURO MARÍ ORDÓÑEZ

Enginyer Agrònom
Universitat Politècnica de València (UPV)
València, Spain

Born on 12.10.1984

Citizen of Spain

Accepted on the recommendation of:

Prof Dr. Olivier Voinnet, examiner
Prof. Dr. David Baulcombe, co-examiner
Prof. Dr. Ueli Grossniklaus, co-examiner
Prof. Dr. Markus Stoffel, co-examiner

2014
Acknowledgements

I would like to acknowledge everyone who has contributed to the success of this thesis, at both scientific and human aspects, inside and outside the lab. Thanks to all of you, the following pages are also yours. Special thanks to:

My parents and family, for their sacrifice, support, understanding and for always believing in me.

Olivier, for his trust, brilliant ideas and discussions (either at the office, restaurant or bar) and for always having his mind and door open to listen to me. It's been an honour to work with you.

The whole Voinnet team members (past and present), in special: Chris, Rebecca, Brande, Peter, Lionel, Ana, Pauline, Nate, Nico, Antonin, Alexis, Stefan and many others who taught me not only science and that made me feel part of a small family, even far away from home (and minimizing the amount of times I forgot my belongings at random places).

Last, but not less important, Amparo, Arnau, Altea, Neus, Jordi, Vlcenta, Mar, Clara, Carles (and the rest of the Valencia crew),... for their unconditional friendship, for being always there, and for teaching me that home is where the people you care are.

Thanks.
Contents

Abstract.............................................................................................................................. 1
Résumé............................................................................................................................... 2
Chapter 1: Introduction...................................................................................................... 5
  1.1. Transposable Elements......................................................................................... 7
  1.2. Structure and Classification of Transposable elements.................................. 9
    1.2.1. Class I elements (Retrotransposons)......................................................... 9
    1.2.1.1. LTR Retrotransposons........................................................................... 10
    1.2.1.2. Non-LTR Retrotransposons................................................................. 12
    1.2.2. Class II elements (DNA transposons)...................................................... 13
    1.2.2.1. Subclass I............................................................................................. 13
    1.2.2.2. Subclass II........................................................................................... 13
    1.2.3. Non-autonomous elements....................................................................... 15
  1.3. Impact of Transposable Elements in Genome Evolution............................. 15
    1.3.1. Functional impact of TE mobilization...................................................... 15
    1.3.2. TEs as source of genomic innovation...................................................... 16
    1.3.3. Domestication of TE genes...................................................................... 17
  1.4. Regulation of Transposable Elements Activity.............................................. 18
    1.4.1. Transcriptional Gene Silencing: Chromatin Modifications.................... 19
      1.4.1.1. DNA methylation.............................................................................. 19
      1.4.1.2. Histone modifications..................................................................... 21
      1.4.1.3. Interaction between chromatin modifications and DNA methylation 22
    1.4.2. Post-Transcriptional Gene Silencing......................................................... 22
      1.4.2.1. General aspects of RNA silencing.................................................... 23
      1.4.2.2. Germline specific RNA silencing: PIWI pathways.......................... 26
    1.4.3. RNA-mediated chromatin modifications............................................... 29
      1.4.3.1. RNA-directed chromatin modifications in animals......................... 30
      1.4.3.2. RNA-directed DNA methylation in plants...................................... 31
  1.5. Aim of the project and biological model of choice......................................... 34
  1.6. References........................................................................................................... 38
Chapter 2: Results.......................................................................................................... 53
  2.1. Reconstructing de novo silencing of an active plant retrotransposon.......... 55
    Abstract.................................................................................................................... 58
    Introduction.............................................................................................................. 59
    Results..................................................................................................................... 60
    Discussion............................................................................................................... 74
    Methods.................................................................................................................. 77
2.2. Mandatory splicing and premature termination of *Copia*-like retrotransposons triggers RNA silencing in *Arabidopsis*.

### References

Supplementary information

Addenum

### Chapter 3: Discussion

3.1. Discussion and perspectives

3.2. References

List of abbreviations
Abstract

Transposable elements (TEs) or “transposons” are genetic parasites found in the DNA of virtually every organism inspected so far. TEs possess the unique ability of changing location within host genomes and to make copies of themselves. These characteristics have made transposons a major source of genetic variability and innovation during evolution, such that TEs comprise, by far, the largest fraction of eukaryotic genomes. However, transposons can also pose a serious threat by interfering with, or disrupting gene expression upon their mobilization. Hence, organisms have developed strategies to control TE activity, which generally include a combination of chromatin modifications and RNA-based processes collectively named “RNA silencing” that prevent transcription (transcriptional gene silencing; TGS) or degrade TE-derived transcripts (posttranscriptional gene silencing; PTGS). At the heart of RNA silencing are 20-30-nt-long RNA molecules (small RNAs) that direct silencing effector complexes to initiate RNA turnover and chromatin modifications in a highly nucleotide sequence-specific manner.

The present study was designed to gain insights into how chromatin- and RNA-based pathways cooperate to recognize and silence newly invading TEs in the model plant *A. thaliana*. We have reactivated and meticulously followed the fate of the single-copy retrotransposon Évadé (*EVD*), representing a class of TEs that copy their RNA into DNA by reverse transcription, to enable their integration into the host genome. Our results show that, upon reactivation of *EVD*, plant cells exploit a unique transposon signature to initiate PTGS. This signature is created by the mandatory requirement for *EVD* to produce excess levels of an indispensable structural protein called Gag. Gag, in turn, protects the *EVD* RNA from PTGS, allowing the TE to successfully colonize the genome. Nonetheless, the steady increase in *EVD* copy number over generations eventually leads to the abnormal accumulation of particular PTGS-related RNA intermediates, allowing the TGS machinery to now take over. TGS-related sRNAs direct DNA methylation first in the body of *EVD*, and, following a few generations, the long-terminal repeats containing the transcriptional signals are affected, resulting in the permanent silencing of the transposon. By reconstructing a transposon burst at an experimental scale, we have thus deciphered the molecular underpinnings of the detection, proliferation and eventual demise of a plant TE.
Résumé

Les éléments transposables (ETs), ou ‘transposons’, sont des parasites génétiques présents dans l’ADN de la plupart des organismes étudiés. Ils possèdent la faculté unique de proliférer au sein des génomes qu’ils envahissent en se recopiant à l’identique, et s’insérant à de nouvelles positions. Ces caractéristiques expliquent le rôle majeur joué par les ETs dans l’évolution des génomes, à laquelle ils contribuent en générant une grande diversité et flexibilité génétiques, au point que les ETs représentent la fraction la plus importante des génomes eucaryotes. Cependant, la prolifération des ETs constitue aussi un danger pour leurs hôtes puisqu’elle est susceptible d’endommager l’intégrité ou l’expression des gènes. Par conséquent, les organismes ont élaboré diverses stratégies pour contrôler les ETs, reposant en général sur une combinaison de modifications de la chromatine et de processus liés à l’ARN collectivement dénommés « RNA silencing » et agissant au niveau de la transcription (« transcriptional gene silencing » ou TGS) et/ou de la stabilité des transcrits des TEs (« post-transcriptional gene silencing » ou PTGS). Au cœur du RNA silencing sont de petits ARNs de 20 à 30 nucléotides qui servent de guides séquence-spécifiques à des complexes protéiques effecteurs initiant la dégradation de l’ARN ou les modifications de la chromatine susmentionnés.

En employant la plante A. thaliana comme modèle expérimental l’étude présentée ici vise à mieux comprendre les mécanismes moléculaires par lesquelles ces deux niveaux de régulation coopèrent dans la reconnaissance et l’inactivation d’ETs actifs qui envahissent le génome de novo. A cette fin, nous avons réactivé et méticuleusement suivi le devenir d’un retrotransposon appelé Évadé (EVD), présent en une seule copie dans le génome d’Arabidopsis. Les retrotransposons copient leur ARN en ADN par transcription reverse de façon à s’intégrer dans le génome hôte. Nos résultats démontrent que la cellule exploite une signature moléculaire unique d’EVD pour initier le PTGS contre ce dernier. Cette signature est inhérente à la nécessité pour EVD, et probablement d’autres ETs, de produire la protéine Gag en excès. Gag, en contrepartie, protège physiquement l’ARN d’EVD contre l’action du PTGS, permettant ainsi sa prolifération au gré de générations successives. Cependant, l’accroissement incessant du nombre de copies d’EVD atteint une limite buttoir correspondant à l’accumulation anormale d’une molécule intermédiaire du
PTGS, indirectement dérivée d’\textit{EVD} lui-même. Cette accumulation permet à la machine du TGS de prendre le relais en méthylant d’abord du cœur de l’ADN d’\textit{EVD}, puis ses régions promotrices après quelques générations, résultant en son inactivation transcriptionnel durable. En reconstruisant un événement de transposition expérimentale nous avons donc éclairci les mécanismes moléculaires qui sous-tendent la détection, la prolifération et l’inactivation finale d’un élément transposable actif.
Chapter 1: Introduction
1.1. Transposable Elements

Transposable elements (TEs) are fragments of DNA with the ability to move throughout genomes by virtue of changing their position or making new copies of themselves. TEs were first discovered by Barbara McClintock in the 1940s and 1950s while studying the mechanism behind the unstable mosaic colour patterns of maize grains in kernels. Her previous experience as a cytogeneticist, detailed microscopic observations and fine genetic experiments, allowed her to uncover the presence of mobile genetic elements (controlling elements or TEs), able to modify the expression of neighbouring genes upon their dissociation or insertion within chromosomes\textsuperscript{1,2}. Her results challenged the static concept of the genome as a mere carrier of genetic information, as she envisioned that “the action of genes had to be and was regulated”\textsuperscript{3}. However, her work was poorly understood and due to continued scepticism directed towards her results and the implications of this, she stopped publishing her research on transposable elements in 1953\textsuperscript{4}. The significance of her discoveries were not recognised until decades later. The work of François Jacob and Jacque Monod in 1961 describing the regulation of the lac operon in bacteria, demonstrating that the synthesis of proteins was largely regulated at the genetic level\textsuperscript{5}, the development of molecular biology and the discovery and characterization of other mobile genetic elements, revealed the importance of McClintock’s research. She was subsequently awarded the Nobel Price for Physiology or Medicine in 1983. However, her concept of TEs as controlling elements and their role in evolution remained poorly understood. The first investigations of genomes revealed that their size did not correlate with the number of genes or complexity of the organism (the C-value paradox)\textsuperscript{6,7}. The fact that much of eukaryotic genomes were full of non-functional DNA “junk” whose repeated nature was largely due to the amplification of transposons led to the concept of “selfish or parasitic DNA” to describe TEs\textsuperscript{8,9}. At present, we have only just begun to appreciate the profound role that TEs have played, and indeed still play, in genome evolution and modulation of its expression.

With the advent of the genomic era, and the development of next-generation sequencing (NGS) technologies, an exponential number of genomes are being sequenced, revealing that TEs are present in all genomes inspected so far and that
they represent the single largest component of the genetic material of most eukaryotes (Table 1). While some organisms display very low amounts of TEs, such as

### Table 1. Transposon content (%) in sequenced genomes of different eukaryotes. The data presented in this table has been obtained from the most recent genome sequencing publications or databases. Color bars (% TE): Blue ≤ 10% < orange ≤ 33% < red ≤ 66% < brown.
as honey bee\textsuperscript{10}, the carnivore plant \textit{U. gibba}, \textsuperscript{45} or yeast\textsuperscript{64}, to date the protozoan parasite \textit{Plasmodium falciparum} (malaria) and the fungi \textit{Ashbya gossypii} are the only organisms found to be devoid of transposons\textsuperscript{62,65}. \textit{A. gossypii} contains a very small genome (9.2 Mb) where only one gene displays similarities to the reverse transcriptase of some transposons, suggesting that at some time during evolution, a process efficiently eliminated most of TEs from the genome\textsuperscript{62}.

1.2. Structure and Classification of Transposable Elements

With the progress of NGS, our knowledge of the structure and composition of genomes has significantly progressed. It has also revealed that TEs display extreme diversity, with thousands or even tens of thousands of different families now known to exist\textsuperscript{66}. Historically, TE classification has been carried out on a species-specific basis, lacking any systematic rules\textsuperscript{65}. The first classification of TE's was proposed in 1989 by Finnegan based on their transposition mechanism: Class I TEs (or retrotransposons), which transpose through a “copy-and-paste” mechanism via an RNA intermediate; Class II TEs (or DNA transposons) where the transposition intermediate is DNA and the transposition mechanism commonly known as “cut-and-paste”\textsuperscript{67}. However, the discovery of non-autonomous elements\textsuperscript{67,68} and a better understanding of their genetic organization as well as the molecular mechanisms underlying their transposition mechanisms have challenged such simple classification\textsuperscript{69}. Recently, in an effort to reconcile the classical classification with the current knowledge of TE biology, Shulman and colleagues have proposed a unified classification system for eukaryotic transposable elements\textsuperscript{65}. The nomenclature used here is based in their classification system with minor additions based on recent findings (Figure 1).

1.2.1. Class I elements (Retrotransposons)

All retrotransposons transpose via an RNA intermediate transcribed from a genomic copy of the TE and reverse transcribed into DNA by a TE-encoded reverse transcriptase (RT). The resulting DNA is then integrated into the genome to generate a new copy, generally inducing the duplication of the targeted sequence for integration (target site duplication, TSD), which can as well be used as a hallmark or footprint of TE insertion\textsuperscript{69}. Due to this “copy-and-paste” mechanism, retroelements
generally represent the major fraction of repetitive sequences in genomes. They can be divided into subclasses based on the presence or absence of long terminal repeats (LTRs), transposition features and organization of their coding sequence.

1.2.1.1. LTR Retrotransposons

LTR-retrotransposons (LTR-RTEs) are very similar to retroviruses, sharing structural and mechanistic features to fulfill their life cycle within cells. They are flanked by LTRs of variable length and typically contain two open reading frames (ORFs): one for GAG, encoding the structural proteins required to form virus-like particles (VLPs), and a second for POL encoding the enzymatic activities required for post translational processing of retroviral proteins, reverse transcription and integration. Generally, POL contains an aspartic protease (PR), integrase (IN), reverse transcriptase (RT), and RNase H (RN) frequently covalently bound to RT. All these elements are usually translated as a single Pol polyprotein subsequently post-translationally processed into the individual components by the Pr, which has the property of self-dissociate from the polypeptide. Similarly to retroviruses, LTR-RTEs also contain sequences for dimerization of the RNA template for RT of genomic RNA (gRNA), packaging into VLPs (Psi) and integration. In addition, they can contain extra ORFs.

LTR-RTEs are divided into superfamilies according to their genetic organization. The two main superfamilies, Copia and Gypsy, differ in the order of IN and RT-RNase within POL (Figure 1). Some TEs such as Bel-Pao, only found in metazoans, have the same structure as Gypsy elements, but their RT differs enough to form a different group. However, for simplicity, they are included here in the Gypsy superfamily. Endogenous retroviruses (ERV) resemble Gypsy LTR-RTE but contain a third ORF encoding an envelope protein (ENV), required for cell-to-cell transmission. As such, ERVs represent infectious elements. It remains unclear if ERVs are Gypsy elements that captured ENV genes during evolution or if they become non-infectious (cell-autonomous) by loosing the proteins required for extracellular mobility. A recent investigation of 38 mammalian genomes revealed that ERVs that have become classical LTR-RTEs through the loss of their ENV ORF, boosted their proliferation in their host genomes by a factor of ~30. This conversion of ERVs into genomic ‘super-spreaders’ happened independently for different elements in different lineages, suggesting that, at least in mammals, ENV loss rather
Figure 1. Classification of transposable elements. TEs are divided into two main categories based on the presence or absence of an RNA intermediate. They are further subdivided into subcases, orders and superfamilies. Superfamilies are named after their most representative element. **LTR**: long terminal repeat; **DIR**: direct inverted repeat; **SDR**: split direct repeat; **TIR**: terminal inverted repeat. **GAG**: nucleocapsid protein; **PR**: aspartic protease; **IN**: integrase; **RT**: reverse transcriptase; **RN**: RNase H; **ENV**: envelope protein; **YR**: tyrosine recombinase; **EN**: endonuclease; **ORF**: open reading frame of unknown function; **TPase**: transposase; **CYP**: cysteine protease; **POL B**: DNA polymerase B; White boxes represents variable number of proteins found within elements of the same superfamily. **P**: plants; **M**: mammals; **F**: fungi; **O**: others.

Adapted from Wicker et. al. 2007, Structure of the plant specific Sirevirus superfamily was adapted from Bousios et. al. 2012.
than gain is a general phenomenon\textsuperscript{80}. In plants, the presence of Gypsy elements with a third ENV-like ORF harbouring surface and transmembrane domains has been documented\textsuperscript{74}, however their functionality is still under debate given that the presence of cell walls in plant cells evidently poses an impediment for cell-to-cell infectivity. Furthermore, to date, no retroviruses have been found to infect plants.

The last superfamily of LTR-RTEs corresponds to the recently discovered Sireviruses (represented by the Opie element), only present in some plants such as maize, rice, Arabidopsis and soybean. They share the same organization as Copia elements, however they also contain an ENV-like ORF encoded after POL together with their own cis-regulatory elements\textsuperscript{81,82}. Sireviruses represent an enigma in plants due to their intriguing characteristics: they display high levels of sequence conservation regardless of the evolutionary distance of their host, and the ENV-like gene is maintained\textsuperscript{83}. Although they have been poorly studied, it is tempting to speculate that they might represent plant “retroviruses” with the ENV-like ORF potentially conferring an infectious potential, perhaps by allowing their transmission by insect vectors.

1.2.1.2. Non-LTR Retrotransposons

As opposed to LTR-RTEs, non-LTR retrotransposons lack canonical LTRs, or are flanked by different types of repeated sequences. Within this class different orders of retroelements are included.

The first order, DIRS-like (Dictyostelium intermediate repeat sequence) elements, main feature is the substitution of the IN by a tyrosine recombinase (YR), hence they integrate in the genome through site-specific DNA recombination instead of direct DNA transfer. Their 3’ and 5’ ends are generally flanked by split direct repeats (SDR) or direct inverted repeats (DIR). They have been found in a broad range of organisms from green algae to animals and fungi\textsuperscript{84,85}.

Penelope (Ple)-like elements lack GAG and PR, and encode a RT that is more similar to telomerases than to retroviral RT, as well as for an endonuclease (EN) similar to bacterial DNA repair proteins. They also posses LTR-like sequences in either a direct or inverted orientation\textsuperscript{86}.

Long interspersed nuclear elements (LINE) encompass multiple families (such as R2, RTE, L1, or I). In general, they contain a single ORF containing both EN and RT, but may also contain a GAG-like ORF (ORF1). However, only I-elements within
this class encode for RNase\textsuperscript{65,87}. \textit{LINE} elements are very abundant in animals, with approximately 20\% of the human genome being composed of \textit{L1} elements\textsuperscript{10}. Their reverse transcription takes place on the DNA by priming the RT with DNA from the open target site\textsuperscript{87}.

Short interspersed nuclear elements (\textit{SINE}), epitomized by the \textit{Alu} element, are non-autonomous elements that rely on \textit{LINE}s for transposition\textsuperscript{88,89}. Although many non-autonomous elements are the result of sequence deletions from ancestral autonomous elements, this is not the case for \textit{SINE}. They contain Pol III promoters and originated from tRNA, 5S or other Pol III transcribed genes\textsuperscript{90}. Since they require reverse transcription from an RNA precursor, they are officially classified as retroelements.

\textbf{1.2.2. Class II elements (DNA transposons)}

DNA transposons transpose through a DNA “cut-and-paste” mechanism not involving any RNA intermediates. There are two main subclasses of DNA transposons depending on the number of DNA strands that are cut during transposition.

\textit{1.2.2.1. Subclass I}

The classical DNA transposons, present in all kingdoms, includes families such as \textit{Mariner}, \textit{hAT}, \textit{Mutator}, \textit{P}, \textit{PiggyBac} and \textit{CACTA}, to name a few. They generally encode for a transposase (TPase) flanked by terminal inverted repeats (TIRs). The TPase recognizes the TIRs and cuts the DNA at both strands to initiate the transfer of the TE into a new location. Some elements, like \textit{CACTA}, can display a second ORF of unknown function\textsuperscript{65}. They can increase their copy number, for example, during DNA replication, by transposing from a region that has already been replicated to another that has not. Consequently, one of the daughter DNA molecules will display the same TE copy number as the mother DNA strand, but in a different location, while the daughter DNA strand will display a duplication of the TE at the original and new site\textsuperscript{91}.

\textit{1.2.2.2. Subclass II}

As opposed to subclass I DNA transposons, subclass II DNA TEs transposition only involves the cleavage of one DNA strand during their replication. The most prominent
1.2.3. Non-autonomous elements

Following the description given by Wicker et al., non-autonomous TEs are defined as “any group of elements that lacks some (or all) of the domains found in autonomous elements, displaying a highly degenerate coding region, or even lacking coding capacity”. They usually appear as result of aberrant transposition, genomic recombination events, or mutations. Some non-autonomous elements can be reactivated and may transpose by parasitizing autonomous elements (e.g. the LTR-RTE BARE2 in barley carries a deletion in the first GAG ATG codon that inactivates its GAG translation capacity, however it parasitizes the homologous BARE1 to fulfil its transposition). The most common non-autonomous TEs are MITEs (miniature inverted-repeat transposable element), short elements flanked by TIRs usually reactivated by class II DNA transposons carrying similar TIRs. Other non-autonomous elements are LARDs (large retrotransposon derivates), SNACs (small non-autonomous CACTA transposon) and TRIMs (terminal repeat retrotransposon in miniature).

1.3. Impact of Transposable Elements on Genome Evolution

As initially observed by McClintock, the mobilization of TEs can have profound consequences on gene expression. In the last few decades, our concept of TEs as mere “junk” or “parasitic DNA” has changed; our knowledge of their transposition mechanisms, detailed analysis of the impact of TEs on gene expression and the study of sequenced genomes has revealed the two faces of transposable elements.
“of good and evil”, because they contribute to shape genomes through evolution but also cause mutations and diseases.

1.3.1. Functional impact of TE mobilization

The ability of TEs to change location or to generate new copies of themselves at different genomic positions intrinsically poses a threat for their host, as they can impact or disrupt gene activity upon transposition. Insertion in the coding sequence can cause transcript truncation by introducing termination and polyadenylation signals, exon interruption, altered splicing behaviour (enhancing, reducing, or generating new splicing patterns) or modification of the coding frame. When integrated nearby the coding region, TE can disrupt the normal expression pattern of genes by disturbing regulatory sequences (e.g. promoter, enhancers) or introducing new ones. Furthermore, not only integration, but also excision can lead to mutations. Helitrons, for example, often capture gene fragments when leaving their original locus, and other DNA transposons can generate allelic variants by imperfect excision when transposing (reviewed in Huang et. al. 201298). In addition, TEs are largely regulated by epigenetic mechanisms through chromatin marks that impose transcriptional silencing99 (see section 1.4). Such silencing marks can spread from new insertions into neighbouring genes, thereby affecting their expression. TEs can even promote mutations and chromosome rearrangements leading to diseases. Alterations caused by transposable elements have been estimated to be responsible for 0.5-1% of human illnesses, in particular cancers100-102. Haemophilia, some types of muscular dystrophy and certain breast and reproductive organ cancers result from aberrant SINE and LINE activities102. In Drosophila, about half of the spontaneous mutations with phenotypic effects are caused by TE insertions. Similarly, mobilization of LTR-RTEs in mouse accounts for 15% of germ-line mutations103 and can explain many spontaneous mutations in plants41. This mutagenic ability of TEs has been extensively exploited in many model organisms as a tool for functional genomic studies, by generating mutants or integrating reporter genes into endogenous loci104-106.
1.3.2. TEs as source of genomic innovation

TE activity and mobilization, however, is not always detrimental. A rough estimation suggests that, depending on the element, TEs transpose at a frequency of $10^{-3}$ to $10^{-5}$ per element per generation, which is higher than the nucleotide-base substitution rate of around $10^{-8}$ to $10^{-9}$ per nucleotide per generation\(^{100}\). Thus, transposons possess the ability to generate higher variability to be selected by evolution than random DNA mutations. Prime examples are found in plants where a number of beneficial traits selected through domestication of crop species are caused by TE insertions. Examples include: tomato, where a gene duplication caused by a transposon resulted in larger fruit shape\(^{107}\); the loss of pigment in white grapevine by the insertion of Grelt (Gypsy-like) in the VvMYBAI gene\(^{108}\); reduction of maize branch number by increased apical dominance due to the insertion of a Copia-like element (hopscotch) upstream of the tb gene\(^{109}\); parthenocarpy (development of fruit without fertilization) in apple by the disruption of MdPI (PISTILLATA) caused by deml (LTR-RTE)\(^{109}\), and many others (an updated list of 51 TE-induced phenotypic traits involved in speciation or domestication of cultivated plants can be found in Vitte et al., 2014\(^{41}\)). Waves of TE transposition or “bursts” have been suggested to play an important role in the speciation of rodents\(^{110}\). In Drosophila, thirteen TE insertions that resulted in regulatory changes of nearby genes have been found crucial to the environmental adaptation of D. melanogaster after its expansion from Africa less than 15,000 years ago\(^{111,112}\). Likewise, many promoters and polyadenylation signals in mammalian genes derive from TEs, and 25% of experimentally characterized human promoters carry TE-derived sequences\(^{113}\).

Transposons have also been proposed to play a relevant role in the evolution of regulatory networks\(^{114}\). Through rounds of TE amplification, their regulatory motifs can be simultaneously recruited at many genomic locations introducing multiple genes into the same network\(^{114,115}\). An additional mechanism by which TEs can wire new regulatory networks is through the production of non-coding small RNAs (sRNAs, see section 1.4.2 for further development on sRNAs pathways), which are important regulators of eukaryotic gene expression. A subclass of sRNAs, the microRNAs (miRNAs), silence gene expression by targeting for degradation or translational repression messenger RNAs (mRNAs) bearing miRNA-complementary sequences. miRNAs are produced from MIRNA genes expressed as independent
transcription units under the form of non-coding primary transcripts invariably containing one or several stem-loop structures\textsuperscript{116,117}. Although the origin of MIRNA genes remains obscure, some have been proposed to arise from gene duplication leading to the formation of inverted repeats (IR), and several lines of evidence also suggest that some may also be of TE-origin\textsuperscript{117}, and indeed, several mammalian, worm and plant miRNA precursors derive, or contain, TE sequences\textsuperscript{118-120}. Some TEs harbour features that make them prone to form hairpin structures with the potential to become MIRNA genes. For example, MITEs carry TIRs (terminal inverted repeats) separated by short sequences, which, upon transcription and intramolecular folding, may form hairpins processed into sRNAs\textsuperscript{117,121}.

As mentioned above, TE silencing through epigenetic modifications can influence the expression of nearby genes. In melon, the insertion of a DNA transposon near \textit{CmWIPI} leads to silencing due to spreading of DNA methylation into the promoter region, while the allele of \textit{CmWIPI} lacking the insertion is normally expressed. Expression of \textit{CmWIPI} results in the development of unisexual male flowers, whereas its silencing causes the development of female flowers\textsuperscript{122}. This illustrates how the generation of an epigenetic (epi)allele after TE insertion has been co-opted for sex determination. A recent study in \textit{Drosophila} shows that upon loss of silencing, genes neighbouring TEs will often display increased expression\textsuperscript{123}. Another way by which organisms have exploited the epigenetic regulation imposed on TEs is to modulate paternal and maternal contribution to gene expression during fertilization or early development. This process is called imprinting, and implies that certain genes are expressed depending of their parent-of-origin\textsuperscript{124}. In \textit{Arabidopsis}, imprinted genes in the endosperm, (the embryo-nourishing tissue generated by the fusion of an haploid sperm cell of paternal origin and the female diploid central cell) are differentially expressed due to different epigenetic marks, and such genes are often associated with TE or TE-derived fragments\textsuperscript{125,126}.

### 1.3.3. Domestication of TE genes

Domestication of transposable element genes refers to the process by which TE proteins or domains are co-opted into functional host proteins. This phenomenon has happened on multiple independent occasions in different organisms. TPases, Gag and Env are the most common TE-encoded proteins to be captured\textsuperscript{114,127,128}.
Transposases have been recurrently domesticated to provide DNA-binding domains (DBD) for many regulatory purposes. In Drosophila, the BEAF-32 insulator protein, required for connecting the chromatin to the nuclear matrix, is derived from the hAT TPase. In Arabidopsis, two transcription factors involved in far-red light perception and signalling (FHY3 and FAR1) are entirely derived from Mutator-like TPases. In mammals and yeast, the pogo-like TPase gave rise to the CENP-B homologues, centromere-binding proteins involved in chromosome segregation. RAG1 and RAG2, which mediate the DNA recombination of variable, diverse, and joining (V(D)J) regions required for the production of the majority of immunoglobulin types in vertebrates, are also derived from DNA transposon TPases.

Many LTR-RTE-derived proteins exert key cellular functions. Telomerases, which maintains the integrity of the chromosomal ends, have been shown to derive from reverse transcriptases. Gag proteins have also been recurrently co-opted for their Zinc finger motifs as transcription factors.

The ENV genes of endogenous retroviruses have been domesticated for several purposes, the most relevant example being the mammalian SYNCYTIN genes that mediate the placental cytotrophoblast fusion, creating the foetal-maternal interface. This domestication event had a major impact on the evolution of mammals, leading to vivipary (development of the embryo inside the mother). Additionally, many ENV genes have been co-opted to confer resistance to exogenous retroviruses via receptor interference.

1.4. Regulation of Transposable Elements Activity

Despite the positive role that can be exerted by transposons, cells must at the same time restrain TE activity to avoid their overt mutagenic effects. In most cases, TEs in genomes are found in a quiescent silent state, generally imposed through epigenetic mechanisms. Epigenetics refers to the changes in gene expression that are inheritable but do not involve changes in the DNA nucleotide sequence. In eukaryotes, this is largely achieved by regulating the accessibility of the transcriptional machinery to DNA. In eukaryotic cells, the large DNA strands containing the genetic information are wrapped around a set of nuclear proteins called histones. This combination of DNA and proteins is called chromatin, and its basic unit is the nucleosome, which is composed of eight core histones wrapped by
147bp of DNA. Chemical modifications of the chromatin, at the DNA or histone level, alter its structure and therefore the accessibility to DNA, thereby regulating gene expression. Hereby chromatin can be divided into two classes: euchromatin, characterized by permissive chromatin modifications allowing gene expression, and heterochromatin, regions that are highly compact (or condensed) in which gene expression is often prevented by transcriptional gene silencing (TGS)\textsuperscript{136}. However, one of the characteristics of epigenetic modifications is their plasticity. Chromatin compaction and TGS can be temporarily reversed, often leading to the reactivation of TEs\textsuperscript{137-140}, such that an array of different mechanisms has been elaborated by hosts to prevent expression and transposition of epigenetically reactivated TEs\textsuperscript{99}. A common feature of the underlying pathways in various organisms is the use of small RNA (sRNA) molecules that guide RNA silencing of TEs at the post-transcriptional level and may also reinforce TGS by directing the establishment of repressive chromatin modifications\textsuperscript{141}.

1.4.1. Transcriptional Gene Silencing: Chromatin Modifications

1.4.1.1. DNA methylation

DNA methylation refers to the addition of methyl (CH\textsubscript{3}) groups to certain nucleotides within the DNA. Although not present in all eukaryotes (no DNA methylation is found in the yeast \textit{S.cerevisiae} and \textit{S.pombe}, the nematode \textit{C.elegans} or the fly \textit{D.melanogaster}), DNA methylation is almost exclusively found on the carbon 5 position of cytosine residues (5-methylcytosine or 5mC)\textsuperscript{135}. In plants and mammals DNA methylation is critical for development and genome stability, as mutants impaired in DNA methylation pathways display a vast array of developmental defects or are lethal\textsuperscript{142-145}. Methylation in the promoter, transcription start site or other regulatory sequences can either attenuate or indeed completely shut down transcription. Although some genes also display gene-body methylation, its role remains largely unknown\textsuperscript{144,146}. In all cases, DNA methylation distribution is heavily biased towards heterochromatic regions and repetitive sequences such as TEs, and its loss leads to the reactivation and, on occasions, mobilization of transposons\textsuperscript{123,144,146,147}. Cytosine residues are mainly methylated in the symmetrical
CG context but in plants methylation also affects CHG and asymmetric CHH sites (where H = A, T or C). Maintenance of DNA methylation patterns mostly relies on CG sites, faithfully transmitted through DNA replication due to the action of the mammalian DNA methyltransferase 1 (Dnmt1) and its plant homolog METHYLTRANSFERSE 1 (MET1). Both enzymes have high affinity for hemi-methylated DNA, and incorporate a methyl group to unmethylated cytosines in the CG context at positions where the complementary dinucleotide is methylated. In plants, maintenance of the other symmetric context, CHG, is operated by the plant specific CHROMOMETHYLASE 3 (CMT3) in cooperation with histone modifications. The CHH context however, cannot be maintained due to the asymmetry of the CHH sequence, and this methylation relies on reiterated de novo DNA methylation at every cycle of DNA replication. Methylation at CHH is effected mainly by DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2) (and sometimes by CMT3) and, like de novo methylation at the other contexts, needs to be directed to specific DNA sequences via mechanisms discussed later in this introduction (see section 1.4.3). In mammals, de novo DNA methylation is exerted by Dnmt3A/B but through different mechanisms than in plants.

DNA methylation is a dynamic process that can be developmentally or environmentally regulated. In mammals, at every generation, the epigenetic status of the gametes is systematically erased upon fertilization and re-established de novo in the embryo after implantation of the blastocyst, rendering methylation patterns in a cell-type specific fashion. In plants, the situation is different, as global DNA methylation patterns are faithfully transmitted to the gametes, although they may undergo reprogramming to a certain degree. For instance, maintenance of DNA methylation is transiently arrested during gametogenesis and in gamete companion cells (nurse cells) this passive loss of epigenetic marks is accompanied by active DNA demethylation. This leads to TE reactivation, which is believed to reinforce the silencing in the gametes by inducing or enforcing de novo DNA methylation (see section 1.4.3). Indeed, the de novo DNA methylation machinery is highly expressed in gametes and during early embryogenesis, where increased levels of DNA methylation and silencing are observed. Environmental cues can also cause loss of DNA methylation, including, in Arabidopsis, antibacterial defense, where maintenance and de novo DNA methylation pathways are down-regulated. This is also accompanied by active DNA
demethylation, resulting in the reactivation of defense-related genes that are often proximal to TEs\textsuperscript{156}.

1.4.1.2 Histone modifications

The protein core of the nucleosome is an octamer composed of two molecules of the H2A, H2B, H3 and H4 histones, although other proteins (histone variants) can be added to define chromatin regions. Histones contain two defined regions, a globular domain, around which DNA is a wrapped, and an unstructured N-terminal domain or “tail” pointing out of the nucleosome core\textsuperscript{157}. These tails, especially those from H3 and H4, can be subjected to wide range of post-translational modifications (PTMs) that include methylation (me), acetylation (ac), ubiquitinylation (ub) or sumoylation (su) of lysine (K); methylation of arginine (R); and phosphorylation (ph) of serine (S) and threonine (T) residues. Some modifications are intrinsically associated with permissive chromatin (transcriptional activation), such as acetylation of lysines (e.g. H3K9ac, H4K5ac...); others are defining features of inactive chromatin (transcriptional repression) such as sumoylation (e.g. H2BK30su). However, additional modifications including methylation of lysines, can have both roles depending on the position of the residue in the tail and the number of modifications\textsuperscript{136}. For instance, in plants, H3K9 di-methylation (H3K9me2) is heavily enriched in pericentromeric heterochromatic regions associated with TEs, although this modification is not specific to heterochromatin \textit{per se}, but rather to repeated DNA elements\textsuperscript{158-160}. On the other hand, H3K4me2, typically associated with euchromatin, can be found on active genes located in heterochromatic regions\textsuperscript{159-161}. The repressive H3K27me3 is exclusively associated with euchromatic genes that are tissue specifically or developmentally regulated in a Polycomb-dependent manner (a family of proteins that can remodel chromatin such that epigenetic silencing takes place)\textsuperscript{162,163}. H3K27me3 is deposited along the entire transcribed region of genes, suggesting a link between the deposition of such repressive mark and transcription. The role of histone repressive modifications in transposon silencing seems conserved, because mutations in the histone H3K9 methyltransferase \textit{Suv39} in mouse, and in its plant homologue \textit{KRYPTONITE} (KYP), result both in the upregulation of TEs\textsuperscript{164,165}.

Maintaining the structure of heterochromatin independently from DNA methylation or direct histone modifications is also important to keep TEs in a
transcriptionally silent state. Recently, members of the MORC ATPases have been shown to be required for maintainance of heterochromatin condensation in Arabidopsis and C. elegans. MORC mutants display reactivation of TEs and silenced transgenes without any loss of DNA methylation or histone modifications\textsuperscript{166}. Stress can also lead to the activation of heterochromatic transcription. For instance, heat stress in Arabidopsis reactivates TEs by heterochromatin decondensation and loss of nucleosomes\textsuperscript{167}.

1.4.1.3. Interaction between chromatin modification and DNA methylation pathways

Both, DNA methylation and repressive histone modifications interact to reinforce and/or propagate the silent state on TEs, as both are highly enriched and overlap at repetitive elements loci\textsuperscript{135,168,169}. Moreover, mutations in those pathways influence each other. The DNA methylation machinery, which partially requires active transcription (section 1.4.3), and histone modifiers have to access highly compacted heterochromatin to maintain the silencing, which, by definition, is poorly accessible. In plants, this access is granted by the ATPase SWI2/SNF2 chromatin remodeler DECREASE IN DNA METHYLATION 1 (DDM1). Within heterochromatin, histone H1 binds to the nucleosome core and to the linker DNA between nucleosomes, condensing chromatin and inhibiting nucleosome mobility and accessibility. DDM1 overcomes the effects of H1, granting the access of DNA methylation machinery to highly condensed chromatin\textsuperscript{170}. Consequently, mutations in DDM1 result in the loss of DNA methylation in heterochromatin and reactivation of TEs\textsuperscript{147}. Conversely, H1 mutants display DNA hypermethylation\textsuperscript{170}. Additionally, KYP recognizes DNA methylation in non-CG contexts to methylate H3K9 and in turn, CMT3 recognizes H3K9me2 through its chromodomain, creating a positive feedback loop for the maintenance of repressive chromatin marks\textsuperscript{148}.

1.4.2. Post-Transcriptional Gene Silencing

RNA silencing refers to the mechanism by which small RNA (sRNAs) guide effector complexes known as RNA-induced silencing complexes (RISC) to complementary transcripts, regulating gene expression. RNA silencing is a very ancient mechanism conserved in most, if not all higher eukaryotes. Recent studies also indicate that
conceptually similar processes also operate in prokaryotes. A key function of RNA silencing is to act as an innate immune system against genomic invaders such as phages, viruses or TEs\textsuperscript{171-175}.

1.4.2.1. General aspects of post-transcriptional RNA silencing pathways

In prokaryotes (bacteria and archaea), virus resistance is acquired by integrating short fragments of the viral genome into dedicated clusters of regularly interspaced palindromic repeats (CRISPRs). Upon transcription, the pre-CRISPR RNA (crRNA) is processed by CRISPR-associated (Cas) proteins into short crRNAs (~57 nt) containing viral sequences that guide effector nuclease complexes to cleave complementary DNA sequences\textsuperscript{174,176,177}. Although conceptually similar, eukaryotic RNA silencing is mechanistically different. The three key protein components of RNA silencing in eukaryotes are:

i. **Dicer** (DCR): A multi-domain ribonucleases that processes double-stranded RNA (dsRNA) into sRNAs duplexes. Depending on the Dicer protein sRNA duplexes are typically 20-25 nt long with 2-nt 3'-overhangs. They generally consist of an RNA binding domain (RBD), a PAZ domain that recognizes the end of the dsRNA precursor, RNase III and helicase domains. The distance between the PAZ and the RNase domains contributes to the final length of the sRNA produced\textsuperscript{178}.

ii. **Argonaute** (AGO): The main effectors of RNA silencing. AGO proteins are the direct binding partners of sRNAs. They can be divided into three clades. The AGO clade, similar to the *Arabidopsis* AGO1; the PIWI clade, homologous to the *Drosophila* Piwi protein, and the WAGO clade (worm-AGO), specific to nematodes. Argonaute proteins are characterized by the presence of four domains. N (amino-terminal), PAZ (PIWI-ARGONAUTE-ZWILLE), MID (middle) and PIWI. The N-terminal domain is required for sRNA loading and unwinding of the small RNA duplex; PAZ and MID domains anchor the 3' and 5' ends of the sRNA, and the PIWI domain, structurally similar to RNase H, provides endonucleolytic activity for cleavage of target RNAs complementary to the loaded sRNA. However only a subset of Argonaute proteins display this specific catalytic activity\textsuperscript{179}.

iii. **RNA-dependent RNA polymerase** (RdRP): Proteins that convert single stranded RNA (ssRNA) into dsRNA serving as a template for DCR in
plants and fungi\textsuperscript{180}. In \textit{C.elegans}, the somatic RdRP, RRF1, uses RNA cleaved by primary siRNAs to produce triphosphated small RNAs directly, without the processing by Dicer\textsuperscript{181}.

Multiple paralogs of these components have arisen through evolution in different organisms, which along with gene duplications and the incorporation of several accessory proteins, have generated the diversity and complexity of RNA silencing pathways in eukaryotes. This is well illustrated by the AGO family, with \textit{C.elegans} containing 27 AGO proteins, \textit{Arabidopsis} containing 10, mouse five, and the fission yeast \textit{S. pombe} containing a single AGO protein. Despite the aforementioned diversity, some of the components are absent in several organisms that still perform RNA silencing. For instance, mammals and flies seem to lack RdRPs. Moreover some RNA silencing pathways do not depend on Dicer to produce sRNAs. Only few unicellular organisms, like the budding yeast \textit{S.cerevisiae}, lack RNA silencing\textsuperscript{180}.

The pathways differ not only in their components, but also in the means by which the sRNAs are generated, their size and the final outcome of silencing. In the canonical RNA silencing pathway (Figure 2), DCR enzymes convert dsRNA into sRNAs that are incorporated into AGO-containing complexes named RNA-induced silencing complexes (RISC) that scan the transcriptome for transcripts that share total or partial sequence complementarity to the sRNA. Following target recognition RISC can induce cleavage and degradation, translational repression (or a combination of both) of mRNAs\textsuperscript{182} (Figure 2). The initial dsRNA substrate can be of different origins. In most eukaryotes a subset of loci is dedicated to generate micro RNAs (miRNA), a class of sRNA primarily devoted to the regulation of endogenous coding mRNAs. miRNAs arise from Pol II transcribed non-coding RNAs that form an imperfect fold-back structure through intra-molecular base-pairing that are first processed into stem-loop precursors (pri- and then pre-miRNAs) and subsequently diced into 21-22-nt long miRNA\textsuperscript{183}. Alternatively, dsRNA can be generated as the result of convergent transcription, intermolecular folding of inverted repeat sequences, viral replication dsRNA intermediates, to name the predominant cases\textsuperscript{183,184}. Furthermore in organisms expressing RdRPs, aberrant transcripts, carrying premature termination codons (PTC), improperly spliced or lacking important modifications such as 3'-polyA tails or 5'-cap (7-methylguanosine; m\textsuperscript{7}G), which are usually degraded by the RNA quality control (RQC) can as well be templates for RdRPs\textsuperscript{185-187}. In all these cases, the dsRNA substrate displays perfect intramolecular
base-pairing and the resulting sRNAs are generically known as small interfering RNAs (siRNAs) to differentiate them from miRNAs\(^{183}\) (Figure 2). In certain circumstances, RdRPs can also act on RISC cleavage products to enable the generation of secondary siRNAs by Dicer\(^ {188}\) (Figure 2). In plants, some non-coding RNAs are solely dedicated to the production of secondary siRNAs following miRNA-directed cleavage followed by RdRP (RDR6) and Dicer-like 4 (DCL4) activity. The derived siRNAs can in turn target additional mRNAs, for this they are known as trans-acting siRNA (tasiRNAs)\(^ {189}\). In mammals, a single DCR (Dcr1) is responsible for the production of both miRNA and siRNA\(^ {180,190}\), as opposed to plants where multiple DCR-like (DCL) proteins have more specific roles. For instance, DCL1 is dedicated to the production of 21-22-nt miRNA, whereas DCL2, DCL3, and DCL4 produce 22-24- and 21-nt siRNAs respectively\(^ {172,191}\). This diversity is not restricted to plants. Recently it was shown in mice that a truncated isoform of Dcr1 is dedicated to the production of siRNAs in the oocyte. Such alternative isoform is the result of an LTR-RTE
insertion within intron 7. While generally spliced out in somatic tissues, expression driven by the TE promoter in oocytes leads to the production of a shortened Dcr1 allele, incidentally providing another example of co-opted TE regulatory elements\textsuperscript{192}.

Many transposons are naturally targeted by RNA silencing. Due to the presence of encoded repeats in antisense orientation (see section 1.2 and Fig. 1) and antisense promoters\textsuperscript{193} their transcripts can easily form dsRNA. Moreover, upon transposition, by landing near promoters, they can generate antisense transcripts or undergo rearrangements that will end in the formation of structured loci generating dsRNA. This is best exemplified by the maize, subclass I DNA transposon \textit{Mutator} (\textit{Mu}), which can be silenced genome-wide by a natural inverted repeat of the element, \textit{Mu}-killer (\textit{Muk}), that becomes potent source of siRNAs\textsuperscript{194}.

\subsection{1.4.2.2. Germline specific RNA silencing: Piwi pathways}

In animals, the germline is isolated early in embryogenesis and specialized RNA silencing pathways have evolved to protect the sperm and oocyte’s genomic integrity from the activity of TEs. Extensively studied in \textit{Drosophila} and \textit{C.elegans}, both of which have involved a different clade of AGO proteins, the PIWI proteins, and their associated 21-35-nt long PIWI-interacting RNAs (piRNAs)\textsuperscript{183}. Besides the heterogeneous size of piRNAs, the pathway is also quite different from canonical RNA silencing described above. As opposed to the miRNA and siRNA pathways, Dicer is dispensable for the production of piRNAs and, instead of double-stranded RNA, single-stranded RNA molecules are used as piRNA precursors. These transcripts are produced from piRNA loci or piRNA clusters\textsuperscript{195}. Additionally, piRNA production involves primary and secondary biogenesis pathways. The steps that lead to the conversion of the piRNA precursor into primary piRNA remain largely unknown, but are nonetheless mandatory to initiate the secondary pathways, dedicated to the amplification of the piRNA-mediated silencing, to ensure that the pool of piRNAs matches the level of expression of the targets (active TEs)\textsuperscript{196}.

In \textit{C.elegans}, most piRNA species (known as 21U RNAs; due to their 21-nt length and 5’U preference)\textsuperscript{197} are individually encoded in the piRNA cluster and expressed as a 26-nt precursor\textsuperscript{198}. Each transcript is then processed to generate the mature primary 21U RNA and loaded into the PIWI (AGO) protein PRG-1 (\textit{Piwi}-related gene 1)\textsuperscript{196,197}. However, in most animals piRNAs are derived from long precursor transcripts (1~100 kb) processed into piRNA populations\textsuperscript{196} (Figure 3).
Figure 3. Germline specific piRNA pathways prevent transposon mobilization. (a) In Drosophila melanogaster, the chromatin environment at dual-strand piRNA clusters licenses piRNA production from these loci. The piRNA RNA clusters are associated with H3K9me3 and the RDC (Rhino, Deadlock and Cutoff) complex. The DEAD box protein UAP56, required for piRNA processing specificity, binds primary transcripts. Primary piRNAs are loaded into the PIWI proteins Piwi and Aub, but only Aub can enter the amplification cycle. The ping-pong amplification cycle promotes the generation of piRNAs complementary to active transposons (TE). Mobilization of transposons into piRNA clusters enhances the repertoire of foreign sequences at such loci. (b) In Caenorhabditis elegans, piRNAs are as well encoded in clusters, however they are individually transcribed into small transcripts that are processed into the mature 21U-RNA. Clusters are strongly increased in TE sequences suggesting that TE mobilization is into 21U-RNA clusters is not relevant to acquire TE specificity. 21U-RNA loaded into PRG1 can recognize exogenous sequences and trigger the production of secondary 22G-RNA through the recruitment of RdRP. 22G-RNA are loaded into the worm-specific AGO proteins (WAGO) to exert diverse function. WAGO1 mediates amplification of 22G-RNA by reinforcing recruitment of RdRP to exogenous sequences and mediating mRNA turnover (PTGS). Nucl: nuclear AGO; Cyto: cytoplasmic AGO (see Fig. 4)
In *Drosophila*, such transcripts arise from clusters of nested transposon fragments generally located in pericentromeric or subtelomeric heterochromatic regions\(^{199}\). Such loci act as transposon traps, capturing TE sequences to generate piRNAs (Figure 3a). The FLAMENCO locus, for example, regulates the activity of Gypsy through the generation of piRNAs\(^{79,200}\). But the recent amplification and silencing of the *P* element better illustrates the role of such loci in silencing TEs. At the beginning of the 21\(^{st}\) century, the DNA transposon *P* spread in wild-type populations of *Drosophila* worldwide, colonizing wild fly stocks. Paternally transmitted active copies of *P* elements cannot be silenced by female flies that have never been exposed to the transposon, leading to a sterility syndrome in F1 progenies termed ‘hybrid dysgenesis’. Surprisingly, as dysgenic hybrids age, fertility is partially restored, coinciding with silencing of the *P* element and the production of piRNAs. Such restoration of fertility was shown to be promoted by the transposition of the TE into piRNA clusters, leading to its final demise\(^{201}\). *Drosophila* germline piRNA clusters displaying dual-strand transcription are identified by chromatin marks and the newly defined RDC (Rhino, Deadlock, and Cutoff) complex. The RDC complex is anchored to piRNA clusters via the heterochromatin protein (HP1) homolog Rhino (Rhi) which binds to H3K9me3 through its chromodomain\(^{202}\) and, in collaboration with the DEAD box RNA binding protein UAP56, licences piRNA precursors for primary piRNA biogenesis\(^{196,203}\). As opposed to *C.elegans*, primary piRNAs produced through this pathway display a broader size distribution ranging from 26- to 30-nt\(^{195}\). Mature primary piRNA are loaded into PIWI proteins (Piwi and Aubergine, Aub, in flies and MILI and MIWI2 in mice)\(^{195,196}\) (Figure 3a).

In flies and mice, secondary piRNA biogenesis in the cytoplasm is initiated by Aub and MILI respectively, driving the “*ping-pong* amplification” cycle\(^{195}\). In *Drosophila*, piRNA loaded Aub can recognize TE or cluster-complementary transcripts and cleave them. The 3’ cleavage fragment is then loaded into another PIWI protein, Ago3, where the cleavage fragment is trimmed and further processed to become a secondary piRNA loaded in Ago3. This RISC complex then targets back the piRNA precursor and the 3’ cleavage product is then incorporated into Aub and processed similarly to again target complementary transcripts, thereby reinitiating the cycle\(^{196}\). Given that the “*ping-pong*” cycle can use TE-derived transcripts, it has been considered that primary piRNA biogenesis allows production of libraries of piRNA
directed against all TEs already “known” (i.e. present in piRNA clusters). In such a scenario, secondary piRNA biogenesis can be considered as an adaptive response, primed by the activation of one particular TE, leading to the reinforcement of silencing of this particular element mediated by the accumulation of secondary piRNAs\(^\text{199}\) (Figure 3a). Very similar processes take place in other species including mouse and zebrafish, where signatures of ping-pong amplification have been detected and linked to TE silencing\(^\text{204,205}\). Despite this, some differences are noticeable. For example, in mice, the source of primary piRNAs is not the piRNA cluster transcript but the TE mRNA, and secondary piRNAs are amplified in the ping pong cycle using cluster antisense transcripts\(^\text{206}\) (Figure 4).

In contrast to the “ping-pong” cycle amplification mechanism, in C.elegans secondary piRNAs are produced through a different mechanism owing to the presence of RdRP in nematodes. The primary 21U RNAs loaded into PRG-1 guide the cleavage of TE and other foreign sequences, recruiting the activity of an RdRP to produce 22G RNAs (secondary piRNAs); importantly endogenous targets (coding genes) are protected from RdRP amplification\(^\text{195,207}\). While the exact mechanism by which the 22G RNAs are generated remains somewhat obscure, they are believed to be Dicer independent as mutations in DCR1, abolishing the production of miRNA and somatic RdRP-dependent secondary siRNAs, do not compromise germline 22G RNAs levels\(^\text{207}\). Secondary 22G RNAs can then be loaded into distinct WAGO proteins to exert different functions. When loaded into WAGO-1 they trigger PTGS and subsequent amplification through RdRP recruitment\(^\text{207}\) (Figure 3b). However, WAGO proteins lack the catalytic domain required for slicing activity, hence, the mechanism by which WAGO-1 engages targets to RNA turnover is still under investigation\(^\text{208}\).

1.4.3. RNA-mediated chromatin modifications

In many organisms, RNA silencing components are required for chromatin modifications. Somewhat counter-intuitively, the maintenance of many repressive chromatin marks do not require RNA silencing (yet it is still possible that the initial signal is RNA dependent). It is believed that redundant sRNA-dependent and sRNA-independent chromatin modifications reinforce and assure proper and stable silencing of transposable elements as well as the assembly of heterochromatic
domains such as those of centromeric regions\textsuperscript{209}. In plants, a prevalent type of RNA-mediated chromatin modification entails the production of small RNAs by the canonical or specialized RNA silencing pathways, and their loading into AGO proteins that scan nascent transcripts in the chromatin. Identification of complementary target sequences triggers the recruitment of chromatin modifications to impose or reinforce transcriptional gene silencing (TGS). Perhaps, the most extreme example occurs in unicellular ciliates, such as \textit{Tetrahymena} and \textit{Paramecium}, where, in order to ensure proper gene expression in the macronucleus during vegetative growth, the germline equivalent micronucleus generates sRNAs, known as scanner RNAs (scRNAs), that promote heterochromatin formation and subsequent elimination of target sequences in the macronucleus\textsuperscript{210}. The fission yeast \textit{Schizosaccharomyces pombe}, containing a single copy of each of the basic RNA silencing machinery components (AGO, DCR, RdRP), exemplifies the underlying principles of the RNA silencing-mediated chromatin modification pathways that commonly operate in other organisms. In \textit{S. pombe}, co-transcriptional silencing by RNA silencing depends on transcription by RNA polymerase II (Pol II) to generate both source and target of siRNAs. During the S phase (the part of the cell cycle in which DNA is replicated), the dilution of heterochromatic marks (H3K9me3) due to the addition of new nucleosomes onto \textit{de novo} replicated DNA, allows low levels of Pol II transcription. It is thought that the heterochromatin protein HP1 not only binds to H3K9me3 but also acts as an RNA binding protein that accepts transcripts from its bound loci and elicits their processing by the RNA silencing machinery\textsuperscript{211}. Thus, HP1 displays a similar role in specifying RNA silencing processing templates as its homologous Rhi in \textit{Drosophila}. RdRP (RDRC) and DCR (Dcr1) then cotranscriptionally process nascent transcripts into siRNAs, which are loaded into an AGO- (Ago1) containing complex named RITS (RNA-induced transcriptional silencing). siRNA-mediated RITS targeting on nascent transcripts recruits the histone H3K9 methyltransferase (Clr4), restoring heterochromatin and TGS\textsuperscript{212} (Figure 4a).

\textbf{1.4.3.1. RNA-directed chromatin modifications in animals}

In animals, a similar mechanism seems to operate to direct chromatin modifications at piRNA targets. In mice, during the ping-pong amplification cycle, MIWI2 is loaded with secondary antisense piRNAs and shuttles between the cytoplasm (where it acts in the ping-pong cycle) and the nucleus, where it triggers \textit{de novo} DNA methylation.
guided by nascent complementary transcripts bound to the chromatin. Even though how the methyltransferases (Dnmt3A/B) are recruited remains obscure, loss of MIWI2 or piRNAs leads to reactivation of TEs due to the absence of DNA methylation\(^{206}\) (Figure 4b). By contrast, in \emph{Drosophila} germinal cells, primary instead of secondary piRNAs guide TGS to transposons in \emph{trans}. TGS is mediated by nuclear Piwi (Figure 3a) through the recruitment of H3K9me3 by a yet unknown mechanism (Figure 4c). Loss of Piwi leads to the derepression of TEs and enhance transcription of TE-neighbouring genes\(^{123}\). In \emph{C. elegans}, both somatic and germline RNA silencing pathways have a nuclear component that promotes TGS of RNA silencing targets. In somatic cells, the Argonaute protein NUCLEAR RNAi DEFECTIVE 3 (NRDE-3) loaded with RdRP-dependent 22G-RNA is translocated into the nucleus where it binds to nascent transcripts and further recruits additional factors (NRDE-1-2-4) that inhibit transcription, by halting Pol II and promoting H3K9me3\(^{213}\). In the germline, WAGO-9, one of the worm germline specific Argonautes, acts similarly to NRDE-3 but, in addition to H3K9me3, it also recruits one of the HP1 orthologs, HPL-2\(^{214,215}\) (Figure 4c). Interestingly, once established in the germline, silencing can be inherited across generations in the absence of the initial silencing trigger and in a PRG-1 independent manner (generally required for 22G-RNA biogenesis, see Figure 3b)\(^{215}\). This phenomenon is called RNA-induced epigenetic silencing (RNAe). While the mechanism still remains unknown, it has been suggested that a combination of histone modifications and HPL-2 (HP1) recruitment sensitizes the silenced locus to preserve multigenerational silencing independently of primary siRNAs\(^{196,216}\).

1.4.3.1. RNA-directed DNA methylation in plants

In addition to the aforementioned MET1 DNA methylation maintenance and the self-reinforcing loop between H3K9me2 and non-CG methylation, plants display an additional pathway to ensure the maintenance of silencing through the DNA methylation at asymmetric sites (CHH). As discuss previously, CHH methylation can’t be self maintained through DNA replication, and such \emph{de novo} DNA methylation mechanisms must be applied continuously to reinforce silencing. This is achieved through a siRNA-mediated mechanism known as ‘RNA-directed DNA methylation’ (RdDM). As with the previously described pathways, it relies on active transcription to generate both the siRNA template and target. However, the reliance of Pol II to enforce silencing pathways has the disadvantage that, as silencing proceeds, the
Figure 4. sRNA mediated chromatin modifications. In multiple organisms the RNA silencing machinery participates into the establishment and maintenance of TGS on transposons (TEs). (a) in fission yeast, dilution of histone silencing marks during the S fase allows Pol II transcription, HP1 protein binds to nascent transcripts facilitating their processing into siRNAs through the RdRP and Dicer (Dcl1) activities. Ago1 bound to siRNAs forms the RNA- induced transcriptional silencing complex (RITS) and binds to nascent transcripts complementary to the siRNA sequence, recruiting the histone methyltransferase Rlt4 to impose silencing. (b) In mouse prenatal testis, MIWI2, besides it role in piRNA amplification, shuttles to the nucleus to guide DNA methylation (5mC) through recognition of nascent transposon transcripts. After birth, when MIWI2 is no longer expressed, MILI
RNA source and target of the silencing effectors are disrupted. In plants, this problem is circumvented by the acquisition of two unique RNA polymerases, Pol IV and Pol V, to initiate siRNA biogenesis and provide the scaffold transcript to recruit the silencing machinery, respectively\(^{150}\). Both polymerases are responsible for mediating RdDM and TGS at a subset of transposons and heterochromatic loci\(^{217}\). Pol IV is believed to be recruited to methylated loci\(^{218,219}\) rich in H3K9m2 and deficient in H3K4me2 through SHAWADEE HOMEODOMAIN HOMOLOG 1 (SHH1), a Pol IV interacting protein that scans the chromatin simultaneously probing for the methylation status of H3 lysine 4 (K4) and lysine 9 (K9) (which are mutually exclusive), displaying higher affinity for H3K9me2 (repressive) than for H3K4me2 (permissive) modifications\(^{220}\). Pol IV transcripts are converted into 24nt-siRNAs\(^{221}\) through the coordinated activities of RDR2 (RdRP) and DCL3. AGO4 (expressed in most Arabidopsis tissues) together with AGO6 and AGO9 are the main effectors of RdDM\(^{222}\). AGO4, loaded with 24-nt siRNAs, interacts with complementary Pol V nascent transcripts\(^{223}\), recruiting DRM2 to methylate cytosines in all sequence contexts\(^{148,149,224}\). The activity of Pol IV/V requires additional factors. Pol IV requires CLASSY1 (CLSY1), a putative chromatin remodelling factor that possibly grants Pol IV access to chromatin or aids in ssRNA processing; Pol V, on the other hand, requires DEFECTIVE IN RNA-DIRECTED DNA METHYLATION 1 (DRD1) another putative remodelling factor\(^{225}\), DEFECTIVE IN MERISTEM SILENCING 3 (DMS3) a structural maintenance of chromosome domain protein\(^{226}\) and RNA-DIRECTED DNA METHYLATION 1 (RDM1) a single-strand methyl-DNA binding protein\(^{227}\). They form a putative

---

(c) Similarly, in Drosophila, Piwi guides H3K9me3 of transposons guided by 19 piRNAs, the role of 29 piRNAs in feeding Piwi is currently unknown. (d) In C. elegans, the germline specific Argonaute WAGO-9 localizes to the nucleus where it recruits NRDE-2 and NRDE-1 to the nascent mRNA. NRDE-1 and NRDE-4 promote H3K9me3 and the recruitment of the HP1 homolog HPL2. In addition, NRDE proteins inhibit Pol II. (e) In Arabidopsis, the plant specific Pol IV is recruited to H3K9me2 rich loci by SHH1. Pol IV works in association with the RDR2 RdRP, to generate dsRNA that is processed into 24-nt siRNAs by the dedicated Dicer protein DCL3. Argonaute proteins belonging to the plant AGO4 clade (AGO4, AGO6 and AGO9) interact through sequence homology to the loaded siRNA with Pol V transcripts. Pol V transcription is assisted by the putative chromatin remodelling DDR complex (DRD1,DMS3 and RDM1), facilitating its association to chromatin. AGO4 interaction with Pol V and its nascent transcripts results in the recruitment of the DNA methyltransferase DRM2 that can methylate cytosines in the symmetric CG and CHG and asymmetric CHH contexts. In some loci, CG methylation can stably be maintained through MET1, while CHG sites can initiate the self reinforcement loop CMT3-KYP, where KYP recognizes methylated non-CG sites to promote H3K9me2 that attracts CMT3 to exert DNA methylation in the CHG context (see section 1.4.1.3).
chromatin-remodeling complex termed DDR (DRD1, DMS3, RDM1), required for Pol V activity, however their exact roles remain unknown\textsuperscript{228,229}.

Although the RdDM pathway \textit{de novo} methylates DNA sequences that cannot be maintained methylated by themselves, it is still a reinforcement mechanism that continuously acts in \textit{cis} to prevent the irreversible loss of the TGS status. How is truly \textit{“de novo”} DNA methylation imposed? Can a Pol II locus be converted into a Pol IV/V target? Endogenous genes are highly recalcitrant to CHG and CHH methylation\textsuperscript{142,230,231} due to the action of INCREASE IN BONSAI METHYLATION 1 (IBM1), a histone demethylase that removes H3K9me2; DEMETER/REPRESSOR OF SILENCING 1 (DME/ROS1) a DNA glycosylase that demethylates DNA by excising methylated cytosines and whose action is facilitated by the histone acetylase INCREASED DNA METHYLATION 1 (IDM1; defining active chromatin, see section 1.4.1.2)\textsuperscript{148}. However, there is evidence indicating that Pol II nascent transcripts can also be used as scaffolds to recruit AGO4\textsuperscript{232,233}. Indeed, in \textit{Arabidopsis} transgenes carrying Pol II promoters get rapidly silenced in trans through the action of the RdDM machinery if they carry sequences homologous to endogenous or artificially generated (i.e. through hairpin DNA constructs) 24-nt siRNAs\textsuperscript{156,234-237}. Nonetheless, no specialized TE-derived siRNA-producing cluster such as those found in \textit{Drosophila} have been found in plants. In fact, most TE-derived sequences correspond to degenerated and truncated insertions, commonly subjected to genomic rearrangements including inverted-duplications, and located in pericentromeric, heterochromatic regions\textsuperscript{70}. Moreover, most plant 24-nt siRNAs are derived from such regions\textsuperscript{219,238}. Therefore, ancient TE remnants might be maintained as genetic memory of past TE bursts to prevent future mobilization events.

\textbf{1.5. Aims of the Project and Biological Model of Choice}

The aim of the project is to gain a deeper understanding of the interactions between transposable elements and their hosts. Although, since their initial discovery by McClintok, a clearer picture has emerged regarding their roles in shaping genome evolution and the mechanisms that control TEs, many questions remain. How does the host recognize and counter the activity of new transposable elements? Are there mechanisms to actively silence such elements? If so, under which circumstances are
such mechanisms visible and experimentally accessible? What are the genetic and epigenetic consequences of de novo genome invasions by TEs? Figure 5 summarizes these and other key questions we have decided to address here by studying the fate of a class-I transposon, the Copia-like LTR-retroelement Évadé, in the model organism Arabidopsis thaliana.

Why an LTR-retrotransposon in plants? Plants are amenable systems to study active transposons. In opposition to animals, plants seem to better tolerate the activity of transposable elements, as plants carrying active TEs do not display impaired fertility or developmental arrest\textsuperscript{147,239-242}. Moreover, plant genomes have been extensively shaped by TEs and one of the most obvious manifestations of this is the large variation in plant genome size\textsuperscript{243}. Among all classes of TEs, retrotransposons constitute by far the major invaders of plant genomes\textsuperscript{244} and among class I transposons (retroelements) in particular, LTR-RTEs alone contribute sizeable fraction of all TEs (Table 2). The fact that LTR-retroelements are closely related to retroviruses makes them an interesting object of study from the epigenetic and RNA silencing perspective, given that during their life cycle they have both an RNA and DNA phase. Accordingly, and as discussed before, LTR-RTEs are under strong epigenetic control (TGS) and can also trigger RNA silencing responses upon their activation (PTGS).

![Figure 5. Aims of the project.](image)

35
Nonetheless, most TEs are highly degenerated such that less than 30% of Arabidopsis LTR-RTEs are estimated to be intact\textsuperscript{245}. Moreover, their high copy number and pericentromeric heterochromatic localization\textsuperscript{243} (potentially preventing expression of euchromatic copies due to sequence homology-based trans-silencing), confound their study: it is difficult to tease apart responses triggered by an active element of interest from the mechanisms imposed on remnants of past mobilization events. Interestingly, Arabidopsis thaliana displays a very compact genome of 130 Mb\textsuperscript{246} (compared to the 2.700 Mb genome of maize\textsuperscript{56}), with a very low TE content (Table 1) and, as opposed to other plant genomes, LTR-RTEs represent a lower fraction of its genome (Table 2). Comparative studies with its relative, Arabidopsis lyrata, have shown that A. thaliana has undergone a strong purification of TE sequences through DNA elimination\textsuperscript{22,247}. Furthermore, detailed analysis of the different TE families present in both species or even between the A. thaliana

<table>
<thead>
<tr>
<th>PLANTS</th>
<th>species</th>
<th>common name</th>
<th>LTR</th>
<th>no LTR</th>
<th>DNA transposons</th>
<th>Non-auton.</th>
<th>Unclassified</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>CONFERNS</th>
<th>thespia</th>
<th>Norway spruce</th>
<th>32.80%</th>
<th>67.20%</th>
<th>n.d.</th>
<th>2.40%</th>
<th>n.d.</th>
<th>14.80%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis thaliana</td>
<td>Thale cress</td>
<td>7.90%</td>
<td>27.40%</td>
<td>n.d.</td>
<td>5.00%</td>
<td>0.30%</td>
<td>n.d.</td>
<td>27.70%</td>
</tr>
<tr>
<td>Arabidopsis lyrata</td>
<td>A. Lyrata</td>
<td>60.00%</td>
<td>n.d.</td>
<td>10.00%</td>
<td>n.d.</td>
<td>25.00%</td>
<td>n.d.</td>
<td>cs</td>
</tr>
<tr>
<td>Helianthus annuus var. parviflorus</td>
<td>Sunflower</td>
<td>25.70%</td>
<td>33.00%</td>
<td>n.d.</td>
<td>14.40%</td>
<td>0.30%</td>
<td>n.d.</td>
<td>15.40%</td>
</tr>
<tr>
<td>Brassica rapa</td>
<td>rapeseed</td>
<td>68.70%</td>
<td>n.d.</td>
<td>7.00%</td>
<td>1.20%</td>
<td>n.d.</td>
<td>4.00%</td>
<td>n.d.</td>
</tr>
<tr>
<td>Cucumis sativus</td>
<td>Cucumber</td>
<td>33.50%</td>
<td>6.30%</td>
<td>39.00%</td>
<td>10.50%</td>
<td>0.00%</td>
<td>n.d.</td>
<td>2.40%</td>
</tr>
<tr>
<td>Cucumis melo</td>
<td>Melon</td>
<td>28.10%</td>
<td>36.70%</td>
<td>9.50%</td>
<td>0.50%</td>
<td>n.d.</td>
<td>21.10%</td>
<td>0.30%</td>
</tr>
<tr>
<td>Solanum tuberosum</td>
<td>Potato</td>
<td>7.00%</td>
<td>27.90%</td>
<td>61.00%</td>
<td>1.30%</td>
<td>0.60%</td>
<td>0.10%</td>
<td>0.30%</td>
</tr>
<tr>
<td>Solanum lycopersicum</td>
<td>Tomato</td>
<td>9.90%</td>
<td>31.30%</td>
<td>56.70%</td>
<td>0.60%</td>
<td>0.30%</td>
<td>0.80%</td>
<td>0.90%</td>
</tr>
<tr>
<td>Cajanus cajan</td>
<td>Pigeonpea</td>
<td>12.04%</td>
<td>22.81%</td>
<td>54.11%</td>
<td>2.15%</td>
<td>0.12%</td>
<td>n.d.</td>
<td>8.77%</td>
</tr>
<tr>
<td>Medicago truncatula</td>
<td>Barbel medici</td>
<td>13.50%</td>
<td>18.60%</td>
<td>47.10%</td>
<td>7.30%</td>
<td>0.20%</td>
<td>n.d.</td>
<td>7.88%</td>
</tr>
<tr>
<td>Glycine max</td>
<td>Soybean</td>
<td>21.20%</td>
<td>50.25%</td>
<td>27.90%</td>
<td>n.d.</td>
<td>0.42%</td>
<td>n.d.</td>
<td>27.10%</td>
</tr>
<tr>
<td>Populus trichocarpa</td>
<td>Poplar</td>
<td>35.40%</td>
<td>37.90%</td>
<td>n.d.</td>
<td>10.00%</td>
<td>n.d.</td>
<td>n.d.</td>
<td>17.00%</td>
</tr>
<tr>
<td>Ricinus communis</td>
<td>Castor bean</td>
<td>9.48%</td>
<td>22.75%</td>
<td>62.12%</td>
<td>0.27%</td>
<td>0.00%</td>
<td>n.d.</td>
<td>3.18%</td>
</tr>
<tr>
<td>Linum usitatissimum</td>
<td>Flax</td>
<td>40.48%</td>
<td>34.02%</td>
<td>1.22%</td>
<td>9.05%</td>
<td>0.00%</td>
<td>n.d.</td>
<td>14.98%</td>
</tr>
<tr>
<td>Fragaria vesca</td>
<td>Strawberry</td>
<td>22.30%</td>
<td>28.80%</td>
<td>18.40%</td>
<td>1.10%</td>
<td>0.30%</td>
<td>n.d.</td>
<td>17.10%</td>
</tr>
<tr>
<td>Malus x domestica</td>
<td>Apple</td>
<td>12.90%</td>
<td>59.50%</td>
<td>n.d.</td>
<td>15.30%</td>
<td>n.d.</td>
<td>n.d.</td>
<td>17.06%</td>
</tr>
<tr>
<td>Pyrus bretschneideri</td>
<td>Pear</td>
<td>63.67%</td>
<td>33.50%</td>
<td>n.d.</td>
<td>1.48%</td>
<td>n.d.</td>
<td>n.d.</td>
<td>3.47%</td>
</tr>
<tr>
<td>Prunus persica</td>
<td>Peach</td>
<td>29.25%</td>
<td>33.70%</td>
<td>13.30%</td>
<td>2.10%</td>
<td>0.12%</td>
<td>26.70%</td>
<td>0.81%</td>
</tr>
<tr>
<td>Vitis vinifera</td>
<td>Grapewine</td>
<td>47.80%</td>
<td>18.40%</td>
<td>16.80%</td>
<td>15.00%</td>
<td>1.90%</td>
<td>2.40%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Theobroma cacao</td>
<td>Cacao</td>
<td>25.10%</td>
<td>17.70%</td>
<td>26.80%</td>
<td>n.d.</td>
<td>n.d.</td>
<td>10.40%</td>
<td>20.30%</td>
</tr>
<tr>
<td>Citrus sinensis</td>
<td>Orange</td>
<td>40.83%</td>
<td>50.44%</td>
<td>1.01%</td>
<td>2.08%</td>
<td>n.d.</td>
<td>n.d.</td>
<td>10.69%</td>
</tr>
<tr>
<td>Heliocalse annua</td>
<td>Sunflower</td>
<td>18.50%</td>
<td>41.20%</td>
<td>13.70%</td>
<td>0.50%</td>
<td>n.d.</td>
<td>0.53%</td>
<td>0.48%</td>
</tr>
<tr>
<td>Phoenix dactylifera</td>
<td>Date palm</td>
<td>63.20%</td>
<td>27.15%</td>
<td>0.02%</td>
<td>0.17%</td>
<td>n.d.</td>
<td>1.20%</td>
<td>0.07%</td>
</tr>
<tr>
<td>Musa acuminata</td>
<td>Banana</td>
<td>58.50%</td>
<td>26.20%</td>
<td>n.d.</td>
<td>12.40%</td>
<td>n.d.</td>
<td>n.d.</td>
<td>2.80%</td>
</tr>
<tr>
<td>Triticum aestivum</td>
<td>Wheat</td>
<td>20.35%</td>
<td>53.23%</td>
<td>4.68%</td>
<td>2.13%</td>
<td>0.00%</td>
<td>n.d.</td>
<td>17.70%</td>
</tr>
<tr>
<td>Hordeum vulgare</td>
<td>Barley</td>
<td>18.80%</td>
<td>27.30%</td>
<td>46.00%</td>
<td>0.35%</td>
<td>0.00%</td>
<td>n.d.</td>
<td>5.30%</td>
</tr>
<tr>
<td>Sorghum bicolor</td>
<td>Sorghum</td>
<td>8.40%</td>
<td>30.70%</td>
<td>48.90%</td>
<td>0.10%</td>
<td>0.25%</td>
<td>n.d.</td>
<td>7.70%</td>
</tr>
<tr>
<td>Setaria italica</td>
<td>Foxtail millet</td>
<td>15.47%</td>
<td>47.70%</td>
<td>0.50%</td>
<td>3.00%</td>
<td>0.30%</td>
<td>n.d.</td>
<td>15.70%</td>
</tr>
<tr>
<td>Brachypodium distachyon</td>
<td>False brome</td>
<td>17.30%</td>
<td>57.10%</td>
<td>1.70%</td>
<td>6.90%</td>
<td>n.d.</td>
<td>12.70%</td>
<td>0.60%</td>
</tr>
<tr>
<td>Oryza sativa var japonica</td>
<td>Rice, japonica</td>
<td>6.30%</td>
<td>30.50%</td>
<td>42.20%</td>
<td>2.00%</td>
<td>1.10%</td>
<td>n.d.</td>
<td>14.70%</td>
</tr>
<tr>
<td>Zea mays</td>
<td>Maize</td>
<td>28.10%</td>
<td>55.30%</td>
<td>5.40%</td>
<td>1.00%</td>
<td>0.00%</td>
<td>n.d.</td>
<td>3.80%</td>
</tr>
</tbody>
</table>

Table 2. Relative contribution (%) of different TE classes to the total TEs found in plant genomes. The data presented in this table has been obtained from the most recent genome sequencing publications and TE families has been reassigned to the different TE classes and subclasses according to the classification presented in Figure 1. Blue to red: higher % of total TE content. n.d: not determined in the original publication.
ecotypes Columbia-0 (Col-0) and Landsberg erecta (Ler), revealed that more than 90% of A. thaliana Copia insertions happened after its divergence from A. lyrata 5-10 MYA million years ago. Furthermore, since the Col-0 / Ler divergence 200,000 years ago, about 200 TEs have been active, indicating that the Copia superfamily has recently colonized A. thaliana and might still be active\textsuperscript{248,249}. In fact, although the total amount of elements is similar between Copia and Gypsy, Copia display a high number of small families (few elements per family), while Gypsy is characterized by a reduced family diversity and a higher proportion of degenerated elements\textsuperscript{250}. The same studies revealed other significant differences between the two major superfamilies. Copia-like elements are not only younger and more diverse, they are generally located further away from the pericentromeric regions, suggesting that they might be under less tight epigenetic control\textsuperscript{250}. Alltogether these various features make Arabidopsis Copia-like elements good candidates to address the various questions that underlie the present thesis.
1.6. References


77. cher, P. L., Bucheton, A. & Pelisson, A. Expression of the Drosophila retrovirus gypsy as
ultrastructurally detectable particles in the ovaries of flies carrying a permissive flamenco allele. 1247-1251 (1997).


415 (2005).


129. Lin, R. *et al.* Transposase-Derived Transcription Factors Regulate Light Signaling in


166. Moissiard, G. et al. MORC Family ATPases Required for Heterochromatin Condensation and

167. Pecinka, A. *et al.* Epigenetic Regulation of Repetitive Elements Is Attenuated by Prolonged Heat Stress in Arabidopsis. *plantcell.org*


171. Maillard, P. V. *et al.* Antiviral RNA Interference in Mammalian Cells. *sciencemag.org*


237. Mette, M. F., Aufsatz, W., van der Winden, J., Matzke, M. A. & Matzke, A. J. M.


Chapter 2: Results
2.1 Reconstructing de novo silencing of an active plant retrotransposon

Reference:


Authorship:

I performed all the laboratory based experiments with the exception of EVD copy number quantification and McrBC-qPCR-based DNA methylation analysis in the ddm1-mutant epiRILs. I wrote the manuscript together with Prof. Dr. Olivier Voinnet.
Reconstructing de novo silencing of an active plant retrotransposon

Arturo Marí-Ordóñez\textsuperscript{1,2}, Antonin Marchais\textsuperscript{1*}, Mathilde Etcheverry\textsuperscript{3*}, Antoine Martin\textsuperscript{3\$}, Vincent Colot\textsuperscript{3} and Olivier Voinnet\textsuperscript{1,2}

\textsuperscript{1} Swiss Federal Institute of Technology (ETH-Z), Department of Biology, Zurich, Switzerland.

\textsuperscript{2} Life Science Zurich Graduate School, Molecular Life Science Program, University of Zurich, Zurich, Switzerland.

\textsuperscript{3} Institut de Biologie de l’Ecole Normale Supérieure (IBENS), CNRS UMR8197 & INSERM U1024, Paris, France.

\textsuperscript{\$} Current address: Institut de Biologie Intégrative des Plantes (IBIP), Biochimie et Physiologie Moléculaire des Plantes, UMR 5004 CNRS/INRA/SupAgro/UM2, F–34060, Montpellier, France.

* These authors contributed equally to this work

Correspondence to: voinneto@ethz.ch
ABSTRACT

Transposable elements (TEs) contribute to genome size, organization and evolution. In plants, their activity is primarily controlled by transcriptional gene silencing (TGS), usually investigated at steady states reflecting how long-established silent conditions are maintained, faithfully re-iterated or temporarily modified. How active, invasive TEs are detected and silenced de novo in plants, however, remains largely unknown. Using inbred lineages of hybrid Arabidopsis epigenomes combining wild-type and mutant chromosomes, we have deciphered the sequence of physiological and molecular events underlying the proliferation and eventual demise of the single-copy, endogenous retrotransposon Évadé (EVD). We also show how this reconstructed TE burst causes widespread genome diversification and de novo epiallelism as possible sources of selectable and potentially adaptive traits.
INTRODUCTION

Transposable elements (TEs) abound in plant genomes, contributing to their size and organization via successive proliferation/deletion “bursts”\(^1,2\). The resulting genetic diversity generates functional variation, upon which selection may act\(^3,4\). TEs may also influence the chromatin state of neighboring genes, often in a developmentally- or stress-responsive manner\(^5\), owing to cellular pathways that restrict their activity primarily via promoter DNA methylation and transcriptional gene silencing (TGS). Several mechanisms account for TGS in Arabidopsis. In DNA methylation, 5-methylcytosine (5mC) maintenance by MET1 and CMT3 copies CG and CHG methylated patterns to daughter DNA strands\(^6\); \textit{de novo} methylation at CG, CHG and CHH sites, by contrast, relies on RNA-directed DNA methylation (RdDM)\(^7\). RdDM entails transcription of TE loci by the plant-specific RNA PolIV; the resulting RNA is copied by RNA-dependent RNA polymerase 2 (RDR2) into double-stranded (ds)RNA, processed by Dicer-like 3 (DCL3) into 24-nt small interfering (si)RNAs; upon loading into the silencing effector ARGONAUTE 4 (AGO4), siRNAs guide \textit{de novo} methyltransferases by recruiting AGO4 to scaffold RNA produced by the distinct RNA PolIV. The methylated DNA is thought to further attract PolIV at TE loci in a self-reinforcing mechanism thought to operate mainly in \textit{cis}\(^7\). Finally, DDM1, a SWI/SNF chromatin-remodeler, required for maintenance, facilitates DNA methyl-transferases access to heterochromatin\(^8\).

Plant TE silencing has been mostly studied at steady states reflecting how long-established silent conditions are maintained (e.g. MET1, DDM1) or faithfully reiterated (e.g. RdDM). Such pre-established states may be also temporarily reversed by stress or during development, correlating with production of TE-derived 21-nt siRNA species, which may backup TE suppression via post-transcriptional gene silencing (PTGS or RNAi)\(^2,8-12\). However, these studies have focused on evolutionary ancient transposon families, which have undergone strong proliferation and degeneration, are highly similar in sequence and mostly clustered in centromeric and pericentromeric regions acting as potential reservoirs for homology-based TE silencing\(^13,14\). Consequently, how new TEs with no homology to the host genome are detected and silenced \textit{de novo} remains largely unknown, as are the genomic and epigenomic impact of their proliferation.

Ideally, investigating this process entails the reconstruction of \textit{de novo} TE invasion/silencing events. Use of TE-based transgenes\(^15\) toward this endeavor has
generated confounding results because of the intrinsic susceptibility of transgenes to RNAi/TGS, of artificial transposition patterns, and of a general lack of evolutionary context. Outbreeding can circumvent these caveats\textsuperscript{16} but is limited in scope due to scarce genetic resources available in wild accessions. Endogenous transposition bursts were reconstructed in Arabidopsis \textit{met1} and \textit{ddm1} single mutants or by their combination with RdDM mutants\textsuperscript{17,18}. Reactivated TEs included members of the evolutionary young, low-copy \textit{ATCOPIA93} family of Long-terminal repeat (LTR) retrotransposons, which proliferate by reverse-transcription (RT) of RNA intermediates. Studying these bursts provided insights into the requirements and genomic consequences of \textit{de novo} TE invasion. However, trans-generational studies, needed to explore the proliferation of reactivated TEs, are hindered by the developmental defects and sterility displayed by inbred \textit{met1} or \textit{ddm1} plants. The continued presence of the causal mutations also intrinsically precludes analyses of \textit{de novo} TE epigenetic silencing following their reactivation\textsuperscript{19}.

Recently, populations of epigenetic recombinant-inbred lines (epiRILs) were derived from crosses between wild-type (WT) and \textit{met1} or \textit{ddm1} homozygous Arabidopsis\textsuperscript{20,21}. Inbreeding epiRILs by single-seed descent in the presence of WT copies of \textit{MET1} or \textit{DDM1} generated collections of lines with mosaic epigenomes consisting of WT and methylation-depleted segments. We reasoned that epiRILs might provide a unique resource to explore trans-generationally the full biology of reactivated, low copy TEs without the complications evoked above.

**RESULTS**

\textit{De novo} invasion of the Arabidopsis genome by a single copy LTR-retroelement

In \textit{met1}-3 inbred plants, the mutation used to generate the original epiRILs population, loss-of-DNA methylation allowed mobilization of an intact \textit{ATCOPIA93} family representative, \textit{Évadé} (EVD)\textsuperscript{18}, defining only two single-insertion loci in \textit{Arabidopsis thaliana} Col-0 genome: \textit{AT5G17125}, a euchromatic locus corresponding to EVD itself, and the pericentromeric, heterochromatic, locus \textit{AT1G34967}, termed \textit{Attrapé} (ATR). Several \textit{met1}-epiRIL lines showed EVD mobilization\textsuperscript{18} and we confirmed that in epiRIL line\#15 ("epi15") EVD reactivation correlated with the unmethylated status of it’s 5’-LTR, inherited from the \textit{met1} parent. By contrast, the \textit{ATR} 5’-LTR displayed WT DNA methylation levels in epi15; moreover, only EVD
transcripts were detected in both *met1* parent and epi15 (Fig.1a-b, Supplementary Fig.1a-c), suggesting that MET1-independent mechanisms restrict *ATR* expression.

*EVD* mobilization was then followed over several generations of epi15-inbreeding, starting from the 8th generation (F8) at which epiRILs were available. Southern analysis showed a significant increase in *EVD* copy number and accumulation of extra-chromosomal DNA from F8 to F11 (Fig.1c). Thus, by overcoming the complications seen in *met1* or *ddm1* inbred plants, and by circumventing...
transgenesis, the epi15 lineage provided the closest approximation to a *de novo* genomic invasion event, thus constituting an ideal material to study the mechanisms underlying the proliferation, detection and eventual demise of the single, euchromatic *EVD* copy of *A. thaliana* Col-0.

**L2-specific EVD expression preserves host integrity and ensures transmission of new insertions through the gamete precursor cells**

Despite the increasing *EVD* copy number, morphological defects were infrequent in the epi15 lineage, as reported for other *met1* epiRILs (Fig.1d-e; Ref.18). To gain better insights into the *EVD* life cycle in its host, we thus investigated *EVD* RNA accumulation patterns in epi15 F10 individuals using *in situ* hybridization (Fig.1f-o). Undetectable in globular-stage embryos (Fig.1f), *EVD* RNA was restricted to the adaxial sub-epidermal (L2) layer of cotyledons at all subsequent stages of embryonic development (Figs 1g-i) until the formation of dormant seeds (Fig.1j). Adaxial L2-specific *EVD* RNA was also detected during post-embryonic growth, mostly in young developing leaves (Fig.1k-l). *EVD* RNA remained undetectable in the shoot apical meristem (SAM) during embryonic (Fig.1g-j) and adult (Fig.1k-o) growth. The SAM, set early in embryogenesis (Fig.1f-g), continuously generates stem cell niches from which all aerial organs derive. Floral meristems also lacked *EVD* RNA (Fig.1m) but it accumulated in the adaxial L2 of young gynoecia, the modified foliar organs bearing the ovules\textsuperscript{22,23} (Fig.1m-n). In particular, *EVD* RNA was detected in the four placental regions ("P", Fig.1n) from each of which one L2 cell differentiates into a megaspore mother cell undergoing meiosis to form the haploid gametophyte\textsuperscript{24} (Supplementary Fig.2a-b). *EVD* RNA was, however, undetectable later in developing ovules (Fig.1o). Unlike in WT plants displaying no hybridization signals, these patterns were seen identically in two other independent *met1* epiRILs and were recapitulated by a transcriptional reporter fusion to the *EVD* 5'LTR promoter (Supplementary Fig.3c). L2-specific *EVD* transcription and meristem exclusion thus rationalized why somatic retrotransposition seemed developmentally inconsequential to the epi15 lineage. The pre-fertilization transience of *EVD* expression in the germline also likely delineated a discrete spatio-temporal window for transmitting new insertions to progenies. Since each gametophyte originates from a distinct placental L2 cell, individual seeds, including from the same fruits (siliques), were predicted to display distinct numbers
and genomic patterns of EVD insertions. This was confirmed by Southern analyses at F10, unraveling an extensive process of genetic diversification (Fig. 1p-q).

**Figure 2. Antiviral PTGS is activated against, and suppressed by, EVD.** (a) EVD copy number analysis by qRT-PCR in WT, epi15-F8 and -F11. (b) qRT-PCR analysis of EVD RNA levels in WT, epi15-F8 and -F11, normalized to ACT2. (c) Northern analysis of EVD siRNAs in WT, epi15-F8 and -F11. Distinct probes mapping EVD genome were used. U6: RNA loading control. (d) Distribution of EVD 21-22-nt deep-sequencing siRNA reads (same RNA as in (c)) onto the EVD genome. y axis: normalized reads per million sequences (rpm); x axis: chromosome 5 coordinates at EVD locus. s, sense; a, antisense. (e) Northern analysis of EVD siRNA levels of p35S:EVD in WT, rdr6 or dcl2/4 backgrounds. epi15-F11 RNA was used as reference; probe set as in (c). miR165: RNA loading control. (f) Northern analysis of EVD siRNAs in AGO1 and AGO2 immunoprecipitates (IP) using the 3’GAG probe (c). miR171 (AGO1), miR393b (AGO1) and miR393b* (AGO2), were probed to validate specific sRNA enrichment into cognate AGOs. U6: control for RNA carry-over from IP cleared extracts. Western analysis of input and IP fractions is shown; Coom: coomassie staining of total protein. (g) Schematic of p35S:EVD and p35S:EVDΔRBD. (h). Northern analyses at 4 dpi in N. benthamiana leaves. Tracks 1-4: GFP silencing suppression assay; co-expression of p35S:GFP with p35S:P19 or p35S:P19m provide positive and negative controls (details in Supplementary Fig. 2). Tracks 5-8: EVD and EVDΔRBD mRNA and siRNA levels upon co-expression with, or without p35S:P19 or p35S:P19m. EtBr, U6: ethidium bromide staining of rRNA, sRNA loading control.
An RDR6-dependent PTGS process intrinsic to EVD primary transcription forms the first host defense layer against EVD

Correlating with increased EVD insertions and expression, epi15 accumulated increasing levels of EVD-derived sRNAs from F8 to F11; by contrast, EVD mRNA and sRNAs were below detection in WT plants (Fig2a-d). EVD-derived sRNAs were in both sense and antisense orientation, predominantly 21-22-nt-long and mostly derived from the GAG 3'-end (3'GAG) in a symmetrical pattern evoking processing of a long dsRNA precursor (Fig.2c-d;25,26). EVD RNA secondary-structure predictions failed, however, to identify such a precursor, suggesting it was perhaps an RT or retrotransposition by-product. This idea was tested with an EVD transgene in which the ubiquitous 35S promoter (35S:EVD) replaced the LTR promoter required for RT and retrotransposition. In five out of ten individual 35S:EVD T1 transformants (WT Col-0), EVD expression correlated with 21-22-nt siRNA production specifically from the 3'GAG region, as in epi15 F8-F11 (Fig.2e, lanes 1-3). This high PTGS incidence and congruence of sRNA patterns suggested that long dsRNA synthesis was intrinsic to EVD transcription and thus likely involved a host-encoded RDR. RDR6 was an obvious candidate, as all of its known functions are coupled to DCL4 and DCL2 action, producing respectively 21- and 22-nt siRNAs27. Introducing 35S:EVD into the rdr6-12 null background indeed dramatically decreased 3'GAG siRNA accumulation (Fig.2e, lanes 4-5); the remaining, primary siRNAs were eliminated in dcl2-dcl4 double mutants (Fig.2e, lanes 5-7). Therefore, EVD was subjected to the same DCL4-, DCL2- and RDR6-dependent siRNA pathway that targets viruses for PTGS28,29.

The GAG nucleocapsid RNA-binding domain enables cis-protection of EVD RNA against PTGS

Although abundant EVD-derived siRNAs were efficiently loaded into their cognate effectors, AGO1 and AGO2, from F8 onward (Fig.2f), EVD expression and mobilization kept increasing from F8 to F11, as if EVD resisted PTGS. Successful plant virus infections, despite abundant virus-derived siRNA accumulation, rely on production of trans-acting viral suppressors of RNA silencing (VSRs;30). To test if EVD encodes such a VSR, we subjected 35S:EVD to a silencing suppression assay
Figure 3. Onset of EVD silencing over multiple generations. (a) qRT-PCR analyses of EVD expression and copy number in successive epi15 generations. EVD mRNA levels were normalized to ACT2 and then to epi15 F8, arbitrarily set to 1. *: the EVD copy number at F12 includes extra-chromosomal DNA owing to active retrotransposition, stopped at F13-14. (b) Northern analysis of LTR and 3’GAG siRNA and EVD mRNA levels in WT and epi15-F8-to-F14. Hybridization to miR165 and EtBr staining provide RNA loading controls. (c) Distribution of 24-nt deep-sequencing siRNA reads onto the EVD genome in WT, epi15 F8, F11 and F14 generations, as in Fig.2D. (d) qRT-PCR analyses of EVD expression and copy number in successive ddm1 epi454 generations from F5 to F15. EVD mRNA levels were normalized to ACT2 and then to epi454 F5, arbitrarily set to 1. *: same as (a) but at F7. (e) Two alternative models for the onset of LTR 24-nt siRNAs. See main text for details. (f) Northern analysis of LTR and 3’GAG siRNAs in ten individual epi15 F11 plants and eight F12 siblings from the single F11 plant #7. U6: RNA loading control.

in N. benthamiana\textsuperscript{31}, in which a transiently expressed GFP transgene is both a trigger and a target of PTGS, achieved within 4 days; co-expression of P19, a 21-nt siRNA-binding VSR, strongly enhances GFP mRNA accumulation in this assay, unlike the siRNA-binding-deficient mutant, P19\textsuperscript{m} (Fig.2h, lanes 1-2; Supplementary Fig.4a-b). GFP mRNA levels remained low upon co-expression of 35S:EVD (Fig.2h, lane 3; Fig.2b) despite production of cognate EVD polyprotein products (Supplementary Fig.5b-c), arguing against an EVD-encoded trans-acting VSR.
Strikingly, EVD transcripts were still detected 4- and up to 8-days post-delivery despite high siRNA accumulation, whereas GFP mRNAs were below detection at these times, even if co-expressed with P19 (Supplementary Fig.4b; Fig.2h, track 3 vs 2). Our search for an EVD-encoded cis-acting factor possibly explaining the resistance of EVD mRNAs to PTGS focused on GAG, which contributes to the structural proteins of virus-like particles (VLPs) required for coating retrovirus/retroelement RNA. In S. cerevisiae, VLP coating protects Ty3 LTR-retrotransposon RNA from endonucleases via the conserved Zn-finger CCHC-motif (Cx2Cx4Hx4C) of the GAG nucleocapsid RNA-binding domain (RBD). The EVD-encoded GAG CCHC-motif was readily identified by sequence alignment (http://www.gydb.org; Supplementary Fig.5e) and removed from 35S:EVD by an in-frame deletion, leading to 35S:EVDΔRBD (Fig.2g). Having confirmed cognate EVDΔRBD polyprotein processing and GAGΔRBD stability (Supplementary Fig.5b-c), 35S:EVDΔRBD was expressed with, or without, P19 or its non-functional allele P19m. Without P19, EVDΔRBD mRNA levels were strongly reduced compared to those of WT EVD, remaining unaltered irrespective of P19 or P19m (Fig.2H, lanes 3-4 and 5-6). By contrast, those from EVDΔRBD reached WT levels upon co-expression of P19, but not P19m (Fig.2h, lanes 7-8). Therefore, the GAG nucleocapsid RBD enables cis-protection of EVD RNA against PTGS; this property likely facilitated EVD proliferation over multiple generations in epi15 despite a strong 21-22-nt siRNA build-up resulting from transcription of new insertions.

Effective EVD silencing occurs at a threshold of ~40 copies and coincides with the onset of abundant LTR-derived 24-nt siRNAs

Trans-generational investigation of EVD’s biology in inflorescences of bulked individuals showed that EVD RNA levels and copy number had attained a peak at F11 and F12, respectively, coinciding with the highest 3’GAG siRNA levels (Fig.3a-b). EVD RNA levels subsequently declined beyond F11, until their dramatic loss at F14, a loss not attributable to reduced EVD insertions that had attained a plateau of ~40 copies/genome from F12 onward (Fig.3a-b). F14 displayed, however, strongly reduced 3’GAG siRNAs (Fig.3b; Supplementary Fig.6a) whose accumulation requires EVD transcription by PolII, suggesting that EVD had undergone TGS at F14; accordingly, F14 coincided with the onset of LTR-derived 24-nt siRNAs (Fig.3b-c). Initially undetectable from F8 to F11, these species also accumulated at F12 and
F13, albeit less than at F14, coinciding with reduced 3’GAG siRNAs (Fig.3b) and EVD insertions reaching ~40 copies genome.

To address if this sudden TGS was epi15-specific or, instead, reflected a bona fide and general feature of EVD biology, we used separate epiRILs in the ddm1 background21, having confirmed EVD mobilization. In both line epi454 and epi439, displaying respectively 33 and 7 EVD copies at generation F8 (Supplementary Fig.6b), abrupt EVD silencing coincided with insertions reaching ~40 copies genome, even though a dissimilar number of generations (respectively 7 and 17) was required to attain this figure in each line (Fig.3d; Supplementary Fig.7b-d). Therefore, potent

Figure 4. ‘All or nothing’ LTR methylation of EVD copies.
(a) Bisulfite sequencing-based methylation analysis at the EVD 5’LTR, as in Fig.1B (details in Supplementary Table 2 and Supplementary Fig.7a).
(b) Schematic of EVD insertions mapped in epi454 F8. Blue triangles: original EVD and ATR loci; red triangles: new insertions, of which ten (A-J) were selected for analysis.
(c) qPCR analysis of EVD expression and copy number in four individuals (1-4) from epi454 F9, F11 and F15. EVD mRNA levels were normalized to ACT2.
(d) Northern analysis of LTR and 3’GAG siRNAs in the F9, F11 and F15 plants used in (c). U6: RNA loading control.
(e-m) McrBC-qPCR-based methylation analysis at the 5’ (LB, left border) and 3’ (RB, right border) LTRs of EVD new insertions depicted in (b), in the epi454 F9, F11 and F15 plants used in (c) and (d). (e) Average methylation at all EVD LTRs inspected; (f-m) methylation levels at LB (5’LTR) and RB (3’LTR) of specific EVD insertions.
TGS of *EVD* occurs at a fixed copy-number threshold (∼40) coinciding with production of LTR 24-nt siRNAs and concurrent loss of all classes of 3’GAG siRNAs.

**3’GAG 21-22-nt or LTR 24-nt siRNAs define one of two mutually exclusive states acquired in individual seeds within each generation**

Figure 3b suggested that distinct siRNA species coexisted in late epi15 generations, as if a progressive trans-generational increase in *EVD* copy number had instigated a build up of 24-nt LTR siRNAs, ultimately triggering TGS and loss of 3’GAG siRNAs at F14 (Fig.3e, model A). An alternative scenario existed, however, in which occurrence of 21-22-nt 3’GAG siRNAs and of 24-nt LTR siRNAs represented either one of two mutually exclusive states (Fig.3e, model B) reflecting differences in *EVD* copy number between individuals of the same generations (Fig.1q). At most generations, including F11-F13, these distinct states would have been confounded by analysis of bulked individuals. This idea was confirmed by analyzing single F12 siblings derived from a unique F11 parent (Fig.3f), validating model B. Therefore, the sharp silencing onset at F14 probably reflected that a vast majority of plants had acquired, by this generation, a ∼40 copy number, which presumably underpins a threshold of overall *EVD* RNA expression level above which TGS is triggered.

**The onset of 24-nt siRNAs triggers LTR methylation and TGS of many individual EVD copies simultaneously**

Bisulfite DNA sequencing of *AT5G17125*, the original Col-0 *EVD* insertion, showed the coincidence of 24-nt siRNAs with LTR methylation at F14, but not F8 or F11, which was even higher than in Col-0 and in the CHH context, hallmark of RdDM (Fig.4a, Supplementary Fig.7a). To test if LTR *de novo* methylation affected all or only some *EVD* copies and correlated with the ∼40 copies number, we exploited whole genome re-sequencing data available for epiRILs 454 and 439 (Gilly et al., submitted). This allowed McrBC-qPCR-based DNA methylation profiling of specific *EVD* insertions randomly spawned and genetically fixed from the epi454 and epi439 F8 generations (10 and 3 insertions respectively; Fig.4b, Supplementary Fig.7b).

In epi454 F9, in which the ∼40 copies figure had been attained, two of four plants tested (#1 and #3) displayed low EVD RNA levels and LTR 24-nt siRNAs exclusively (Fig.4c-d), correlating with LTR methylation at all *EVD* insertions inspected (Figs.4e-f, 4h); in contrast, the two other F9 plants (#2 and #4) accumulated only 3’GAG-derived siRNAs, had higher *EVD* expression (Fig.4c-d) and
low or no LTR methylation (Fig. 4e, 4g, 4i). At F11 and F15, displaying low or undetectable EVD RNA levels, 100% of plants contained LTR 24-nt siRNAs exclusively (Fig. 4c-d), with LTR methylation at all EVD copies inspected (Fig. 4e, 4j-m). Analyses of epi439 provided strikingly similar results (Supplementary Fig. 7b-d), which, combined with those of Figure 3, indicate that EVD silencing entails a PTGS-to-TGS shift whereby many copies accumulated over generations simultaneously undergo an "all-or-nothing" methylation process via LTR 24-nt siRNAs.

3′GAG 24-nt siRNA production and 3′GAG DNA methylation precede the PTGS-to-TGS shift underlying the final demise of EVD

What mechanism could possibly connect the 40-copy number to the PTGS-to-TGS shift in EVD silencing? We first considered that the gain in EVD copy number across generations had favored the emergence, by position effects or rearrangements, of one or a few peculiar EVD loci producing disproportionate amounts of 24-nt siRNAs with potent trans-TGS properties. This idea, however, was ruled out in segregation analyses of an F14 backcross (Supplementary Fig. 8a). Further mining of deep-sequencing data revealed that at F11, but not F8, 3′GAG 21-22-nt siRNAs were mixed with a much smaller proportion of 24-nt siRNAs, the level of which was strongly reduced at F14, when TGS of EVD had been achieved (Fig. 3c). Thus, as for the 21-22-nt species, 3′GAG 24-nt siRNA accumulation was positively correlated with increased EVD copy number and PolII-dependent transcript levels. At F11, these siRNAs were detected in immunoprecipitates (IPs) of AGO4, the cognate effector of 24-nt siRNAs; this IP signal had dropped by F14 and a LTR 24-nt siRNA signal was detected instead, correlating with loss-of PolII-dependent transcription and TGS of EVD (Fig. 5a).

The processing patterns of 21-22-nt and 24-nt 3′GAG siRNAs were strikingly superimposable at F11 (Fig. 5b). The most abundant 24-nt siRNAs indeed invariably corresponded to 2-3-nt-longer versions of 21-nt species (Fig. 5b, Supplementary Fig. 8b), implying their processing from the same RDR6-dependent dsRNA precursor, albeit by distinct DCLs. Because amplified 24-nt siRNAs are usually produced by DCL3 from RDR2-, but not RDR6-dependent dsRNA, we first considered that the DCL3 protein levels had increased from F8 to F11, potentially granting DCL3 access to RDR6 products. Western analyses disproved this idea, however, also confirming unaltered DCL4 levels (Fig. 5c). Therefore, the most parsimonious explanation entails
Figure 5. RDR6-DCL3-dependent RdDM triggers EVD TGS. (a) Northern analysis of LTR and 3’GAG siRNAs in AGO4 IPs. siR1003 (AGO4), and miR171 (AGO1) were probed to validate specific sRNA enrichment into AGO4. U6: control for RNA carry-over from IP cleared extracts. Western analysis of input and IP fractions is shown. Coom: Coomassie staining of input protein. (b) Co-mapping of 21-22-nt (black) and 24-nt (red) deep-sequencing siRNA reads in epi15-F11, as in Figure 2d-3c. Inset: sequence alignment of the five most abundant 24-nt siRNAs (peaks 1-5) with overlapping 21-nt siRNA reads. (c) Western analysis of endogenous DCL3 and DCL4 in WT, epi15-F8, F11 and F14. Coom: as in (a) (d,f) Bisulfite methylation analysis at EVD 5’ and 3’GAG regions (details in Supplentary Table 3). Primers could not differentiate EVD from ATR, thus WT values are for both elements and include new EVD insertions in epi15. (e,g) Dot-plot methylation pattern of individual clones used in (d,f). Filled, empty circles: methylated, unmethylated cytosine. (h) Northern analysis of EVD siRNA and mRNA levels in segregating populations of an epi15 F11 cross to the dcl2/dcl4 double mutant used as female (♀) or male (♂) progenitors. Mutant vs WT status of DCL4 and DCL2 is indicated alongside average copy number (EVD# ± SD).
that DCL3 could effectively access the EVD-derived dsRNA at F11 due to the exceptional concentration build-up of this molecule, which likely overwhelmed the normal DCL4/DLC2 processing machinery. This build-up can be inferred from a matching, dramatic increase in 3’GAG siRNA levels from F8 to F11 (Figs. 2d, 5a). As anticipated from their loading into AGO4, 3’GAG 24-nt siRNAs induced de novo methylation of 3’GAG DNA in all cytosine contexts at F11; also as expected, this methylation was absent at F8 (Fig. 5d-g). Therefore, coinciding with the copy number threshold of ~40, PTGS of EVD had acquired a DNA-level component via de novo 3’GAG DNA methylation.

3’-to-5’ methylation spreading intrinsic to EVD transcription likely initiates TGS

RDR6-DCL3-dependent 3’GAG de novo methylation at F11 and beyond likely constituted a key epigenetic trigger for the eventual demise of EVD. This predicted that introducing a low EVD copy number into the dcl2-dcl4 double mutant background would accelerate the TGS onset by enabling RDR6-dependent dsRNA to be processed solely by DCL3, as predicted from the hierarchical interactions linking siRNA-generating DCLs. We tested this idea with individuals in which an introgressed dcl2-dcl4 double mutation and F11-derived active EVD insertions were allowed to segregate. Either WT or double-homozygotes mutant for both DCL4 and DCL2 were selected on the basis of their EVD insertion number ranging from 20 to 30 copies/genome; pools of two plants with the same genotype where then analyzed. As expected, all plants with the WT background accumulated the EVD RNA and 3’GAG siRNAs exclusively (Fig. 5h, lanes 5, 6). Strikingly, neither the 3’GAG siRNAs nor the EVD RNA accumulated in plants with the dcl2-dcl4 mutant genotype, which, by contrast, displayed only LTR-derived 24-nt siRNAs (Fig. 5h tracks 4, 7). Therefore, allowing the sole action of DCL3 on EVD-derived dsRNA was sufficient to trigger rapid and potent TGS of a low number of EVD copies.

A second implication for a key role for 3’GAG siRNAs in TGS onset was the existence of a 3’GAG-to-5’LTR methylation spreading process, supported by a gain

---

displayed by pools of two plants with the same genotype. (i) Northern analysis of 35S-promoter and 3’GAG EVD siRNA and EVD mRNA levels in p35S:EVD plants in the WT, rdr6 or dcl2/4 backgrounds. tasiR255 and siRNA1003 detection were used to validate cognate mutant backgrounds.
in DNA methylation in all cytosine contexts on the 5’GAG region at F14, but not F11 (Fig.5d-e). If 3’-to-5’ methylation spreading was sufficient to initiate TGS, it was predicted to impair EVD transcription regardless of the nature of its regulatory sequences, an idea tested with the 35S:EVD lines (Fig.2e). Indeed, EVD transcription was abolished upon introduction of 35S:EVD in the dcl2-dcl4 double mutant, but not in the rdr6 single-mutant; this coincided with abundant 35S promoter-derived 24-nt siRNAs, the likely trigger of transgene TGS (Fig.5i).

Genomic invasion and epigenetic silencing direct expression changes and de novo epiallelism at EVD-proximal endogenous loci

The “all-or-nothing” methylation unveiled in Figure 4 suggested that LTR 24-nt siRNAs could trans-silence active EVD copies, which we tested in reciprocal crosses between WT, epi15 F11 and F14. When introduced into WT, F14-derived copies remained silenced in F1 progenies, in which lower levels of LTR 24-nt siRNAs correlated with the expected halved EVD copy number (Fig.6a; lanes 3 and 7 vs 9; Supplementary Fig.8a). By contrast, F11-derived copies remained transcriptionally active in crosses to WT, displaying reduced EVD RNA levels that correlated with decreased 3’GAG siRNAs, further underscoring the link between EVD transcription levels and RDR6-dependent siRNA production (Fig.6a; lanes 2 and 4 vs 5). Strikingly, crossing F11 to F14 resulted in silencing of all F11-derived EVD copies, as EVD RNA and 3’GAG 21-22-nt siRNAs were below detection in F1 progenies, demonstrating the potent trans-silencing abilities of LTR 24-nt siRNAs (Fig.6a; lanes 6 and 8).

One of the Copia93 remnants in the Col-0 genome is a solo LTR displaying >88% DNA sequence similarity to that of EVD, and embedded within the RPP4 promoter (AT4G16860). Bisulfite sequencing showed that, as in Col-0, the solo_RPP4 DNA was unmethylated in both epi15 F8 and F11. However, coinciding with the onset of 24-nt LTR siRNAs, it became increasingly methylated from F14 onward (Fig.6b-c, Supplementary Fig.9), correlating with reduced RPP4 expression, induced by treatments with the salicylate analog benzothiadizole (BTH;Fig.6d). To investigate the impact of newly inserted, active EVD copies on gene expression, we focused on two homozygous insertions (I and B) mapped in epi454 F8 (Fig.4d) within AT1G68420 and downstream of AT5G25550, respectively. Both full-length EVD insertions caused a marked increased in transcription at both loci in leaves of F9
individuals, displaying active EVD transcription and no LTR methylation (Fig.6e-h, 4e). By contrast, in F9 plants carrying LTR-methylated and silenced EVD copies, transcription at the disrupted AT1G68420 locus and expression of the EVD-proximal AT5G25550 locus were low, as also seen in F15, in which EVD was silenced in 100% of plants (Fig.6e-h, 4e). While, AT1G68420 illustrates gene disruption by EVD, the example of AT5G25550 shows how trans-acting LTR 24-nt siRNAs produced during the onset of EVD TGS could trigger functionally relevant epiallelism at an EVD-proximal locus existing before, or formed during, genome invasion.

Figure 6. Genetic and epigenetic consequences of EVD mobilization. (a) Northern analysis of LTR and 3’GAG siRNA and EVD mRNA levels in F1 progenies of crosses between WT, epi15 F11 and -F14 plants. EtBr, U6: ethidium bromide staining of rRNA and sRNA loading control. (b) Northern analysis of EVD-derived siRNA accumulation in epi15-F8, -F11, -F14 and -F15. U6: RNA loading control. (c) Bisulfite sequencing-based methylation analysis at the Copia93-derived solo-LTR within the RPP4 (AT4G16860) promoter, as in Figure 1b (details in Supplementary Fig.9 and Supplementary Table 4). (d) qRT-PCR analysis of RPP4 expression in WT, epi15-F8, -F11, -F14 and -F15 plants treated with benzothiadizole (BTH) or with mock. RPP4 mRNA levels were normalized to ACT2 and then to WT mock-treated samples, arbitrarily set to 1. (e-f) Expression levels of AT1G68420 (e) in WT and ddm1 parental lines or in epi454 F9 and F15 carrying the new insertion I (Figure 4d) in the unmethylated or methylated state (f) assessed by McrBC-qPCR as in Fig.4E. (g-h) Same as in (e-f), but with AT5G25550, carrying the 3’-proximal EVD insertion B (Figure 4b). AT1G68420 and AT5G25550 mRNA levels were normalized to those of AT5G13440, AT2G36060 and AT4G29130. Blue lines: probes for qRT-PCR.
DISCUSSION

Most studies of plant TEs have involved the transcriptional reactivation of highly proliferated and predominantly heterochromatic transposon remnants, whereas rare studies of intact, active elements have intrinsically precluded analyses of TE re-silencing. Here, we have resolved these conundra through the unique use of epiRILs. By erasing its DNA methylation status in met1/ddm1 backgrounds, we have awakened an endogenous single-copy, functional retroelement, and subsequently followed the regain of its epigenetic silencing in a near-WT genome, effectively reconstructing the sequence of events underlying a retrotransposon burst (Fig.7, Supplementary Fig.10). This exercise unraveled remarkable features of EVD’s biology that undoubtedly underpin the success of its proliferation over multiple generations. First, the L2 confinement and meristematic exclusion of EVD likely explain the paucity of phenotypic aberrations associated with its host’s vegetative growth and seed-set. This preservation of host fitness and L2-specific gametophytic transmission patterns likely explain the extraordinary reservoir of genetic diversity constituted in successive offsprings. A second unexpected underpinning of EVD’s invasive success is its resistance to PTGS. The finding that the RDR6, DCL2-4 and AGO1-2 antiviral pathway constitutes the first defense layer against EVD rationalizes previous observations that TE-derived 21-nt siRNA production is a common response to transcriptional reactivation of TEs in ddm1 mutant or stressed plants\textsuperscript{9-12,19}. The use of GAG-dependent protection rather than direct suppression of this PTGS response by EVD might be explained by the fact that trans-acting VSRs, as deployed by most viruses, concurrently interfere with microRNA functions, impacting plant development and fertility. This would be unfavorable to EVD, which, unlike infectious plant viruses, co-evolves with its host and must preserve its integrity.

Paradoxically, EVD’s innocuity to PTGS likely constitutes the very foundation of its eventual demise, invariably initiated beyond a threshold of ~40 copies/genome, regardless of the epiRILs genetic backgrounds or initial copy-number. Our data suggest that this threshold underlies the limit of PolII-dependent EVD transcript accumulation (itself determined by the number of active copies) beyond which RDR6-dependent dsRNA becomes accessible to DCL3, in addition to DCL4 and DCL2, action; the ensuing AGO4-loaded 3’GAG 24-nt siRNAs guide de novo methylation of 3’GAG DNA as a premise to the eventual TGS of EVD. The key underlying notion
that distinct DCLs may share a continuum of action and templates under circumstances of exceptionally high dsRNA accumulation fully concurs with, and biologically rationalizes, numerous observations that abundant viral- or hairpin-derived dsRNA normally diced into 21-22-nt siRNAs effectively becomes a potent substrate for DCL3 in dcl4-dcl2 double mutant Arabidopsis. We rule out that 3′GAG 24-nt siRNAs derive from the classical PolII-RDR2 pathway because (i) their abundance was strictly correlated with active, PolII-dependent EVD transcription: it was high at F11 and dramatically reduced at F14 upon TGS onset; (ii) these species shared a near-identical processing pattern with the PolII- and RDR6-dependent 21-22-nt 3′GAG siRNAs; and (iii) introducing the rdr6 mutation into 35S:EVD transgenic plants strongly reduced their accumulation without production of 24-nt species, as expected if a parallel RDR2-dependent pathway was in operation.

The access of DCL3 to RDR6-dependent dsRNA, which itself derives from PolII transcripts, unveils a previously unexplored RdDM facet explaining how PTGS might acquire a DNA-level component, and is consistent with earlier work on

Figure 7. Rise and demise of EVD (a) In early proliferation stages, a low EVD copy number accounts for moderate transcription and dsRNA synthesis by RDR6 (1). Processing of dsRNA by DCL4 and DCL2 produces 3′GAG 21-22-nt siRNAs (2) loaded into AGO1/2 (3) that can only partly degrade the EVD RNA (4) due to its protection by GAG as part of putative viral-like particles (VLPs; 5). (b) Because single seeds acquire distinct patterns and numbers of EVD insertion over generations, an increasing amount of progeny plants will attain the ∼40 copies threshold beyond which saturating dsRNA levels allow processing, by DCL3, of 3′GAG 24-nt siRNAs loaded into AGO4 (6). These induce de novo methylation of the 3′GAG DNA (7), which, in turn, initiates antisense transcription, possibly by PolIV/PolV, and coordinated spreading of methylation towards the LTR (8). (c) Eventual epigenetic silencing of EVD is achieved over a few subsequent generations resulting in TGS maintained at the LTR by RdDM (9). The ensuing 24-nt LTR siRNAs display potent cis- and trans-silencing properties (10).
transgene silencing\textsuperscript{39-42} and more recent studies. Hence, a similar RDR6-AGO1-AGO2 pathway guides DNA methylation of intergenic regions and recently evolved pseudogenes: as with \textit{EVD}, this methylation is also associated with 21-nt and 24-nt siRNA production and is sensitive to mutations in \textit{AGO4}, but not \textit{RDR2}\textsuperscript{43,44}. Furthermore, comprehensive methylome analyses of many Arabidopsis silencing mutants\textsuperscript{45} revealed that of all mutants not involved in canonical RdDM, \textit{rdr6} and the related \textit{rdr1} displayed the strongest hypo-methylation affecting all cytosine contexts. RDR6-associated methylation was linked to PolII-dependent transcription of active genes rather than PolIV/PoLV-dependent transcription of TE remnants/repeats. Moreover, as with \textit{EVD}, it coincided with 21- and 24-nt siRNA production, but with a clear predominance of the former species, contrasting with the near-exclusive 24-nt siRNAs accumulating at canonical RdDM target loci.

\textit{De novo} 3’GAG methylation eventually activated, via 3’-5’ spreading, RdDM at the 5’ LTR of \textit{EVD}, leading to its final demise via TGS. However, unlike in most cases of RdDM affecting TE remnants in \textit{cis}, the ensuing LTR 24-nt siRNAs displayed remarkable \textit{trans}-silencing abilities, as originally observed in transgene silencing studies\textsuperscript{46-51}. While the 3’-5’ spreading mechanism involved in this process awaits further clarification, it is tempting to propose that the initial 3’GAG methylation recruits antisense transcription driven by PoLIV, known to guide AGO4 to active gene promoters \textit{via} transcription of long non-coding RNAs\textsuperscript{52,53}. Our work finally shows how trans-silencing by 24-nt LTR siRNAs can establish functional, \textit{de novo} epiallelism at \textit{EVD}-proximal genes. Owing to its intrinsic reversibility and flexibility, this additional layer of epigenetic control may further optimize plant evolution and adaptation as a result of TE proliferation, in full agreement with McClintock’s visionary concept of ‘controlling elements’\textsuperscript{1}. 
METHODS

Plant Material.
Arabidopsis thaliana ecotype Columbia-0 (Col-0) was used in this study as WT, 
met1-3, dcl2-1, dcl4-2, rdr6-12, ddm1 mutants and the met1- or ddm1-derived 
epiRILs have been previously described\textsuperscript{20,54-59}, propagation of the epiRILs was 
performed by single seed descent procedures.

Plasmid Construction and Transformation.
EVD constructs were generated using single site Gateway\textsuperscript{®} technology (Invitrogen), 
primer sequences can be found in Supplementary Table 5. GAG-POL ORF from 
EVD chromosome 5 insertion was amplified from BAC F2K13 by PCR using Phusion 
HF DNA Polymerase (Thermo Scientific) and cloned into pJET1.2 (Thermo 
Scientific). Deletion of the GAG RBD was carried out by Site-Directed Mutagenesis 
following the QuickChange\textsuperscript{™} protocol developed by Stratagene. HA and FLAG tags 
were added by PCR. EVD and EVD\textsuperscript{ΔRBD} clones were recombined into pDONRP1P2. 
LR recombination was performed using the 35S promoter containing vector 
pB7WG2,0 (https://gateway.psb.ugent.be/\textsuperscript{60}) and the resulting construct introduced 
into Agrobacterium tumefaciens GV3101. Arabidopsis transformation was carried out 
by the floral dip method\textsuperscript{61}. At least ten transgenic lines were analysed. pLTR:GUS- 
GFP transgenic plant was a gift from Dr. Lionel Navarro. Transient expression 
assays were performed as previously described\textsuperscript{31}.

In situ hybridization (ISH).
EVD-ISH probe DNA was cloned into the pGEM\textsuperscript{®}-T Easy Vector (Promega). Antisense EVD Digoxigenin-(DIG)-labelled riboprobe was in vitro synthesized from 
pGEM-EVD-ISH plasmid T7 RNA polymerase (Promega) and DIG RNA Labelling Mix 
(Roche). Samples were fixed in FAA (3.2% formaldehyde, 5% acetic acid, 50% 
ethanol) twice by vacuum infiltration for 30 min. prior to dehydration in ethanol and 
embedding in Paraplast Plus\textsuperscript{®} (McCornick Scientific). 10 \textmu m sections were generated 
using a Leica microtome and mounted on Polysine\textsuperscript{TM} (Menzel-Glaser) slides. In situ 
hybridization was performed as described by Jeff Long 
(http://www.its.caltech.edu/~plantlab/protocols/insitu.htm). Slides were hybridized
with approximately 40-45 ng of DIG-labelled probe overnight at 52°C. Immunological detection of the DIG-labelled probes was performed using a DIG Nucleic Acid Detection Kit (Roche).

**Bisulfite Sequencing-based DNA Methylation Analysis.**
Genomic DNA was extracted from *A. thaliana* inflorescences using the DNeasy Plant Mini Kit (Qiagen). Bisulfite treatment was performed using the EpiTect Bisulfite Kit (Qiagen). Bisulfite treated DNA was then amplified by PCR. Primer design and PCR cycling conditions were performed following the recommendations in Henderson et al., 2010\(^{62}\). PCR fragments were cloned into pGEM\(^®\)-T Easy (Promega) and individual colonies were sequenced. Sequences were analysed with Kismeth software (http://katahdin.mssm.edu/kismeth/revpage.pl\(^{63}\)) to obtain the percentage of methylated sites for each sequence context. Wilson score interval was used to find the 95% confidence interval of these percentages. Results shown were obtained from two independent experiments.

**McrBC-qPCR-based DNA Methylation Analysis.**
McrBC digestion was performed on DNA from young leaves; RT-qPCR was performed on equal amounts of digested and undigested DNA. Results were expressed as percentage of loss of molecules after McrBC digestion. Error bars represent standard deviation from two independent experiments involving duplicate PCRs each.

**Real-Time qPCR.**
*EVD* copy number quantification and expression analysis were conducted by Real Time (RT)-qPCR on a LightCycler\(^®\)480 II apparatus (Roche) using SYBR\(^®\) Green qPCR kit (Eurogentech) and gene-specific primers (Supplementary Table 5). For *EVD* copy number quantification, genomic DNA was extracted from *A. thaliana* with DNeasy Plant Mini Kit (Qiagen). *EVD* genomic copies were subjected to absolute quantification analysis and normalized with the single-locus gene *ACTIN2* (*ACT2, AT3G18780*). For RT-qPCR, total RNA was extracted with RNeasy Plant Mini Kit (Qiagen). 2 µg of total RNA was reversed transcribed into cDNA using Maxima\(^®\) First Strand cDNA Synthesis Kit (Thermo Scientific) for relative quantification. Results were normalized using *ACT2*. For RT-qPCR analysis of genes nearby new EVD
insertions (Fig. 6), cDNA was generated using only oligo-dT to prime the RT. Error bars represent standard deviation from three technical replicates of two biological replicates.

**RNA Blot Analysis.**
Total RNA was extracted from *A. thaliana* and *N. benthamiana* with Trizol reagent (Invitrogen). Unless otherwise indicted, 10 plants of the different epiRILs generations were bulked for analysis such that we always used the same number of plants per experiment per generation. 5-10 plants of the transgenic lines were used for RNA analysis. For RNA gel blot analysis of high-molecular weight (HMW) RNA, 10-15 µg of total RNA were resolved on denaturing 1.2% agarose gels with 2.2 M formaldehyde, capillary transferred to Hybond™-NX membrane (GE Healthcare) and cross-linked by UV irradiation. Ethidium bromide staining before transfer was used to confirm loading. For low-molecular weight (LMW) RNA blot analysis, LMW RNA was isolated as described before and 5-10 µg were resolved on a denaturing 17.5% polyacrylamide/urea gel, transferred to membrane by electroblotting and chemically cross-linked. Radiolabelled probes for detection of *EVD* and *GFP* RNA and siRNAs were made by random priming reactions using the Prime-a-gene kit (Promega) in the presence of α-32P-dCTP (Hartmann Analytic). The templates used were PCR products (Supplementary Table 5). DNA oligonucleotides complementary to miRNAs, tasiRNAs, *cis-acting* siRNAs or U6 snRNA were end labelled with γ-32P-dATP using T4 PNK (Thermo Scientific).

**DNA Blot Analysis.**
Unless otherwise indicted, 10 plants of the different epiRILs generations were bulked for genomic DNA extraction using the CTAB method and 1 µg of DNA was digested overnight with *Ssp*I (Thermo Scientific), separated by gel electrophoresis in 1% agarose gel and transferred to Hybond™-NX membrane (GE Healthcare) and cross-linked by UV irradiation. 32P-labelled DNA probes were generated by random priming using *EVD* PCR products as template (Supplementary Table 5). Detection and analysis of *EVD* band pattern was performed using ImageQuant TL 7.0 software (GE Healthcare).
Protein Analysis.
Total proteins were extracted from 0.1-0.3 g of A. thaliana and N. benthamiana tissues ground in liquid nitrogen and homogenized with 0.6 ml of extraction buffer (0.7 M sucrose, 0.5 M Tris-HCl pH 8, 5mM EDTA pH 8, 0.1 M NaCl, 2% [v/v] 2-mercaptoethanol) containing 2 µM proteasome inhibitor MG-132 (Calbiochem®, Millipore) and Protease Inhibitor Cocktail (cOmplete®, Roche) following manufacturer’s instructions and mixed for 5 min. An equal volume of water-saturated Tris-buffered phenol was added and mixed for 5 min. Phases were separated by 10 min. centrifugation (12000g at 4°C). Proteins were precipitated from the phenol phase by addition of 5 vol. of 0.1 M ammonium acetate in methanol and incubated at -20°C overnight. After 15 min. centrifugation (5000g at 4°C), precipitate was washed with ammonium acetate in methanol and resuspended (3% [v/v] SDS, 62.3 mM Tris-HCl pH 8, 10% [v/v] glycerol). 100-300 ng of total protein were resolved on SDS-PAGE, transferred by electroblotting onto Inmobilon®-P PVDF membrane (Millipore) and incubated with antibodies in PBS with 0.1% Tween-20 and 5% non-fat dried milk following standard western blot procedures. After secondary antibody incubation with HRP-conjugated goat anti-rabbit (Sigma), detection was performed by ECL Western Blotting Detection Kit (GE Healthcare). Equal loading was verified by coomassie staining of the membrane after western blotting. Affinity-purified antibodies were used at indicated dilutions: AGO1 (1/8000), AGO2 (1/4000), AGO4 (1/2000), DCL3 (1/1000), HA (1/5000), FLAG (1/7000). Anti-serum to DCL4 was used at 1/5000 and to GFP at 1/15000.

AGO Immunoprecipitation.
A. thaliana inflorescences were ground in liquid nitrogen and homogenized in 1ml of IP buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 10% [v/v] glycerol, 0.1% Nonidet P-40), containing 2 µM proteasome inhibitor MG-132 (Calbiochem®, Millipore) and Protease Inhibitor Cocktail (cOmplete®, Roche) following the manufacturer’s instructions, for 1 h at 4°C. Cell debris was removed by 30 min. centrifugation (12000g at 4°C). Protein was quantified from extracts to normalize total protein inputs. Extracts were precleared by 30 min. incubation with 10 µl of Protein A-agarose beads (Sigma) at 4°C (10 rpm) and removed by 1 min. centrifugation (12000 g at 4°C). Clarified lysate was incubated with antibodies for 2 h at 4°C (10 rpm), and for an additional 1 h with 10 µl of Protein A-agarose beads. Beads were collected by
centrifugation (4000g at 4°C) and immunoprecipitates washed three times (20 min each) with 1 ml of IP buffer. For RNA analysis, immune complexes were subjected to Trizol extraction (Invitrogen). Aliquots of input, flow-through (before washes) and IP fractions were subjected to protein analysis by western blot to assess efficiency of immunoprecipitation. For AGO1 immunoprecipitation antibody was used at 1/1000 dilution, and for AGO2 and AGO4 at 1/500.

**Antibodies.**

Peptides used to raise rabbit polyclonal antibodies for AGO1, AGO2, AGO4, DCL3 and DCL4 were previously described\(^{43,68-70}\). Efficiency of affinity-purified antibodies from AGO1, AGO2, AGO4, DCL3 antisera and antiserum to DCL4 were previously validated by comparing protein detection between wild type plants and the corresponding mutant\(^{43,71}\). Antiserum to GFP was a gift of Dr. D. Gilmer. Affinity purified HRP-coupled anti-HA and anti-FLAG antibodies (Sigma) were used for detection of tagged EVD proteins.

**BTH Treatment.**

To induce RPP4 expression, a 0.5 mM Benzothiadizole (BTH, Sigma) solution in water was sprayed on 2 week-old plants. Rosette leaves were harvested for RNA extraction 24h after treatment\(^{72,73}\).

**Protein alignments.**

Retroviral and retrotransposon GAG protein sequences were retrieved from the Gypsy Database (http://www.gydb.org\(^\text{34}\)). Sequence alignments were performed with the Jalview software\(^{74}\) by clustalW Multiple Sequence Alignment.

**Deep sequencing.**

Total RNA (10 µg) was extracted from Arabidopsis inflorescences with Trizol reagent (Invitrogen) and processed into sequencing libraries using adapted Illumina protocols and sequenced at Fasteris (http://www.fasteris.com, Switzerland) using the HiSeQ 2000 sequencer. Reads were mapped against the TAIR9 version of the A. thaliana genome using the Bowtie software\(^{75}\) (options: -e 50 -a --best --strata -m 5000 --
nomaqround -y) together with SAMtools\textsuperscript{76} and BEDTools\textsuperscript{77} suites. In-house R-cran scripts were implemented to display the mapping.

**Accession codes.**

Raw and processed data sets of sRNAs are deposited at NCBI’s Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE43412.

**AKNOWLEDGEMENTS**

We thank members of the Voinnet laboratory for fruitful discussions and critical reading of the manuscript. C. Brosnan is specially thanked for his intellectual input and support all throughout this project. We are indebted to D. Meyer for help with *in situ* hybridization. We thank L. Navarro (Istitute de Biologie de l'Ecole Normale Supérieure, IBENS) for providing the LTR:GUS transgenic lines. We thank D. Gilmer (Institute de Biologie Moléculaire des Plantes) for providing antiserum to GFP. This project was supported by a core grant from ETH-Z and a grant from the Swiss National Foundation, Genetics of miRNA Action and Biogenesis (31003A_132907). A.M.-O. is supported by a PhD fellowship from the ETH-Z foundation.

**AUTHOR CONTRIBUTIONS**

O.V. and A.M.-O conceived and designed the experiments, helped by V.C. on experiments involving *ddm1*-mutant epiRILs. A.M.-O., M.E and A. Martin performed the experiments. O.V., V.C., A.M.-O and M.E. analyzed the data. A. Marchais performed computer and statistical analyses. O.V and A.M.-O wrote the manuscript.

**COMPETING FINANCIAL INTERESTS**

The authors declare no competing financial interests.
REFERENCES


51. Daxinger, L. et al. A stepwise pathway for biogenesis of 24-nt secondary siRNAs and


SUPPLEMENTARY INFORMATION

Supplementary Figure 1: EVD vs. ATR expression and LTR methylation.
Supplementary Figure 2: Arabidopsis gynoecium and ovule development.
Supplementary Figure 3: EVD in situ hybridization in WT and two additional epiRILs.
Supplementary Figure 4: Suppression of silencing assay in N. benthamiana.
Supplementary Figure 5: GAG RNA binding domain conservation and EVD polyprotein product production in EVD constructs.
Supplementary Figure 6: Mapping of 21-22-nt siRNAs to EVD in epi15 F14 and EVD copy number in epi454 and 438 F8.
Supplementary Figure 7: Bisulfite sequencing dot-plots of EVD LTR in epi15 generations and EVD copy number and DNA methylation in epi439 generations.
Supplementary Figure 8: Correlation of LTR 24-nt siRNAs abundance with EVD copy number and 21-24-nt siRNAs mapping correlation on EVD.
Supplementary Figure 9: Bisulfite sequencing dot-plots of RPP4 Solo-LTR in epi15 generations.
Supplementary Figure 10: EVD life cycle in Arabidopsis.
Supplementary Table 1: Figure 1 and Supplementary Figure 1 bisulfite sequencing raw data.
Supplementary Table 2: Figure 4 and Supplementary Figure 7 bisulfite sequencing raw data.
Supplementary Table 3: Figure 5 bisulfite sequencing raw data.
Supplementary Table 4: Figure 6 and Supplementary Figure 9 bisulfite sequencing raw data.
Supplementary Table 5: List of primers.
Supplementary Figure 1: EVD vs ATR expression and LTR methylation.

(a) Restriction fragment length polymorphism (RFLP) analysis of RT-PCR products shown in Figure 1a. Genomic DNA (gDNA), and PCR products obtained from F11O6 BAC (comprising chromosome 1 ATR, AT1G34967, copy) and F2K13 BAC (comprising chromosome 5 EVD, AT5G17125, copy), were digested with BglII to determine whether transcripts originated from EVD or ATR. (b) Bisulfite sequencing-based methylation analysis at the EVD and ATR 5’LTR in WT, met1 and epi15 F8. The 5’LTR of AT5G17125 (EVD) and AT1G34967 (ATR) was specifically amplified using a primer inside the LTR and a specific primer outside of the LTR 5’ end. Cytosine methylation levels are given at CG, CHG and CHH sites. Error bars represent the 95% confidence limits given by the Wilson score interval (see Supplementary Table 1 for details). (c) Dot-plots depiction of bisulfite sequencing data obtained for individual clones used to generate the results shown in Figure 1b, Supplementary Figure 1b and Supplementary Table 1. Filled circles indicate methylated cytosines and empty circles unmethylated cytosines.
Supplementary Figure 2: Arabidopsis gynoecium and ovule development. (a) Scheme of a cross-section of the Arabidopsis gynoecium. The gynoecium is formed by the congenital fusion of two carpels (modified foliar organs that bear the ovules) through their lateral margins (cmm, yellow). Placental tissues (p, pale yellow), from which ovules arise (see b), differentiate from the cmm at the intersections of lateral and medial domains21-23. ab, Abaxial lateral domain (green); ad, adaxial lateral domain (blue); abaxial medial domain (orange). Adapted from ref. 23. (b) Ovule development in Arabidopsis. Ovule primordia emerge from the placentas, where a somatic cell from the sub-epidermal L2 cell layer (in red) differentiates into the megaspore mother cell (MMC) that undergoes meiosis, forming a syncytium (sync) of haploid nuclei, of which three degenerate, leaving a single haploid product: the female megaspore (FM). The FM then gives rise to the female gametophyte. Adapted from ref. 24.
Supplementary Figure 3: *EVD in situ* hybridization in WT and two additional epiRILs. (a) *In-situ* hybridization study of *EVD* transcript accumulation, using an antisense probe, in developing embryos of WT and epi15 F10. Longitudinal sections of embryos at globular- (1-2), heart- (3-4), mid-torpedo- (5-6), late-torpedo- (7-8) and mature green-stage (9-10). (b) *In-situ* hybridization study of *EVD* transcript accumulation, using an antisense probe, in various tissues of WT, epi15 F10 and two other epiRILs expressing *EVD*: epi12 (F9) and epi35 (F9) previously described in ref.18. Longitudinal section of the SAM region from a young Arabidopsis rosette (1-4). Longitudinal sections of flower bud (5-7). Gynoecium transversal sections at floral developmental stage 8 (8-11). Longitudinal sections of ovule primordia at floral developmental stage 12 (12-15). (c) Longitudinal section of plant rosettes at the SAM region of non-transgenic WT and pLTR:GUS-GFP transgenic WT plants, following GUS-staining. Bars = 50 μm for (a) 1-4, (b) 12-15; 100 μm for (a) 5-10, (b) 1-7 and (c); 10 μm (b) 8-11.
Supplementary Figure 4: Suppression of silencing assay in *N. benthamiana*. (a) Suppression of silencing assay by Agrobacterium-mediated transient expression in *N. benthamiana* leaves. *Agrobacterium tumefaciens*, infiltrated into plant leaves mediates transfer of transgenes into plant cells, allowing ectopic expression of the genes of interests, but it rapidly decreases 3-4 days post infiltration (dpi) due the onset of PTGS. This feature of the system has been routinely used to characterize proteins with VSR activity by co-infiltrating the protein of interest with a GFP reporter transgene acting both as a trigger and a target of PTGS. In the scheme, when infiltrated alone, GFP protein levels (visualized under UV illumination) are reduced due to the action of PTGS; however co-infiltration with the VSR P19, prevents the onset of PTGS in the infiltrated area allowing significantly higher levels of GFP accumulation. P19 acts as a head-to-tail homodimer that specifically binds 21-nt small RNA duplexes, preventing their loading into AGO. This property is abolished in the point mutant P19m, which dimerizes but is unable to bind small RNAs. (b) Analysis of p35S:EVD transient expression in *N. benthamiana*. The potential suppression of silencing activity of EVD was tested by co-infiltration of EVD constructs (Figure 2g) with the p35S:GFP reporter construct. p35S:GFP co-infiltrated with p35S:P19 or its mutant version p35S:P19m were used as controls for positive and negative suppression activity, respectively. Northern analysis of High-molecular weight (mRNA) and low-molecular weight RNA (siRNA) was carried out on samples extracted from infiltrated leaves at 4 and 8 dpi. Ethidium bromide (EtBr) staining of ribosomal RNA and detection of U6 are shown as loading controls.
Supplementary Figure 5: GAG RNA binding domain conservation and EVD polyprotein product production in EVD constructs. (a) Alignment of GAG amino acid sequences from representative members of Retroviridae and Pseudoviridae (Ty3/Gypsy and Ty1/Copia). The bottom sequence (black arrow) corresponds to the EVD-encoded GAG protein. Coloured columns in the alignment indicate residues with conserved physicochemical properties. The black box displays the consensus sequence logo of highly conserved residues corresponding to the RNA-binding motif (C_{x2}C_{x4}H_{x4}C). (b) Scheme of p35S:HA-EVD-FLAG and p35S:HA-EVD_{ΔRBD}-FLAG constructs, see methods for details. (c) Accumulation of the EVD polyprotein N-terminal and C-terminal products from p35S:HA-EVD-FLAG and p35S:HA-EVD_{ΔRBD}-FLAG was assessed by Western analysis of total proteins extracted from N. benthamiana-infiltrated leaves at 4 dpi. Equal loading was verified by Coomassie staining of the membranes after blotting.
Supplementary Figure 6: Mapping of 21-22-nt siRNAs to EVD in epi15 F14 and EVD copy number in epi454 and 438 F8. (a) Deep sequencing analysis of 21-22-nt siRNAs mapping to EVD in epi15 F14. y axis: number of normalized small RNA reads per million sequences (rpm); x axis: Arabidopsis chromosome 5 coordinates corresponding to the EVD locus found in Col-0 (TAIR 9). s, sense; as, antisense. (b) EVD copy number in epi454 and 439 F8 calculated as the total EVD insertions inferred from the number of heterozygous and homozygous insertions obtained by sequencing (Gilly et al, submitted) including the original EVD and ATR loci.
Supplementary Figure 7: Bisulfite sequencing dot-plots of EVD LTR in epi15 generations and EVD copy number and DNA methylation in epi439 generations. (a) Dot-plots depiction of bisulfite sequencing data obtained for individual clones used to generate the results shown in Figure 4a and Supplementary Table 2. Filled circles indicate methylated cytosine and empty circles unmethylated cytosine. (b) Scheme of Arabidopsis chromosomes showing EVD insertions found in the ddm1-derived epi439 line at F8. The original EVD and ATR loci are marked with blue triangles, red triangles indicate new transposition events. The three new insertions were selected for further analysis. (c) EVD copy number measured by qPCR in individual plants of epi439 F17. EVD copy number was not determined (ND) in the F9 and F15 generations. Error bars represent standard deviation from two independent experiments involving duplicate PCRs each. (d) McrBC-qPCR-based methylation analysis at the LTRs of EVD new insertions shown in (b) in the epi439 F9, -F15 and -F17 individual plants used in (c). Error bars represent standard deviation from duplicate PCRs in c-d.
Supplementary Figure 8: Correlation of LTR 24-nt siRNAs abundance with EVD copy number and 21-24-nt siRNAs mapping correlation on EVD. (a) To explain the PTGS-to-TGS shift in EVD silencing, we considered the possible emergence of discrete EVD loci with peculiar features allowing production of 24-nt LTR siRNAs with potent trans-TGS activities. This hypothesis had two predictable and interrelated implications: (i) such loci would produce disproportionate amounts of 24nt siRNAs independently of EVD copy numbers and, (ii) these loci would be easily segregated away from an F14. We therefore tested these ideas here. (1) EVD copy number measured by qPCR in WT, epi15 F14 and the F1 progeny of the F14 X WT cross. (2) Same as in (1) but in distinct plants of the F2 progeny, grouped according to their respective copy numbers. Error bars represent standard deviation from two independent experiments involving triplicate PCRs each. (3) Northern analysis of EVD LTR 24-nt siRNA levels in F2 progeny plants containing similar numbers of EVD insertions as determined in (2). U6 detection is shown as loading control. The results clearly show that (i) all F2s, with no exception, produce 24nt siRNAs and (ii) 24nt-LTR siRNA levels strictly correlate with the copy number of segregating EVD copies, making the emergence of the hypothesized, specific EVD loci highly unlikely. (b) Pearson correlation coefficient (Y-axis) between the 5’-end of all 24mers and all 21-mers mapping the 2,000 first nucleotides of EVD with the computation of virtual a lag from -10 to +10 nucleotides (X-axis) between the first nucleotide of all 21-mers and all 24-mers. This allows to assess if DCLs process the 24-mers and the 21-mers at the same position of a shared dsRNA substrate.
Supplementary Figure 9: Bisulfite sequencing dot-plots of *RPP4* Solo-LTR in epi15 generations. Dot-plots of bisulfite sequencing from individual clones for results shown in Figure 6b and Supplementary Table 4. Filled circles indicate methylated cytosine and empty circles unmethylated cytosine.
Supplementary Figure 10: EVD life cycle in Arabidopsis. (a) EVD expression pattern ensures host integrity through developnet and transmission of new insertions to the next generation. Once reactivated, adaxial L2-specific EVD expression (in red) restricts potential transposition to somatic tissue. Meristem (harbouring the stem cells from which aerial organs derive; M) exclusion will then allow the plant to develop without the accumulation of EVD new insertions. However, L2 expression and transposition in gynoecia (flower female reproductive organ), particularly at the placental regions (P) which give rise to the ovules (ov), from which a L2 cell differentiates into
the haploid female megaspore (FM), will allow transmission of EVD new insertions (see Supplementary Fig.2 for more details). Therefore, after fertilization, and embryo development, each seed will display different number and pattern of EVD insertions. (b) Rise and demise of EVD. In early proliferation stages (A), a low EVD copy number (light blue; i) accounts for moderate transcription and dsRNA synthesis by RDR6 (1). Processing of dsRNA by DCL4 and DCL2 produces 3’GAG 21-22-nt siRNAs (2) loaded into AGO1/2 (3) that can only partly degrade the EVD RNA (4) due to its protection by GAG as part of putative viral-like particles (VLPs; 5). Because single seeds acquire distinct patterns and numbers of EVD insertion over generations, an increasing amount of progeny plants will attain the 40 copies threshold (dark blue; ii) beyond which saturating dsRNA levels allow processing, by DCL3, of 3’GAG 24-nt siRNAs loaded into AGO4 (6), these induce de novo methylation of the 3’GAG DNA (7), which, in turn, initiates antisense transcription, possibly by PolIV/PoIV, and coordinated spreading of methylation towards the LTR (8). Eventual epigenetic silencing of EVD is achieved over a few subsequent generations (red, iii) resulting in TGS maintained at the LTR by RdDM (9). Ensuing 24-nt LTR siRNAs display potent cis- and trans-silencing properties (10).
Supplementary Table 1: Figure 1 and Supplementary Figure 1 bisulfite sequencing raw data.

Raw data from the bisulfite sequencing analysis in Figure 1b and Supplemtary Figure 1. The number of CG, CHG and CHH sites as well as the number of sequenced clones are indicated. Wilson score intervals are shown for a confidence level of 0.95.

Supplementary Table 2: Figure 4 and Supplementary Figure 7 bisulfite sequencing raw data.

Raw data from the bisulfite sequencing analysis in Figure 4a and Supplemtary Figure 7. The number of CG, CHG and CHH sites as well as the number of sequenced clones are indicated. Wilson score intervals are shown for a confidence level of 0.95.

Supplementary Table 3: Figure 5 bisulfite sequencing raw data.

Supplementary Table 4: Figure 6 and Supplementary Figure 9 bisulfite sequencing raw data.

Raw data from the bisulfite sequencing analysis in Figure 1b and Supplemtary Figure 1. The number of CG, CHG and CHH sites as well as the number of sequenced clones are indicated. Wilson score intervals are shown for a confidence level of 0.95.
### Supplementary Table 5. List of Primers

(Non-converted bisulfite primers sequence is indicated in brackets)

<table>
<thead>
<tr>
<th>EXPERIMENT</th>
<th>PRIMER</th>
<th>SEQUENCE (5'-3')</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EVD RT-PCR/RFLP</strong></td>
<td><strong>EVD RT F</strong></td>
<td>GATCCGGTGAGACCTCAATACTTTT</td>
</tr>
</tbody>
</table>
|                             | **EVD RT R**    | TACTCCCTCCCTCAAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGA
<table>
<thead>
<tr>
<th>439 EVD-ChIIR R</th>
<th>CAACATTCTAGGTTTCTTGC</th>
</tr>
</thead>
<tbody>
<tr>
<td>439 EVD-ChIV F</td>
<td>TCCTCCACCCAAAAATCTGAG</td>
</tr>
<tr>
<td>439 EVD-ChIV R</td>
<td>GGCTTGTTGTTTGGCAAGTGAT</td>
</tr>
<tr>
<td>EVD-LB R</td>
<td>ACTCTGGTCATTGTTGACCTTG</td>
</tr>
<tr>
<td>EVD-RB F</td>
<td>CCAAGACCAAGGAACTCTCTC</td>
</tr>
<tr>
<td>454 EVD-A qF</td>
<td>ATGAAAATATTGGCGGTTCAT</td>
</tr>
<tr>
<td>454 EVD-A qR</td>
<td>TGAGAATTCTGCTCTAGATTAGTG</td>
</tr>
<tr>
<td>454 EVD-B qF</td>
<td>GCTCAAGGCGACTTGTGAAAG</td>
</tr>
<tr>
<td>454 EVD-B qR</td>
<td>ACTTTTAAGGAACTGTGTTGAG</td>
</tr>
<tr>
<td>454 EVD-C qF</td>
<td>CATAGCAGTCTGATCATCAGAGAG</td>
</tr>
<tr>
<td>454 EVD-D qF</td>
<td>TTGGCTGGAAGATGAGTTC</td>
</tr>
<tr>
<td>454 EVD-D qR</td>
<td>ACCAGAGATATAGAATCTGATGTGAAGTG</td>
</tr>
<tr>
<td>454 EVD-E qF</td>
<td>CGGTTAAGACGAGGCTGAGTG</td>
</tr>
<tr>
<td>454 EVD-E qR</td>
<td>AGGACTGACGTCACACCTA</td>
</tr>
<tr>
<td>454 EVD-F qF</td>
<td>CTTTCCGATCGACTTCAAT</td>
</tr>
<tr>
<td>454 EVD-F qR</td>
<td>TCATTCTCATATGTTGTTGAG</td>
</tr>
<tr>
<td>454 EVD-G qF</td>
<td>CTGTACATTGTTAATTTTATCCATCTTG</td>
</tr>
<tr>
<td>454 EVD-G qR</td>
<td>AAATCAACAACATATGAAATGATCAAGT</td>
</tr>
<tr>
<td>454 EVD-H qF</td>
<td>GAAGCTTATTGCAGGCTTCA</td>
</tr>
<tr>
<td>454 EVD-H qR</td>
<td>TTTACATTGTTAGCTTCTGG</td>
</tr>
<tr>
<td>454 EVD-qF</td>
<td>TTGAGCTTTACATTATAGAAGACAGGGTAG</td>
</tr>
<tr>
<td>454 EVD-J qF</td>
<td>AAAACCTGGTACCTCTGTCGAG</td>
</tr>
<tr>
<td>454 EVD-J qR</td>
<td>TCAAAAAACTCAGAGGCTAGAG</td>
</tr>
<tr>
<td>439 EVD-ChI qF</td>
<td>AAACGAGTCCAGAATAGAAGCA</td>
</tr>
<tr>
<td>439 EVD-ChI qR</td>
<td>AGACTTCTCCTGTTGATGCAAGTC</td>
</tr>
<tr>
<td>439 EVD-ChIII qF</td>
<td>TGAGAGTGGATTCTGGTGTAGA</td>
</tr>
<tr>
<td>439 EVD-ChIII qR</td>
<td>CTGAGCCCAAAAGGAAAGTC</td>
</tr>
<tr>
<td>439 EVD-ChIV qF</td>
<td>CGTACCAAACATAAGGAGAACCC</td>
</tr>
<tr>
<td>439 EVD-ChIV qR</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>5'GAG Bisulfite PCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>EVD-5GAG BS F</td>
</tr>
<tr>
<td>EVD-5GAG BS F</td>
</tr>
<tr>
<td>EVD-5GAG BS F</td>
</tr>
<tr>
<td>EVD-5GAG BS F</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3'GAG Bisulfite PCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>EVD-3GAG BS F</td>
</tr>
<tr>
<td>EVD-3GAG BS F</td>
</tr>
<tr>
<td>EVD-3GAG BS F</td>
</tr>
<tr>
<td>EVD-3GAG BS F</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RPP4 qRT-PCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>RPP4-qRT F</td>
</tr>
<tr>
<td>RPP4-qRT R</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RPP4 SOLO LTR Bisulfite PCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>RPP4-SLTR BS F</td>
</tr>
<tr>
<td>RPP4-SLTR BS F</td>
</tr>
<tr>
<td>RPP4-SLTR BS F</td>
</tr>
<tr>
<td>RPP4-SLTR BS F</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AT1G68420 qRT-PCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT1G68420 qF</td>
</tr>
<tr>
<td>AT1G68420 qR</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AT5G25550 qRT-PCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT5G25550 qF</td>
</tr>
<tr>
<td>AT5G25550 qR</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>qRT-PCR Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT2 qF</td>
</tr>
<tr>
<td>ACT2 qR</td>
</tr>
<tr>
<td>AT5G13440 qF</td>
</tr>
<tr>
<td>AT5G13440 qR</td>
</tr>
<tr>
<td>AT2G36060 qF</td>
</tr>
<tr>
<td>AT2G36060 qR</td>
</tr>
<tr>
<td>AT4G29130 qF</td>
</tr>
<tr>
<td>AT4G29130 qR</td>
</tr>
</tbody>
</table>
ADDENDUM

After the questions raised on the integrity of figures 2h, 5h and 6a and the following institutional investigation, I provide here the results of the re-probing of the original membranes given to the external investigation commission at ETH-Zurich to prove the reproducibility of the data. The report presented here was also provided to the Journal editors.
Case 1: Figure 2h

High-molecular weight northern blot probed against EVD mRNA

Scan of original film

After enhanced contrast

Figure 2h from NG 2013

Concerned blot

Used for figure
Case 2: Figure 5h

Low-Molecular Weight Northern Blot probed against EVD-LTR derived siRNAs

1) Original membrane re-probed against LTR siRNAs (phosphorimager)

3) Original scan of autoradiography on film after probing against tasiRNA 255 used for figure 2h

2) Original membrane re-probed against GAG siRNAs (phosphorimager)

4) Superposition of 1 and 3
5) Original membrane re-probed against tasiRNA 255 (phosphorimager)

6) Superposition of 5 and 3
Case 3: Figure 6a

Low-Molecular Weight Northern Blot probed against EVD-LTR derived siRNAs

High-Molecular Weight Northern Blot probed against EVD mRNA

1) Original membrane re-probed against LTR siRNAs (phosphorimager)

2) Original membrane re-probed against GAG siRNAs (phosphorimager)

3) Original membrane (HMW-NB) re-probed against EVD mRNA (phosphorimager)
2.2 Mandatory splicing and premature termination of Copia-like retrotransposons triggers RNA silencing in Arabidopsis.

Reference:

Manuscript in preparation

Authorship:

I performed all the laboratory-based experiments with the help of Stefan Oberlin under my supervision. I wrote the draft of the manuscript, corrected by Prof. Dr. Olivier Voinnet.
Mandatory splicing and premature termination of Copia-like retrotransposons triggers RNA silencing in Arabidopsis

Arturo Mari-Ordóñez¹,², Stefan Oberlin¹, Alexis Sarrazin¹ and Olivier Voinnet¹,²

¹ Swiss Federal Institute of Technology (ETH-Z), Department of Biology, Zurich, Switzerland.

² Life Science Zurich Graduate School, Molecular Life Science Program, University of Zurich, Zurich, Switzerland.

Correspondence to: voinneto@ethz.ch
ABSTRACT

Transposable elements (TEs) are endogenous, host-encoded, molecular parasites able to move in their host genome. Early recognition of active TEs to silence them is therefore important to minimize their harmful effects upon transposition. RNAi is the most widespread defense mechanism to silence TEs among eukaryotes, however, very little is known about the TE features that allow cells to recognize them and trigger RNAi in plants. Here, we show that the source of siRNAs from the Ty1/Copia-type retrotransposon Évadé (EVD) is a prematurely terminated, short, spliced transcript solely dedicated to the production of its major structural component, Gag. Both, splicing and features of GAG RNA, modulate the ability of this transcript to trigger RNA silencing, establishing a link between the biology of TEs and their recognition by the host RNA silencing machinery. We also show that splicing is a well conserved feature of Copia-like elements in Arabidopsis.
INTRODUCTION

Mobile genetic elements known as transposable elements (TEs) are genomic parasites that rely on the host transcription and translation machinery to fulfill their life cycle\(^1\). Since their discovery by McClintock at the beginning of the second half of last century\(^2\), transposons have been found in nearly all genomes, contributing to their plasticity and evolution\(^3-9\). However, due to their mobile nature, TEs can also pose a threat by disrupting gene function\(^4\). Hence, tight regulation of TEs is important to avoid their deleterious effects. In eukaryotes, DNA methylation and/or histone modifications are the most common epigenetic marks deposited on chromatin to impose transcriptional gene silencing (TGS) on TEs, thereby preventing their expression\(^10,11\). However, TGS can be reset during development\(^12\) or altered by environmental cues including biotic and abiotic stress, leading to TE reactivation\(^13-15\). In addition, some transposons are poised to actively respond transcriptionally to certain stresses\(^16,17\).

In plants and other organisms, post-transcriptional gene silencing (PTGS) is a widespread strategy to contain transcriptionally reactivated TEs and thereby safeguard genome integrity during critical developmental stages including germline development\(^18,19\). In PTGS, 20-to-35-nt-long RNA molecules are loaded into Argonaute (AGO) or AGO-like effector proteins to guide TGS and/or PTGS of complementary DNA sequences or RNA molecules, respectively\(^20\). In metazoans, PIWI-associated small RNAs (piRNAs) silence transposons in the germline\(^18\). In \textit{D. melanogaster}, dedicated genomic clusters (such as \textit{FLAMENCO}) act as “genetic traps" for active TEs, which, upon antisense transcription are converted into piRNAs via the still poorly understood mechanisms of primary piRNA synthesis\(^8,21,22\). Although such adaptive immune devices against TEs do not exist in plants, TE reactivation is often concomitantly accompanied by the production of active TE-derived siRNAs 21-24-nt in length\(^13,23-26\) that are thought to maintain TE RNAs at a low level. Therefore, plant genomes must be somehow able to discriminate the reactivated TE transcripts from the pool of cellular mRNAs, and target them for silencing.

One possible mechanism of TE recognition has been recently proposed upon analysis of decrease in DNA methylation (\textit{ddm1}) Arabidopsis mutants. In this model, TE transcripts might be targeted by canonical micro-RNAs (miRNAs) or epigenetically-activated miRNA (eamiRNA) that possibly co-evolved with some
TEs$^{27}$. The model posits that the initial miRNA-guided cleavage of TE transcripts promotes de novo synthesis of TE-derived siRNAs via the action of RNA-dependent RNA polymerase 6 (RDR6) and DCL4/DCL2. Nonetheless, the alleged widespread nature of this mechanism has been questioned, given that multiple RNA quality control (RQC) mechanisms present a robust barrier to the pervasive access of RDR6 to miRNA cleavage fragments in plants$^{28}$. Additionally, canonical plant miRNAs usually display highly confined and cell-autonomous expression patterns that are unlikely to commonly overlap spatially with the tissues or cell types undergoing stress- or developmentally-triggered TEs reactivation$^{28}$. Thus, the use of the ddm1 mutant background, in which TEs are transcriptionally reactivated in the whole plant, probably provides an overestimated picture of the haphazard interactions that possibly link canonical miRNAs to TE transcripts. A more conceivable scenario, albeit much more restricted in its scope, entails that the genomic proliferation of TEs during “TE-burst” events might be accompanied by TE copy rearrangements that form miRNA-like precursors$^{29,30}$ and thus undergo the same genetic and epigenetic control as the TE they arose from$^{27}$. As a result, the ensuing eamiRNA and the founder TE family would be epigenetically reactivated in a coordinated manner and would be naturally confined within the same spatial and temporal niches. This model, however, requires the target TE to have previously colonized the host genome and, therefore, likely applies mostly to evolutionary ancient transposon families$^{28}$. The model fails to explain how young or newly invading TEs with no sequence homology to the genome are detected by the plant silencing machinery$^{28}$.

To address this question we focused on the Ty1/Copia-like long terminal repeat (LTR) retrotransposon Évadé (EVD), member of the young, low-copy, ATCOPIA93 family in A. thaliana$^{31}$. We have previously shown that the epigenetic reactivation of EVD is first antagonized by a host PTGS response via the production of 21-22-nt-long siRNAs through the coupled activities of RDR6 and DCL4/DCL2$^{32}$. Strikingly, the capacity of EVD to promote PTGS was found independent of its ability to transpose: transgene constructs harboring the EVD coding sequence under the control of the ubiquitous 35S promoter (35S:EVD) but lacking elements required for reverse transcription and genome integration, triggered the same response$^{32}$. Furthermore, endogenous EVD and 35S:EVD displayed a strikingly similar and specific pattern of siRNAs restricted to a narrow window within the 3'-proximal region of the Gag open-reading frame (ORF), which encodes the structural components
required for virus-like particle (VLP) formation\textsuperscript{32}. We thus decided to explore which mechanism(s) could possibly link EVD transcription to the initiation of 21-22-nt siRNA biogenesis as the earliest event underpinning EVD's recognition by the host.

**RESULTS**

**EVD-derived siRNAs are DCL1-independent**

Transposon-derived siRNAs that accumulate in wild-type pollen or in mutants impaired in epigenetic silencing such as *decrease in DNA methylation* (*ddm1*)\textsuperscript{24,25} are RDR6-, DCL4- and AGO1-dependent\textsuperscript{33}. These resemble trans-acting siRNA (tasiRNAs) and other amplified endogenous siRNAs whose biogenesis is triggered by miRNA-directed cleavage under rare and controlled circumstances\textsuperscript{34,35}. Creasey *et al.* recently showed that synthesis of some transposon-derived siRNAs is indeed initiated by DCL1-dependent miRNAs\textsuperscript{27}. Most EVD siRNAs are lost in the *rdr6-12* mutant background\textsuperscript{32} (Fig. 1a) and EVD transcripts have been predicted as potential targets for the DCL1-dependent miR-833 and miR-171 (Ref. 27; Supplementary Fig. 1). We thus inspected the available small RNA libraries from *ddm1-2, ddm1-1 rdr6-15* and *ddm1-2 dcl1-11*\textsuperscript{27} to investigate if EVD-derived siRNA production is also initiated by one or several miRNAs. However, no reduction in EVD siRNA accumulation was detected in *ddm1-2 dcl1-11* compared to *ddm1-2* tissues (Supplementary Fig. 1a-c), although, as expected, the vast majority of these molecules was eliminated in *ddm1-11 rdr6-15* double mutants (Supplementary Fig. 1a,d), supporting our previous findings in *met1* recombinant inbred lines\textsuperscript{32}. Therefore, a mechanism distinct from miRNA-directed RNA cleavage must account for the recruitment of RDR6 onto EVD transcripts.

**The EVD Gag protein levels but not those of full-length EVD mRNAs increase in rdr6 mutants**

A remarkable feature of the epigenetically-reactivated EVD is its ability to protect itself from the initial host PTGS response as part of viral-like particles, owing to the Gag nucleocapsid RNA-binding capacity\textsuperscript{32}. Nonetheless, the mere accumulation of Gag-derived 21-22-nt siRNAs implies that a subset of unprotected EVD transcripts must be used as substrates for RDR6\textsuperscript{32} to initiate the PTGS response. In 35S:EVD plants, the EVD full-length transcripts (fl-EVD) showed no significant increase in the *rdr6* mutant compared to WT background, despite the near-complete loss of EVD.
siRNAs in *rdr6* (Fig. 1a). To test whether the lack of siRNAs in *rdr6* affected the *EVD* protein rather than RNA levels, we raised an antibody against Gag (see Materials and Methods for more details). Western blot analysis confirmed the specificity of the antibody, which detected a unique ~35kDa product from the reactivated endogenous *EVD* in the *met1*-derived epi15 lineage (epi15 F11)\(^{31,32}\). The same 35kDa protein was also detected in transgenic WT plants expressing 35S:*EVD*, but not in non-transgenic WT plants, in which the endogenous *EVD* is transcriptionally silenced (Supplementary Fig. 3). Strikingly, in 35S:*EVD* plants, the Gag protein levels were 5 to 10-fold higher in the *rdr6* than in the WT background, in stark contrast to the near-unchanged fl-*EVD* transcript levels (Fig.1a-b). This discrepancy could be explained if the main output of *EVD* siRNA action was at the translational level\(^{36}\), but this idea contradicted our previous observations that the *EVD* mRNA levels strongly decrease upon disruption of the Gag RNA-binding motif required for *EVD* cis-protection against PTGS. Moreover, these are restored to WT levels upon co-expression of a viral suppressor of RNA silencing (VSR)\(^{32}\). An alternative hypothesis, therefore, was that a GAG-producing *EVD* RNA species distinct from the fl-*EVD* genomic transcript, acted both as a specific trigger and target of PTGS.

**A short GAG-derived transcript is stabilized in *rdr6***

Expression of the compact genomes of retrotransposons is constrained by the necessity to regulate both the timing and amplitude of TE-encoded protein production\(^{37}\). A conundrum faced by LTR-retroelements, in particular, is the production of the structural Gag protein, required for VLP assembly, in much higher quantities than the catalytic products of Pol. The above results in *rdr6* thus prompted us to investigate if a previously uncharacterized *EVD* sub-genomic transcript, specialized in Gag production, could constitute the primary trigger of *EVD* PTGS. To identify this putative RNA, we compared by Northern analysis the respective accumulation of *EVD* transcripts and siRNAs along the *EVD* coding sequence in 35S:*EVD* plants with either a WT or *rdr6* background (Fig.1c). In addition to the fl-*EVD* mRNA detected, as expected, to similar levels in both genetic backgrounds, a shorter, ~1500-nt-long transcript was detected exclusively in *rdr6* plants using probes covering the Gag and 5’-proximal region of the integrase ORF (Fig. 1d-f); probes located downstream of this region could only detect the fl-*EVD* signal (Fig. 1g-h). Strikingly, the detection of *EVD* siRNAs with the same probe set was inversely
**Figure 1. Enhanced EVD GAG protein and RNA levels in the rdr6 mutant.**

(a) Northern and Western analysis of EVD siRNAs, mRNA and Gag protein levels in WT, epi15 F11 and 35S:EVD in WT or rdr6 backgrounds. tasiR255 detection is used to validate cognate mutant backgrounds. miR172 and ACTIN2 (ACT2): RNA loading control for low- and high-molecular weight Northern blots. Coom: Coomassie staining of input proteins.

(b) Comparison of 35S:EVD Gag protein levels in WT or rdr6 background by Western analysis using different amounts of total protein input (µg of protein).


(c) Probes mapping to EVD genome used in (d-h).

(d-h) Northern analysis of EVD mRNA and siRNAs in WT, epi15 F11 and 35S:EVD in WT or rdr6 backgrounds using probes depicted in (c). Black triangle indicates the presence of additional GAG mRNA. (i) Loading controls for (d-h) as in (a). (j) Location of PCR probes used in (k-l). Red arrow: cDNA gene-specific priming site (GSP) used for reverse transcription (RT) in (k).

(k) Absolute quantification of EVD mRNA molecules by qRT-PCR in total RNA from WT, epi15 F11 and 35S:EVD in WT or rdr6 backgrounds reverse transcribed using a EVD GSP(j). qPCR was carried using probes depicted in (j). Inset: Relative (%) amounts of EVD RNA molecules using GAG probe over POL probes (IN and RT).

(l) Same as (k) but using random hexamers (RH) to prime reverse transcription.

Correlated with the ability to detect the short EVD transcript in 35S:EVD rdr6 compared to 35S:EVD WT and epi15 F11 plants (Fig. 1g-i). These results supported the notion that a short EVD GAG transcript (sh-GAG) likely forms a specific template for EVD siRNA production upon recruitment of RDR6.

To confirm the existence and explore the relative abundance of this putative transcript, we performed absolute quantification of EVD RNAs by real-time PCR.
(qRT-PCR) using the cRNA equivalent method\textsuperscript{38}. This method directly compares the cDNA amplification of a given mRNA to a standard curve established with known amounts of this same RNA synthetized \textit{in vitro} (the cRNA) and subjected to the same reverse transcription (RT) conditions (see Materials and Methods). Using the fl-EVD cRNA as a standard, \textit{EVD} transcripts were quantified with three different probes corresponding to the 5’-proximal Gag region (GAG), the integrase (IN) and the reverse-transcriptase (RT), respectively (Fig. 1j). Total RNA was reverse-transcribed into cDNA using a gene-specific primer (GSP-RT) located at the 3’-end of the \textit{EVD} coding sequence (Fig. 1j), or using random hexamers (RH-RT). GSP-RT, which only allows amplification of the fl-EVD, revealed little variation, among the sample tested, in the amount of \textit{EVD} mRNA using either of the three different probes, with the GAG-to-IN and GAG-to-RT ratio remaining lower than 4% (Fig. 1k), in accordance with the expression of the three domains as a single mRNA (fl-EVD). As expected from our previous findings, \textit{rdr6} mutation had minor impact on the accumulation of fl-EVD leading to a mild 10% increase (Fig. 1k and Supplementary Fig. 4a). However, RH-RT, permitting amplification of any possible RNA derived from \textit{EVD}, yielded increased GAG-to-IN and GAG-to-POL ratios. This is specially accused in the \textit{rdr6} background where GAG levels were ~60% higher than those of IN and RT (Fig. 1l). Moreover, as opposed to the 50% increase in GAG levels in \textit{rdr6} compared to WT, IN and RT displayed very little variation (Fig. 1l and Supplementary Fig. 4b). These results confirmed that an \textit{EVD} sh-GAG transcript existed besides the fl-EVD. Such short transcript was strongly stabilized in 35S:\textit{EVD} \textit{rdr6} samples, while little or no effect was observed in the full-length mRNA (IN and RT), reinforcing our hypothesis that sh-GAG is the siRNA template and explaining the disproportionate increase in Gag protein levels in the \textit{rdr6} mutant background.

\textbf{3’-end termination and polyadenylation of the \textit{EVD} sh-GAG transcript}

Northern analysis of total versus polyadenylated RNA (polyA+) isolated from 35S:\textit{EVD} \textit{rdr6} plants revealed that a significant fraction of sh-GAG is polyadenylated, which is also the case of the fl-EVD transcript, as expected (Fig. 2a). We then used this feature \textit{i} to confirm that transcription of sh-GAG RNA is a \textit{bona fide} feature of \textit{EVD} and not an artifact of the 35S:\textit{EVD} construct, and \textit{ii} to map the 3’-end of this transcript by 3’-RACE. Total RNA from epi15 F11 (expressing the reactivated endogenous EVD) and 35S:\textit{EVD} plants was reverse transcribed using an oligo(dT)-
Figure 2. Polyadenylation status and mapping of sh-GAG 3' ends.
(a) Analysis of EVD transcripts polyadenylation by Northern blot of total and PolyA+ purified RNA from 35S:EVD in the rdr6 background. Empty triangle: full-length EVD mRNA (fl-EVD), Black triangle: sh-GAG mRNA. ACT2: positive control for the polyA+ purification. EtBr: ethidium bromide staining of rRNA control for non-polyadenylated (polyA-) RNA carry-over in the purified polyA fraction. (b) 3'RACE-PCR on EVD sh-GAG mRNA in WT, epi15 F11 and 35S:EVD in WT or rdr6 backgrounds. The same reaction was carried in the absence of cDNA (-cDNA) as control for reagent contamination. (c) Mapping of sh-GAG 3'ends from epi15 F11 and 35S:EVD in WT and rdr6 background sequenced 3'RACE-PCR products (b) to EVD coding sequence. Positions in nucleotides (nt) are given relative to the first ATG (Met) codon of EVD. Girdled numbers indicate the number of clones sequenced mapping to indicate positions. Black arrow indicates the position of primer used in 3'RACE-PCR.

anchored adaptor. RACE-PCR, performed with primers located within the Gag ORF, yielded a PCR product of the same expected size in all samples (Fig. 2b). Thus, the sh-GAG RNA is produced and terminated at the same position from both the endogenous EVD and its transgenic, non-reverse-transcribed version. Upon cloning and sequencing of ~50 RACE-PCR products, the 3' ends of sh-GAG were precisely mapped to the beginning of the IN ORF, with the majority of clones terminating at position 1500 of the EVD coding sequence (Fig. 2c). Intriguingly, no conventional polyadenylation signal (AATAAAA) was found surrounding this region, although a non-canonical CTTAAA motif was detected at position 1414 (data not shown) and will require further investigation via point-mutagenesis. Therefore, sh-GAG defines a highly abundant polyadenylated subgenomic RNA that accumulates in parallel to the fl-EVD transcript.

Partial splicing of EVD transcripts generates a Gag-only producing mRNA
Sequencing of the 3’-RACE-PCR products revealed a 296-nt deletion in all the clones inspected. The missing sequence, located at the 3’-end of GAG and
comprising most of the protease (PR) domain, is flanked by GT and AG di-
nucleotides at the left and right border, respectively, strongly suggesting a splicing
event\textsuperscript{39} (Fig. 3a). The \textit{EVD} genomic sequences flanking the deletion (CCG/GTA--
CAG/GT) are indeed highly similar to the consensus sequences of donor (DS) and
acceptor (AS) splicing sites in Arabidopsis\textsuperscript{40}, which was further confirmed using the
Netgene2 splice site predictor (http://www.cbs.dtu.dk/services/NetGene2/). RT-PCR
employing primers flanking the presumptive \textit{EVD} intron (EVD\textsubscript{i}; Fig. 3b) indeed
generated a large and a small amplicon, which, when sequenced, corresponded to
the unspliced and predicted spliced versions of \textit{EVD}, respectively (Fig.3c and data
not shown). Therefore, only a fraction of \textit{EVD} transcripts undergoes splicing, which
was further confirmed by comparing RNA-seq data obtained from WT, \textit{met1-3} and
\textit{epi15} plants (generations F8 and F11) where reads corresponding to the splicing
junction were mixed with a higher proportion of intron-derived reads (Supplementary
Fig. 5).

Remarkably the \textit{EVD} AS and DS sites are present in different coding frames,
such that splicing generates a 2-nt frame-shift creating a stop codon two amino acids
downstream of the splice junction (Fig.3a, d). Consequently, spliced \textit{EVD} transcripts
can be only translated into Gag, a situation similar to that observed with the plant
\textit{Ogre} and \textit{BARE1} LTR-retro-transposons, which also contain an intron in the PR
domain. Moreover, at least in the latter case, splicing also causes a frame-shift in the
POL domain (PR, IN, RT-RNase) generating a stop codon shortly after the 3’-end of
the \textit{Gag} ORF. This configuration provides the clear advantage of allowing the
independent production of much more structural Gag protein relative to the catalytic
components of the POL domain, all encoded by the same fl-\textit{EVD} mRNA and
produced as a polyprotein.

\textbf{sh-GAG, but not fl-EVD, transcripts are spliced}

Strikingly, a PCR screen of three hundred independent 3’-RACE-PCR clones (thus,
corresponding to sh-GAG) employing two intron-flanking primers revealed that 100%
of the clones tested corresponded to spliced sequences (data not shown). To
investigate if fl-\textit{EVD} and sh-GAG are differentially spliced, we designed primers to
specifically amplify either spliced (spl \textit{GAG}) or unspliced (u-spl \textit{GAG}) \textit{EVD} transcripts
(Fig.3e) and performed absolute quantification of \textit{EVD} transcripts using the cRNA
equivalent method described in a previous section. The cRNA standard used in these
Figure 3. Splicing of sh-GAG EVD transcript. (a) Alignment of EVD genomic DNA (gDNA) and consensus sequences from 45-50 3’RACE-PCR clones from epi15 F11 and 35S:EVD in WT and rdr6 background. Conserved splice donor and acceptor (SD, SA) sites are marked, as well as intron flanking 5’GT and 3’AT dinucleotides (in red). Frame 1: Coding frame of unspliced EVD. Frame after splicing: splice-induced frameshift. *: First stop codon (TGA) generated by frameshift. Positions in nucleotides (nt) are given relative to the first ATG (Met) codon of EVD. (b) Representation of primers flanking EVD intron (EVDi) used to diagnose EVD splicing. (c) RT-PCR analysis of EVD splicing in WT, epi15 F11 and 35S:EVD in WT or rdr6 backgrounds using cDNA primed with RH. Empty triangle: unspliced; black triangle: spliced. gDNA is used as size control for unspliced EVD. ACT2: loading control. RT-: gDNA contamination control. (d) Schematic representations of the unspliced, spliced fl-EVD and spliced sh-GAG EVD transcripts. Black dash line: splice donor (SD) and acceptor (SA) sites. Red dashed line: Exon junction (EJ) after splicing. Red squares: Stop codons generated by the splice-induced frameshift. (e) Diagram of primers used to quantify total GAG sequence containing EVD transcripts (GAG) and unspliced (u-spl GAG) or spliced (spl GAG) transcripts. (f) Absolute quantification of total (GAG total), spliced (splGAG) and unspliced (u-splGAG) EVD mRNA molecules by qRT-PCR in total RNA from epi15 F11 and 35S:EVD in WT or rdr6 backgrounds reverse transcribed using a EVD GSP (as in Fig. 1j,k). qPCR was carried using primers depicted in (e) (g) Same as (f) but using random hexamers (RH) to prime reverse transcription.

experiments corresponded to the full-length spliced EVD mRNA. A GSP-RT analysis (only detecting fl-EVD) showed that >95% of the fl-EVD is contributed by unspliced transcripts, whose levels differed only marginally between 35S:EVD WT and 35S:EVD rdr6 samples (Fig.3f and Supplementary Fig.6a). Nonetheless, the remaining low levels (3-4%) of spliced fl-EVD must be sufficient to account for
integration of spliced genomic copies, which we detected upon the reactivation of EVD in the epi15 lineage (Supplementary Fig.7 and accompanying text). By contrast, the RH-RT analysis (allowing amplification of both fl-EVD and sh-GAG) revealed that the increased GAG RNA levels seen in rdr6 where almost exclusively contributed by a gain in spliced RNAs, while the unspliced transcripts remained at similar levels (Fig. 3g and Supplementary Fig.6b). The most straightforward interpretation of these results is that EVD produces two main transcripts: (i) a full-length unspliced transcript that most likely serves as the main genomic RNA template for reverse transcription and polyprotein translation (distinguished by alternative 5’ ends, see Supplementary Fig. 7 and accompanying text), and (ii) a subgenomic, spliced, prematurely terminated and polyadenylated transcript, shGAG, which is likely destined for the sole production of Gag following the generation of a stop codon after the splice junction.

The sh-GAG RNA is the source of EVD siRNAs
The strong and specific impact of rdr6 on the stability of sh-GAG and Gag protein production strongly suggested that the sh-GAG transcript is the main source of Gag protein and EVD-derived siRNAs. Given that sh-GAG, unlike fl-EVD, is spliced, we thus tested the accumulation of intron-derived siRNAs (Fig. 4a) in epi15F11 and 35S:EVD Arabidopsis plants. While the sh-GAG RNA and 21-22-nt siRNAs were readily detected by Northern analysis employing two distinct exon probes, they were barely detectable if intron-specific probes of similar length were used instead; as expected, both exon- and intron-probes could detect fl-EVD (Fig. 4b-e). Consistent with these results, a strong drop in siRNA reads observed in the pattern of EVD-derived siRNAs (previously established in epi15 F1132) was found to overlap specifically with the intron (Fig. 4f). Furthermore, mapping the siRNA reads to the spliced EVD RNA as opposed to its genomic DNA allowed the identification of abundant and previously unmapped reads covering precisely the splice junction (Fig. 4g and Supplementary Fig. 8). These results thus identify the spliced sh-GAG transcript as the specific and near-exclusive source of EVD-derived siRNAs following its epigenetic reactivation or under conditions of its constitutive transcription.
Figure 4. sh-GAG RNA is the source of EVD siRNAs. (a) Schematic representation of exonic and intronic probes used for Northern analysis. (b-d) Northern analysis of EVD mRNA and siRNAs in WT, epi15 F11 and 35S:EVD in WT or rdr6 backgrounds using exonic and intronic probes depicted in (a). (e) miR172 and ACTIN2 (ACT2): RNA loading control for low- and high-molecular weight Northern blots. (f-g) Distribution of EVD 21-22-24-nt deep-sequencing siRNA reads from epi15 F11 onto unspliced (f) or spliced (g) EVD. y axis: total read count, (top: sense, bottom: antisense); x axis: coordinates at EVD coding sequence, position 1 corresponds to the first ATG (Met) codon of EVD. Positions of splice donor (SD) and acceptor (SA) sites, exon junction (EJ) and mapped sh-GAG 3’ end are indicated by dashed lines. Unique reads mapping to the EJ onto spliced EVD (g) are highlighted in red.

**EVD intron quality and splicing efficacy influence sh-GAG siRNA levels**

Collectively, the results obtained thus far suggested that splicing somehow singularizes the sh-GAG from the fl-EVD transcript to become specifically targeted by the plant PTGS machinery. We thus reasoned that altering EVD splicing might uncover features linking the two processes, and we accordingly engineered several mutant versions of EVD to pursue this idea. First, the intron was entirely removed from the EVD coding sequence, creating EVDΔi. Second, the EVD sequence was point-mutated at well-characterized positions known to influence splicing\(^{39,40}\), in order to increase or decrease its efficiency. To enhance splicing (EVD\(_{ES}\)), the A/T content of the intron was increased from 65.9 to 70.3% and the U1 snRNA-binding site at the 5’ splice site was mutated to match the conserved consensus sequence. To reduce splicing (EVD\(_{RS}\)), we mutated the conserved :GT and AG: di-nucleotides at the 5’ and
3’ splice sites, respectively (Fig. 5a and Supplementary Fig. 9a). All these changes were made so as to minimally perturb the authentic amino acid sequence of the EVD protein products (Supplementary Fig. 9b). These constructs were then expressed under the 35S promoter using an Agrobacterium-mediated transient assay in N. benthamiana leaves. Semi-quantitative RT-PCR using primers flanking the intron (Fig. 3b) was then performed to ascertain effects on splicing (Fig. 5b). Consistent with the observations made in Arabidopsis, mostly unspliced transcripts were amplified upon expression of the unmodified EVD (EVD<sub>WT</sub>), yielding marginal amounts of the short amplicon corresponding to the spliced mRNA. As expected, only the shorter fragment was amplified upon EVD<sub>Δi</sub> expression owing to the full intron deletion, whereas expression of EVD<sub>ES</sub> and EVD<sub>RS</sub> coincided with respectively enhanced and reduced accumulation of the spliced amplicon relative to EVD<sub>WT</sub> (Fig. 5b).

The above constructs were then expressed in leaves of WT or RDR6-deficient (RDR6i) N. benthamiana plants, in which NbRDR6 is silenced by RNAi via an inverted-repeat transgene. Northern analyses revealed that in WT N. benthamiana the levels of sh-GAG RNA produced from EVD<sub>WT</sub> were comparatively lower than those observed in 35S:EVD transgenic Arabidopsis with a WT background, presumably reflecting the strong NbRDR6 activity typically manifested in the transient expression assay. Nonetheless, a clear increase in sh-GAG levels coinciding with a loss of sh-GAG-derived siRNAs was observed in RDR6i N. benthamiana, whereas the levels of fl-EVD transcripts remained high and nearly unchanged in WT compared to RDR6i leaves (Fig. 5c; lanes 1,5). These observations were highly reminiscent of the results obtained in 35S:EVD Arabidopsis with the WT versus rdr6 mutant background (Figs. 1d-f and 4b-d), suggesting that the same sh-GAG-induced PTGS mechanism operates in both Arabidopsis and N.benthamiana systems. This inference is further supported by our previous demonstration that EVD<sub>WT</sub> generates the same specific pattern of sh-GAG siRNAs in both systems.

As predicted from the RT-PCR analysis, EVD<sub>ES</sub> and EVD<sub>RS</sub> respectively produced increased and decreased levels of sh-GAG RNA and Gag protein compared to EVD<sub>WT</sub> (Fig. 5c,d; lanes 3-4, 7-8). However, EVD<sub>ES</sub> displayed reduced levels of siRNAs compared to EVD<sub>WT</sub> (Fig. 5e; lanes 1,3) while EVD<sub>RS</sub> showed a marked increase in these species (Fig. 5e; lanes 1,4). The full-length EVD mRNA was noticeably less stable in the EVD<sub>RS</sub>-treated samples, most likely due to the
Figure 5. **EVD splicing influences siRNA levels and premature termination.** (a) Schematic representation of wild type and mutated EVD clones with altered splicing behaviour. (b) RT-PCR analysis using cDNA primed with RH of EVD splicing mutants transiently expressed by agroinfiltration in WT N. benthamiana. Mock infiltrated N. benthamiana is shown as control for any possible interference with putative endogenous EVD-like elements reactivated by infiltration. BAR: selection marker BASTA RESISTACE gene encoded within the same binary vector is used as loading control. RT-: gDNA contamination control. (c) Northern analysis of EVD splicing mutants mRNA in leaves of WT and RDR6i N. benthamiana. (d) Western analysis of EVD splicing mutants Gag protein levels from same infiltrated leaves as in (c). (e) Northern analysis of EVD splicing mutants siRNAs from the same infiltrated leaves as in (c,d). (f) Schematic representation of mutated snRNA U1 binding site EVD. (g-i) EVDmU1 mRNA, protein and siRNA analysis as in (c-e). Empty triangle: fl-EVD; black triangle: sh-GAG; red triangle: shifted unspliced sh-GAG due to intron retention. EtBr, Coom, U6: Ethidium bromide staining of rRNA, Coomassie staining of input proteins and U6 snRNA are shown as loading controls.

Reduced levels of Gag, normally required for its stabilization. These results suggest that the intron quality and splicing efficiency are important features in the initial recognition of EVD by the host silencing machinery: while efficient splicing seems to reduce the incidence of PTGS, suboptimal splicing seems to enhance it.

**Splicing enables premature termination and polyadenylation of sh-GAG**

Surprisingly, the EVD_{Δi}-treated samples accumulated almost exclusively the sh-GAG transcript and concurrently displayed dramatically enhanced Gag protein levels compared to EVD_{WT}-treated samples in WT N. benthamiana leaves. This effect was further accentuated in RDR6i leaves due to the loss of sh-GAG siRNAs (Fig. 5c,d; lanes 2,6). These results were consistent with the idea that sh-GAG, but not fl-EVD,
contributes the main bulk of Gag protein owing to the creation of a stop codon downstream of the splice junction. A rationale for the unbalanced accumulation of sh-GAG versus fl-EVD transcripts in the EVD\textsubscript{Δi}-treated samples came from previous findings that, besides its role in defining 5'-splice sites through RNA-RNA basepairing\textsuperscript{44}, the U1 snRNA also suppresses premature cleavage and polyadenylation (PCPA) from nearby cryptic polyadenylation signals found in pre-mRNAs, a function independent of its role in splicing per se\textsuperscript{45,46}. We thus considered the possibility that unspliced EVD might be normally protected from PCPA due to U1 binding in the intron. Conversely, by removing the U1 binding site, splicing would favor termination of the sh-GAG mRNA, a situation artificially created in the EVD\textsubscript{Δi} construct by the complete removal of the intron.

To test the above hypothesis, the sequence of the U1 binding site in the EVD intron was debilitated by point mutations; the resulting construct, EVD\textsubscript{mU1}, generated no detectable spliced product in the transient assay (Supplementary Fig. 9c-d). Despite this inability to splice, the accumulation of sh-GAG remained disproportionally higher than that of fl-EVD in EVD\textsubscript{mU1}-treated samples, as in the EVD\textsubscript{Δi} treatment. Additionally, in Northern analysis, the sh-GAG produced from EVD\textsubscript{mU1} displayed a slower electrophoretic mobility compared to the sh-GAG generated by the EVD\textsubscript{WT}, EVD\textsubscript{Δi} and EVD\textsubscript{ES} constructs, consistent with the retention of the 296–nt-long intron (Fig. 5f,g). Further agreeing with intron retention in the EVD\textsubscript{mU1}-derived sh-GAG, the Gag protein was in that case below detection levels, most likely because the termination codon required to form the Gag ORF was not created. Therefore, in the absence of the U1 binding site, PCPA was not inhibited, causing the unspliced sh-GAG RNA to prematurely terminate. Transposed into the context of the WT EVD transcription, these results thus suggest that the unspliced fl-EVD is naturally protected against PCPA while, on the other hand, splicing and consequential loss of U1 binding favors PCPA and the proper termination of the sh-GAG transcript. This readily explains why the vast majority of spliced EVD RNA is contributed by sh-GAG as opposed to fl-EVD transcripts (Fig. 3g and 4b-d).

**Cis-elements in the sh-GAG 3'-UTR facilitate the host PTGS response**

The extent of siRNA production following expression of EVD\textsubscript{Δi} was similar to that triggered by expression of EVD\textsubscript{WT}. Accordingly, sh-GAG accumulation was strongly stabilized in RDR6i leaves expressing EVD\textsubscript{Δi} (Fig. 5c,e; lanes 2,6). However the
EVDΔi-derived sh-GAG transcript is produced independently of splicing, suggesting that elements within the sh-GAG mRNA itself further facilitate the onset of PTGS independently of splicing. To identify these putative PTGS determinants, the sh-GAG transcript was divided into its principal components, and their ability to trigger silencing was evaluated by transient expression in WT versus RDR6i N. benthamiana. sh-GAG transgene constructs were thus engineered with or without intron (Δi) in combination with or without the sh-GAG 3’-UTR (Δu). A NOS terminator (tNOS) was included in all constructs and a stop codon introduced in those constructs carrying an intron, to allow Gag protein production from unspliced transcripts (Fig. 6a).

Northern analysis revealed that several distinct transcripts were produced from the different constructs, reflecting the various combinations of intron retention and/or transcriptional termination at the GAG 3’UTR or at the tNOS (Fig. 6b). We found that the Gag protein levels were consistently higher when the intron was removed (Fig. 6c).
6c; lanes 1,2 and 5,6 vs 3,4 and 7,8) and that the mRNA levels were consistently increased in the absence of the sh-GAG 3'UTR (GAGWT vs. GAGΔu and GAGΔi vs. GAGΔΔu; Fig. 6b). Accordingly, more siRNAs were produced from constructs harboring the sh-GAG 3’-UTR (Fig. 6d; lanes 1-4). Unlike the results obtained thus far by transient or transgenic expression of full-length EVD constructs, the presence of an intron in the minimal GAG transgene constructs promoted production of 24-nt siRNAs, which unlike the 21-nt siRNA species, were still detected in the RDR6i background (Fig. 6d; lanes 1-2 versus 5-6). Thus, these 24-nt siRNAs were most likely processed by NbRDR2, and their persistence in RDR6i leaves probably explained why mRNA accumulation from GAGWT and GAGΔu was maintained at similar low levels in WT and RDR6i plants. The NbRDR2 effect triggered by the presence of the intron in the minimal constructs used in these experiments mimics those previously obtained when U1 binding sites are introduced into GFP45. By contrast, accumulation of mRNAs derived from the intron-less GAGΔi and GAGΔΔu constructs was readily increased, with the highest levels achieved upon expression of the 3’-UTR-less GAGΔΔu. Hence, a yet uncharacterized feature of the sh-GAG 3’-UTR clearly displays an intrinsic ability to attract the silencing machinery. This notion was further substantiated by preliminary results of a pilot experiment in which the fusion of the sh-GAG 3’-UTR to a GFP reporter transgene was sufficient to significantly enhance siRNA production from this reporter in the transient assay (data not shown).

**Splicing and sh-GAG formation are conserved among Arabidopsis Copia-like elements**

The original copia element of Drosophila splices the entire POL coding sequence to generate a subgenomic RNA solely dedicated to Gag production47, yet it remains unknown if this feature is shared by other members of the Ty1/Copia superfamily of flies. Likewise, to our knowledge, EVD in Arabidopsis (this study) and BARE1 in barley48 are the only plant Copia-like elements reported so far to produce splice variants in order to increase the Gag-to-POL ratio. We thus decided to address whether splicing is a general feature of plant Copia elements and, by extension, if production sh-GAG-like RNAs might represent a widespread source of TE-derived siRNAs upon their epigenetic reactivation.
Figure 7. Conservation of \textit{EVD} splicing, intron and PCPA features among \textit{Copia}-like elements in \textit{Arabidopsis}. (a) Consensus splice behaviour of \textit{Arabidopsis} \textit{Copia}-like elements reactivated in \textit{met1}. Plotted relative densities and position of donor and acceptor splice sites along the full-length transposon sequences. Predicted and averaged core domains of all spliced elements are depicted below (LTR: long terminal repeat; GAG: GAG nucleocapsid; PR: Protease; IN: Integrase; RVT: Reverse transcriptase; RNase domain was not predicted but its located between IN and RVT). (b) Boxplots showing length (nt) of all \textit{Arabidopsis} and active spliced \textit{Copia}-like retroelement introns based on junction reads extracted from \textit{met1} RNAseq. $p$-value based on Wilcoxon rank-sum test with continuity correction. Box plot shows: median (bold line), second to third quartile (box) and upper and lower intron length ranges (whiskers). (c) Scatterplot of GAG to POL expression ratio (GAG:POL) over splicing for all active spliced \textit{Copia}-like TEs in \textit{met1} (n = 28). For each TE GAG:POn ratio was calculated as the number of reads covering the Integrase domain divided by those mapping the GAG domain. Splicing levels were assessed similarly by comparing reads mapping to the intron and to GAG.

To accurately and reliably annotate \textit{Arabidopsis} \textit{Copia}-like elements, sequences obtained from the TAIR database\textsuperscript{49} were subjected to a series of filters. First, all elements deviating by more than 10\% in sequence length from that of the reference family (as annotated in Repbase\textsuperscript{50}) were discarded. A second filter selected only those elements displaying more than 90\% sequence similarity to the corresponding family reference sequence. Each of the selected element was finally individually inspected to validate the presence of repeated LTR sequences at both the 3'- and 5'-ends. To establish a list of expressed transposons required for the classification of spliced and unspliced elements, we conducted differential RNA deep-sequencing analyses in \textit{Arabidopsis} plants with a WT or a \textit{met1} mutant background, in which many TEs, including \textit{EVD}, are epigenetically reactivated transcriptionally.
Of the selected annotated Copia-like elements, 60 were found transcriptionally active in the *met1* background. RNA reads corresponding to splice junctions were identified based on the fact that their ends mapped to two distinct genomic locations separated by a gap flanked by GT and AG dinucleotides. For each splice junction, a confidence value was calculated by dividing the amount of junction reads by the amount of total reads mapping to the element. Expressed Copia-like elements were classified as spliced if this confidence value was equal or larger than the one of the experimentally validated EVD. Conversely, active elements were deemed unspliced if the confidence value was below the EVD-based threshold. Despite the high stringency of this approach (deliberately applied to reduce the occurrence of false positives), 34 out of the 60 expressed Copia-like elements were found to splice (Supplementary Table 1). This figure is likely conservative because unspliced elements were generally less expressed than spliced ones (Supplementary Fig. 10) and, hence, corresponding low abundant junction reads could have possibly escaped our analysis. Moreover, TEs with an ability to splice displayed no evident phylogenetic relationships (Supplementary Fig. 11) and elements belonging to the same family could be classified as either spliced or unspliced (e.g. COPIA27, COPIA38, COPIA43, Supplementary Fig. 11 and Supplementary Table 1). Therefore, splicing among Copia-like elements might be even more common than what our analysis suggested.

Using LTRdigest, the core domains (LTR, GAG, PR, IN, RT) of LTR-retrotransposons could be predicted for 28 of the 34 spliced elements. Transposon length and position of each domain were then averaged to establish a “model Copia-like retroelement” onto which splice donor and acceptor sites were plotted. Remarkably, the resulting density-distribution of splice events revealed that not only splicing, but also its position between the GAG and Protease/Integrase domains, is a highly conserved feature of Copia-like elements in *Arabidopsis* (Fig. 7a). The introns of spliced elements also displayed an unusual median length of 291-nt compared to the 104-nt length of *Arabidopsis* introns (Fig. 7b) and, in all elements analyzed, splicing was predicted to cause a frame-shift in the downstream sequences (Supplementary Table 1). In EVD, splicing causes the frame-shift that creates the required stop codon for Gag translation and also potentially promotes PCPA required to generate the sh-GAG subgenomic mRNA. To investigate if the latter feature is generally shared by spliced Copia-like elements, we set to analyze the impact of
splicing on fl-TE vs. sh-GAG ratios. Investigating PCPA events would ideally entail the construction of Parallel Analysis of RNA Ends (PARE) libraries, however, splicing cannot be simultaneously assed on such libraries. Instead, as illustrated by EVD, GAG-to-POL (IN and RT) levels are diagnostic of PCPA (Fig. 1l and 5c). Therefore, GAG-to-POL was quantified using met1 RNA-seq data as a GAG-to-IN ratio between reads covering IN to those covering GAG. Besides, splicing can be intuitively detected on RNA-seq data graphic representations as a drop or gap in intron sequences. Hence, splicing was quantified as a ratio between intron and GAG coverage. The resulting plot revealed that both events, splicing and PCPA are as well linked in other TEs. Elements with higher levels of splicing also display reduced POL expression (Fig. 7c). Furthermore a close investigation of the data uncovered the existence of Copia “supersplicer” elements, in which splicing and PCPA at GAG happens so efficiently that no POL is transcribed (Supplementary Fig. 12). The conservation of splicing and coupled PCPA among other members of the Copia superfamily, suggests that the events leading to the recognition of EVD by the RNA silencing machinery might generally account for triggering siRNA responses on TEs. Unfortunately, the lack of met1 small RNA libraries prevents us from establishing a final link between the conservation of splicing among Copia-like elements, production of sh-GAG, and triggering host RNA silencing, however we will address this issues in the future.

**DISCUSSION**

Multiple mechanisms account for the recognition of TEs by host RNA silencing machineries\(^\text{52}\). In flies, a prevalent, adaptive process entails the integration of active TEs into dedicated transposon-trapping loci with distinctive genomic and epigenetic features, in which de novo inserted TEs are converted into TE-derived piRNAs\(^\text{53}\). In plants, one mechanism involves the creation, by active TEs, of unusual genomic DNA rearrangements that favor the formation of double-stranded (ds)RNA or aberrant (ab)RNA templates for the synthesis of siRNAs or, as shown recently, eamiRNAs\(^\text{27,54}\). Thus, both in flies and plants, the formation of TE-derived sRNA-producing loci can provide a genetic memory against future transposition events. However, these processes inevitably entail the prior colonization of the genome by TEs, a window during which hosts are potentially vulnerable to the detrimental effects of TEs. Thus, the question arises as to if and how hosts can actively recognize and
defend themselves against new invasive TEs with little or no homology to their genomes. Based on recent analyses of Arabidopsis ddm1 mutants, in which many TEs are reactivated in many tissues, a model was proposed whereby serendipitous interactions between evolutionary old miRNAs and reactivated TEs would trigger the RDR6-dependent production of TE-derived siRNAs. However, transposed in the context of WT plants, the model shows several limitations that make it probably inapplicable to many classes of transposons\textsuperscript{28}, including, as shown here, the Ty1/Copia-like LTR-retrotransposon EVD. This led us to propose that alternative, innate and TE-intrinsic mechanisms must additionally underlie the first wave of siRNA production during \textit{de novo} TE invasions or upon their epigenetic reactivation.

In plants, RNA silencing is among the first barriers to be opposed to viruses through the recognition of their inherent ability to produce dsRNA as an obligatory step of the viral life cycle\textsuperscript{55}. Our previous work showed that the same antiviral PTGS pathway constitutes the primary defense line against EVD upon its reactivation. We now show, here, how unique features of TE biology stimulate this pathway, thereby allowing the discrimination of TE transcripts from endogenous mRNAs. By exploiting the fact that transgene constructs transcribing the EVD coding sequence produce the same pattern of siRNAs as the reactivated WT EVD itself, we have unveiled the existence of a specific RDR6 substrate that fully explains the creation and narrow distribution of EVD-derived siRNAs, almost exclusively centered on the GAG coding sequence. This substrate, sh-GAG, is a previously uncharacterized, spliced subgenomic mRNA exclusively dedicated to the production of the Gag protein; sh-GAG accumulates in both 35S:EVD and epi15 F11 plants as an intrinsic and most likely mandatory feature of Copia retrotransposon biology. We show that two intertwined and distinctive features of sh-GAG account for its ability to trigger PTGS: the quality and efficacy of its splicing as well as \textit{Cis}-elements present in the sh-GAG mRNA itself. We also provide evidence that formation of sh-GAG-like transcripts is a common feature of plant Copia-like elements that is likely to provide a distinctive and widespread signature for the initiation of PTGS by this large class of plant TEs.

The RNA of retrotransposons, like those of retroviruses, serve two distinct and mutually exclusive purposes: (i) as part of VLPs, genomic-size RNAs must be reverse-transcribed into new cDNA copies to be integrated into the host genome while (ii) concomitantly, protein translation must somehow proceed\textsuperscript{1}. Furthermore, the compaction and small size of TE genomes imposes strong constraints on TE
gene expression, often resulting in the use of polycistronic mRNAs expressed from the same promoter. In this context, a significant challenge for TEs is to produce an excess of Gag nucleocapsid (the major structural component) over Pol protein (encompassing the Pr, IN and RT-RNase enzymes) in order to successfully assemble VLPs, in which reverse-transcription takes place. Across kingdoms, different retro-elements have evolved distinct strategies to enable GAG to be more efficiently translated than POL, the most common of which relies on ribosomal frameshifting, where POL is generally in the ±1 frame relative to GAG. A second widespread mechanism relies on stop codon read-through for those TEs in which GAG and POL are in the same frame.56 Interestingly, most plant TEs, and in particular the Copia superfamily, encode GAG and POL as a single transcriptional unit but display none of the above-mentioned translational recoding mechanisms. The striking prevalence of single-ORF genomic organization in plant TEs suggests that strong negative selection is exerted against such mechanisms. We recently showed how nonsense-mediated decay (NMD), whereby mRNA displaying premature or internal termination codons are selectively degraded, constitutes a major and pervasive restriction pathway against plant RNA viruses. Being subjected to the same genome expression constraints as viruses, TEs are also most likely scrutinized by the plant NMD machinery, which may strongly prevent their use of translational recoding processes.

In the rdr6 mutant background, increased levels of Gag protein were almost exclusively contributed by stabilization of the sh-GAG mRNA which is, without exception, always spliced. As a consequence of a splice-induced frameshift, a newly generated stop codon allows proper and exclusive Gag translation. However, this solution for enhanced Gag expression would come at a strong cost in the context of the full-length EVD RNA, because it would place the downstream ~4kb out-of-frame, resulting in a genomic transcript with an abnormally long 3’-UTR and a multitude of premature termination codons (PTC; Fig. 3d), two prevalent NMD-activating features. Thus, the avoidance of NMD and possibly other RNA quality control (RQC) pathways may well underpin the necessary premature termination and polyadenylation of the spliced sh-GAG transcript by EVD. However, a strong termination/polyadenylation signal would concurrently pose a threat to the transcription of downstream sequences from the fl-EVD, mandatory for both reverse-transcription and POL translation/processing. An apparent solution to this conundrum
is to encode a PCPA signal that is mostly or only functional/accessible after splicing, as suggested by the unforeseen ability of EVΔ to produce sh-GAG nearly exclusively. Nonetheless, the process that couples splicing to premature termination creates itself a second complication, because splicing removes parts of the Pr domain, and must, therefore, remain relatively inefficient in order to simultaneously accommodate transcription and productive translation of the unspliced fl-EVD.

We propose that the snRNA U1 plays a major role in balancing these processes by inhibiting the cryptic PCPA signal. Indeed, deletion or mutation of the U1 binding site to impair its interaction with the EVD RNA invariably resulted in accumulation of sh-GAG at the strong expense of the fl-EVD transcript. Enhancing U1 binding led to increased splicing, by contrast, agreeing with independent roles of U1 in splicing and PCPA inhibition45,46. Noteworthy, the EVD U1 binding site diverges from the consensus U1 sequence, a feature shared by many of the other TEs found in this study to display a similar splicing behavior to EVD. Strikingly, the identified “supersplicer” TEs display, on the contrary, higher sequence homology to the consensus U1 binding site (Supplementary Fig. 13). Thus, the alternative role of U1 in inhibiting PCPA seems to be conditioned by the quality of its pairing to the 5′-exon-intron junction, and this may have strong consequences on the biological cycle of the TEs concerned. On the one hand, perfect homology will lead to efficient splicing and shGAG accumulation/Gag production to the detriment of fl-EVD transcription and translation, resulting in dead-end non-autonomous elements epitomized by the ‘supersplicers’. A weaker interaction, on the other hand, will reduce splicing and concomitantly facilitate U1-mediated inhibition of PCPA, thereby allowing production of both Gag via sh-GAG, and POL via fl-EVD, granting retro-transposition and genomic proliferation as illustrated with EVD. Additionally, similarly to retroviruses, multiple elements within EVD could be responsible for splicing regulation37 and requires further investigation.

The above considerations on TE gene expression strategies are important, because our results strongly suggest that the efficiency of splicing and the quality of the intron play a crucial role in triggering silencing trough the recruitment of RDR6. Enhancing the splicing of EVD leads to increased levels of sh-GAG mRNA and Gag protein, yet the siRNA levels are reduced. Conversely, decreasing the frequency of splicing by lowering the quality of the intron lessens the sh-GAG levels but boosts siRNA production. Transposed in the context of the Arabidopsis Copia-like elements
evoked above, these results predict that dead-end ‘supersplicer’ TEs would be naturally immune to PTGS, while active, EVD-like TEs will trigger PTGS owing to the mandatory nature of their inefficient splicing. Ongoing siRNA deep-sequencing from met1 versus WT Arabidopsis should help assessing these predictions.

Recently, Madhani and coworkers proposed that the spliceosome acts as a transposon sensor by surveying the transcriptome for anomalous splicing events in the yeast C. neoformans. Their results suggest that introns derived from TE remnants, due to their unusual length and disposition to carry suboptimal splicing signals, tend to stall splicing in ways that signal the recruitment of an RdRp-containing complex named ‘SCANR’, eventually leading to siRNA production. Other observations also have raised the possibility that stalled spliceosomes are involved in TE-derived piRNA biogenesis. In Drosophila, piRNAs arise from exonic and intronic sequences of piRNA precursor loci, and mutations in the fly splicing and nuclear exporting factor UAP56 impair piRNA production. Recent work indicates that Rhino, a fly homolog of the yeast heterochromatin protein 1 (HP1), in cooperation with the cap-binding protein CUTTOFF (Cuff) and UAP56 licenses piRNA cluster transcripts for piRNA biogenesis by suppressing splicing in ways that can apparently distinguish piRNA precursor transcripts from endogenous mRNAs.

Although our work with EVD suggests that a similar mechanism might be operating in plants, there are notable differences. First, the lack of intron-derived siRNAs indicates that the spliced mRNA rather than splicing intermediates (as in flies and yeast) is the main target of RDR6 in plants, consistent with previous comparative genome-wide studies of 21-22-nt siRNAs accumulation in mutants displaying enhanced gene-derived siRNAs in Arabidopsis. Nonetheless, we cannot formally rule out that a low level of intron retention in sh-GAG constitutes the initial trigger for siRNA production, which would then be amplified from comparatively much more abundant spliced sh-GAG templates. Second, transient expression of the sh-GAG mRNA, produced completely independently of splicing (EVDΔi and sh-GAGΔi constructs), is itself sufficient to promote RDR6-dependent siRNA biogenesis. Preliminary transient expression analyses strongly suggest that sh-GAG 3'-UTR elements are key in enhancing siRNA levels. However, studies of stable transgenic Arabidopsis lines are needed to avoid the intrinsic and confounding propensity of the N. benthamiana transient expression assay to trigger silencing. While awaiting the
results of ongoing sh-GAG 3'-UTR dissection experiments and transgenesis, we temptatively propose the following model.

Similarly to yeast, the inefficiently spliced and long intron of EVD would cause enhanced retention of the spliceosome, an event sensed by the RQC machinery to “label” the processed (spliced) mRNA to become a substrate for RDR6. RDR6 recruitment or action would then be facilitated by features (at the sequence/structure levels) present in the sh-GAG 3'-UTR. An interaction between RQC and RDR6 has been well documented in Arabidopsis, as both compete for the same RNA substrates. For instance, RDR6 has high affinity for aberrant RNA lacking a 5'-Cap, 3'-polyA tail, or displaying NMD-activating features. RDR6 action is normally outcompeted by the RQC to prevent such endogenous aberrant mRNAs from becoming source of siRNAs, which could otherwise have strong detrimental effects. Nonetheless, when RQC becomes overwhelmed, or its components genetically suppressed, aberrant RNAs may become substrates of RDR6 action and subsequent siRNA biogenesis. It is tempting to speculate that as yet uncharacterized RQC component(s) or pathway, perhaps related to the ‘SCANR’ complex of C. neoformans, monitor the spliceosome to identify anomalous, suboptimal, or unconventional splicing events to instruct RDR6 to act on the resulting mature mRNAs.

Our genome-wide identification of many additional spliced Copia-like retrotransposons shows that the presence of an intron after GAG, the splice-induced frameshift and PCPA are widespread features that ensure an optimal Gag:Pol stoichiometry. Additional conserved features such as the obligatory weakness and abnormal TE intron length (about three times longer than introns present in Arabidopsis coding genes) makes scrutinizing splicing a potent strategy to discriminate TE transcripts from endogenous coding mRNAs, and to selectively promote PTGS on the former. Evidently, a way for TEs to escape such control mechanism would be to display efficient splicing. However, as illustrated by the identified “supersplicers” and discussed before, this would readily abolish the TE’s ability to transpose. Alternatively, splicing could be avoided altogether. This is exemplified in S. pombe by the LTR-retrotransposon Tf1, which displays a single ORF but achieves Gag molar excess through post-translational degradation of Pol products. While it remains unknown if such a strategy is employed by non-spliced TEs in Arabidopsis, EVD can indeed produce Gag without splicing through the
integration of rare, spliced full-length copies. These copies, although non-autonomous due to the lack of the Pr domain (removed by splicing), would be molecularly equivalent to EVD_{Δi}, and solely transcribed into the sh-GAG mRNA. They could, in principle, constitute autonomous reservoirs for Gag production possibly used in trans by other, full-length elements, which predicts that such loci might be under positive selection. The evolutionary young Gypsy-like element Ogre has been shown to also integrate spliced copies and many retroviruses integrate spliced RNA in infected cells.

Unique characteristics of splicing seem to be exploited in yeast and flies to direct the RNAi silencing machinery onto TE remnants or TE-derived precursor transcripts, respectively. Here, we demonstrate that inefficient splicing is actively used by plants to discriminate and contain the proliferation of de novo invading or epigenetically reactivated, intact TEs. This is achieved by converting specific TE-derived subgenomic transcripts into well-defined bulks of 21-22-nt siRNAs. By providing a biological rationale for TEs to harbor unusual introns and to produce such PTGS-trigger mRNAs, our study establish, for the first time, a link between the biology of active TEs and the ability of plants to recognize them as parasitic elements in a completely innate manner.
METHODS

Plant Material.

*Arabidopsis thaliana* ecotype *Columbia-0* (Col-0) was used in this study as WT, *met1*-3, *rdr6*-12 mutants, the *met1*-derived epiRIL#15 (epi15), 35S:EVD transgenic lines and the *Nicotiana benthamiana* RDR6i line have been previously described\(^\text{32,42,71-74}\)

Plasmid Construction and Transformation

*EVD* splicing mutants and sh-GAG constructs were generated using multiple site Gateway\textsuperscript{®} technology (Invitrogen), primer sequences can be found in Supplementary Table 2. Template for PCR amplification was 35S:EVD. PCR was performed using Phusion HF DNA Polymerase (Thermo Scientific) Mutations in *EVD* intron were introduced by PCR. Deletion of the intron was achieved by separately amplifying flanking sequences and fusing them by overlapping PCR. PCR products were recombined into pDONR221. LR recombination reactions were performed using the resulting plasmids and a 35S promoter in pDONR4-1r and the destination binary vector pB7m34GW (https://gateway.psb.ugent.be/\(^\text{75}\)) the resulting construct introduced into *Agrobacterium tumefaciens* strain GV3101. Transient expression assays were performed as previously described\(^\text{43}\).

RNA Blot Analysis.

Total RNA was extracted from *A. thaliana* and *N. benthamiana* with Trizol reagent (Invitrogen). Unless otherwise indicated, 10 *Arabidopsis* plants or two *N. benthamiana* leaves from two independently agroinfiltrated plants were bulked for analysis. For RNA gel blot analysis of high-molecular weight (HMW) RNA, 10-15 µg of total RNA were resolved on denaturing 1.2% agarose gels with 2.2 M formaldehyde, capillary transferred to Hybond\textsuperscript{TM}-NX membrane (GE Healthcare) and cross-linked by UV irradiation\(^\text{76}\). Ethidium bromide staining before transfer was used to confirm loading. For low-molecular weight (LMW) RNA blot analysis, RNA was either subjected to RNA fractionation step and 500 ng of LMW RNA or 5-10 µg total RNA was loaded. LMW RNA was isolated from total RNA by precipitating HMW RNA in 10% PEG 8000 ,500 mM NaCl for 30 min on ice and collecting HMW RNA by centrifugation (10 min, 10000g at 4°C). LMW RNA was precipitated from the supernatant following standard
nucleic acid precipitation procedures. RNA was resolved on a denaturing 17.5% polyacrylamide/urea TBE buffered gel, transferred to membrane by electroblotting and chemically cross-linked\textsuperscript{77}. Radiolabelled probes for detection of EVD and ACT2 mRNA and siRNAs were made by random priming reactions using the Prime-\textit{a}-gene kit (Promega) in the presence of $\alpha$-$^{32}$P-dCTP (Hartmann Analytic). The templates used were PCR products (Supplementary Table 2). DNA oligonucleotides complementary to miRNAs, tasiRNAs, or U6 snRNA were end labelled with $\gamma$-$^{32}$P-dATP using T4 PNK (Thermo Scientific).

**Absolute Quantification of EVD DNA and RNA by Real-Time qPCR.**

\textit{EVD} copy number quantification and expression analysis were conducted by Real Time (RT)-qPCR on a LightCycler$^{\textregistered}$ 480 II apparatus (Roche) using SYBR$^{\textregistered}$ Green qPCR kit (Eurogentech) and gene-specific primers (Supplementary Table 2). For \textit{EVD} copy number quantification, genomic DNA was extracted from \textit{A. thaliana} with DNeasy Plant Mini Kit (Qiagen). \textit{EVD} genomic copies were subjected to absolute quantification analysis by using the standard curve method and normalized with the single-locus gene \textit{ACTIN2} (\textit{ACT2}, \textit{AT3G18780}). Standard curves were generated using serial dilutions of known concentrations of plasmid containing full length \textit{EVD}_{wt}, \textit{EVD}_{\Delta i} or \textit{ACT2} cDNA. Error bars represent standard deviation from three technical replicates of two biological replicates consistent of bulks of 10 plants each.

For absolute quantification of \textit{EVD} RNA molecules, total RNA was extracted with RNaseasy Plant Mini Kit (Qiagen) followed by DNaseI (Thermo Scientific) treatment. Absolute quantification was performed following the copy (c)RNA standard curve method\textsuperscript{38}. In brief, endogenous \textit{EVD} containing both LTRs, its mutated \textit{EVD}_{\Delta i} version and \textit{ACT2} cDNA were cloned into the pGEM-T vector (Promega) in sense orientation to the T7 promoter. 2 \textmu g of plasmid were linearized using Ncol (Thermo Scientific) and 1 \textmu g used as input for \textit{in vitro} transcription of cRNA standards using T7 RNA polymerase (Thermo Scientific) following manufacturer's instructions. After DNaseI treatment and subsequent purification of cRNA (standard Phenol/Chlorophorm/Isoamil RNA isolation) the integrity and the size of \textit{in vitro} synthetized cRNAs was checked on a 1% agarose gel using as reference a high molecular range RNA ladder (Thermo Scientific). Accurate quantification of cRNA concentration was achieved by spectrophotometry measurements on 3 serial
dilutions (1:10) of cRNAs. RNA concentration was converted into number of RNA molecules/µl using the following formula:

\[ N \text{ (molecules/µl)} = \left( \frac{C}{K} \right) \times 182 \times 5 \times 10^{13} \]

Where:

\[ C = \text{concentration of cRNA (µg/µl)} \]
\[ K = \text{size of the RNA (bp)} \]

Standard curves were made by performing serial dilutions (1:10) containing from $10^{11}$ to $10^{4}$ cRNA molecules/µl in the presence of 1 µg/µl yeast tRNA (Sigma) and stored as $10^{11}$ to $10^{4}$ cRNA molecules/µg RNA stocks at -80°C. For cDNA synthesis same procedure was always applied to standard cRNAs and *Arabidopsis* RNA samples. 1 µg of cRNA standard curve stocks and total RNA were reversed transcribed (RT) into cDNA with RevertAid® First Strand cDNA Synthesis Kit (Thermo Scientific) using either random hexamers or gene specific primers (GSP) (Supplementary Table 2) following manufacturer’s instructions. GSP-RT was always conducted in the presence of ACT2 GSP to provide an endogenous normalizator. Each primer was added to a final concentration of 20 pmol. cRNA standards were subjected to three technical replicates to generate in run standard curves for each PCR probe. Following qPCR total EVD and ACT RNA molecules were quantified and expressed as cRNA equivalents. A correction factor obtained from normalizing ACT2 cRNA equivalent values relative to those of WT was applied to EVD values to normalize input RNA. For all error bars represent standard deviation from three technical replicates of two biological replicates consistent of bulks of 10 plants each.

**Purification of polyA+ RNA from total RNA**

Polyadenylated RNA was purified from total *Arabidopsis* RNA by affinity purification with dC_{10}T_{30} oligonucleotides covalently linked to polystyrene-latex particles using the Oligotex mRNA kit (Qiagen) following manufacturers mini-prep spin column protocol. In brief, total RNA was extracted as described above and 10 µg were used for polyA+ purification. The resulting elute (polyA+ RNA) from the resin was analysed by HMW Northern together with 10 µg of input total RNA.
Mapping of EVD 3’ and 5’ mRNA ends.

The mapping of transcription start and ends was performed by 5’ RNA ligation-mediated (RLM) amplification of cDNA ends (RACE; 5’RLM-RACE) and 3’RACE using the FirstChoice RLM-RACE kit (Ambion, AM1700) following manufacturers instructions and primers depicted in the text and Supplementary Table 2. Following RACE-PCR, total PCR products were purified (GeneJet PCR purification kit, Thermo Scientific) and cloned into pJET1.2 (Thermo Scientific). 50 clones were sequenced.

Protein Blot Analysis.

Total proteins were extracted from 0.1-0.3 g of A. thaliana and N. benthamiana tissues ground in liquid nitrogen and homogenized with 0.6 ml of extraction buffer (0.7 M sucrose, 0.5 M Tris-HCl pH 8, 5mM EDTA pH 8, 0.1 M NaCl, 2% [v/v] 2-mercaptoethanol) containing 2 µM proteasome inhibitor MG-132 (Calbiochem®, Millipore) and Protease Inhibitor Cocktail (cOmplete®, Roche) following manufacturer’s instructions and mixed for 5 min. An equal volume of water-saturated Tris-buffered phenol was added and mixed for 5 min. Phases were separated by 10 min. centrifugation (12000g at 4°C). Proteins were precipitated from the phenol phase by addition of 5 vol. of 0.1 M ammonium acetate in methanol and incubated at -20°C overnight. After 15 min. centrifugation (5000g at 4°C), precipitate was washed with ammonium acetate in methanol and resuspended (3% [v/v] SDS, 62.3 mM Tris-HCl pH 8, 10% [v/v] glycerol). 100-300 ng of total protein were resolved on SDS-PAGE, transferred by electroblotting onto Inmobilon®-P PVDF membrane (Millipore) and incubated with antibodies in PBS with 0.1% Tween-20 and 5% non-fat dried milk following standard western blot procedures76. After secondary antibody incubation with HRP-conjugated goat anti-rabbit (Sigma), detection was performed by ECL Western Blotting Detection Kit (GE Healthcare). Equal loading was verified by coomassie staining of the membrane after western blotting. Affinity-purified Gag antibody was used at 1/2000 dilution.

Antibodies.

Peptide antibodies for EVD Gag were raised in rabbits against QETHEESEQAGSSKG Gag peptide and affinity purified from antisera following Eurogentec (Eurogentec SA; http://www.eurogentec.com) standard program. The efficiency of purified antibodies was validated by comparing protein detection by
western blot between WT plants, epi15 F11 (expressing endogenous EVD) and 35S:EVD transgenic plants in the WT background (Supplementary Fig. 3).

**Deep sequencing and bioinformatic applications.**

For RNA-seq, total RNA (10 µg) was extracted from inflorescences from two pools of 10 Arabidopsis plants using the RNeasy plant extraction kit (Qiagen) and subsequent DNaseI (Thermo Scientific) treatment. Library preparation, ribosomal depletion, sequencing of stranded, pair-end 2x 100bp reads on Illumina HiSeq 2000 sequencer and trimming of the data was accomplished by the Functional Genomic Centre Zurich (FGCZ). Reads were mapped with *ab initio* splice site discovery using TopHat (V.2.0.4)\(^78\). The Integrative Genomics Viewer (IGV)\(^79\) was used for visualization of mapped RNA-seq reads on *Arabidopsis* genome. Read counting was performed with RSEM\(^80\) and differential analysis using DESeq\(^81\). Retrotransposon domains were predicted with LTRdigest\(^51\). In-house Python and R-cran scripts allowed for selection of junction reads mapping to *Copia*-like elements and classification of splice and unspliced elements. Visualization of the consensus splice behaviour, scatterplot and boxplots are based on the R package ggplot2\(^82\). Sequence reads were aligned against the Arabidopsis genome (TAIR10) using MUMmer v3.0 software\(^83\). Only small RNA sequences with perfect matches over their entire length (15-35 nt) were further analyzed. Protein and DNA sequence alignments were performed with the Jalview software\(^84\) by clustalW Multiple Sequence Alignment. Phylogenetic tree of *Copia*-like elements active in *met1*, based on TE full sequence were build in CLC Genomics Workbench (http://www.clcbio.com/products/clc-genomics-workbench/) using the Neighbour-Joining method.

**Accession codes.**

WT (Col-0), *ddm1*-2, *ddm1*-2 *dcl*-1-11 and *ddm1*-2 *rdr6*-15 data sets\(^27\) and *met1*-derived epi15 F11 siRNA library\(^32\) were obtained from NCBI’s Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/), accession number GSM1278497, GSM1278498, GSM1278500, GSM1278501 and GSM1062221 respectively. RNA-seq raw and processed data from WT, *met1*-3, epi15 F8 and F11 will be deposited upon submission of the final version of the manuscript.
AKNOWLEDGEMENTS
We thank the members of the Voinnet team for fruitful discussions, in special Gregory Schott for technical assistance characterizing the EVD Gag antisera and purified antibody.

AUTHOR CONTRIBUTIONS
O.V. and A.M.O conceived and designed the experiments. A.M.O and S.O performed the experiments. O.S. and A.S performed computer and statistical analysis. O.V., A.M.O and S.O analyzed the data. A.M.O and O.V. wrote the manuscript.
REFERENCES

18. Klattenhoff, C. & Theurkauf, W. Biogenesis and germline functions of piRNAs. *Development*
37. Leblanc, J., Weil, J. & Beemon, K. Posttranscriptional regulation of retroviral gene expression: primary RNA transcripts play three roles as pre-mRNA, mRNA, and genomic RNA. *Wiley


70. Houzet, L., Morichaud, Z. & Mougel, M. Fully-spliced HIV-1 RNAs are reverse transcribed with similar efficiencies as the genomic RNA in virions and cells, but more efficiently in AZT-treated cells. *Retrovirology* 4, 30 (2007).
SUPPLEMENTARY INFORMATION

Supplementary Figure 1: Predicted miRNAs targeting EVD.
Supplementary Figure 2: EVD siRNA are RDR6-dependent but DCL1 independent.
Supplementary Figure 3: Specificity of EVD Gag polyclonal antibody.
Supplementary Figure 4: Enhanced short GAG transcript accumulation in rdr6.
Supplementary Figure 5: Splicing of endogenous EVD.
Supplementary Figure 6: Increased relative amounts of spliced sh-GAG in rdr6.
Supplementary Figure 7 and Supplementary Text: Spliced full-length EVD RNA is competent for genomic integration.
Supplementary Figure 8: Mapping of siRNA reads to EVD splice junction.
Supplementary Figure 9: Alignment of DNA and protein sequences from EVD splicing mutants.
Supplementary Figure 10: Expression levels of spliced and unspliced Copia-like elements in met1.
Supplementary Figure 11: Phylogeny of active Arabidopsis Copia-like elements in met1.
Supplementary Figure 12: Copia-like “supersplicer” TEs are mainly transcribed into sh-GAG RNA.
Supplementary Figure 13: Conservation of U1 binding site and splicing behaviour.
Supplementary Table 1: List of active Arabidopsis Copia-like elements in met1 and their splicing behaviour.
Supplementary Table 2: List of primers.
Supplementary Figure 1. Predicted miRNAs targeting EVD. Predicted miRNA targeting EVD positioned along the distribution of EVD 21-22-24-nt deep-sequencing siRNA reads from epi15 F11 on EVD. y axis: total read count, (top: sense, bottom: antisense); x axis: coordinates at EVD coding sequence, position 1 corresponds to the first ATG (Met) codon of EVD. Red arrow marks the position of the potential target sites.
Supplementary Figure 2. EVD siRNA are RDR6-dependent but DCL1 independent. (a-d) Distribution of EVD deep-sequencing siRNA reads from WT (a), ddm1 (b), and the double mutants ddm1 dcl1 (c) and ddm1 rdr6 (d) on EVD. Blue: 21-nt; Green: 22-nt; Red: 24-nt y axis: total read count, (top: sense, bottom: antisense); x axis: coordinates at EVD coding sequence, position 1 corresponds to the first ATG (Met) codon of EVD.
Supplementary Figure 3. Specificity of EVD Gag polyclonal antibody. Western analysis of EVD Gag polyclonal antibody specificity tested on 50 µg of total protein extracted from WT Col-0, epi15 F11 and 35S:EVD in WT Arabidopsis plants.
Supplementary Figure 4. Enhanced short GAG transcript accumulation in *rdr6*. Relative (%) amounts of EVD RNA molecules from 35S:EVD in *rdr6* to 35S:EVD in WT using GAG, IN and RT probes. cDNA was primed with EVD GSP (a) or with random hexamers (b).
Supplementary Figure 5. Splicing of endogenous EVD. Screenshot of Integrated Genome Browser (IGV) displaying mapped RNA-seq reads onto EVD from met1, epi15 F8 and F11. Number of junction reads mapped and their spanning location is indicated. Dotted line indicates reads scale annotated in brackets (left). x axis: Chromosome 5 coordinates (TAIR 10). EVD domains are represented below.
Supplementary Figure 6. Increased relative amounts of spliced sh-GAG in *rdr6*. Relative levels (%) of spliced (splGAG) and unspliced (u-splGAG) EVD molecules relative to total EVD (see Fig. 3f,g) in epi15 F11 and 35S:EVD in WT and *rdr6* backgrounds. cDNA was primed with EVD GSP (a) or with random hexamers (b).
Supplementary Figure 7. Spliced full-length EVD RNA is competent for genomic integration. (a) Schematic representation of EVD and position of primer used for fl-EVD 3'RACE (red arrow). (b-c) 5'RLM-RACE-PCR of spliced (b) and unspliced (c) EVD transcripts using primers that specifically anneal to splice or unspliced mRNA (depicted in Fig. 3e) in WT, epi15 F11 and 35S:EVD in WT or rdr6 backgrounds. The same reaction was carried in the absence of cDNA (±cDNA) as control for reagent contamination. (d) 3'RACE PCR of fl-EVD transcripts amplified from position shown in (a). (e) Mapping of spliced (dashed line) and unspliced 5'ends (red) and fl-EVD 3'ends (blue) from epi15 F11 sequenced 5'RLM-RACE (b,c) and 3'RACE PCR products to EVD sequence. Positions in nucleotides (nt) are given relative to the first ATG (Met) codon of EVD. The number of clones
Endogenous EVD produces spliced and unspliced transcripts from different TSS

Transposons, like retroviruses must regulate the fate of their transcripts to be used as mRNA for protein production and/or as a genomic RNA (gRNA) to be reversed transcribed into new DNA copies\textsuperscript{37,85}. Alternative use of transcription start (TSS) and termination (TTS) sites is one mechanism used to generate different pools of transcripts containing or not LTR sequences required for reverse transcription of the gRNA\textsuperscript{48,85,86}. The similar relative abundance of the different EVD transcripts between endogenous EVD in epi15 F11 and 35S:EVD in WT (Fig. 1l and Supplementary Fig. 4) readily indicated that EVD splicing behaviour is independent of the promoter driving its expression. Nonetheless, we sought to investigate if spliced and unspliced transcripts shared the same 5’ and 3’ ends. 5’RLM-RACE was performed using two different sets of primers to alternatively amplify 5’ends from spliced or unspliced RNA (Fig. 3e). On the other hand, 3’ends from fl-EVD transcripts were amplified by 3’RACE from the end of the RNase domain (Supplementary Fig. 7a. Total RACE-PCR products (Supplementary Fig. 7b-d) were cloned and ~50 clones sequenced. 5’RACE-PCR on spliced EVD RNA yielded a single band in all EVD expressing samples that mapped to the described 35S promoter TSS\textsuperscript{87} (data not shown) or to the LTR of EVD, displaying a dominant start position 259-nt upstream of the ATG (Supplementary Fig. 7e). 5’ ends from unspliced RNA in 35S:EVD transgenic plants shared the same 35S TSS than spliced mRNA (Supplementary Fig. 7b,c and data not shown). However, epi15 F11 yielded two prominent bands (Supplementary Fig. 7b). The biggest amplicon (11/50 clones) mapped to the same TSS as spliced EVD,

sequenced mapping to each position is indicated in brackets. PBS: Reverse transcription priming binding site for tRNA (Met) U3: LTR promoter region, only transcribed from 3’LTR; R: mRNA repeated region, transcribed from 5’ and 3’ LTR; U5: LTR 3’region, only transcribed from 5’LTR. (f) Absolute quantification of total, spliced (spl) and unspliced (u-spl) EVD mRNA molecules containing 3’R sequences (LTR+) by qRT-PCR in total RNA from WT, met1 and epi15 generations (F8, F11, F14) reverse transcribed using a EVD GSP located in the LTR R region (e). qPCR was carried using primers located at the GAG region depicted in Fig. 3e. (g) Relative levels (%) of spliced fl-EVD transcripts containing 3’R sequences in met1 and epi15 F8-F11. (h,i) PCR analysis of uspliced and spliced EVD copies in gDNA using primers depicted in Fig. 3e. (h) gDNA from WT, epi15 F11 and 35S:EVD in WT or rdr6 backgrounds (i) gDNA from WT, met1 and epi15 generations (F8, F11, F14). cDNA of 35S:EVD in WT is used as for amplification specificity and size control. The same PCR reactions were carried in the absence of DNA (-gDNA) as control for reagent contamination. (j) Absolute quantification of EVD total, unspliced (u-spl) and spliced (spl) copies present in the genome of WT, met1 and epi15 generations (F8, F11, F14).
located in the 5'LTR. The rest, mapped to EVD 5'UTR ~80-nt upstream of the ATG (Supplementary Fig. 7e). These results indicate that all of spliced EVD mRNA contains 5'LTR sequences known to be required for the reverse transcription, specially the tRNA^{MET} binding site required to prime reverse transcription\textsuperscript{85} (primer binding site, PBS; Supplementary Fig. 7e). However, spliced RNA, mainly corresponding to the sh-GAG transcript, is terminated at the IN domain, and therefore will fail to be reverse transcribed. On the other hand, only a fraction of unspliced RNA (11/50), corresponding to the fl-EVD transcript necessarily used as gRNA, does contain 5'LTR sequences. Hence, the dual role of fl-EVD transcript as gRNA and Pol mRNA can potentially be abstracted by their 5’ ends, given that transcripts lacking 5'LTR and PBS sequences will not be competent for gRNA. It is also noticeable that sequences required for gRNA packaging (Psi) into Gag containing virus-like particles (VLPs) in retroviruses are generally located within the LTR or 5'UTR\textsuperscript{88}, meaning that sh-GAG transcripts could compete the gRNA for Gag binding. However, EVD Psi sequence remains unknown, but the fact that 35S:EVD mRNA can be co-immunopurified with Gag (data not shown), and mutations that disrupt Gag RNA binding ability lead to enhance sensitivity of the construct to silencing\textsuperscript{32}, strongly suggest that Psi, or other packaging signals, are present in the coding sequence of EVD. Further experiments using purified recombinant Gag protein\textsuperscript{89,90} are being carried to answer such questions.

**Full-length spliced EVD RNA can be integrated in the genome**

In order to be competent for reverse transcription into cDNA, the gRNA must contain in both 5’ and 3’ ends an mRNA repeated region (R), present in the LTRs between the 5’TSS and the 3’TTS\textsuperscript{85}. The above results prone us to investigate in detail the 3’ends of fl-EVD transcripts to find such R element in EVD. 3’RACE-PCR yielded two main amplicons (Supplementary Fig. 7d). In contrast to 5’RACE-PCR products the 3’ends were slightly less defined, 6/49 clones mapped before or shortly after the beginning of the 3’LTR, the rest showed that most of EVD 3’ends locate towards the end of the LTR (Supplementary Fig. 7e). Hence most of fl-EVD transcripts contain 3’LTR sequences. Combining 5’ and 3’ RACE results allowed us to map a 31-nt long R domain in EVD by looking for LTR sequences present at both ends (Supplementary Fig. 7e).

The results shown in Fig. 3f and Supplementary Fig. 6a revealed that a minor
fraction (<3%) of fl-EVD is spliced. Having shown that all of spliced RNA contains 5’LTR R sequences and most of full-length transcripts do similarly at their 3’end, we set to investigate if spliced fl-EVD was qualified for reverse transcription. We first validated that spliced fl-EVD RNA indeed contained 3’R sequence (LTR+) by performing quantitative RT-PCR analysis using a primer matching the R domain to prime reverse transcription of endogenous EVD from met1 and epi15 plants from different generations (F8, F11, F14). Levels of unspliced (LTR+ u-spl) and spliced (LTR+ spl) full-length transcript were quantified and compared to those of total fl-EVD as described before using primers shown in Fig. 3e. Independently of the level of endogenous EVD expression, most of the fl-EVD containing the 3’ R region appeared to be unspliced, while only a minute fraction (~1%) was spliced (Supplementary Fig. 7f,g). However, these low levels were sufficient to promote integration of spliced EVD copies in epi15 F11 plants as we were able to amplify spliced EVD from epi15 genomic DNA but not from 35S:EVD transgenic plants, lacking LTR sequences (Supplementary Fig. 7h). This integration of spliced EVD seems to be of stochastic nature given that, despite spliced fl-EVD containing the R domain is produced in all of epi15 generations carrying active copies, from met1 to epi15 F11, integration of the spliced ones could only be detected from F11 to onward generations (Supplementary Fig. 7i), even in the F14 plants where EVD is not longer active (32, Supplementary Fig. 7f). Additionally, quantification of genomic EVD copy number illustrated that a single integration event of spliced EVD was present in epi15 F11 and carried over inbreeding to the F14 generation (Supplementary Fig. 7j). Similarly, the gypsy-like Ogre element, recently active in the S. latifolia genome, displays around 3% of spliced genomic copies66. The integrated copies of such low abundant full-length spliced RNA species will not be autonomous anymore since in both cases, EVD and Ogre, the Pr domain is absent in the spliced transcript. However a potential role for such non-autonomous copies will be discussed later in the main text.
Supplementary Figure 8. Mapping of siRNA reads to EVD splice junction. Alignment of small RNA-seq reads (arrows) from epi15 F11 mapping to EVD exon junction after splicing (highlighted in red in Fig. 4g). Abundance of reads is indicated. Top: sense; bottom: antisense.
Supplementary Figure 9. Alignment of DNA and protein sequences from *EVD* splicing mutants. (a) DNA sequence alignment of WT and mutant *EVD* introns. 5' and 3' splicing sites (ss) is underlined. Box: U1 snRNA binding site. (b) Alignment of conceptual translation of DNA sequences shown in (a). (c) 5' splice sites DNA sequence alignment of *EVD*<sub>Δi</sub> with WT and mutant *EVD* clones. (d) RT-PCR analysis using cDNA primed with RH of *EVD* splicing mutants in (c) transiently expressed by agroinfiltation in WT *N. benthamiana*. Empty triangle: unspliced; black triangle: spliced. Intron sequences are highlighted in blue, deleted sequence in *EVD*<sub>Δi</sub> is highlighted in red. Mutated nucleotides or aminoacids altered by such mutations are shown in red.
Supplementary Figure 10. Expression levels of spliced and unspliced Copia-like elements in met1. Boxplot representation of the number of met1 RNA-seq reads for unspliced (n = 26) and spliced (n = 36) Copia-like elements normalized to TE length (RPK). p-value based on Wilcoxon rank-sum test with continuity correction. Box plot shows: median (bold line), second to third quartile (box) and upper and lower expression ranges (whiskers).
Supplementary Figure 11. Phylogeny of active Arabidopsis Copia-like elements in met1. Phylogenetic tree of Copia-like elements identified as transcriptionally active by differential analysis of met1 to WT RNA-seq data based on the RT domain using the Neighbor-Joining method. In blue: splicing elements; in red: non-splicing elements. EVD is boxed.
Supplementary Figure 12. *Copia*-like “super-splicer” TEs are mainly transcribed into sh-GAG RNA. Screenshot of Integrated Genome Browser (IGV) displaying mapped RNA-seq reads from *met1* onto “super-splicer” TEs along with the Pfam (http://pfam.xfam.org) predicted GAG, IN and RT core domains and stop codons (red squares). Number of junction reads mapped and their spanning location is indicated. Dotted line indicates reads scale annotated in brackets (left). X axis: corresponding chromosomal coordinates (TAIR 10).
Supplementary Figure 13. Correlation between U1 pairing with its binding site and splicing behaviour. Screenshot of Integrated Genome Browser (IGV) displaying mapped RNA-seq reads from met1 onto TEs displaying different splicing behaviour and sequence alignments between consensus U1 snRNA binding site and those of each TE. x axis: corresponding chromosomal coordinates (TAIR 10).
Supplementary Table 1. List of active *Arabidopsis* Copia-like elements in met1 and their splicing behaviour. Active Copia-like elements are identified by their *Arabidopsis* TE identifier, AGI number and the family they belong to. Classification of spliced and unspliced as described in the text. Splicing-induced frameshift was assessed by inspecting if splicing donor (SD) and acceptor (SA) were in the same coding frame. When not (leading to frameshift) appearance of stop codons was manually validated by in silico translation of spliced sequences.

<table>
<thead>
<tr>
<th>TE identifier</th>
<th>AGI number</th>
<th>Family</th>
<th>Splicing</th>
<th>DS/AS in frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT1TE30195</td>
<td>AT1G26990</td>
<td>ATCOPIA36</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT1TE35480</td>
<td>AT1G30340</td>
<td>ATCOPIA73</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT1TE41580</td>
<td>AT1G34967</td>
<td>ATCOPIA83</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT1TE43975</td>
<td>AT1G60400</td>
<td>ATCOPIA43</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT1TE58075</td>
<td>AT1G47650</td>
<td>ATCOPIA38A</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT1TE59745</td>
<td>AT1G48680</td>
<td>ATCOPIA49</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT1TE62440</td>
<td>AT1G50810</td>
<td>ATCOPIA36</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT1TE70490</td>
<td>AT1G57640</td>
<td>ATCOPIA22</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT1TE71950</td>
<td>AT1G58889</td>
<td>ATRE1</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT1TE72060</td>
<td>AT1G59265</td>
<td>ATRE1</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT1TE72970</td>
<td>AT1G60200</td>
<td>ATCOPIA5</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT1TE86360</td>
<td>AT1G70010</td>
<td>ATCOPIA11</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT2TE06390</td>
<td>AT2G04135</td>
<td>ATCOPIA28</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT2TE08785</td>
<td>AT2G05240</td>
<td>ATCOPIA66</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT2TE12115</td>
<td>AT2G06830</td>
<td>ATCOPIA67</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT2TE12470</td>
<td>AT2G06950</td>
<td>ATCOPIA68</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT2TE12510</td>
<td>AT2G06950</td>
<td>ATCOPIA8A</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT2TE13385</td>
<td>AT2G07420</td>
<td>ATCOPIA20</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT2TE17870</td>
<td>AT2G10760</td>
<td>ATCOPIA32B</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT2TE26230</td>
<td>AT2G14930</td>
<td>ATCOPIA9</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT2TE26230</td>
<td>AT2G14930</td>
<td>ATCOPIA9</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT2TE27820</td>
<td>AT2G15650</td>
<td>ATCOPIA36</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT2TE28160</td>
<td>AT2G15870</td>
<td>ATCOPIA38</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT2TE28826</td>
<td>AT2G15920</td>
<td>ATCOPIA38</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT2TE29325</td>
<td>AT2G16000</td>
<td>ATCOPIA38B</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT2TE29450</td>
<td>AT2G16670</td>
<td>ATCOPIA70</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT2TE30650</td>
<td>AT2G20460</td>
<td>ATCOPIA72</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT2TE43735</td>
<td>AT3G29510</td>
<td>ATCOPIA27</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT3TE61000</td>
<td>AT3G42883</td>
<td>META1</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT3TE62940</td>
<td>AT3G43640</td>
<td>ATCOPIA43</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT3TE63765</td>
<td>AT3G43955</td>
<td>ATCOPIA62</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT3TE68900</td>
<td>AT3G57775</td>
<td>ATCOPIA81</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT4TE09800</td>
<td>AT4G04280</td>
<td>ATCOPIA59</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT4TE19470</td>
<td>AT4G07810</td>
<td>ATCOPIA53</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT4TE25590</td>
<td>AT4G09955</td>
<td>ATCOPIA49</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT4TE28870</td>
<td>AT4G10990</td>
<td>ATCOPIA17</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT4TE28860</td>
<td>AT4G16870</td>
<td>ATCOPIA4</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT4TE28860</td>
<td>AT4G16870</td>
<td>ATCOPIA4</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT4TE44080</td>
<td>AT4G17450</td>
<td>ATCOPIA1</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT4TE49990</td>
<td>AT4G20130</td>
<td>META1</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT5TE00230</td>
<td>AT5G01185</td>
<td>ATCOPIA27</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT5TE00870</td>
<td>AT5G07685</td>
<td>ATCOPIA18</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT5TE20395</td>
<td>AT5G17125</td>
<td>ATCOPIA83</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT5TE23385</td>
<td>AT5G18197</td>
<td>ATCOPIA89</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT5TE23385</td>
<td>AT5G19165</td>
<td>ATCOPIA90</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT5TE32355</td>
<td>AT5G25615</td>
<td>ATCOPIA27</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT5TE43215</td>
<td>AT5G32484</td>
<td>ATCOPIA27</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT5TE47100</td>
<td>AT5G4965</td>
<td>ATCOPIA15</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT5TE48985</td>
<td>AT5G55555</td>
<td>ATCOPIA24</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT5TE50260</td>
<td>AT5G59935</td>
<td>ATCOPIA18A</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT5TE50380</td>
<td>AT5G59665</td>
<td>ATCOPIA81</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT5TE54905</td>
<td>AT5G38035</td>
<td>ATCOPIA87</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT5TE56585</td>
<td>AT5G39095</td>
<td>ATCOPIA25</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT5TE56645</td>
<td>AT5G39155</td>
<td>ATCOPIA83</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT5TE63610</td>
<td>AT5G43800</td>
<td>END0VR1</td>
<td>spliced</td>
<td>no</td>
</tr>
<tr>
<td>AT5TE65370</td>
<td>AT5G49295</td>
<td>ATCOPIA21</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT5TE65640</td>
<td>AT5G45116</td>
<td>ATCOPIA23</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT5TE79225</td>
<td>AT5G54203</td>
<td>ATCOPIA22</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT5TE81385</td>
<td>AT5G55875</td>
<td>ATCOPIA13</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>AT5TE83215</td>
<td>AT5G57126</td>
<td>ATCOPIA49</td>
<td>unsPLICED</td>
<td>-</td>
</tr>
<tr>
<td>EXPERIMENT</td>
<td>PRIMER</td>
<td>SEQUENCE (5'→3')</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------</td>
<td>--------</td>
<td>------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACT2 cDNA cloning</td>
<td>cACT_F</td>
<td>GAAAGGCCGCAAAATATAAGGTTG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cACT_R</td>
<td>AGCDAAAAATGAAGACTATATAAACCC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cDNA GSP primers</td>
<td>ACT2_GSP</td>
<td>AGCDAAAAATGAAGACTATATAAACCC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EVD-R_GSP</td>
<td>CTTATCTGTGATGATGAACCTCTCCAT</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LTR_R_GSP</td>
<td>AGACAACAAAGAAAGATATGGGTTG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RT and qPCR probes</td>
<td>qGAG-F</td>
<td>TTTAGCCGCGCGTGGTTGGAAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>qGAG-R</td>
<td>AACCTTGGGTCGAACGGTCCT</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>qIN-F</td>
<td>CCGGAGAAGAAAGAGAGCAAGC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>qIN-R</td>
<td>AATGTCGCGTTCCTGGTTTG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>qRT-F</td>
<td>ACATATGCTGCGACTGACATG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>qRT-R</td>
<td>AGGTGCCTGTGAGAACAAATTG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon1 Probe</td>
<td>EVD_Ex-F</td>
<td>CCGCTAAACAAAGAGAATGCTTG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EVD_Ex-R</td>
<td>CCGAATCTTGGGTAGCTGAAAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intron Probe</td>
<td>EVD_In-F</td>
<td>GATGCATCTTTCATCAACAAAGAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EVD_In-R</td>
<td>CGAAATAACATACACATTTAGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon2 Probe</td>
<td>EVD_Ex2-F</td>
<td>TATTAGACCGGGAAGAATTTTGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EVD_Ex2-R</td>
<td>CACCCGAAATGACTGATAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3' RACE</td>
<td>3GA-G</td>
<td>ACATTGGGTCTGCGAGGAAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3GA-rn</td>
<td>ACCGGCCCTTATTGCTGGAAAA</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3RNout</td>
<td>GAAATGCTACACAGGACAGAATC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5' RLM-RACE</td>
<td>5 sgl GAG-R (out)</td>
<td>TTTCTTCCCCGCTCAATATCCGGAT</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5 ugl GAG-R (out)</td>
<td>TTTCTTCCCCGCTCAATATCCGGAT</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5GA-rn</td>
<td>AATCTCGCGTCAAGAGCTGTC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cloning EVD1</td>
<td>npv_in-F</td>
<td>CTTCTTGGGTCTGCGAGGAAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>npv_in-R</td>
<td>GACTTGGAAGACGACTCATCAAGCTTCTGCAATCTCCATGAAAATTTTGGGTTGGAAGAATGTTATG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cloning EVD2</td>
<td>EVD_ES1_F</td>
<td>TATTAGATCAAGAAGATAGATAGATAGACATACTGATAGATAGAGACCCGCAATCTAACAATACTGATGAGTTCTGATAGAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EVD_ES1_R</td>
<td>TGAAAATCCTTTGAAATCTTTGTTAATCTTCTCAATCATCGTACTGCTTCAATTCCTGAAATGAGCTATTTCCAAATTTCCAAATTTCC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EVD_ES2_F</td>
<td>AGAGACAAATAATAGCTACAAATTAAATGCTCTCTGCTGATGAGTTCGTTCTTCCCTC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EVD_ES2_R</td>
<td>AGAGACAAATAATAGCTACAAATTAAATGCTCTCTGCTGATGAGTTCGTTCTTCCCTC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EVD_ES1_F</td>
<td>AGAGACAAATAATAGCTACAAATTAAATGCTCTCTGCTGATGAGTTCGTTCTTCCCTC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EVD_ES1_R</td>
<td>AGAGACAAATAATAGCTACAAATTAAATGCTCTCTGCTGATGAGTTCGTTCTTCCCTC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EVD_ES2_F</td>
<td>AGAGACAAATAATAGCTACAAATTAAATGCTCTCTGCTGATGAGTTCGTTCTTCCCTC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EVD_ES2_R</td>
<td>AGAGACAAATAATAGCTACAAATTAAATGCTCTCTGCTGATGAGTTCGTTCTTCCCTC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EVD_ES1_F</td>
<td>AGAGACAAATAATAGCTACAAATTAAATGCTCTCTGCTGATGAGTTCGTTCTTCCCTC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EVD_ES1_R</td>
<td>AGAGACAAATAATAGCTACAAATTAAATGCTCTCTGCTGATGAGTTCGTTCTTCCCTC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cloning EVD3</td>
<td>GAGSpDonor_F</td>
<td>GAGATGCATTGGGTTGAAAGAATGATAGATAGATAGATAGAGACCCGCAATCTAACAATACTGATGAGTTCTGATAGAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GAGSpDonor_R</td>
<td>CGAAATAACATACACATTTAGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GAGSpAcceptor_R</td>
<td>CTTCTTGGGTCTGCGAGGAAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GAGSpAcceptor_F</td>
<td>CTTCTTGGGTCTGCGAGGAAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cloning EVD4</td>
<td>EVD_RS1_F</td>
<td>GATGTGTGGGGTTGAAAGAATGATAGATAGATAGATAGAGACCCGCAATCTAACAATACTGATGAGTTCTGATAGAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EVD_RS1_R</td>
<td>CGAAATAACATACACATTTAGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EVD_RS2_F</td>
<td>GATGTGTGGGGTTGAAAGAATGATAGATAGATAGATAGAGACCCGCAATCTAACAATACTGATGAGTTCTGATAGAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EVD_RS2_R</td>
<td>CGAAATAACATACACATTTAGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EVD_RS3_F</td>
<td>GATGTGTGGGGTTGAAAGAATGATAGATAGATAGATAGAGACCCGCAATCTAACAATACTGATGAGTTCTGATAGAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EVD_RS3_R</td>
<td>CGAAATAACATACACATTTAGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EVD_RS1_F</td>
<td>GATGTGTGGGGTTGAAAGAATGATAGATAGATAGATAGAGACCCGCAATCTAACAATACTGATGAGTTCTGATAGAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EVD_RS1_R</td>
<td>CGAAATAACATACACATTTAGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cloning GAG+/−UTR</td>
<td>UTR+_Av3b</td>
<td>GGGGACCACATCTGACACAAGAAAGATGCTGAGGTTTTCTAATTTGAAAGAATACGATGTTG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>uspUTR_Av3b</td>
<td>GGGGACCACATCTGACACAAGAAAGATGCTGAGGTTTTCTAATTTGAAAGAATACGATGTTG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>UTR+_Av1b</td>
<td>GGGGACCACATCTGACACAAGAAAGATGCTGAGGTTTTCTAATTTGAAAGAATACGATGTTG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>UTR_EcoR</td>
<td>GGGGACCACATCTGACACAAGAAAGATGCTGAGGTTTTCTAATTTGAAAGAATACGATGTTG</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Supplementary Table 2. List of primers
Chapter 3: Discussion
In order to increase our understanding of *de novo* transposon (TE) silencing in the plant model organism *Arabidopsis thaliana*, we have focused on the young, low copy LTR-retrotransposon *Évadé* (*EVD*). Detailed genetic and molecular analysis of the different phases of *EVD*’s biology across multiple generations—from its epigenetic reactivation, genome colonization to its eventual TGS—have unraveled novel features of TE biology on the one hand, of plant RNA silencing mechanisms on the other, as well as many unsuspected interconnections linking the two processes. Our findings with *EVD* can not only be readily extended to other TEs, but also have broader implications in terms of gene regulation and evolution at both the genomic and epigenomic levels.

### 3.1. DISCUSSION AND PERSPECTIVES

*Multiple layers of silencing restrain TE reactivation*

Our results, and those of others, have shown that *EVD* is readily reactivated when DNA methylation is impaired by interfering with maintenance pathways in *met1* or *ddm1* mutants. However, this reactivation does not apply to all members of the *COPIA93* family. *Attrapé* (*ATR*), for instance, despite displaying >99% sequence homology to *EVD* and having apparently retained a potential for autonomy (no apparent nucleotide substitutions between *EVD* and *ATR* leads to aminoacid changes in core domains) does not mobilize, nor does it become transcriptionally active, in *met1, ddm1* mutants or their derived epigenetic recombinant inbreed lines (epiRILs). However, both *ATR* and *EVD* loose DNA methylation to a similar extent (Chapter 2.1), probably reflecting that additional layers of transcriptional gene silencing (TGS) are selectively imposed upon *ATR*. Indeed, we observed that in WT plants, *ATR* displays twice the amount of CHG and CHH methylation compared to *EVD*. Furthermore, while *EVD* is located in an euchromatic region, *ATR* is inserted in a pericentromeric, heterochromatic region rich in H3K9me2, suggesting that additional histone-, rather than DNA-level, modifications might restrain *ATR* transcriptional reactivation in the absence of DNA methylation. Consistent with this idea, introducing the *kyp* mutation (H3K9 methyltransferase) into the *met1* background leads to a massive mobilization and increased copy number of *COPIA93* elements, although it remains to be determined if this increase is contributed solely by *EVD* or, additionally, by *ATR* or other related *COPIA93* elements. In fact, different
TGS mutants (e.g. met1, ddm1, pol IV, pol V, cmt3, drm2, kyp) impact overlapping and non-overlapping subsets of TEs, further substantiating the view that multiple layers of epigenetic regulation are differentially imposed upon TEs\textsuperscript{3-10}. Epistatic and positive feed-back loops among TGS components facilitate reinforcement mechanisms linking DNA methylation to histone modifications pathways, such as the ones driven by KYP and CMT3 on one hand, and SHH1 and Pol IV\textsuperscript{11} on the other. Non-epistatic interactions, by contrast, act as a “double lock”, ensuring that temporal down-regulation of one pathway does not lead to a genome-wide reactivation of TEs. Recently, it has been shown that, on many TEs, RNA-dependent DNA methylation (RdDM) and symmetrical DNA methylation maintenance (CG and CHG) act independently, and mutations on both pathways are required to reactivate elements under this “double lock” control\textsuperscript{4}. Environmental cues can lead to temporal or prolonged loss of DNA methylation or heterochromatin silencing, allowing some, albeit not all, transposons to transcribe, and sometimes mobilize\textsuperscript{12,13} by escaping control from the mechanisms mentioned above. Such ‘escape’ events might become useful to spawn genetic diversity that may serve as a reservoir for adaption to the ongoing stress\textsuperscript{13}, a situation mimicked by EVD in the epi15 lineage, discussed in depth later in this Chapter.

**Functional and developmental constrains to TE mobilization and transgenerational inheritance of new insertions**

Despite hundreds of TEs become transcriptionally active in ddm1 or met1 mutants\textsuperscript{14,15}, only a few have been showed to actually mobilize\textsuperscript{16}. These include members from all transposable element classes such as COPIA21, GP3 (Gypsy), MU1 (Mutator), CACTA or VANDAL. Rare TEs are apparently concurrently transcriptionally reactivated and transpose in both backgrounds, including EVD or CACTA, although this picture might be underestimated given that it is mostly drawn from sporadic phenotypes induced by transposition into coding genes\textsuperscript{1,17,18}. Consequently, many active transposons that preferentially transpose into gene-poor regions would have escaped our attention. Nonetheless, our transcriptome analysis of met1 mutants (which, again, probably provides only a partial view of reactivated TEs) revealed that most of the expressed TEs contain small deletions, PTC (premature termination codons), or are likely too degenerated to be functional. Hence, transcriptional reactivation certainly does not warrant mobilization although it
remains possible that many TEs, despite being non-autonomous, are still under positive selection that allows them to parasite/help related autonomous elements. **BARE** elements illustrate this phenomenon in barley: the defective *BARE2* carries a small deletion that impairs GAG translation; however it carries encapsidation signals allowing its incorporation into *BARE1* VLPs in order to be reverse-transcribed and integrated into the genome\textsuperscript{19}. Surely, genome resequencing of multiple methylation-deficient Arabidopsis mutants or their derived epiRILs should shade better light on the extent to which TEs are effectively mobile in this species.

However, even the competence for transposition does not necessarily warrant the efficient colonization of the host genome. Transposition in somatic tissues, for instance, would be mostly inconsequential to the host and an evolutionary dead-end for the TE, as new copies would not be transmitted to the next generations. All plant tissues arise from meristems containing stem cells that continuously develop into new organs and provide a source of incessant adaptation to the environment, enabling substantial phenotypic plasticity\textsuperscript{20}. A sudden burst of transposition in one leaf, even in early-stage leaf primordia (which would create a large clone of mutated cells), would only have a low impact on the normal development and physiology of the plant. One solution for plant TEs is to transpose into the meristem itself, from which reproductive organs will eventually emerge. However, plants, unlike animals, do not set apart their germline early during development: reproductive organs differentiate late from apical meristems. Thus, accumulation of TE-induced mutations in meristems throughout the plant development would eventually impair fitness and, in extreme cases, might even cause sterility. This would negatively impact the TE itself, because, being a genomic parasite, its success is tightly linked to the success of its host.

**EVD** provides a remarkable example of how TEs might balance germline transmission and host fitness cost through development. The confined L2-specific expression of **EVD** has three major implications: (i) it ensures meristem exclusion of potentially detrimental TE insertions by confining **EVD** expression to the leaves and, (ii) given that ovules differentiate from L2-cells located in the carpels (modified foliar organs that harbor the female gametophytes) it guarantees the successful transmission of new insertions; in addition, (iii) as ovules arise from independent cells, each carries a different landscape of **EVD** insertions. A further advantage of this scheme is that **EVD**’s transposition in diploid precursor cells would in principle
allow selection against detrimental insertions in the following haploid stage, after meiosis. Combinations of insertions that would seriously compromise basic cellular functions would thus be counter-selected before fertilization. Upon fertilization, new insertions having survived this proposed ‘haploid selection’ would become hemizygous, allowing an additional round of “fitness testing” before the next generation. The overall outcome of this scheme, illustrated in Chapter 2.1, is the generation of an immense pool of genetic diversity. Hence, considering that a single Arabidopsis plant can generate up to 10,000 seeds, each with its own history of EVD insertions at every generation, the raw genetic material generated and potentially available for selection is simply staggering. This might be of special relevance for autogamous (self-fertilizing) plants, like A. thaliana: colonization of new environments or prolonged exposure to stress might facilitate the reactivation of TEs as a source of genetic variability, otherwise scarce in inbred plant populations. Moreover, TE bursts can lead to new expression patterns with adaptive potential. This is exemplified by the Copia-like element Onsen, which carries heat response elements in its LTR. Upon induction and mobilization of Onsen by heat-shock, genes carrying new insertions nearby their promoter may themselves become heat inducible due to the cooption of Onsen regulatory elements, potentially contributing to new genetic schemes for heat tolerance. Likewise, the L2-specific elements present in the EVD LTR might provide tissue-specificity to genes that become proximal to new EVD insertions during the genomic proliferation phase.

**Complex transcriptional regulation of gene expression in Copia-like elements**

Our work has uncovered a feature of the biology of LTR-retrotransposons that had remained largely overlooked thus far but might be at the very core of their detection by hosts: the necessity to produce subgenomic transcripts in order to optimize the Gag-to-Pol ratio required for the successful assembly of VLPs. In EVD, and it seems in many other Copia-like elements, this is achieved through alternative splicing coupled to PCPA (premature cleavage and polyadenylation), which generates a short (sh)-GAG mRNA (sh-GAG). The rationale for this splicing strategy has been already extensively discussed in Chapter 2.2. Perhaps a more profound notion that emerges from these analyses, still, is the realization that EVD most likely uses the sh-GAG strategy primarily to avoid recognition by the non-sense mediated decay (NMD) machinery of its host, by maintaining all of its ORFs within the same, single frame.
Coupled splicing and PCPA provides the transposon with the ability to produce the sh-GAG mRNA and to simultaneously create stop codons that ensure the proper and efficient translation of Gag while minimizing the length of the ensuing, out-of-frame 3'-UTR (covering part of Pol). Thus, unlike many plant RNA viruses, WT EVD might be largely insensitive to mutations in the NMD pathway that normally stabilize transcripts harboring PTCs or long 3'UTRs, including many viral RNA genomes.

The balance between splicing-coupled PCPA and intron retention, required for production of full-length transcripts, must be also carefully maintained. On the one hand, lack of splicing would generate only full-length mRNAs most likely unable to produce the required molar excess of Gag for proper VLP formation. On the other hand, excessive splicing would inevitably lead to another dead-end by promoting the sole production of sh-GAG, as illustrated with the “super-splicer” Copia elements. The apparent solution to maintain the necessary equilibrium relies in the combined properties of the intron and U1 small nuclear (sn)RNA binding site. “Low quality” introns are less efficiently spliced than “bona fide” canonical introns, owing to features that partially depend on the length of the intron and U1 complementarity to 5' splice sites (ss). U1 is involved in the first steps of intron recognition by specifying 5'-ss, and the degree of complementarity to U1 influences splice site strength and splicing efficiency. Once U1 has identified the 5'-ss, the spliceosome scans the intron for proximal 3' splice sites, a process usually impinged by long introns causing, as a result, intron retention.

In addition, U1 can inhibit PCPA independently of its role in splicing. A study in human cultured cells using antisense morpholino-oligonucleotides (AMO) as competitors for U1 binding showed that at dosages not sufficient to inhibit splicing, the functional reduction of U1 resulted in the failure of cells to produce full-length mRNAs. This effect is due to co-transcriptional PCPA at the first polyadenylation signal (PAS) available after canonical and non-canonical U1 binding sites. Some mammalian viruses rely on U1-mediated PCPA inhibition to regulate the levels of different gene products in early and late infection phases. In plants, U1 seems to play a similar role: addition of U1 binding sites to the 3'-end of a GFP reporter mRNA inhibits transcriptional termination and polyadenylation, leading to read-through over terminator sequences. Our results strongly suggests that the combination of U1-mediated repression of PCPA and reduced splicing efficiency account for the regulation of EVD’s genome expression, and possibly of the genomes of many other
retroelements. Indeed, (i) EVD and other elements with similar splicing behavior display low complementarity to U1 snRNA, allowing a fraction of the transcripts to remain unspliced, whereas (ii) improving U1 binding site complementarity in EVD results in enhanced splicing, and elements with higher or almost perfect complementarity to U1 snRNA are more efficiently spliced. Furthermore, (iii) deletion of the intron carrying the U1 binding site or point mutation thereof, leads to PCPA. We thus predict that reducing the capacity of U1 to bind EVD will result in enhanced sh-GAG levels and that, conversely, introducing U1 binding sites into intron-less EVD constructs should restore their ability to produce full-length transcripts, two hypothesis that we are currently pursuing experimentally.

An additional aspect of the regulation of EVD via splicing involves the possible auto-regulation of this process by Gag, an idea not discussed in Chapter 2.2. Indeed, many retroviral proteins have been shown to influence splicing in order to regulate the relative levels of retroviral proteins at different stages of the infection\textsuperscript{31}. Thus far, our analyses of the comparative steady states levels of sh-GAG versus fl-EVD reflect mostly an equilibrium reached upon the establishment of a successful Gag-to-Pol molar ratio by EVD. These analyses are, however, poorly informative of the possible dynamic processes required for establishing this equilibrium, mechanisms that are probably too transient and subtle to be appreciated experimentally, at least in authentic genomic invasion contexts.

It is in fact conceivable that during the early stages of EVD transcription, only (or mostly) spliced sh-GAG is produced, generating enough Gag protein to provide protection to the bulk of fl-EVD against PTGS. A fraction of this bulk would also need to become available to Pol translation, a process that, we propose, might be facilitated by the binding of Gag to EVD nascent transcripts in a manner that could reduce splicing to favor the transcription/stabilization of fl-EVD. Although our preliminary subcellular localization data (data not shown) indicate that Gag is mostly cytoplasmic, other retroviral Gag proteins were found to be at least partly nuclear and their binding to retroviral RNA was only evident when nuclear export was inhibited\textsuperscript{32,33}. The use of drugs inhibiting nuclear export, like leptomycin B, should help elucidating if this is also the case for EVD Gag. We speculate that the presumptive binding site for EVD should be located at the intron, cryptic PAS or in sequences located downstream of the sh-GAG polyadenylation site to guarantee that only unspliced full-length transcripts are incorporated into VLPs for reverse-
transcription. This idea would somewhat conflict with the current views of retroviral encapsidation, where Gag binding sites (or packaging signals) are located in the 5'-UTR shortly after the LTR\textsuperscript{32}. However, our results show that EVD must carry some Gag binding element(s) in the coding sequence, since 35S:EVD constructs lacking native LTRs and UTRs are protected by Gag from PTGS and since the mRNA produced from them can be co-immunoprecipitated with Gag. One way to reconcile the models is to propose the existence of both UTR- and ORF-based Gag binding sites. Given that fl-EVD transcripts display two alternative transcription start sites, one including most of the 5'LTR and a second located 80-nt upstream of GAG, we propose that Gag binding to ORF sequences will protect (in addition to potentially regulate splicing) fl-EVD transcripts from PTGS; if the transcript contains 5'-LTR sequences, additional Gag binding to the 5'-UTR might favor VLP formation. In this scenario, transcripts lacking the 5'-LTR would not be encapsidated and would be solely dedicated to translation into Pol. This model is further supported by the fact that all sh-GAG RNAs contain 5'-LTR sequences. If those were sufficient to allow encapsidation, many VLPs would end up containing short transcripts not competent for reverse transcription and unsuited for Gag translation. In an effort to map such sequences and understand the potential role of Gag RNA binding activity in EVD biology, we are currently attempting to purify recombinant Gag protein to perform electrophoretic mobility assays (EMSA).

**Roles of RNA quality control in initiating RNA silencing on TEs and other foreign genetic elements**

Plants lack piRNA clusters or CRISPR-like systems to mount adaptive immune responses against transposons and viruses. They must therefore rely on purely innate processes to fend off these parasites. It is now well-accepted that the epigenetic reactivation of TE transcription in plants is accompanied by the production of 21-22-nt small siRNAs, a hallmark of post-transcriptional gene silencing (PTGS) evocative of the processing of viral dsRNA by DCL4 and DLC2 during antiviral defense\textsuperscript{16}. Primary antiviral silencing is often amplified through the action of host RDRs, chiefly RDR6, that provide further dsRNA templates to DCLs\textsuperscript{34}; however, the features attracting RDRs to viral transcripts and even the nature of these transcripts remain largely ill-defined, a situation also encountered in the case of transposons\textsuperscript{16}.
As discussed in Chapter 2.2, the models whereby ancient or TE-derived miRNAs might initiate RDR6-dependent secondary siRNAs\textsuperscript{35} fail to explain how young, invasive TEs are recognized and targeted for PTGS\textsuperscript{36}. Our study of EVD sheds significantly light on this question by suggesting how the unique set of events leading to the production of sh-GAG via alternative splicing and PCPA is most likely scrutinized by RNA quality control (RQC) to help distinguishing EVD RNAs from endogenous transcripts. This eventually results in the recruitment of the PTGS machinery to EVD, with sh-GAG being the main, if not exclusive, template for RDR6 action.

RQC pathways scan the transcriptome to prevent mRNAs carrying aberrant features from being translated. These aberrations include the absence of a 3’-polyA tail or 5’-cap, the presence of PTC or unusually long 3- UTRs (both triggers and targets of NMD), intron retention, and the absence of translation termination signals\textsuperscript{37-41}. The tight relationship between RQC and RNA silencing has been now convincingly established in plants, as mutations in several RQC pathway components have been retrieved in forward and reverse genetic screens for altered RNA silencing. Hence, mutations that impair RQC involved in monitoring mRNA capping, polyadenylation, 3’-end processing or splicing result in enhanced virus and transgene silencing because the pool of aberrant RNA derived from these foreign genetic elements becomes more available to the RNA silencing machinery in the corresponding RQC mutant than it would in the WT background\textsuperscript{41-43}. The general “security level” afforded by RQC in WT plants probably explains why endogenous mRNAs are generally protected from entering RNA silencing pathways: aberrant transcripts or miRNA cleavage products are probably intercepted and degraded by RQC before they have the opportunity to be scrutinized/accessed by RDRs. Consistent with this idea, some endogenous loci effectively become source of siRNAs in RQC mutants\textsuperscript{37-39,41,44}.

Based on our results, we propose that an additional RQC pathway that surveys the spliceosome is able to discriminate TE-derived transcripts and to promote their incorporation into the PTGS pathways through the recruitment of RDR6. Although further validation in transgenic Arabidopsis is needed, our preliminary results uncovered a strong connection between the quality of the EVD intron and the ability to trigger PTGS in the N. benthamiana transient assay, suggesting that, indeed, EVD’s recognition is initiated through the spliceosome. In
the yeast *C. neoformans* and in *Drosophila* a similar mechanism has been proposed to license specific transcripts to become source of sRNAs. In *C. neoformans*, many loci corresponding to centromere or TE-derived intergenic and genic centromere-like sequence are sources siRNAs corresponding to intronic and exonic sequences. The silencing-trigger transcripts are characterized by the presence of introns that are longer than the average, displaying enhanced retention time and a higher tendency to stall in spliceosomes. A kinetic competition between splicing and an RdRP-containing complex (known as SCANR) is believed to direct such transcripts for siRNA production. One key component of this system is the debranching enzyme (Dbr1), which, by resolving the lariat structure at the branching point might allow access to the RdRP. Similarly to what we have shown with EVD, mutations in splicing sites lead to increased accumulation of siRNAs in the yeast system\(^{45}\). In flies the system seems to operate in a more directed manner, owing to the specialization of piRNA clusters. Hence, very recently, mutations in components of the RDC (Rhino, Deadlock and Cutoff) complex (see Chapter 1, Fig. 3a) were shown to impair piRNA biogenesis and to concurrently increase splicing of piRNA clusters. Conversely, tethering Rhino (Rhi) to transgenes that display bidirectional transcription (and hence mimic piRNA clusters) inhibits their splicing and triggers piRNA biogenesis from these artificial loci\(^{46,47}\). The authors proposed that binding of Rhi to piRNA clusters through an interaction with H3K9me3 allows the recruitment of additional components of the RDC complex in ways that interfere with the splicing machinery, ultimately resulting in the stalling of spliceosomes and the licensing of stalled RNAs for primary piRNA biogenesis\(^{46}\). Given the TE-based origin of such clusters, it is entirely possible that TE-derived introns acting as a general signal for PTGS initiation (as seen in yeast with TE remnants and, as suggested by us in plants with active TEs) have been co-opted by the specialized piRNA clusters to optimize piRNA biogenesis via spliceosome stalling. In any case, our findings together with those in yeast and flies clearly identify aberrant splicing as a cross-kingdom feature that initiates and perhaps stimulates sRNAs biogenesis. Moreover, our results further suggest that the mechanism likely evolved originally as a dynamic and purely innate process used to singularize actively mobilizing TE mRNAs from the pool of cellular mRNAs.

Despite the above striking similarities observed across kingdoms, there are also noticeable differences as discussed in Chapter 2.2. First, unlike in Drosophila,
preexisting chromatin marks or heterochromatin-associated proteins unlikely underpin the process in plants because it was triggered by both endogenous EVD and 35S:EVD (independently of the position of the transgene in the multiple lines analyzed) through either a loss-of-DNA methylation or de novo introduction into the genome. A second distinguishing feature of the plant system compared to the yeast and fly systems is that very small amounts of siRNA arise from the intron itself, indicating that the mature sh-GAG RNA is most likely the main RDR6 template for dsRNA synthesis. A third distinctive characteristics of the plant-based mechanism (although this feature has not been specifically investigated in the other models) is that the sh-GAG 3’UTR is, by itself, sufficient to enhance siRNA production, at least in the transient assay system. We thus propose a two-step model for the initiation of EVD PTGS, as schematized in Chapter 2.2.

Step 1. As seen in yeast, the prolonged retention of EVD RNA in the spliceosome would recruit an unknown silencing component that “marks” the resulting mRNA to be later accessed by RDR6. RNA helicases are good candidates to play such a marking role because these enzymes have been shown to participate in almost every aspect of RNA biology and several have been isolated in genetic screens for RNA silencing factors. It is thus possible that RNA helicases that facilitate splicing by resolving stalled spliceosomes are maintained hooked onto the resulting mRNA and attract RDR6 in the cytoplasm. In Arabidopsis, the RNA helicase SDE3 (SILENCING DEFECTIVE 3) contributes to RNA silencing amplification by RDR6, and at least a fraction of it is nuclear-localized. A role for SDE3 in assisting splicing of TE-introns and facilitating RDR6 recruitment might be supported by our preliminary observation that the mRNAs of TEs identified as “splicers” in our genomic survey (e.g. AT5G35935; COPIA18A) are indeed up-regulated in sde3 mutants. However, given the diversity of RNA helicases, many such uncharacterized enzymes might similarly recruit RdRPs to TE RNAs.

Step 2. RDR6 recruitment to EVD is facilitated by additional elements present in the sh-GAG 3’-UTR. A first hypothesis to explain this phenomenon is that, given the weakness of the cryptic PAS, sh-GAG transcripts might be inefficiently polyadenylated (PolyA), a feature known to attract RDR6. Even only a small fraction of transcripts lacking a PolyA tail might be sufficient to initiate an RDR6 amplification cascade that would eventually lead to dsRNA synthesis from all (i.e. PolyA and non-PolyA) sh-GAG transcripts. Our ongoing quantification of PolyA RNA
levels using 3'-RNA-ligation mediated (RLM)-RACE should soon clarify this point. Substituting the non-canonical PAS found in sh-GAG for a canonical PAS and monitoring the effect on sh-GAG siRNA levels will also help us to understand the role and contribution of polyadenylation (or lack thereof) to EVD silencing. A second, non-mutually exclusive hypothesis entails that secondary structures present in the sh-GAG 3'-UTR might also contribute to the recruitment of RDR6. We favor a structure over a specific nucleotide sequence motif, given that stimulation of silencing by sh-GAG-like mRNAs is likely a shared feature of diverse TEs, divergent in sequences but similar in their transcription strategy. Secondary structures, like small stem-loops present at the 3'-ends of viral RNAs participate in the recognition of gRNA templates by viral RdRPs\(^{50,51}\). Such structures have also been shown to stabilize mRNAs lacking a polyA tail in many organisms, in particular those of histones\(^{52,53}\). Hence, in mammalian lysates, substituting a polyA tail for a 3'-end stem-loops can indeed stabilize reporter mRNAs and promote their translation. Interestingly, in organisms displaying endogenous RdRP activities, including C. elegans, fission yeast and plants, histone mRNAs display polyA tails, although some still contain stem-loop structures for regulatory purposes\(^{54-56}\). In Xenopus, histone mRNAs are generally deprived of polyA tails but become polyadenylated in oocytes\(^{57}\), which, incidentally, have been recently shown to specifically display an RdRP activity\(^{58}\). This striking correlation makes it thus conceivable that plant TEs such as EVD use histone-like secondary structures to stabilize non- or poorly-polyadenylated subgenomic transcripts but, that, as an unwanted consequence, these stem-loops are recognized by endogenous RdRPs. Ongoing in-depth analyses of the EVD sh-GAG 3'UTR and structural alignments among sh-GAG-like mRNAs from other TEs sharing the same transcriptional behavior as EVD will help addressing this issue.

Evidently, the experiments proposed above should be soon complemented by the deep-sequencing of sRNA in met1 mutants, in order to conclusively determine if TEs with similar splicing behavior display resembling siRNA distribution patterns. A subset of identified Copia-like elements belonging to the ‘non-splicers’, EVD-like ‘splicers’, and “super-splicers” have also been cloned to test directly their ability to trigger (or not) the anticipated siRNA responses when stably expressed from the 35S promoter in Arabidopsis. Our prediction is that the pressure from RDR6 will generally maintain the Gag levels below the threshold required for functional VLP assembly, but without necessarily affecting fl-TE transcript accumulation. Following TE
mobilization in met1 rdr6 mutant plants could be particularly informative in this respect.

**PTGS evasion as a tenet of successful TE expansion**

One of the striking features of EVD, uncovered in our study, is its ability to protect its gRNA from PTGS through Gag-RNA interactions. Identifying sh-GAG as the main source of siRNAs reinforces the arguments developed in Chapter 2.1, since fl-EVD can only be targeted by sh-GAG-derived siRNAs if the RNA-binding domain of Gag is mutated. As also explained in the same Chapter, the Cis-acting solution of EVD to evade rather than actively suppress silencing (as would most exogenous viruses) has been most likely driven by the necessity to maintain host fitness by minimizing the impact of silencing trans-suppression on endogenous miRNA- and siRNA-regulated developmental programs. Recent findings have provided more support to this hypothesis. The Hiun (Hi) element (VANDAL21 family) contains an anti-silencing factor within its transposase, which induces loss of CHG and CHH DNA methylation at silenced copies across the genome⁵⁹. However, this anti-silencing device is only effective on VANDAL21 copies, such that the detrimental consequences of a genome-wide loss of DNA methylation are avoided.

Some elements may still have retained a potential to trigger strong developmental effects upon their reactivation or following their horizontal transfer. For instance, the Brassica napus SB1.7 SINE element, when overexpressed in Arabidopsis, strongly interacts with HY Ponastic leaves 1 (HYL1), a double-stranded RNA binding protein (DRB) required for DCL1 processing of many pri-miRNA into mature miRNAs. Out-competing pri-miRNAs from HYL1 results in a general decreased accumulation of miRNAs and upregulation of their targets.⁶⁰. However, given that the natural SINE element activity is conditioned to that of LINE elements, their expression pattern might be tightly controlled and their mRNA levels generally below the threshold required for HYL1 saturation. Another example is provided in Drosophila, where introducing active P elements causes the reactivation of many other elements, leading to abnormal ovary development and sterility through as yet unknown mechanisms⁶¹.
A hierarchical continuum of action of plant DCL proteins provides a mechanistic link between PTGS and TGS

The generally silent and quiescent state of TEs in genomes indicates that, despite their ability to suppress or evade RNA silencing, TGS can be eventually re-installed on active elements and their derived copies. Early studies in *Drosophila* showed that some TEs are subject to copy number limitation. The *LINE* element *I* is present in all *Drosophila* strains in about 20-30 defective heterochromatic copies per haploid genome. In addition, most strains also contain 10-15 copies of *I* in euchromatic regions. Some strains lacking the euchromatic copies are susceptible to *I* mobilization when new copies are introduced but these stop transposing when a figure of 10-15 euchromatic copy is achieved again\(^{62,63}\). Although the mechanism underlying this phenomenon remains elusive, it clearly illustrates how hosts may somehow monitor TE copy number and modulate the active/non-active status of these parasites accordingly. In *Arabidopsis*, our observation that *EVD* is restricted to ~40 copies/genome reinforces the concept that TE amplification is tolerated to a certain degree before complete silencing. As discussed in Chapter 2.1, the underlying mechanism in Arabidopsis apparently relies on the combination of two key aspects: (i) the ability of *EVD* to protect itself from PTGS via Gag, and (ii) the ability of plant DCLs to act on each other’s substrates, with distinct affinities\(^{64,65}\).

The increase in *EVD* copy number across generations indeed leads to enhanced levels of sh-GAG RDR6-derived dsRNA, itself manifested by a steady increase in the corresponding siRNA levels in the epi15 inbred lineage. Normally processed by the PTGS proteins DCL4 and DCL2, the overwhelming levels of dsRNA accumulating over generations reaches a saturation threshold that now grants DCL3 access to the RDR6-dependent dsRNA. The consequence of this saturation event are important, because DCL3 is the main producer of 24-nt siRNAs that, once loaded into AGO4, direct *de novo* DNA methylation. In the case of *EVD*, this leads to methylation of a stretch of DNA that matches to near perfection the sh-GAG mRNA, which, we have proposed, is incidentally the main template of RDR6. Given the absence of preexisting chromatin marks on *EVD*, the only possible scaffold RNAs to be used by the 24-nt GAG siRNA-loaded AGO4 are Pol II nascent transcripts. A previously uncharacterized RDR6 – DCL3 – AGO4 – Pol II pathway thus allows the establishment of an RNA-level-to-DNA-level switch in silencing, whereby PTGS is converted into TGS. Once the first wave of GAG methylation is
installed, we infer that the system is instructed, probably through the recruitment of additional chromatin marks attracted by non-CG methylation, to spread the silencing to the LTR promoter. We speculate that this spreading entails antisense transcription by Pol IV or Pol V attracted by chromatin modifications deposited de novo onto the GAG stretch of DNA. The end result of this process is the permanent installation of TGS on the LTR through the canonical RdDM pathway, diagnosed by the synchronized transcriptional shutdown of all EVD copies and the concomitant production of LTR-derived 24-nt siRNAs. We have now introduced pol IV and pol V mutations in epi15 plants an, in parallel, are establishing tools in the normal epi15 background enabling chromatin immunoprecipitation (ChIP) of the two polymerases before, during and after the onset of TGS, to unravel the possible contribution of each component to the initiation, spreading and maintenance of RdDM on EVD.

Additional demonstration that true de novo methylation can be installed at Pol II transcribed loci through the mechanism proposed here will require validation of physical interactions between Pol II and AGO4, as has been shown for Pol V\textsuperscript{66}, and the characterization of novel factors involved in such a process. The precise nature of the mechanisms leading to the final installation of TGS also requires further investigation in order to elucidate, in particular, how the Pol IV/V system is eventually recruited precisely to the EVD LTR. This will entail the study of EVD’s transcriptional behavior, copy number and siRNA production in mutants known to impair the recruitment of either Pol IV, such as the recently identified shh1\textsuperscript{9}, Pol V (e.g. drd1)\textsuperscript{67} or the histone methyltransferases associated with the process (suvh2, suvh9)\textsuperscript{68}. We will also investigate the effects of mutant backgrounds impaired in the establishment of chromatin modification loops (e.g. cmt3)\textsuperscript{11}. Furthermore, given the near-exclusive symmetrical pattern of cytosines found in the EVD LTR, the epi15 inbred lineage has the potential to evolve further into a pure MET1- or DDM1-mediated maintenance system, and thus, to become independent of RdDM. To investigate this possibility, we are currently maintaining the propagation of the epi15 lineage well over the F14 generation documented in Chapter 2.1. According to the above hypothesis, EVD copies might be maintained silent in the absence of abundant LTR 24nt-siRNAs in future generations. In parallel, introducing EVD into backgrounds that prevent the installation of TGS (e.g. dcl3 or rdr2 mutants), will give crucial insights into (i) the extent of transposition that can be physically supported by the Arabidopsis genome and (ii) whether alternative pathways exist to impose chromatin-level silencing onto
proliferating TEs independently of RdDM (e.g. DNA repair mechanisms triggered by excessive EVD-induced DNA damage)\textsuperscript{69}.

A recent study conducted in \textit{ddm1 rdr6}, has corroborated that the POL II - RDR6 pathway drives the initiation of silencing on active TEs. However, in contrast to our findings, the authors state that the 21-22-nt siRNAs direct RdDM as they are the most abundant siRNA species derived from the reactivated TE loci\textsuperscript{70}. However, we have shown that the RDR6-DCL3-dependent 24nt-siRNAs are much lower in steady state abundance compared to the DCL2/DCL4 products. Nonetheless, they are efficiently loaded into AGO4 and readily detectable via immunoprecipitation, experiments not conducted in the above-mentioned study. Moreover, DCL2/DCL4 are dispensable for, and even detrimental to, the silencing of EVD since both the endogenous and 35S:EVD switch to promoter RdDM-dependent TGS much more rapidly in the \textit{dcl2 dcl4} double mutant than in the WT background. Incidentally, the results with 35S:EVD also suggest that methylation spreading is not an LTR-intrinsic process, but can affect other types of promoters. To our knowledge, no existing data shows a correlation between the levels 24-nt siRNA and their ability to trigger DNA methylation and, in fact, extremely low levels of 24-nt siRNAs are sufficient to induce DNA methylation at the flowering locus FWA, for instance\textsuperscript{71}.

As discussed in Chapter 2.1, further lines of evidence supporting the widespread occurrence of the RDR6-DCL3-directed RdDM pathway are found in a recent study of genome-wide DNA methylation conducted in 86 mutants affecting RNA silencing or chromatin modifications. The study shows that mutations in POL II or RDR6 lead to decreased DNA methylation at many loci generally associated with 21-22-nt siRNAs. Yet careful inspection of these loci shows that, without exception, the 21-22-nt species are always mixed with variable proportions of 24-nt siRNAs\textsuperscript{6}. We thus consider it likely that the same RDR6-DCL3 pathway that targets EVD might promote RdDM at these endogenous loci through a similar saturation mechanism as documented in our study. A clear proof-of-concept is provided by the most obvious candidates represented by TAS loci, which are well known producers of siRNAs in an RDR6-DCL2/DCL4 dependent manner\textsuperscript{72}. TAS loci have also been shown to undergo cytosine methylation interpreted, as in the above-mentioned study, as the direct result of 21-22-nt siRNA action\textsuperscript{73}. Yet, our inspection of all Arabidopsis TAS loci shows again their invariable association with lower levels of 24-nt siRNAs species. The RDR6-dependent dsRNA produced from tasiRNA precursor transcripts is known
to undergo alternative processing into differently sized siRNA in single or double *dcl* mutant backgrounds similar to our findings with *EVD*. Our *in silico* analysis of the available data has indeed revealed that, consistent with our observation at *EVD* loci, TAS loci display enhanced levels and a broader physical extent of DNA methylation (Fig 1) in *dcl4* and *dcl2 dcl4* mutants.

Figure 1. Enhanced DNA methylation in *dcl2 dcl4* at loci processed by RDR6. WT and *dcl2 dcl4* DNA methylation profiles at trans-acting siRNA generating loci. Data was obtained from Stroud et al., 2013. Red: CG; blue: CHG; Green: CHH. x-axis: coordinates on Arabidopsis genome (TAIR 10); y-axis: single resolution C methylation levels in sense (+) and antisense (-) strands.

The above examples might constitute a mere snapshot of a much more profound and dynamic silencing mechanism, since our study of *EVD* illustrates how *de novo* 3’GAG methylation could further activate, via 3’-5’ spreading, RdDM at the 5’ LTR, ultimately triggering TGS. Unlike in most cases of RdDM affecting TE remnants/repeats in *cis*, the ensuing LTR 24-nt siRNAs displayed remarkable trans-silencing abilities. This overall process might affect many types of loci, including transgenes, but it would be mostly, if not only, evident in trans-generational studies, as conducted with *EVD*. Consequently, the same mechanism could well underlie several epigenetic phenomena only manifested after multiple generations in plants. These phenomena include “transgressive RNAi” where F2 progenies display siRNA populations and RNAi-based phenotypes not seen in either parent or in the F1, as well as the onset of seemingly spontaneous epimutations in highly inbred Arabidopsis. The same mechanism could, in principle, provide plants with a means
to correct abnormally elevated gene expression triggered by prolonged stress exposure, for instance. High transcriptional activity would render the involved loci more prone to aberrant RNA production and to RQC saturation, promoting their access by RDR6 to initiate PTGS which, under persistent stress conditions, could possibly cause their eventual TGS through the mechanisms uncovered here with EVD.

**Contribution of TE bursts to the epigenetic landscape of plant genomes**

As in the genome invasion phase described earlier in this discussion, mutualism is also an anticipated feature of the epigenetic silencing phase of EVD biology. Hence, genetic drift caused by TGS could decrease the excessive load of active EVD copies, but also allow the selection, for gene regulation purposes, of epigenetic modules derived from EVD drift remnants. This notion is illustrated with the solo-LTR found in the promoter of RPP4, which belongs to a class of disease resistance genes whose constitutive activation decreases plant fitness\(^7^8\). In this context, *de novo* epiallelism resulting from the dynamic onset of EVD silencing might be favorable in adding a rheostatic component to RPP4 expression. Interestingly, biotic and abiotic stress causes transient *de novo* de-methylation and temporal dampening of RdDM at TEs and repeats\(^1^2^,7^9\). A similar pathogen-induced release of RPP4 solo-LTR methylation could thus help adjusting the RPP4-mediated defense response to the stress duration, upon which epigenetic silencing could be re-installed at the locus. Prolonged stress might also reactivate intact, yet silent EVD copies\(^8^0\), possibly prompting novel retrotransposition episodes similar to those created here in *ddm1* or *met1* epiRILs. The dynamic and flexible control afforded by epiallelism is expected to be particularly pronounced in epiRILs, given the sheer number and diversity of insertions accumulated during inbreeding and the extraordinary trans-acting potency of LTR 24-nt siRNAs. Grasping the full extent and significance of this phenomenon both at the genomic and ecological scales will undoubtedly shed further light on the impact of TEs on plant evolution and adaptation.
3.2. REFERENCES


70. Nuthikattu, S. et al. The Initiation of Epigenetic Silencing of Active Transposable Elements Is


<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>5mC</td>
<td>5-methylcytosine</td>
</tr>
<tr>
<td>ac</td>
<td>Acetylation</td>
</tr>
<tr>
<td>ACT2</td>
<td>ACTIN2</td>
</tr>
<tr>
<td>AGO</td>
<td>ARGONAUTE</td>
</tr>
<tr>
<td>AS</td>
<td>Splicing acceptor site</td>
</tr>
<tr>
<td>ATR</td>
<td>Attrapé</td>
</tr>
<tr>
<td>Aub</td>
<td>Aubergine</td>
</tr>
<tr>
<td>Cas</td>
<td>CRISPR-associated</td>
</tr>
<tr>
<td>cDNA</td>
<td>copy DNA</td>
</tr>
<tr>
<td>CENP</td>
<td>CENTROMERE-BINDING PROTEIN</td>
</tr>
<tr>
<td>Cir4</td>
<td>CRL4-like protein</td>
</tr>
<tr>
<td>CLSY1</td>
<td>CLASSY1</td>
</tr>
<tr>
<td>CMT3</td>
<td>CHROMOMETHYLASE 3</td>
</tr>
<tr>
<td>CRISPR</td>
<td>Cluster of regularly interspaced palindromic repeats</td>
</tr>
<tr>
<td>cRNA</td>
<td>copy RNA</td>
</tr>
<tr>
<td>crRNA</td>
<td>CRISPR-RNA</td>
</tr>
<tr>
<td>CYP</td>
<td>Cysteine protase</td>
</tr>
<tr>
<td>DBD</td>
<td>DNA-binding domain</td>
</tr>
<tr>
<td>DCL</td>
<td>DICER-like</td>
</tr>
<tr>
<td>DCR</td>
<td>DICER</td>
</tr>
<tr>
<td>DDM1</td>
<td>DECREASE IN DNA METHYLATION 1</td>
</tr>
<tr>
<td>DDR</td>
<td>DRD1-DMS3-RDM1</td>
</tr>
<tr>
<td>DIR</td>
<td>Direct inverted repeat</td>
</tr>
<tr>
<td>DIRS</td>
<td>Dictostelium intermediate repeat sequence</td>
</tr>
<tr>
<td>DME</td>
<td>DEMETER</td>
</tr>
<tr>
<td>DMS3</td>
<td>DEFECTIVE IN MERISTEM SILENCING 3</td>
</tr>
<tr>
<td>DNMT1</td>
<td>DNA METHYLTRANSFERASE</td>
</tr>
<tr>
<td>dpi</td>
<td>days post-inoculation</td>
</tr>
<tr>
<td>DRD1</td>
<td>DEFECTIVE IN RNA-DIRECTED DNA METHYLATION</td>
</tr>
<tr>
<td>DRM2</td>
<td>DOMAINS REARRANGED METHYLTRANSFERASE 2</td>
</tr>
<tr>
<td>DS</td>
<td>Splicing donor site</td>
</tr>
<tr>
<td>dsRNA</td>
<td>Double-stranded RNA</td>
</tr>
<tr>
<td>eamiRNA</td>
<td>epigenetically-activated miRNA</td>
</tr>
<tr>
<td>EJ</td>
<td>Exon junction</td>
</tr>
<tr>
<td>EMSA</td>
<td>electrophoretic mobility assay</td>
</tr>
<tr>
<td>EN</td>
<td>Endonuclease</td>
</tr>
<tr>
<td>ENV</td>
<td>Envelope protein</td>
</tr>
<tr>
<td>epi15</td>
<td>epiRIL #15</td>
</tr>
<tr>
<td>Epiallele</td>
<td>Epigenetic allele</td>
</tr>
<tr>
<td>epiRIL</td>
<td>Epigenetic recombinant-imbred line</td>
</tr>
<tr>
<td>ERV</td>
<td>Endogenous retrovirus</td>
</tr>
<tr>
<td>ES</td>
<td>enhanced splicing</td>
</tr>
<tr>
<td>EtBr</td>
<td>Ethidium bromide</td>
</tr>
<tr>
<td>EVD</td>
<td>Evadé</td>
</tr>
<tr>
<td>EVDi</td>
<td>EVD intron</td>
</tr>
<tr>
<td>F(n)</td>
<td>n Generation</td>
</tr>
<tr>
<td>FAR1</td>
<td>FAR-RED-IMPAIRED RESPONSE 1</td>
</tr>
<tr>
<td>FHY3</td>
<td>FAR-RED ELONGATED HYPOCOTYLS 3</td>
</tr>
<tr>
<td>fl-EVD</td>
<td>Full-length EVD</td>
</tr>
<tr>
<td>FM</td>
<td>Female megaspore</td>
</tr>
<tr>
<td>GAG</td>
<td>retroviral capsid/nucleocapsid protein(s)</td>
</tr>
<tr>
<td>GFP</td>
<td>Green fluorescence protein</td>
</tr>
<tr>
<td>gRNA</td>
<td>genomic RNA</td>
</tr>
<tr>
<td>GSP</td>
<td>Gene-specific primer</td>
</tr>
<tr>
<td>GSP-RT</td>
<td>Gene-specific primer-primed reverse transcription</td>
</tr>
<tr>
<td>H1/2/H3/H4</td>
<td>HISTONE 1/2/3/4</td>
</tr>
<tr>
<td>HMW</td>
<td>High-molecular weight</td>
</tr>
<tr>
<td>HP1</td>
<td>HETEROCHROMATIC PROTEIN 1</td>
</tr>
<tr>
<td>HPL-2</td>
<td>HP1-like protein 2</td>
</tr>
<tr>
<td>HYL1</td>
<td>HYPOASTIC LEAVES 1</td>
</tr>
<tr>
<td>IBM1</td>
<td>INCREASE IN BONSAI METHYLATION</td>
</tr>
<tr>
<td>IDM1</td>
<td>INCREASED DNA METHYLATION</td>
</tr>
<tr>
<td>IN</td>
<td>Integrase</td>
</tr>
<tr>
<td>IP</td>
<td>Immunoprecipitation</td>
</tr>
<tr>
<td>IR</td>
<td>Inverted repeat</td>
</tr>
<tr>
<td>K</td>
<td>Lysine</td>
</tr>
<tr>
<td>KYP</td>
<td>KRYPTONITE</td>
</tr>
<tr>
<td>LARD</td>
<td>large retrotransposon derivates</td>
</tr>
<tr>
<td>LINE</td>
<td>Long interspersed nuclear elements</td>
</tr>
<tr>
<td>LMW</td>
<td>Low-molecular weight</td>
</tr>
<tr>
<td>LTR</td>
<td>Long terminal repeat</td>
</tr>
<tr>
<td>LTR-RTE</td>
<td>Long terminal repeat retrotransposon</td>
</tr>
<tr>
<td>m7G</td>
<td>7-methylguanosine</td>
</tr>
<tr>
<td>me</td>
<td>Methylation</td>
</tr>
<tr>
<td>Met</td>
<td>Methionine</td>
</tr>
<tr>
<td>MET1</td>
<td>METHYLTRANSFERASE 1</td>
</tr>
<tr>
<td>MID</td>
<td>Middle</td>
</tr>
<tr>
<td>miRNA</td>
<td>Micro RNA</td>
</tr>
<tr>
<td>MMC</td>
<td>Megaspore mother cell</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger RNA</td>
</tr>
<tr>
<td>mU1</td>
<td>mutated U1</td>
</tr>
<tr>
<td>N</td>
<td>Amino-terminal</td>
</tr>
<tr>
<td>NB</td>
<td>Northern blot</td>
</tr>
<tr>
<td>NGS</td>
<td>Next-generation sequencing technologies</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>NMD</td>
<td>Nonsense-mediated decay</td>
</tr>
<tr>
<td>NRDE</td>
<td>NUCLEAR RNAi DEFECTIVE</td>
</tr>
<tr>
<td>nt</td>
<td>Nucleotide</td>
</tr>
<tr>
<td>ORF</td>
<td>Open reading frame</td>
</tr>
<tr>
<td>PAS</td>
<td>Polyadenylation signal</td>
</tr>
<tr>
<td>PAZ</td>
<td>PIWI-ARGONAUTE-ZWILLE</td>
</tr>
<tr>
<td>PCA</td>
<td>Premature cleavage and polyadenylation</td>
</tr>
<tr>
<td>ph</td>
<td>Phosphorylation</td>
</tr>
<tr>
<td>pRNA</td>
<td>PIWI-interacting RNA</td>
</tr>
<tr>
<td>PLE</td>
<td>Penelope</td>
</tr>
<tr>
<td>POL</td>
<td>DNA polymerase</td>
</tr>
<tr>
<td>PR</td>
<td>Aspartic protease</td>
</tr>
<tr>
<td>PRG1</td>
<td>PIWI-RELATED GENE</td>
</tr>
<tr>
<td>PTC</td>
<td>Premature termination codon</td>
</tr>
<tr>
<td>PTGS</td>
<td>Posttranscriptional gene silencing</td>
</tr>
<tr>
<td>PTM</td>
<td>Posttranslational modification</td>
</tr>
<tr>
<td>R</td>
<td>Arginine</td>
</tr>
<tr>
<td>RACE</td>
<td>Rapid amplification of cDNA ends</td>
</tr>
<tr>
<td>RAG</td>
<td>RECOMBINATION-ACTIVATING GENE</td>
</tr>
<tr>
<td>RBD</td>
<td>RNA-binding domain</td>
</tr>
<tr>
<td>RDC</td>
<td>RHINO-DEADLOCK-CUTOFF</td>
</tr>
<tr>
<td>RdDM</td>
<td>RNA-directed DNA Methylation</td>
</tr>
<tr>
<td>RDM1</td>
<td>RNA-DIRECTED DNA METHYLATION 1</td>
</tr>
<tr>
<td>RDR</td>
<td>RNA-DEPENDENT RNA POLYMERASE</td>
</tr>
<tr>
<td>RdRC</td>
<td>RNA-dependent RNA polymerase complex</td>
</tr>
<tr>
<td>RdRP</td>
<td>RNA-dependent RNA polymerase</td>
</tr>
<tr>
<td>RH</td>
<td>Random hexamers</td>
</tr>
<tr>
<td>RH-RT</td>
<td>Random hexamers-primed reverse transcription</td>
</tr>
<tr>
<td>Rhi</td>
<td>Rhino</td>
</tr>
<tr>
<td>RISC</td>
<td>RNA-induced silencing complex</td>
</tr>
<tr>
<td>RITS</td>
<td>RNA-induced silencing complex</td>
</tr>
<tr>
<td>RN</td>
<td>Rnase H</td>
</tr>
<tr>
<td>RNAi</td>
<td>RNA interference</td>
</tr>
<tr>
<td>ROS1</td>
<td>REPRESSOR OF SILENCING 1</td>
</tr>
<tr>
<td>rpm</td>
<td>reads per million</td>
</tr>
<tr>
<td>RQC</td>
<td>RNA quality control</td>
</tr>
<tr>
<td>rRNA</td>
<td>Ribosomal RNA</td>
</tr>
<tr>
<td>RS</td>
<td>reduced splicing</td>
</tr>
<tr>
<td>RT</td>
<td>Reverse transcriptase</td>
</tr>
<tr>
<td>RTE</td>
<td>Retrotransposon</td>
</tr>
<tr>
<td>S</td>
<td>Serine</td>
</tr>
<tr>
<td>SAM</td>
<td>Shoot apical meristem</td>
</tr>
<tr>
<td>SB</td>
<td>Southern blot</td>
</tr>
<tr>
<td>SDE3</td>
<td>SILENCING DEFECTIVE 3</td>
</tr>
<tr>
<td>SDR</td>
<td>Split direct repeat</td>
</tr>
<tr>
<td>sh-GAG</td>
<td>Short GAG</td>
</tr>
<tr>
<td>SHH1</td>
<td>SHAWADEE HOMOLOG 1</td>
</tr>
<tr>
<td>SINE</td>
<td>Short interspersed nuclear elements</td>
</tr>
<tr>
<td>siRNA</td>
<td>Small-interfering RNA</td>
</tr>
<tr>
<td>SNAC</td>
<td>Small non-autonomous CACTA elements</td>
</tr>
<tr>
<td>snRNA</td>
<td>small nuclear RNA</td>
</tr>
<tr>
<td>spl</td>
<td>spliced</td>
</tr>
<tr>
<td>sRNA</td>
<td>small RNA</td>
</tr>
<tr>
<td>ssRNA</td>
<td>Single-stranded RNA</td>
</tr>
<tr>
<td>su</td>
<td>Sumoylation</td>
</tr>
<tr>
<td>T</td>
<td>Threonine</td>
</tr>
<tr>
<td>tasiRNA</td>
<td>Trans-acting small-interfering RNA</td>
</tr>
<tr>
<td>TE</td>
<td>Transposable element</td>
</tr>
<tr>
<td>TGS</td>
<td>Transcriptional gene silencing</td>
</tr>
<tr>
<td>TIR</td>
<td>Terminal inverted repeat</td>
</tr>
<tr>
<td>Tpase</td>
<td>Transposase</td>
</tr>
<tr>
<td>TRIM</td>
<td>Terminal repeat retrotransposon in miniature</td>
</tr>
<tr>
<td>TSD</td>
<td>Target site duplication</td>
</tr>
<tr>
<td>u-spl</td>
<td>un-spliced</td>
</tr>
<tr>
<td>ub</td>
<td>Ubiquitylation</td>
</tr>
<tr>
<td>UPF1</td>
<td>UP-FRAMESHIFT 1</td>
</tr>
<tr>
<td>VLP</td>
<td>Virus-like particle</td>
</tr>
<tr>
<td>VSR</td>
<td>Viral suppressor of RNA silencing</td>
</tr>
<tr>
<td>WAGO</td>
<td>Worm-AGO</td>
</tr>
<tr>
<td>WB</td>
<td>Western blot</td>
</tr>
<tr>
<td>WT</td>
<td>Wild-type</td>
</tr>
<tr>
<td>YR</td>
<td>Tyrosine recombinase</td>
</tr>
<tr>
<td>Δi</td>
<td>intron deletion</td>
</tr>
<tr>
<td>Δu</td>
<td>UTR deletion</td>
</tr>
</tbody>
</table>