

The turbulent life of Sirevirus retrotransposons and the evolution of the maize genome: more than ten thousand elements tell the story

Alexandros Bousios^{1,*†}, Yiannis A. I. Kourmpetis², Pavlos Pavlidis³, Evangelia Minga¹, Athanasios Tsaftaris^{1,4} and Nikos Darzentas^{1,†}

¹Institute of Agrobiotechnology, Centre for Research and Technology Hellas, Thessaloniki 57001, Greece,

²Laboratory of Bioinformatics, Wageningen University, 6708 PB Wageningen, The Netherlands,

³The Exelixis Lab, Scientific Computing Group, Heidelberg Institute for Theoretical Studies, D-69118 Heidelberg, Germany, and

⁴Department of Genetics and Plant Breeding, Aristotle University of Thessaloniki, Thessaloniki 54006, Greece

Received 2 August 2011; revised 27 September 2011; accepted 29 September 2011; published online 8 November 2011.

*For correspondence (fax +30 231 0498270; e-mail alexandros.bousios@gmail.com).

†These authors contributed equally.

SUMMARY

Sireviruses are one of the three genera of *Copia* long terminal repeat (LTR) retrotransposons, exclusive to and highly abundant in plants, and with a unique, among retrotransposons, genome structure. Yet, perhaps due to the few references to the Sirevirus origin of some families, compounded by the difficulty in correctly assigning retrotransposon families into genera, Sireviruses have hardly featured in recent research. As a result, analysis at this key level of classification and details of their colonization and impact on plant genomes are currently lacking. Recently, however, it became possible to accurately assign elements from diverse families to this genus in one step, based on highly conserved sequence motifs. Hence, Sirevirus dynamics in the relatively obese maize genome can now be comprehensively studied. Overall, we identified >10 600 intact and approximately 28 000 degenerate Sirevirus elements from a plethora of families, some brought into the genus for the first time. Sireviruses make up approximately 90% of the *Copia* population and it is the only genus that has successfully infiltrated the genome, possibly by experiencing intense amplification during the last 600 000 years, while being constantly recycled by host mechanisms. They accumulate in chromosome-distal gene-rich areas, where they insert in between gene islands, mainly in preferred zones within their own genomes. Sirevirus LTRs are heavily methylated, while there is evidence for a palindromic consensus target sequence. This work brings Sireviruses in the spotlight, elucidating their lifestyle and history, and suggesting their crucial role in the current genomic make-up of maize, and possibly other plant hosts.

Keywords: maize, transposable elements, plant genome evolution, Sirevirus, epigenetics, comparative genomics.

INTRODUCTION

The maize (*Zea mays*) B73 inbred line was recently fully sequenced by the Maize Sequencing Consortium (Schnable *et al.*, 2009), because of the importance of maize as a staple food source but also its intricate genomic landscape. With a nuclear content of 2300 Mb, it is the largest and most complex genome sequenced to date, mostly comprising transposable elements (TEs) (approximately 85%), and in particular long terminal repeat (LTR) retrotransposons (>75%) (Baucom *et al.*, 2009a). The intense activity of such elements has been identified as the main driver for the

dramatic genome expansion during the last 3 million years (Myr) (SanMiguel *et al.*, 1998), and, concurrently with the rapid removal through homologous and illegitimate recombination (Shirasu *et al.*, 2000; Ma and Bennetzen, 2004; Ma *et al.*, 2004), for the vast differences in the composition of intergenic regions between the Mo17 and B73 inbred lines (Brunner *et al.*, 2005), or between maize haplotypes (Fu and Dooner, 2002; Wang and Dooner, 2006).

Research has shown that this remarkable TE activity has not been constant throughout evolution. Retrotransposon

families have undergone amplification bursts in different time periods within the last 3 Myr in maize (Liu *et al.*, 2007; Kronmiller and Wise, 2008) and related species (Baucom *et al.*, 2009b; Choulet *et al.*, 2010). Moreover, intense proliferation further back in evolutionary time has almost certainly occurred but is difficult to discern, as the half-life of LTR retrotransposons is only a few million years (SanMiguel *et al.*, 1998; Ma *et al.*, 2004).

The resulting LTR retrotransposon diversity in maize is immense, with >400 families described (Baucom *et al.*, 2009a), generating the typical organization of large grass genomes with seas of nested elements surrounding gene islands (Schnable *et al.*, 2009; Choulet *et al.*, 2010). High-copy number families in maize preferentially accumulate either in heterochromatin near gene-rich regions (*Copia* superfamily) or in large gene-poor heterochromatic blocks of pericentromeric areas (*Gypsy* superfamily). On the other hand, less abundant families and other TEs bias their integration near or within genes (Liu *et al.*, 2007; Baucom *et al.*, 2009a).

The LTR retrotransposons impact not only on the host genome size and organization, but also affect gene function, regulation and evolution. Like *Helitron* DNA transposons (Morgante *et al.*, 2005; Yang and Bennetzen, 2009), they can capture and mobilize gene fragments (Wang *et al.*, 2006), with >400 potential cases recently reported in maize (Baucom *et al.*, 2009a). Their mobility can trigger large-scale chromosomal rearrangements with dramatic changes in gene order and synteny, whilst their integration in the local environment can interrupt or alter gene expression and function. Furthermore, LTR retrotransposons harbor numerous *cis*-regulatory elements, thereby providing a profuse source of building blocks for the rewiring of host regulatory networks (Feschotte, 2008). As a consequence of their tremendous ability to induce change, the vast majority of TEs are kept quiescent by epigenetic regulation (Slotkin and Martienssen, 2007), although this state can be suddenly reversed by means of stress and 'genomic shock' (McClintock, 1984).

The International Committee on the Taxonomy of Viruses (ICTV) has classified Sireviruses (together with the Pseudo- and Hemiviruses) as one of the three genera of the *Copia* superfamily (Boeke *et al.*, 2006). Notably, Sireviruses are the only LTR retrotransposons that have exclusively proliferated within plant genomes (Peterson-Burch and Voytas, 2002). Moreover, it is the only *Copia* genus for which phylogenetic data strongly suggest a monophyletic origin (Bousios *et al.*, 2010) (Figure 1a) and whose members often contain an envelope gene (Havecker *et al.*, 2005). Examples include the *Opie* and *Ji* families that account for nearly a fifth of the maize genome (Baucom *et al.*, 2009a), the abundant *Osr8* family in rice (McCarthy *et al.*, 2002) and the envelope-containing *Hopie* in maize and *Maximus* in barley and wheat (Wicker and Keller, 2007). Sireviruses are

unique among LTR retrotransposons also in terms of their genome structure (Gao *et al.*, 2003; Bousios *et al.*, 2010). They contain a plethora of short but highly conserved motifs within their otherwise extremely diverse genome (Figure 1b), regardless of the evolutionary distance of their monocot or eudicot hosts. The motifs are found in key non-coding regions, critical for the regulation, reverse transcription and integration of the element, and possibly for its virulence capacity through the activation of the envelope gene.

Despite these intriguing characteristics and their abundance among plant TEs (Havecker *et al.*, 2004), there have been only a few publications on Sireviruses, and scarce reference to the Sirevirus-like origin of some elements. As a result, coherent information on the extent of Sirevirus infiltration in species such as maize, rice, soybean and other fully sequenced genomes is virtually absent. This is exemplified in the recently proposed unified classification system for eukaryotic TEs (Wicker *et al.*, 2007), where classification at the genus level is missing (as opposed to the ICTV), possibly due to the difficulty in correctly assigning TE families into genera. However, looking at the conserved genome structure of Sireviruses, we recently showed (Darzentas *et al.*, 2010) that it is possible to accurately assign members of diverse families to this genus, and hence uncovered previously obscured information for the integrative impact of a whole TE genus on host genomes.

In the current study, we conducted a systematic and multilayered analysis of >10 600 intact and approximately 28 000 degenerate maize Sireviruses discovered by the MASiVE algorithm (Mapping and Analysis of SireVirus Elements) in the very large genome of their host. The phylogenetic complexity of the maize Sirevirus lineage was uncovered, revealing that Sireviruses are in fact the only *Copia* genus that has successfully proliferated in maize, currently making up approximately 90% of the *Copia* population. Many new copies from all Sirevirus families have accumulated during the last 600 000 years, while smaller subfamily-specific amplification explosions have occurred further back in evolutionary time. We discerned the deletion rates of Sireviruses by studying fragmented elements and solo LTRs to discover that *Opie* and *Ji* are much more rapidly removed from the genome than previously suggested. Their chromosomal distribution differs from the general *Copia* preference for pericentromeric areas of plant genomes, by residing closer to genes than expected by random. Evidence is provided for a consensus palindromic target sequence and preferred integration zones within the genomes of TEs, whereas the Sirevirus LTRs and their flanking domains appear to be heavily methylated. Our findings describe how Sireviruses infiltrated the maize genome and provide insights into the impact of their turbulent life on the evolution of their plant hosts.

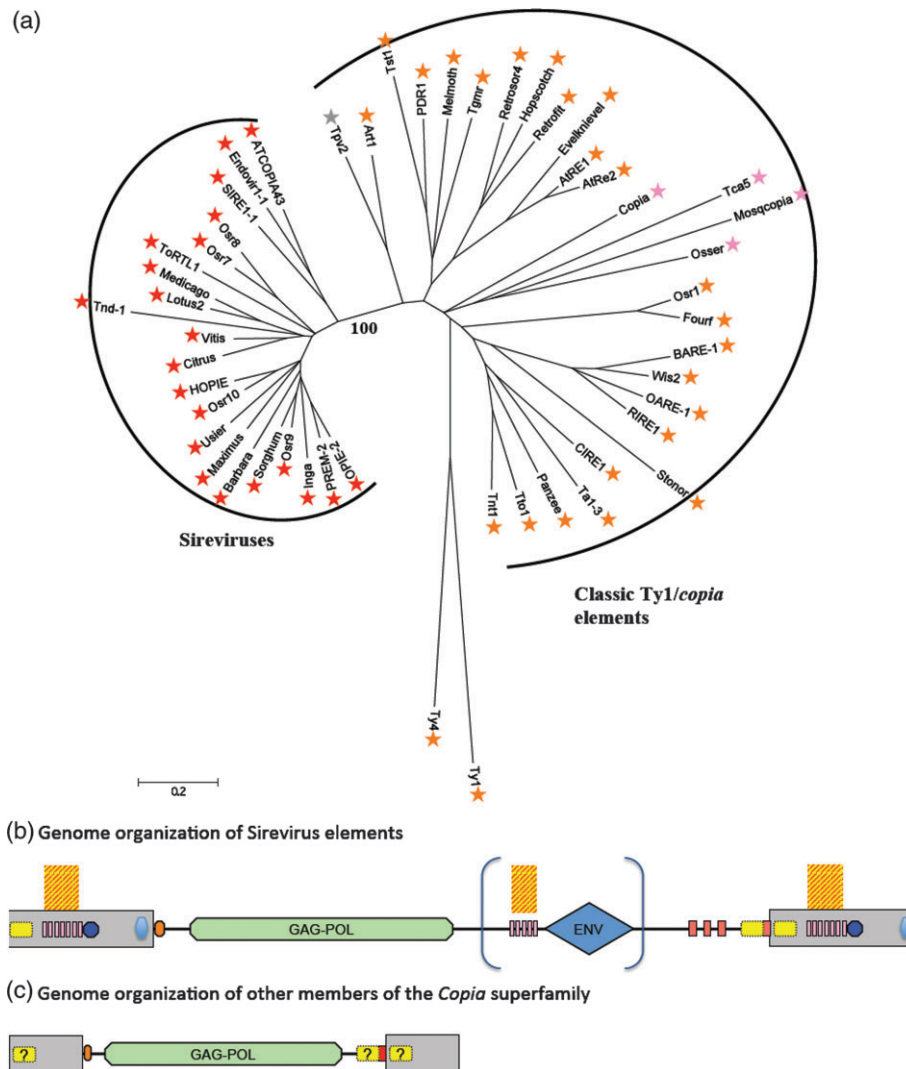


Figure 1. Phylogenetic and genome structure analyses within the *Copia* superfamily (Figure adapted from Bousios *et al.*, 2010).

(a) Exemplars from all three *Copia* genera, i.e. Sirevirus (red star), Hemivirus (pink star) and Tpv2 (grey star) that is of unknown classification according to the ICTV (Boeke *et al.*, 2006), were used for the construction of the reverse transcriptase (RT) phylogenetic tree. Sireviruses form a separate branch, which is supported with 100% confidence by the bootstrap analysis – there is no phylogenetic basis for separating the other two genera (herein referred as ‘classic’ elements following Havecker *et al.*, 2005).

(b, c) Genome organization of Sireviruses and classic elements. The *gag-pol* polyprotein is shown in green and the envelope gene (if present) as a light blue diamond. The inverted repeat (IR) arms (yellow) surround the internal domain/3′ long terminal repeat (LTR) junction of all Sireviruses and few classic elements. The outmost 5′ side of the junction is occupied by the terminal polyurine tract octamer (PPT, in red), which precisely borders the IR left arm. The upstream PPTs cluster within the proximal 1000 bp to the junction. The palindromic and putative *cis*-regulatory repeated motifs (RMs) (pink) are located within the first 200–700 bp of the Sirevirus LTRs upstream of a highly conserved TATA box (blue circle), and at the 5′ side of the envelope gene (when present). The RM clusters define the borders of CpG islands (orange bars). The 5′ LTR/internal domain junction harbors the conserved primer binding site (PBS, orange box), while the C-rich integrase signal (light blue hexagon) is located 20–30 bp upstream of the 3′ terminus of the Sirevirus LTRs. The genome size difference between Sireviruses and classic elements is approximately drawn to scale.

RESULTS

Remarkable abundance, diversity and familial relationships of Sireviruses in the maize genome

MASiVe (Darzentas *et al.*, 2010) is based on identifying, in a stepwise manner, some of the highly conserved motifs of the Sirevirus genome, thus eventually building full-length elements with high sensitivity (Figure S1 in Supporting

information). Its application to the first version of the maize B73 genome (<http://www.maizesequence.org/>) yielded 10 619 intact elements (Table 1). To compare this initial result with existing work, we studied the overlap of their chromosomal coordinates with all annotated TEs from 601 families generated by the Maize Transposable Element Consortium (MTEC) (Schnable *et al.*, 2009) (see Figure S2 and Methods S1). The analysis showed that MASiVe

Table 1 Properties of the Sirevirus families identified in the maize genome

Family ^a	FL	solo	frag	FL:solo	FL: frag	solo: frag	Avg age (Myr)	Avg length	
								FL	LTR
<i>Opie</i>	5310 +1780 ^b	2028	9826	2.6	0.5	0.2	0.90	9117	1254
<i>Ji</i>	4865 +772 ^b	2421	11 377	2.0	0.4	0.2	0.94	9519	1271
<i>Jienv</i> ^c	175 +175 ^b	103	469	1.7	0.4	0.2	0.71	12 123	1534
<i>Giepum</i> ^c	143	180	698	0.8	0.2	0.3	0.76	12 666	1469
<i>Hopie</i> ^c	74 +31 ^b	149	478	0.5	0.2	0.3	1.03	11 696	1675
<i>Dijap</i> ^d	15 +6 ^b	38	28	0.4	0.5	1.4	2.0	10 783	1525
Other	37	19	59	1.9	0.6	0.3	1.59	10 449	1279
Total	10 619	4938	22 935	2.2	0.5	0.2	0.92	9424	1273

^aFamily assignment was based on the *RT* phylogenetic analysis.

^bNumber of previously unidentified elements.

^cEnvelope-containing families.

^d*Dijap* is located within the envelope branch of the *RT* tree (Figure 2); however, it lacks the respective gene, and branches with the non-envelope families in the *INT*- and LTR-derived trees (Figures S4 and S6).

LTR, long terminal repeat; FL, full-length; solo, solo LTRs; frag, fragmented elements.

recovered 75% of the MTEC full-length elements that belong to families that overlapped with at least one MASiVE element (5674/7554), but also contributed approximately 50% new elements to the same pool (3618) (Figure S3). In fact, nearly 90% of the MTEC elements (of the aforementioned families) that MASiVE failed to match (1880) were actually reported as problematic and thus filtered out by the algorithm. These sequences are indeed of Sirevirus origin; however, they most likely do not represent intact copies. Such data could now be used to assist in TE annotation of the maize genome.

Phylogenetic analyses based on the reverse transcriptase (*RT*) (Figure 2) and integrase (*INT*) genes (Figure S4) of the MASiVE elements, of 154 maize *Copia* exemplars including 16 for *Ji* and 17 for *Opie* (<http://maizetdb.org/>), and of known Sireviruses and other (non-Sirevirus) classic *Copia* elements (Bousios *et al.*, 2010), showed that the vast majority (>99%) of MASiVE elements congregate within one major branch together with all known Sireviruses. Based on the topology of exemplars within this branch, a number of internal branches were accordingly annotated to reveal a plethora of families and subfamilies with different population and amplification characteristics. The *Opie* and *Ji* families comprise the bulk of Sireviruses with 5310 and 4865 intact elements respectively (Table 1), although both form several well-defined subfamilies (Figure 2). After searching for the presence of the envelope gene (see Methods S1), we discovered that the envelope-containing elements cluster in a single, less populated branch: within this, two distinct *Giepum* families were identified, hereafter named separately *Giepum* and *Hopie* due to the presence of

the *Hopie* exemplar (Bousios *et al.*, 2010) in the latter, and one *Ji*-related family with a single *Ji* exemplar (hereafter named *Jienv*), which was unexpected since *Ji* elements are typically devoid of an envelope gene. The different genome lengths of the Sirevirus families (Table 1) supported their classification, while these findings were consistent across all trees (Figure S5). Evidently then, with no specific *a priori* knowledge, our approach managed to identify in one step members of all known Sirevirus families, but also of previously obscured ones: *Giepum*, which is now split in two distinct families – *Giepum* and *Hopie*; *Jienv*, whose genome characteristics and phylogeny show that it is unrelated to *Ji* despite the presence of a *Ji* exemplar in the branch; and *Dijap*, a low copy number family.

Comparing the number of intact elements of the families our analysis placed in the Sirevirus genus with the respective figures of the original genome annotation (Table 2 in Baucom *et al.*, 2009a), 2764 new full-length elements are now reported, the majority of which belong to the *Opie* family (1780, Table 1). Moreover, using the estimation of Baucom *et al.* (2009a) for the proportion of the maize genome that is occupied by each TE family (including intact and fragmented elements, remnants etc.), then combined, the three most abundant *Copia* families (*Ji*, *Opie* and *Giepum*) are in fact Sireviruses that collectively take up approximately 21% of the host genome and approximately 90% of the *Copia* population. Arguably, Sireviruses is the only genus of the *Copia* superfamily that has successfully colonized and proliferated in maize.

Finally, the maize genome harbors another abundant type of Sirevirus with elements lacking the genes necessary for

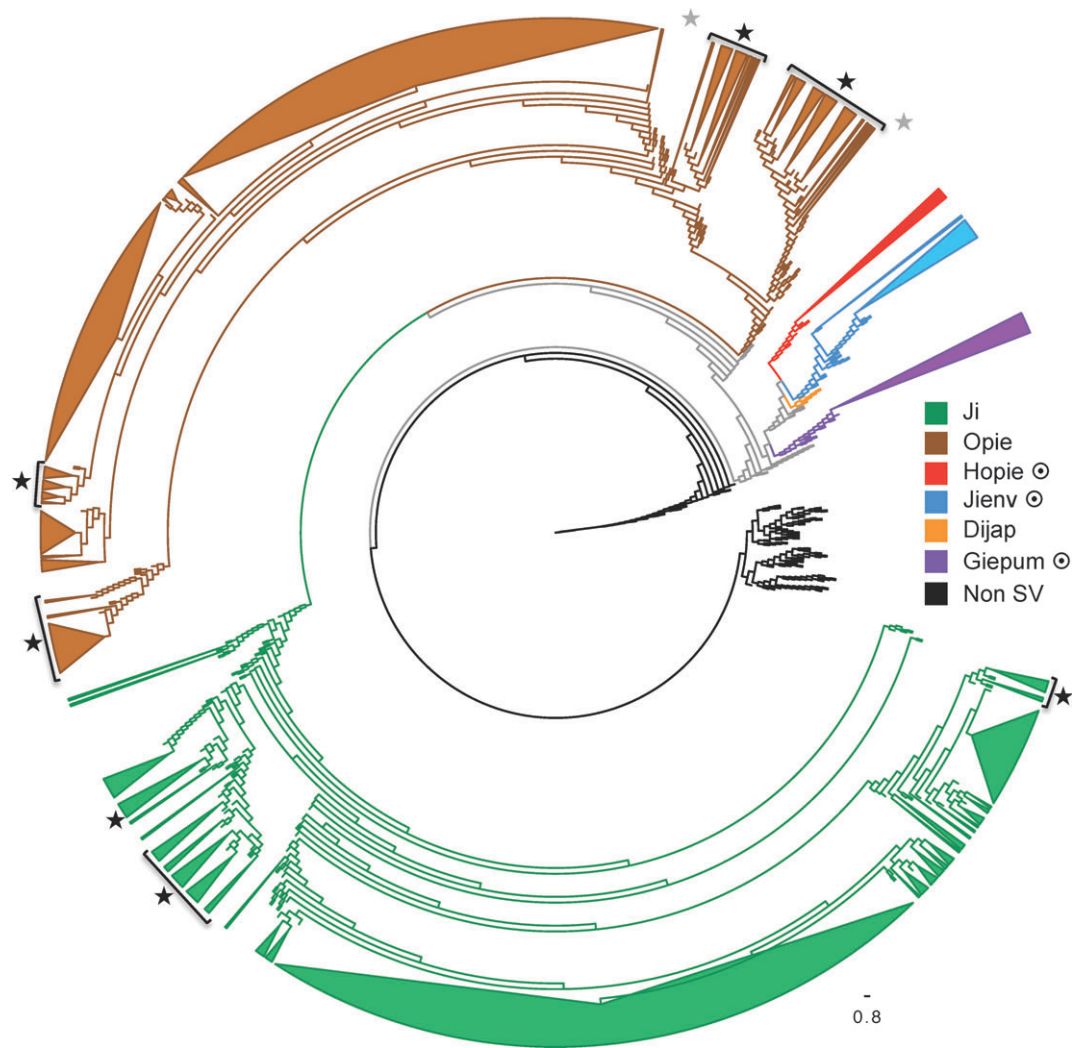


Figure 2. Tree of the full-length Sirevirus population based on the RT gene.

Opie and *Ji* comprise the majority of Sireviruses. The *Ji* subfamily branching with the other envelope-containing elements is now called the *Jienv* family (light blue). The *Giepum* family is split into two different families with distinct genome characteristics (Table 1), now called *Giepum* (purple) and *Hopie* (red), due to the presence of the *Hopie* exemplar (Bousios *et al.*, 2010) in the latter branch. The width of branches with more than 10 elements was artificially and proportionally increased (see Experimental procedures). The circle next to the name of some Sirevirus families indicates the presence of the envelope gene. Although *Dijap* is located within the envelope-containing branch, it lacks the respective gene (see Table 1). Black stars highlight abundant branches that do not contain any of the *Ji* or *Opie* exemplars from the maize TE database (<http://maizetdb.org/>). In contrast, grey stars represent *Ji* or *Opie* exemplars located in branches with only few Sireviruses; SV, Sireviruses.

transposition. We discovered that the three exemplars of the *Ruda* family (<http://maizetdb.org/>) contain the distinctive Sirevirus genome structure (Figure 1b), but completely lack a coding domain, resembling the structure of a specific TE type called 'large retrotransposon derivatives' (LARDs) (Kalendar *et al.*, 2004). *Ruda* consists of 568 intact elements according to Baucom *et al.* (2009a), and relates to the *Opie* family based on our analysis (Figure S6), although, as evident by its RLX (retrotransposon like unknown) prefix, the lack of genes had not allowed its classification to either *Copia* or *Gypsy* until now. This highlights the sensitivity of

the structural-based approach of MASiVE in such cases, and allows us to speculate that, by retaining the Sirevirus *cis*-signals and conserved motifs, these elements managed to amplify successfully in a non-autonomous way using the functions of their autonomous counterparts in *trans*.

Distinctive patterns of birth and death of Sireviruses shape the maize genome

Sequence divergence between the LTRs of each element and the application of a published formula for dating retrotransposon insertions (Ma and Bennetzen, 2004) provided

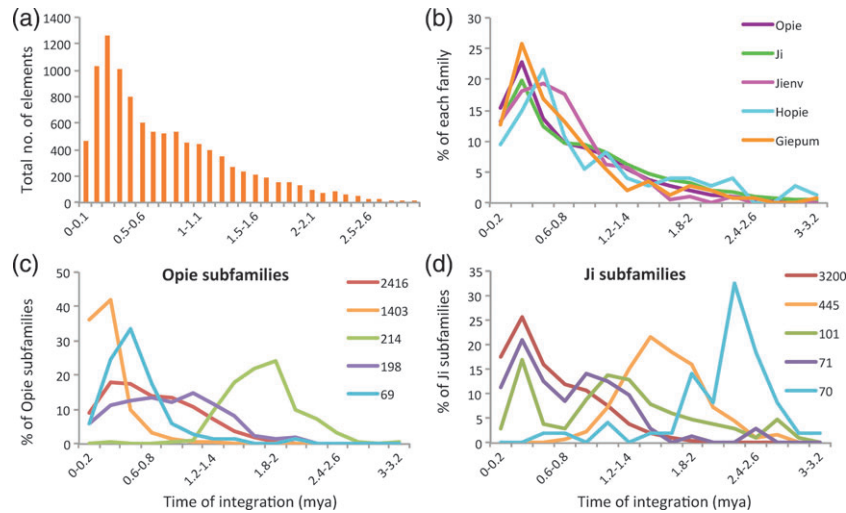


Figure 3. Sirevirus time of integration during the last 3 Myr.

(a) For all maize Sireviruses (in 0.1-Myr bins).

(b) For each Sirevirus family (in 0.2-Myr bins).

(c, d) As (b) but for the five most abundant *Opie* and *Ji* subfamilies (represented and color-coded by the number of their members) respectively. Family and subfamily assignment was derived from the RT phylogenetic analysis.

insights into the lifetime of Sireviruses in the maize genome. The ancestry of the lineage in maize dates back at least approximately 11 Myr, around the time when the maize progenitors diverged from the common ancestor of sorghum, with 95% of transpositions occurring during the last 3 Myr (Figure 3a), hence shaping the genome mainly after the allotetraploidization event approximately 5 million years ago (Mya) (Swigonova *et al.*, 2004). The Sirevirus age distribution is the outcome of two opposing forces, birth through transposition and death by mutations followed by deletion from the genome. The constant increase in the number of elements younger than 3 Myr from all Sirevirus families (Figure 3b) represents the dynamic balance between these two mechanisms. The observed accumulation of many new copies the last 600 000 years (Figure 3b) especially for most *Opie* and *Ji* subfamilies (Figures 3c,d), possibly corresponds to an intense amplification phase experienced by Sireviruses. However, as it is difficult to separately assess the efficiencies of the birth and death processes, the current age distribution may also be the result of dysfunctional removal mechanisms during the same time period, or even the outcome of a relatively constant birth and death ratio. Nearly 500 new insertions have taken place in the past 100 000 years (Figure 3a), of which most belong in the second most abundant *Opie* subfamily (Figure 3c). It is very likely that these elements have been mobile during the domestication of maize from its wild grass progenitor *teosinte*, a process that started 10 000 years ago in Mexico (Doebley *et al.*, 2006). Intriguingly, apart from the in sync amplification pattern of all families, some smaller *Opie* and *Ji* subfamilies underwent distinct and rapid proliferation

bursts further back in evolutionary time (Figures 3c,d and S7). One possible explanation for these activity differences in elements that ultimately belong in the same Sirevirus family may be a subtle, yet crucial, variation in their *cis*-regulatory loci that ultimately respond to diverse external stimuli.

Sireviruses impose their evolutionary force on the maize genome not only by their transposition but also by their disintegration, the combined result of which has directly affected the current genome size and organization of their host. In total, 4938 solo LTRs and 22 935 fragmented (with one LTR) Sireviruses were identified (Table 1), excluding severely truncated elements and remnants. The results indicate that solo LTR formation of the *Opie* and *Ji* families is considerably more frequent than earlier calculations based on smaller sections of the maize genome have suggested (SanMiguel *et al.*, 1996; Liu *et al.*, 2007; Kronmiller and Wise, 2008), with a current estimate of approximately 2.2 (compared with approximately 15.0) intact elements for every solo LTR. Furthermore, *Giepum* and *Hopie* have more solo LTRs than full-length copies. Overall, we uncovered significant differences in the full-length to solo LTR ratio across Sirevirus families, and particularly within *Opie* or *Ji* subfamilies, as well as in the ratio of their full-length to fragmented elements (Tables 1 and S1). In general, solo LTR abundance is concomitant with high numbers of fragmented copies and *vice versa*, as reflected by the same ratio of solo LTRs to fragmented elements across all (sub) families.

Considering this high intensity of solo LTR formation, which critically may also delete the intervening genomic

sequence between two highly similar elements (Bennetzen, 2002), and that Sireviruses (as part of the maize *Copia* population) reside in gene-rich areas, it is possible that their removal has aided the extensive (>50%) elimination of gene redundancy following the hybridization of the two maize progenitors approximately 5 Mya (Ilic *et al.*, 2003; Lai *et al.*, 2004).

Sireviruses reside and accumulate in gene-rich chromosomal areas of maize

It has been previously shown that the *Copia* distribution in maize is biased towards euchromatic regions of the chromosome arms, contrasting with the *Gypsy* abundance in pericentromeric heterochromatin (Baucom *et al.*, 2009a; Schnable *et al.*, 2009). The herein observed non-random preference of Sireviruses for gene-rich areas (Figures 4a–d and S8, and Methods S1), strongly suggests that the *Copia* distribution corresponds well to the distribution of Sireviruses. It also implies that throughout their life cycle Sireviruses have been critical players in the current organization, spacing and make-up of gene islands. In general, LTR retrotransposons are excluded from inserting within or very close to genes, due to the detrimental effects on host fitness. This pattern was observed in the distribution of Sireviruses, which avoid the proximal-to-genes 2 kb region (binomial test, P -value < 0.001 for intact and fragmented elements, and 0.021 for solo LTRs) (Figure S9). By contrast, however, Sireviruses do tend to locate closer to genes than expected by a random distribution hypothesis (30.7–52.9 kb average distance to the nearest gene, respectively; randomization test on the chromosomal location of Sireviruses, P -value < 0.001), perhaps due to their inability to successfully penetrate pericentromeres, or their genuine preference for gene-rich areas. Sporadically, during the transposition events some Sireviruses have landed in pericentromeric regions. These elements are significantly older than average (Pearson correlation coefficient -0.80 , P -value < 0.001) (Figure S10), possibly as a result of the overall slower rate of TE removal from these gene- and recombination-poor areas (Ma and Bennetzen, 2006). By contrast, younger elements are more abundant at the chromosome arms.

The million years of co-evolution of Sireviruses and the maize genome can be traced along the chromosomes. There are chromosomal niches that contain different relative proportions of intact, solo LTRs or fragmented Sireviruses (Figure 4e). An interesting observation is that an area on the arm of chromosome 4 appears to concomitantly lack genes and any Sirevirus feature (black star in Figure 4), suggesting that the local chromatin structure may not have been a favorable environment for either. Finally, we managed to capture the footprints of elements only very recently transposed (i.e. Sireviruses with >98% full-length identity), thereby uncovering the most recent activity hotspots on the genome, such as a section of chromosome 2 (Figure 4g).

Reconstituted Sirevirus pre-integration sites reveal a complementary mirror consensus target sequence

Sequence analysis of the narrow genomic neighborhood of Sireviruses indicated that the vast majority of intact elements (approximately 90%) generated 5-bp long target site duplications (TSDs) upon their integration into new chromosomal locations. This information was used to reconstitute and study the nucleotide composition of the Sirevirus pre-integration sites. Although specific nucleotide motifs were not detected even with advanced pattern discovery algorithms, a complementary mirror compositional pattern is evident, with the mirror being right in the middle of the TSD (Figure 5a). This palindromic pattern continues for a few bases in the flanking domain around the TSD, resulting in a clearly AT-rich sequence context (Figure 5d). More specifically, the nucleotide composition within and at the vicinity of the TSD differs significantly from the composition of the surrounding 50 bp of each side (chi-square tests, P -values for all positions < 0.001). Apart from the consensus target sequence for all Sirevirus families, *Opie* and *Ji* exhibit distinct nucleotide biases (Figures 5b,c). Figure 5e directly highlights such differences, the most apparent of which are, interestingly, located in the flanking domain outside the TSD: the pre-integration site of *Ji* is more AT-rich at the outer positions -3 , -2 , 2 , 3 with GC-rich internal positions -1 and 1 , while the *Opie* pre-integration site shows a GC-rich composition at positions -3 , -2 , 2 , 3 .

Sireviruses have preferred integration zones within TE genomes

Maize mostly comprises LTR retrotransposons (>75% of the whole genome), the abundant families of which primarily accumulate inside other LTR retrotransposons (Baucom *et al.*, 2009a). Consequently, Sireviruses are expected to mainly target other TEs including their own genomes, although it is not known whether they have any preferred integration zones. We investigated this by comparing the upstream and downstream genomic sequences of each Sirevirus with a combined sequence dataset of our Sireviruses and the maize TE database. Overall, 6032 elements were found to have inserted within the genomes of various TEs, the majority of which (approximately 70%) belong to the four most abundant LTR retrotransposons, i.e. the *Copia* *Ji* and *Opie* Sireviruses, and the *Gypsy* *Huck* and (to a lesser extent) *Cinful zeon* families. Although the general distribution of *Gypsy* elements is located in pericentromeric regions, *Huck* elements partially exhibit an *Opie/Ji*-like behavior (Baucom *et al.*, 2009a), which explains why they are often being targeted by Sireviruses for landing. For another 4038 intact Sireviruses it was not possible to assign a specific TE family to their integration site, nevertheless they all appear to reside in mosaic TE-rich domains.

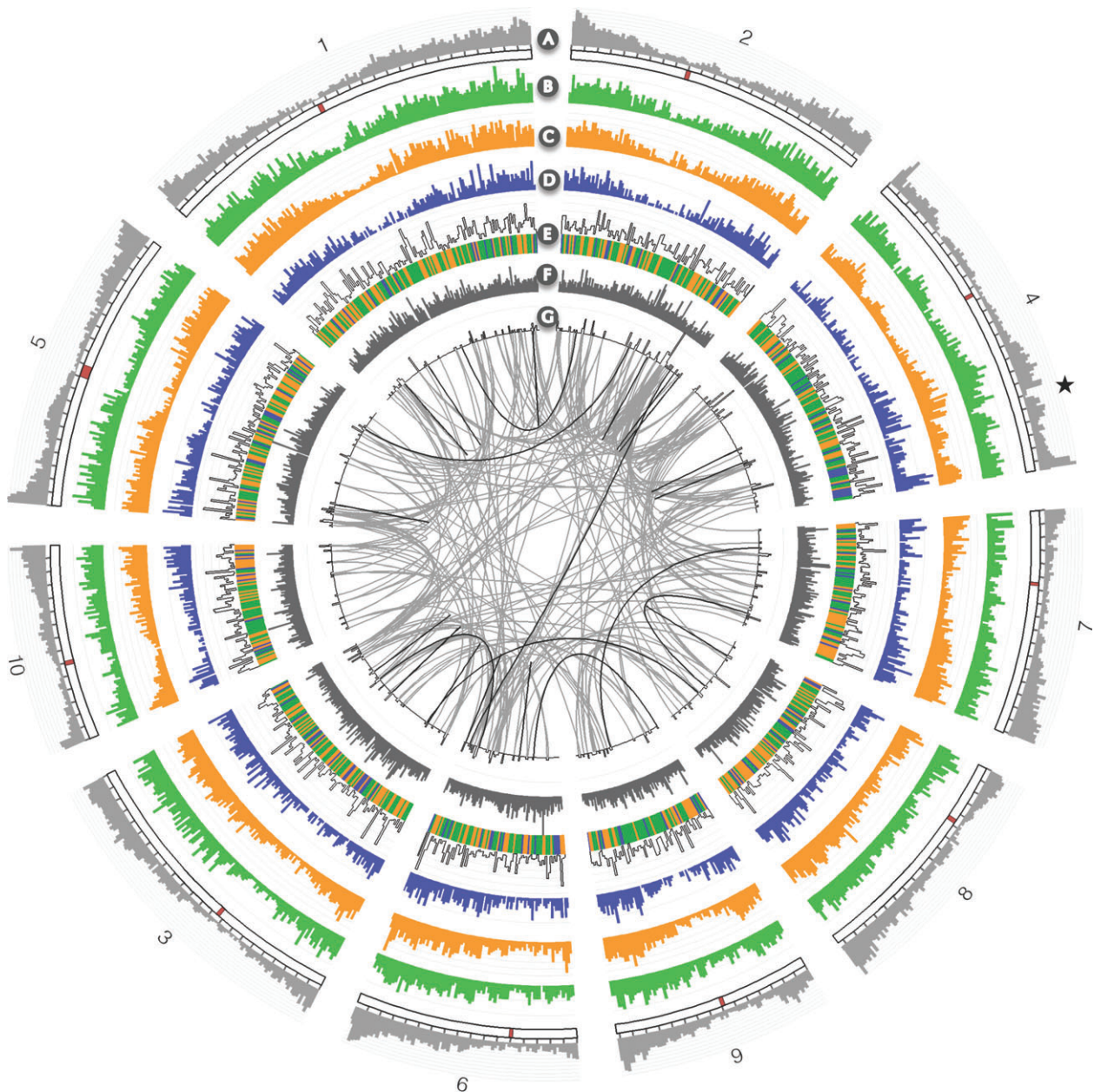


Figure 4. The maize B73 reference genome and the distribution, abundance and dynamics of the Sirevirus genus. Concentric circles show different aspects of the genome related to full-length Sireviruses and their deletion derivatives, using frequency histograms and a 2 Mb non-overlapping window.

(a) Chromosomes (with centromeres shown in red) and maize genes (grey), (b) intact Sireviruses (green), (c) fragmented elements (orange), (d) solo LTRs (blue), (e) most abundant of the (b–d) data (color-coded) after normalizing the frequencies with the corresponding maximum frequency of each Sirevirus type, and difference from the second most abundant shown as a histogram, (f) age, (g) recent Sirevirus mobility (grey lines; black lines indicate duplication of larger chromosomal segments that contain the Sirevirus element) and frequency histogram. The black star on chromosome 4 indicates the domain that appears to concomitantly lack genes and any Sirevirus feature. Visualized with Circos (Krzywinski *et al.*, 2009) (see Methods S1).

Previous analyses of maize LTR retrotransposons have shown a fivefold bias towards inserting within LTRs (Bennetzen, 2000). The results herein indicate that Sireviruses insert neither uniformly across the target genomes nor only within the LTRs, but have various preferred zones (Figure S11). In particular, they tend to target the internal

genome of *Opie* elements, the 5' LTR/internal domain junctions of *Ji* and *Cinful zeon* and the 3' LTR/internal domain junction of *Huck*. Also, there is a distinct lack of Sireviruses inserting within their own 3' LTR and its junction with the internal genome (chi-square test for uniformly distributed integration sites in the LTR retrotransposon

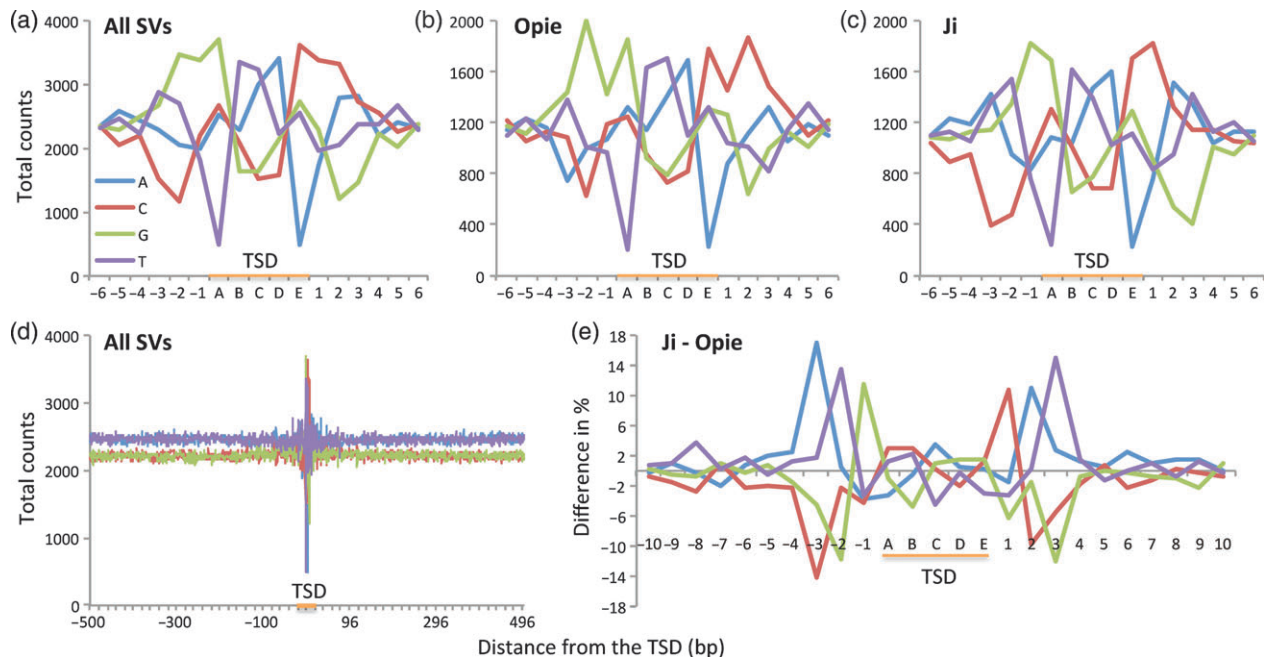


Figure 5. Nucleotide composition of Sirevirus pre-integration sites.

The target site duplication (TSD) pentamer is highlighted in orange and indicated in letters with the midpoint marked with the letter C: (a) 9370 Sirevirus elements, (b) 4641 *Opie* and (c) 4337 *Ji* elements, (d) similar to (a) with the flanking domains on each side of the TSD extending for 500 bp, (e) difference in per cent between the nucleotide composition of the pre-integration sites of *Ji* and *Opie* elements with peaks in the positive y-axis (top) pointing to *Ji*-specific bases, and in the negative y-axis (bottom) *Opie*-specific ones; SVs, Sireviruses.

genomes, *P*-values < 0.001 for *Huck*, *Ji* and *Opie* and 0.007 for *Cinful zeon*.

Sirevirus LTRs are hotspots of DNA methylation

The histone modifications and DNA methylation profiles of maize have been the focus of recent investigations (Zhang *et al.*, 2008; Liu *et al.*, 2009; Li *et al.*, 2010), offering insights into the spatial and temporal chromosomal landscapes of its dynamic epigenome. To investigate their interplay with Sireviruses, we obtained the epigenetic datasets for maize shoots and roots produced by the Deng lab (Wang *et al.*, 2009), including three modifications associated with open chromatin structure (H3K36me3, H3K9ac, H3K4me3) and two repressive marks (H3K27me3 and DNA methylation). The results of our coordinate overlap analysis show that the distributions of the three open chromatin marks on the Sirevirus genome and flanking sequences differ significantly compared with that of DNA methylation (Figure 6), whilst no clear trend was found for the other repressive mark, H3K27me3. The distribution profiles do not depend on Sirevirus phylogeny nor behave in a tissue-specific manner.

Generally, it appears that the Sirevirus LTRs and their immediate flanking sequences are highly methylated (binomial tests, all *P*-values < 0.001) and hence may retain a condensed chromatin structure obstructing their transcription. Although the long DNA methylation reads (average of 177 bp) do not allow identification of specific LTR domains

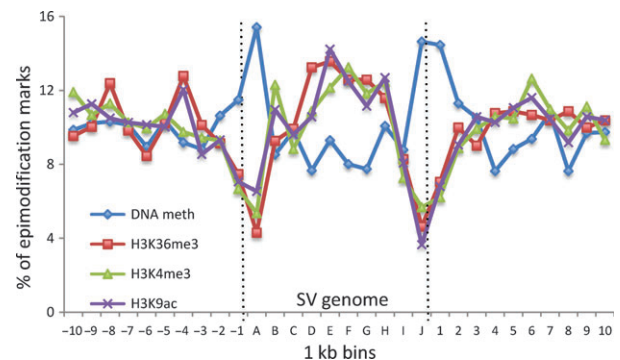


Figure 6. Distribution profile of epigenetic modifications on Sireviruses and their flanking regions.

The x-axis represents the Sirevirus genome (average length approximately 10 kb) and 10 kb of up- and downstream regions, in 10 1-kb bins. The y-axis shows the frequencies of the epigenetic marks in relation to their total number. The vertical dashed lines define the Sirevirus genome and the borders with the flanking domains; SV, Sirevirus.

(if any) that are preferentially methylated, we hypothesize that the highly conserved *cis*-regulatory RMs (Figure 1b), exactly due to their putative functional role (Bousios *et al.*, 2010), may be the targets of DNA methylation. In contrast, the internal genome of Sireviruses remains hypomethylated compared with the LTRs, and enriched in marks indicative of activate chromatin. Although one would expect the host to

exert the control on retrotransposons through their compact regulatory areas, i.e. their LTRs, which are therefore highly methylated, it is not clear why their internal gene-containing section is marked by signals that promote an open chromatin structure.

DISCUSSION

The tree of life of maize Sireviruses: an ancient and complex history

Analyzing the tens of thousands of Sireviruses in maize has revealed a plethora of diverse families and subfamilies with distinct characteristics, some placed inside the genus for the first time, and which have all been particularly active within the last 3 Myr, colonizing the distal parts of the chromosomes. Of these, *Opie* and *Ji* reached extremely high copy numbers, while the envelope-containing populations (*Jienv*, *Giepum* and *Hopie*) were not as efficient in infiltrating the genome. The non-autonomous *Ruda* proliferated successfully despite lacking coding capacity, while the origin of the *Jienv* family, how/when it acquired the envelope gene, and if it is indeed related to the *Ji* family are intriguing questions that may provide novel insights into the evolutionary history of retrotransposons in maize. In parallel to their high insertional activity, the disintegration rate of Sireviruses is intense; in fact, it may be much more intense than our current ratio estimates of approximately 2.2 intact elements for every solo LTR, since many cases of partially deleted LTRs were probably missed. Nevertheless, based only on these solo LTRs, approximately 50 Mb has been deleted from the maize genome, albeit this number is probably a serious underestimate.

Overall, a high-resolution picture of the chromosomal distribution, mobility and life cycle of Sireviruses in the maize genome has been obtained. Sireviruses avoid pericentromeres and prefer to reside near gene-rich areas, spacing them into gene islands by inserting continuously within themselves. To a large extent, Sireviruses keep, or are kept, a safe distance from genes avoiding physical interactions. Simultaneously, they are also being recycled by host mechanisms, which have possibly removed flanking sequences and genes as a by-product.

Sireviruses is an ancient retrotransposon genus that has co-evolved with its plant hosts at least since the split of monocots and eudicots (Bousios *et al.*, 2010), although the extent of the infiltration in each plant lineage is not yet known. Based on the results presented here, where Sireviruses comprise the vast majority of *Copia* elements and a fifth of the maize genome, and on the Sirevirus origin (Bousios *et al.*, 2010) of abundant elements in wheat, barley and rice (McCarthy *et al.*, 2002; Wicker and Keller, 2007), it is anticipated that Sireviruses have successfully colonized and amplified within grass genomes in general. To truly appreciate the evolution, diversity and impact of Sireviruses on

their diverse hosts, whole-genome comparative analyses are needed. Especially within the *Zea* lineage where Sireviruses have massively proliferated after the allotetraploidization event approximately 5 Mya (Swigonova *et al.*, 2004), analyses of maize haplotypes and the impending whole genomes of the Mo17 inbred line and the wild relative *Palomero toluqueno* landrace will shed light into some of the mechanisms that have promoted the 'a-maize-ing' genomic variation (Fu and Dooner, 2002; Wang and Dooner, 2006; Mackay, 2009).

Temporal and spatial patterns of Sirevirus colonization of maize chromosomes

Sireviruses have been actively colonizing the maize genome for at least the past 3 Myr. The large number of elements younger than 600 000 years (Figure 3a,b) suggests a period of intense transposition for all families and/or a decrease in the rate of Sirevirus removal by host mechanisms, or even a reasonably steady balance of both. Apart from some smaller bursts further back in time, this in sync amplification pattern of all families of the Sirevirus genus differs from previous analyses that showed multiple activity peaks spanning the last 3–4 Myr to have led to the accumulation of LTR retrotransposons in maize (Liu *et al.*, 2007; Kronmiller and Wise, 2008) and other grasses (Wicker and Keller, 2007; Paterson *et al.*, 2009; Choulet *et al.*, 2010). It will be interesting to analyze whether the total populations of the highly abundant *Gypsy* families, especially *Huck* and *Cinful zeon*, exhibited a temporally similar pattern of activity. If so, this universal accumulation of maize LTR retrotransposons would suggest that silencing-based control or indeed retrotransposon removal mechanisms were hampered for the past 0.6 Myr. Irrespective of this, based on the current paucity of transpositionally active plant (and maize) LTR retrotransposons, it appears that the host repressing mechanisms have been recently restored (Feschotte *et al.*, 2002). Hence, the second most abundant *Opie* subfamily may be among the few that have managed to escape silencing, as evident by the high frequency of its very young (<100 000 years old) members.

In contrast to such findings in maize, it was recently shown that the wheat genome was shaped by LTR retrotransposons through a plethora of amplification waves for 3 Myr, a burst 1.4 Mya and a mainly pericentromeric activity for the past 0.5 Myr (Choulet *et al.*, 2010). During the same time Sireviruses in maize have been inserting in gene-rich regions (Figure 4). Additionally, insertion age estimation of the *Osr8* and *Osr10* Sireviruses in rice placed their expansion during a long period approximately 0.5–3 Mya with the center of activity being >1 Mya (Wicker and Keller, 2007). Therefore, it appears that intensive Sirevirus colonization of their plant hosts has not been limited to the time period that corresponds to their proliferation in maize, but has probably occurred at different times in different hosts.

Large-scale analyses in the complete genomes of rice (The International Rice Genome Sequencing Project, 2005), brachypodium (The International Brachypodium Initiative, 2010), sorghum (Paterson *et al.*, 2009), soybean (Schmutz *et al.*, 2010) and Arabidopsis (Peterson-Burch *et al.*, 2004), have shown that LTR retrotransposons of both *Gypsy* and *Copia* superfamilies, and especially their high copy number members, typically cluster near pericentromeric regions away from genes and subtelomeric areas. The opposite preference of Sireviruses in maize could point to distinctive interactions between the Sirevirus integration machinery and chromatin configurations. Histones form a code made up of numerous epigenetic modifications (Kouzarides, 2007), combinations of which can generate a plethora of chromatin states suitable for TE integration specificities, as has been recently demonstrated for retrotransposons in Arabidopsis (Mirouze *et al.*, 2009; Tsukahara *et al.*, 2009). The identification of empty pre-integration Sirevirus sites in maize could provide excellent case studies for investigating the spatial and temporal histone decorations before and after the insertion of a new Sirevirus element. It is not clear what features of the Sirevirus genome, if any, recognize and/or physically interact with histone epigenetic marks. Perhaps, the conserved and unique C-termini of the Sirevirus *INT* gene (Peterson-Burch and Voytas, 2002) participates in the binding process in a similar (and crucial) manner to the stress-responsive respective domain of the yeast *Ty5* element that stabilizes *INT* binding to components of heterochromatin (Dai *et al.*, 2007).

The near-genes distribution of Sireviruses may be a universal characteristic of the genus across plants. Apart from Arabidopsis, other fully sequenced species have a higher ratio of *Gypsy* to *Copia* elements (4.9:1 for rice, 3.3:1 for brachypodium, 3.7:1 for sorghum and 2.3:1 for soybean) than maize (1.6:1) (Paterson *et al.*, 2009). In conjunction with preliminary data from our lab indicating the presence of only a few Sireviruses in those small genomes, this might make it difficult to discern such specificities. On the other hand, in plants with large genomes composed primarily of LTR retrotransposons, Sireviruses may have been a major driver of this expansion and comprise a significant part of the *Copia* superfamily, as shown here for maize, and suggested for barley and wheat with the Sirevirus *Maximus* lineage (Wicker and Keller, 2007), for soybean with *Sire1* (Laten *et al.*, 2003), for sugarbeet with *Cotzilla* (Weber *et al.*, 2010), for lotus (Holligan *et al.*, 2006) and for other plant species (Havecker *et al.*, 2004).

It is also possible that the palindromic consensus target sequence (Figure 5) interacts with the Sirevirus machinery during integration. We do, however, speculate that it does not determine the wider chromosomal location, rather it assists with the successful completion of the insertion process once the new copy has been navigated to the favorable chromatin environment via the respective histone

code. Possibly then, the element is guided to specific regions within apparently AT-rich environments (Figure 5d) to bind at the most similar sequence motif to the palindrome. Recently, similar palindromic consensus target sequences have been identified for several DNA transposons, including *Mu* and *Tourist*-like elements in maize (Liu *et al.*, 2009; Zerjal *et al.*, 2009) and the *Harbinger3_DR* family in zebrafish (Kapitonov and Jurka, 2004), suggesting that such a sequence structure may indeed positively aid TE binding and integration.

Possible implications for maize genome function

New TE copies can trigger local changes in chromatin status by inserting in gene-rich areas, or place their own *cis*-acting elements in close proximity to the gene regulatory network (Feschotte, 2008), a likely scenario in maize where Sireviruses show a near-gene insertion bias. Recent evidence also showed that several maize TE families, including *Opie* and *Giepum*, still have a considerable transcriptional activity in some tissues (Vicient, 2010). Although plant LTR retrotransposons are typically silenced *in vivo* (Feschotte *et al.*, 2002), the 'leaking' transcription of some Sirevirus elements, or their inherent (retrotransposon) capacity for production of aberrant RNA transcripts (Faulkner *et al.*, 2009), may be actively shaping the functional output and effect of the maize transcriptome. Our data show that the *cis*-regulatory regions on the Sirevirus LTRs are heavily methylated (Figure 6), suggesting that many elements are indeed kept transcriptionally inactive. However, under stress conditions the dynamic DNA methylation patterns can be suddenly modified, which would significantly loosen the host control on Sireviruses. To this end, it was recently shown that many *Opie* and *Ji* subfamilies are up-regulated when the RNA-directed DNA methylation (RdDM) silencing pathway is disturbed (Jia *et al.*, 2009), which indicates the potential of Sireviruses to reactivate given the opportunity. Combined, these findings emphasize the impact that the intense Sirevirus lifestyle may have had, and still has, on maize gene diversification, control and function.

This work brings Sireviruses into the spotlight of the 'maize genomic map', revealing in great detail their impact on the organization, composition and evolution of their host genome. We hope our insights will pave the way for future research to deepen our understanding of this interaction and its effect on maize diversification, and also of the true evolutionary depth of Sireviruses and their intricacies across the plant kingdom.

EXPERIMENTAL PROCEDURES

Data

Version 1 of the *Zea mays* B73 genome sequence (RefGen_V1) was downloaded from <http://www.maizesequence.org/> in June 2010, along with evidence-based and *ab initio* predicted maize genes. Non-Sirevirus retrotransposons and other TEs were either available

in-house (Bousios *et al.*, 2010) or downloaded from the maize TE database (<http://maizetdb.org/>).

Identification and initial analysis of Sireviruses: the MASiVE pipeline

The MASiVE pipeline was applied as described in Darzentas *et al.* (2010) (see Figure S1 and Methods S1 for more information).

Phylogeny

For the estimation of phylogeny, we first created non-redundant sets of the *RT* and *INT* sequences, and of the first 200 bases of the 5' LTR of: (i) Sireviruses, (ii) exemplars from the maize TE database, and (iii) other LTR retrotransposons used in Bousios *et al.* (2010), using CD-HIT-EST from the CD-HIT package (Li and Godzik, 2006) with a 90% identity threshold (at the default word length of 8). We required alignments to cover at least 90% of the lengths of each sequence pair (–aL and –aS both at 0.9), and used the non-default, slower but more accurate strategy that places sequences in the most similar cluster and not the first one that meets the thresholds (–g 1). The final set of sequences resulted from the selection of each cluster's representative, and the removal of all sequences with more than five consecutive N characters. The procedure yielded 577 *RT*, 729 *INT* and 717 LTR representative sequences. We used the MAFFT algorithm (Kato *et al.*, 2005) to calculate pairwise distances with the Needleman–Wunsch global alignment algorithm (globalpair) and output the tree (treeout). Finally, we annotated tree leaves with the number of members behind the cluster representatives (also indicated by artificially and proportionally increasing the width of branches with more than 10 members by one leaf for each 10 members) and with the names of the exemplars and known elements. Trees were visualized with Figtree v.1.3.1 for Mac OS X (<http://tree.bio.ed.ac.uk/software/figtree/>).

Solo LTRs and fragmented elements

Efficient annotation of degraded retrotransposons is a troublesome and imperfect process (Ma *et al.*, 2004; Vitte and Bennetzen, 2006). We therefore only identified solo LTRs and fragmented Sireviruses with one intact LTR. We first masked the genomic sequence to avoid re-detecting the LTRs of full-length elements. Masking was performed by the exact matching of all (including incomplete) MASiVE elements with Vmatch (<http://www.vmatch.de/>). BLASTn was used to detect statistically very significant ($E\text{-value } 1 \times 10^{-180}$) hits of the entire LTR population. Strict alignment length and overlap criteria were employed, since we required that the BLAST alignment began and ended within 50 bp of the MASiVE LTR with the largest percentage identity. To assist with the classification of the hit as solo or fragmented, we used the existing (from the original MASiVE run) multiple PPT signatures and also discovered all instances with up to three mismatches of the PBS signature used in MASiVE. Coordinates-based linking of the BLAST hits and the two signatures provided adequate information to annotate each BLAST hit as a 3' LTR of a fragmented Sirevirus if a multiple PPT signature was present within 50 bp of the beginning of the BLAST hit or as a 5' LTR if the PBS signature was within 50 bp of the end of the BLAST hit, and as a solo LTR otherwise. Finally, we transferred the age and phylogeny of the corresponding MASiVE LTR to the new solo LTR or fragment as an informative approximation.

Time of integration

Time of integration (Mya), or LTR retrotransposon age (Myr), was calculated by aligning the LTR pair of each element using MAFFT, gathering data on mismatches and indels, and subsequently

applying the LTR retrotransposon age formula with a substitution rate of 1.3×10^{-8} mutations per site per year (Ma and Bennetzen, 2004).

Proximity of Sireviruses to maize genes

For the random distribution hypothesis we investigated the distribution of the distances between the 10 619 full-length Sireviruses and their neighboring genes. We fixed the positions of the genes and randomized the positions of the MASiVE data assuming that they are randomly distributed across each chromosome. We did not allow overlapping between the MASiVE data and genes (including their 2-kb flanking regions). We performed a simulation experiment with 1000 iterations to construct the background distribution of the mean distance of the MASiVE data to the nearest gene. Finally, we tested for significance comparing the mean distance calculated from the observed data with the background distribution obtained from the simulation.

Target site duplication analysis

To study the composition of TSDs, we captured 10 bp on either side of each element, and through preliminary sequence analysis discovered that Sireviruses generate, almost exclusively, 5-bp TSDs. Consequently, when the flanking pentamers were identical, then the sequence was considered informative and kept as a whole after removing one of the two pentamers and the Sirevirus element (thus forming a pseudo pre-integration site).

Sirevirus insertion site neighborhoods

To investigate the extent of Sirevirus integration into other TEs, we extracted 1000-bp long sequences upstream and downstream of all elements and calculated their similarities, through BLASTn and an E-value threshold of 1×10^{-4} , to the complete set of Sireviruses and TEs from the maize TE database (<http://maizetdb.org/>). Sequences from both sets that contained more than 10 consecutive N characters were removed. We excluded *Opie* and *Ji* elements from our dataset with lengths more or <1 kb of the average length of the respective family (9117 bp for *Opie*, 9519 bp for *Ji*; see Results). Parsing the BLASTn results, we demanded that any aligned regions between a 1000mer and a TE were at least 25 bp long with more than 80% identity and within 20 bp of the insertion site (all thresholds based on preliminary analysis). The proximity to the insertion site controlled for the complex nested pattern of TEs that could generate multiple hits of different TEs in different positions on the 1000mers. Furthermore, before assigning the best-scoring TE to the insertion site, we demanded to find the same TE on both sides of the insertion site, reconstituting a realistic pre-integration site with the two parts of the same element. Finally, to accommodate insertions that occurred very close to the beginning or end of a TE, we allowed for a TE to feature only in the upstream or downstream region if the insertion occurred within 50 bp of the beginning or end of that TE and the alignment length was >100 bp.

Overlap to epigenetic modifications

To study the overlap of epigenetic modification marks with our set of intact Sireviruses, we retrieved the respective datasets from the Deng lab (Wang *et al.*, 2009). These included four histone modifications marks (H3K4me3, H3K9ac, H3K27me3 and H3K36me3) and DNA methylation patterns for maize roots and shoots. Since the provided coordinates were on a bacterial artificial chromosome (BAC) collection, we used the appropriate maize B73 genome assembly file to map those to the B73 RefGen_v1 maize sequence; this process resulted in almost half of the marks being lost due to

the set of BACs ultimately used in the genome assembly. To detect coordinate overlaps we used the Perl programming language. We analyze in Methods S1 a problem in relation to the structure of LTR retrotransposons that is raised by the MAQ (Mapping and Assembly with Quality) application (Li *et al.*, 2008) to map epimodification sequencing reads.

Statistical analysis

All statistical tests were conducted using the R language (<http://www.r-project.org/>).

Availability

Links to data files, a downloadable version of the MASiVe algorithm, the Sirevirus and monocot-specific hidden Markov model (HMM) of the envelope gene, and other general information, are all freely available at <http://bat.ina.certh.gr/research/TEs/>.

ACKNOWLEDGEMENTS

We thank Dr Kostas Stamatopoulos for reading and improving the manuscript. This project was supported by the FP6 BioSapiens Network of Excellence (LSHG-CT-2003-503265).

SUPPORTING INFORMATION

The following supporting information is available for this article online:

Figure S1. The MASiVe algorithm pipeline.

Figure S2. Length distribution of *Opie* and *Ji* elements identified by MASiVe and MTEC in maize chromosome 1.

Figure S3. Venn diagram of the overlap of MASiVe-detected Sireviruses to MTEC long terminal repeat (LTR) retrotransposons.

Figure S4. Tree of the full-length Sirevirus population based on the *INT* gene.

Figure S5. Sirevirus sub/families based on the reverse transcriptase (*RT*-), integrase (*INT*-) and long terminal repeat (LTR)-derived trees.

Figure S6. Tree of the full-length Sirevirus population based on the long terminal repeat (LTR) domain.

Figure S7. Time-confined amplification bursts of Sirevirus subfamilies.

Figure S8. Spatial and temporal distribution of the most abundant *Ji* and *Opie* subfamilies in the maize chromosomes.

Figure S9. Proximity of Sireviruses to maize genes.

Figure S10. Sirevirus age and abundance in relation to the distance to the centromeres.

Figure S11. Integration preferences of Sireviruses within the genome of the four most abundant long terminal repeat (LTR) retrotransposon families in maize.

Table S1. Disintegration rates of Sirevirus subfamilies.

Methods S1. Identification and initial analysis of Sireviruses: the MASiVe pipeline; overlap of MASiVe Sireviruses with MTEC full-length elements; sequence domains in the Sirevirus genomes; overlap to epigenetic modifications; visualization with Circos.

Please note: As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

REFERENCES

- Baucom, R.S., Estill, J.C., Chaparro, C., Upshaw, N., Jogi, A., Deragon, J.M., Westerman, R.P., SanMiguel, P.J. and Bennetzen, J.L. (2009a) Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *Plos Genet.* **5**, e1000732.
- Baucom, R.S., Estill, J.C., Leebens-Mack, J. and Bennetzen, J.L. (2009b) Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome. *Genome Res.* **19**, 243–254.
- Bennetzen, J.L. (2000) Transposable element contributions to plant gene and genome evolution. *Plant Mol. Biol.* **42**, 251–269.
- Bennetzen, J.L. (2002) Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica*, **115**, 29–36.
- Boeke, J., Eickbush, T., Sandmeyer, S. and Voytas, D. (2006) *Index of Viruses – Pseudoviridae* (2006). ICTVdB – The Universal Virus Database, version 4. Columbia University, New York: USABuchen-Osmond C 2006.
- Bousios, A., Darzentas, N., Tsiftaris, A. and Pearce, S.R. (2010) Highly conserved motifs in non-coding regions of Sirevirus retrotransposons: the key for their pattern of distribution within and across plants? *Bmc Genomics*. **11**, 89.
- Brunner, S., Fengler, K., Morgante, M., Tingey, S. and Rafalski, A. (2005) Evolution of DNA sequence nonhomologies among maize inbreds. *Plant Cell*, **17**, 343–360.
- Choulet, F., Wicker, T., Rustenholz, C. *et al.* (2010) Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *Plant Cell*, **22**, 1686–1701.
- Dai, J.B., Xie, W.W., Brady, T.L., Gao, J.Q. and Voytas, D.F. (2007) Phosphorylation regulates integration of the yeast Ty5 retrotransposon into heterochromatin. *Mol. Cell*, **27**, 289–299.
- Darzentas, N., Bousios, A., Apostolidou, V. and Tsiftaris, A.S. (2010) MASiVe: mapping and analysis of SireVirus elements in plant genome sequences. *Bioinformatics*, **26**, 2452–2454.
- Doehle, J.F., Gaut, B.S. and Smith, B.D. (2006) The molecular genetics of crop domestication. *Cell*, **127**, 1309–1321.
- Faulkner, G.J., Kimura, Y., Daub, C.O. *et al.* (2009) The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.* **41**, 563–571.
- Feschotte, C. (2008) Transposable elements and the evolution of regulatory networks. *Nat. Rev. Gen.* **9**, 397–405.
- Feschotte, C., Jiang, N. and Wessler, S.R. (2002) Plant transposable elements: where genetics meets genomics. *Nat. Rev. Gen.* **3**, 329–341.
- Fu, H.H. and Dooner, H.K. (2002) Intracellular violation of genetic colinearity and its implications in maize. *Proc. Natl Acad. Sci. USA*, **99**, 9573–9578.
- Gao, X., Havecker, E.R., Baranov, P.V., Atkins, J.F. and Voytas, D.F. (2003) Translational recoding signals between gag and pol in diverse LTR retrotransposons. *RNA*, **9**, 1422–1430.
- Havecker, E.R., Gao, X. and Voytas, D.F. (2004) The diversity of LTR retrotransposons. *Genome Biol.* **5**, 225.
- Havecker, E.R., Gao, X. and Voytas, D.F. (2005) The siveviruses, a plant-specific lineage of the Ty1/copia retrotransposons, interact with a family of proteins related to dynein light chain. *Plant Physiol.* **139**, 857–868.
- Holligan, D., Zhang, X.Y., Jiang, N., Pritham, E.J. and Wessler, S.R. (2006) The transposable element landscape of the model legume *Lotus japonicus*. *Genetics*, **174**, 2215–2228.
- Ilic, K., SanMiguel, P.J. and Bennetzen, J.L. (2003) A complex history of rearrangement in an orthologous region of the maize, sorghum, and rice genomes. *Proc. Natl Acad. Sci. USA*, **100**, 12265–12270.
- Jia, Y., Lisch, D.R., Ohtsu, K., Scanlon, M.J., Nettleton, D. and Schnable, P.S. (2009) Loss of RNA-dependent RNA polymerase 2 (RDR2) function causes widespread and unexpected changes in the expression of transposons, genes, and 24-nt small RNAs. *Plos Genet.* **5**, e1000737.
- Kalendar, R., Vicient, C.M., Peleg, O., Anamthawat-Jonsson, K., Bolshoy, A. and Schulman, A.H. (2004) Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. *Genetics*, **166**, 1437–1450.
- Kapitonov, V.V. and Jurka, J. (2004) Harbinger transposons and an ancient HARBI1 gene derived from a transposase. *DNA Cell Biol.* **23**, 311–324.
- Katoh, K., Kuma, K., Toh, H. and Miyata, T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518.
- Kouzarides, T. (2007) Chromatin modifications and their function. *Cell*, **128**, 693–705.
- Kronmiller, B.A. and Wise, R.P. (2008) TEneST: automated chronological annotation and visualization of nested plant transposable elements. *Plant Physiol.* **146**, 45–59.
- Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645.

- Lai, J.S., Ma, J.X., Swigonova, Z. *et al.* (2004) Gene loss and movement in the maize genome. *Genome Res.* **14**, 1924–1931.
- Laten, H.M., Havecker, E.R., Farmer, L.M. and Voytas, D.F. (2003) SIRE1, an endogenous retrovirus family from Glycine max, is highly homogenous and evolutionarily young. *Mol. Biol. Evol.* **20**, 1222–1230.
- Li, W.Z. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
- Li, H., Ruan, J. and Durbin, R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858.
- Li, H., Freeling, M. and Lisch, D. (2010) Epigenetic reprogramming during vegetative phase change in maize. *Proc. Natl Acad. Sci. USA*, **107**, 22184–22189.
- Liu, R.Y., Vitte, C., Ma, J.X., Mahama, A.A., Dhlwayo, T., Lee, M. and Bennetzen, J.L. (2007) A GeneTrek analysis of the maize genome. *Proc. Natl Acad. Sci. USA*, **104**, 11844–11849.
- Liu, S.Z., Yeh, C.T., Ji, T.M., Ying, K., Wu, H.Y., Tang, H.M., Fu, Y., Nettleton, D. and Schnable, P.S. (2009) Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. *PLoS Genet.* **5**, e1000733.
- Ma, J.X. and Bennetzen, J.L. (2004) Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl Acad. Sci. USA*, **101**, 12404–12410.
- Ma, J.X. and Bennetzen, J.L. (2006) Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. *Proc. Natl Acad. Sci. USA*, **103**, 383–388.
- Ma, J.X., Devos, K.M. and Bennetzen, J.L. (2004) Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res.* **14**, 860–869.
- Mackay, T.F.C. (2009) A-maize-ing Diversity. *Science*, **325**, 688–689.
- McCarthy, E.M., Liu, J.D., Lizhi, G. and McDonald, J.F. (2002) Long terminal repeat retrotransposons of *Oryza sativa*. *Genome Biol.* **3**, RESEARCH0053.
- McClintock, B. (1984) The significance of responses of the genome to challenge. *Science*, **226**, 792–801.
- Mirouze, M., Reinders, J., Bucher, E. *et al.* (2009) Selective epigenetic control of retrotransposition in Arabidopsis. *Nature*, **461**, 427–U130.
- Morgante, M., Brunner, S., Pea, G., Fengler, K., Zuccolo, A. and Rafalski, A. (2005) Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat. Genet.* **37**, 997–1002.
- Paterson, A.H., Bowers, J.E., Bruggmann, R. *et al.* (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature*, **457**, 551–556.
- Peterson-Burch, B.D. and Voytas, D.F. (2002) Genes of the Pseudoviridae (Ty1/copia retrotransposons). *Mol. Biol. Evol.* **19**, 1832–1845.
- Peterson-Burch, B.D., Nettleton, D. and Voytas, D.F. (2004) Genomic neighborhoods for Arabidopsis retrotransposons: a role for targeted integration in the distribution of the Metaviridae. *Genome Biol.* **5**, R78.
- SanMiguel, P., Tikhonov, A., Jin, Y.K. *et al.* (1996) Nested retrotransposons in the intergenic regions of the maize genome. *Science*, **274**, 765–768.
- SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y. and Bennetzen, J.L. (1998) The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20**, 43–45.
- Schmutz, J., Cannon, S.B., Schlueter, J. *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.
- Schnable, P.S., Ware, D., Fulton, R.S. *et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.
- Shirasu, K., Schulman, A.H., Lahaye, T. and Schulze-Lefert, P. (2000) A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Res.* **10**, 908–915.
- Slotkin, R.K. and Martienssen, R. (2007) Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Gen.* **8**, 272–285.
- Swigonova, Z., Lai, J.S., Ma, J.X., Ramakrishna, W., Liaca, V., Bennetzen, J.L. and Messing, J. (2004) Close split of sorghum and maize genome progenitors. *Genome Res.* **14**, 1916–1923.
- The International Brachypodium Initiative (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, **463**, 763–768.
- The International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.
- Tsukahara, S., Kobayashi, A., Kawabe, A., Mathieu, O., Miura, A. and Kakutani, T. (2009) Bursts of retrotransposition reproduced in Arabidopsis. *Nature*, **461**, 423–U125.
- Vicient, C.M. (2010) Transcriptional activity of transposable elements in maize. *Bmc Genomics*, **11**, 601.
- Vitte, C. and Bennetzen, J.L. (2006) Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. *Proc. Natl Acad. Sci. USA*, **103**, 17638–17643.
- Wang, Q.H. and Dooner, H.K. (2006) Remarkable variation in maize genome structure inferred from haplotype diversity at the bz locus. *Proc. Natl Acad. Sci. USA*, **103**, 17644–17649.
- Wang, W., Zheng, H.K., Fan, C.Z. *et al.* (2006) High rate of chimeric gene origination by retroposition in plant genomes. *Plant Cell*, **18**, 1791–1802.
- Wang, X.F., Elling, A.A., Li, X.Y. *et al.* (2009) Genome-wide and organ-specific landscapes of epigenetic modifications and their relationships to mRNA and small RNA transcriptomes in maize. *Plant Cell*, **21**, 1053–1069.
- Weber, B., Wenke, T., Frommel, U., Schmidt, T. and Heitkam, T. (2010) The Ty1-copia families SALIRE and Cotzilla populating the Beta vulgaris genome show remarkable differences in abundance, chromosomal distribution, and age. *Chromosome Res.* **18**, 247–263.
- Wicker, T. and Keller, B. (2007) Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. *Genome Res.* **17**, 1072–1081.
- Wicker, T., Sabot, F., Hua-Van, A. *et al.* (2007) A unified classification system for eukaryotic transposable elements. *Nat. Rev. Gen.* **8**, 973–982.
- Yang, L.X. and Bennetzen, J.L. (2009) Distribution, diversity, evolution, and survival of Helitrons in the maize genome. *Proc. Natl Acad. Sci. USA*, **106**, 19922–19927.
- Zerjal, T., Joets, J., Alix, K., Grandbastien, M.A. and Tenaillon, M.I. (2009) Contrasting evolutionary patterns and target specificities among three Tourist-like MITE families in the maize genome. *Plant Mol. Biol.* **71**, 99–114.
- Zhang, W.L., Lee, H.R., Koo, D.H. and Jiang, J.M. (2008) Epigenetic modification of centromeric chromatin: hypomethylation of DNA sequences in the CENH3-associated chromatin in Arabidopsis thaliana and maize. *Plant Cell*, **20**, 25–34.