

Considerations and complications of mapping small RNA libraries to transposable elements

Alexandros Bousios¹, Brandon S. Gaut², Nikos Darzentas³

¹ School of Life Sciences, University of Sussex, Brighton, UK

² Department of Ecology and Evolutionary Biology, UC Irvine, Irvine, CA 92697, USA

³ Central European Institute of Technology, Masaryk University, Brno, Czech Republic

Author for Correspondence:

Alexandros Bousios; School of Life Sciences, University of Sussex, Brighton, UK;
+441273678480; alexandros.bousios@gmail.com

Running title: Mapping small RNAs to Transposable Elements

Keywords: transposable elements, small RNA, high-throughput sequencing, siRNA, RNA-seq, genome mapping, annotation, bioinformatics

ABSTRACT

The advent of high-throughput sequencing (HTS) has revolutionized the way in which epigenetic research is conducted. Often coupled with the availability of fully sequenced genomes, millions of small RNA (sRNA) reads are mapped to regions of interest and the results scrutinized for clues about epigenetic mechanisms. However, this approach requires careful consideration in regards to experimental design, especially when one investigates repetitive parts of genomes such as transposable elements (TEs), and especially when such genomes are large as is often the case in plants. Here, to shed light on the challenges of mapping sRNAs to TEs, we focus on the 2,300Mb maize genome, of which >85% is derived from TEs. We compare various methodological strategies that are commonly employed in TE studies. These include choices for the reference dataset, the normalization of multiple mapping sRNAs, and the selection among different types of sRNA metrics. We further examine how these choices influence the relationship between sRNAs and the critical feature of TE age, and explore and contrast their effect on low copy regions (exons) and other popular HTS data (RNA-seq). Finally, based on our analysis, we share a series of take-home messages to help guide TE epigenetic studies specifically, but our conclusions may also apply to any work that involves mapping and analysis of HTS data.

INTRODUCTION

Across eukaryotes, epigenetic pathways contribute to diverse functions, including gene regulation and transposable element (TE) silencing (1). Small RNAs (sRNAs) are a key component of these pathways. Numerous studies have investigated the biogenesis and functional roles of sRNAs, with most focusing on the molecular mechanisms that underlie these processes (for recent reviews see (2-4)). Some of these studies have utilized high-throughput sequencing (HTS) technologies, which generate vast numbers of sRNA reads. This capacity of HTS has facilitated the identification of novel sRNA classes, the quantification and comparison of sRNA expression profiles across tissues, and the discovery of genomic loci that map large volumes of sRNAs. These tasks have been aided by numerous computational tools, most of which have been tailored to study micro RNAs (miRNAs) (5-11), with fewer offering comprehensive identification, quantification and visual-based support for all sRNA types (12-17).

Even with these tools, significant challenges remain in the handling and interpretation of HTS sRNA data. One major challenge concerns the sequencing depth of libraries during differential expression analysis across tissues; this topic, however, has been thoroughly investigated (e.g., (6,15,18)) and will not be discussed further.

Another important challenge, which we will address here, stems from the fact that some sRNAs map to unique locations (U_sRNAs) of a reference genome, while others align equally well to multiple locations (M_sRNAs). The handling of the latter is a major concern, as it impacts downstream analyses (15), and is as yet practically unresolved as

different studies (reviewed in Johnson et al. 2016) and related sRNA analysis tools have used different approaches. For example, the NiBLS method allows multiple mapping without any kind of normalization for the number of mapping locations (19), the SiLoCo tool of the UEA sRNA Toolkit weights each read by its repetitiveness in the genome (20), the segmentSeq package of Bioconductor allocates each M_sRNA only once to a predefined locus even if it maps to more than one places within this locus or indeed across the genome (13), Novoalign (<http://www.novocraft.com>) excludes M_sRNAs, and bowtie (21) and bwa (22) randomly place each M_sRNA to a single locus under their default settings (22). Finally, a recently updated version of ShortStack allocates M_sRNAs to single loci based on the densities of U_sRNAs (23).

The importance of M_sRNAs and of their handling may be dependent on the component of the genome under investigation; for instance, due to their repetitive nature, TEs are likely to map many M_sRNAs, which unavoidably complicates TE-related studies. This effect may be especially prominent in plants, because of their large genomes (the average size of a diploid angiosperm is ~6,400Mb) and the fact that most of plant DNA has originated from TEs (24). This point is exemplified by contrasting data from the unusually small genome of *Arabidopsis thaliana* (which is only 125Mb and contains ~24% TE-derived DNA) and the larger - but still small, relative to the angiosperm average - genome of maize (2,300MB, ~85%). sRNA mapping studies have shown that <25% of *A. thaliana* TEs are mapped solely by M_sRNAs (25), but this proportion raises to >72% for maize TEs (26). Hence, careful consideration of M_sRNAs is crucial for understanding epigenetic processes in genomes like maize. The challenges of mapping sRNAs to TEs are further exacerbated by the fact that accurate TE identification is a notoriously difficult task (27). To simplify the problem, previous studies have often used TE exemplars (28-30), which are a consensus of many TE sequences representing a single TE family or subfamily. The use of exemplars may be pragmatic, but it likely reduces the analysis resolution compared to examining whole populations of annotated TEs.

Here we attempt to systematically address the complex, but understudied, issue of analyzing sRNAs in the context of TEs, because the impact of their treatment on analyses is presently unclear. To better assess different approaches, we focus on the maize genome and the most abundant *Copia* and *Gypsy* Long Terminal Repeat (LTR) retrotransposon families. We perform standard sRNA mapping using existing HTS data from three different tissues, but vary several features of the analyses, such as i) the reference dataset, which ranges from whole genome TE annotations to TE exemplars, ii) the treatment of M_sRNAs, which ranges from various normalization options to their complete exclusion, and iii) the sRNA metrics, i.e. consideration of distinct sequences or their abundances. Figure 1 depicts the methodological matrix of our work, along with many of the terms that we use throughout the study. We then comment on the effect of some of these choices on the relationship of mapping with other TE features such as TE age, with low

copy regions of the maize genome, or when using HTS mRNA data. We conclude by sharing our insights as take-home messages to help guide researchers in epigenetic analyses of TEs or mapping of HTS data to large and complex genomes.

		maize genome		TEdb exemplars
TE datasets	Copia	Ji	3,285 [6,283]	15
		Opie	2,926 [6,881]	17
		Giepum	102 [221]	1
	Gypsy	Huck	2,460	20
		Cinful-zeon	6,276	41
		Flip	483	7
sRNAs		tissue	leaf tassel ear	
		length	21nt 22nt 24nt	
		metric	species expression	
		location	unique (U) multiple (M)	
normalization		un-weighted weight by...		
		genome		TEdb
		...locations		
mapping		species species/nt expression/nt		
locations for M_sRNAs mapping to...		our TEs* exons other M_sRNAs		our TEs* *across TEdb
		*across genome & within annotated TEs		

Figure 1. A matrix of the terms, data and analyses used in this study.

RESULTS & DISCUSSION

Reference datasets: TE exemplars vs. annotated TE populations

It is common practice in TE epigenetic studies to examine only full-length TEs. But how do inferences vary as a function of the reference dataset? In this study, we compared mapping patterns of 21, 22 and 24 nucleotide (nt) sRNAs of leaf, tassel and ear tissues between two TE datasets. The first was a comprehensively annotated dataset of Sirevirus LTR retrotransposons of the *Copia* superfamily in maize (31). Sireviruses encompass the three most abundant *Copia* families in maize, namely *Ji*, *Opie* and *Giepum*. *Ji* and *Opie* each constitute ~10% of the genome, and *Giepum* represents another ~1.2% (32,33). We used 3,285 *Ji*, 2,926 *Opie* and 102 *Giepum* full-length elements that were recently analyzed for their epigenetic patterns (26) (Figure 1). To add to this dataset, we devised a pipeline (see Methods) to identify full-length elements of the three most abundant *Gypsy* families, including *Huck* (10.1% of the genome), *Cinful-zeon* (8.2%) and *Flip* (4.2%) (32) (Figure 1). We annotated 2,460, 6,276 and 483 copies respectively. The second reference dataset consisted of exemplar TE sequences. We downloaded all maize TE exemplars from <http://maizetedb.org>. The number of exemplars for the six *Copia* and *Gypsy* families ranged from one to 41 consensus sequences (Figure 1).

Exemplars miss a substantial fraction of sRNA information

We began by first examining the total number of distinct sRNA sequences (termed “sRNA species” hereafter) mapped to each family. An initial observation was that there is a much lower number of sRNAs (3-fold decrease on average) that mapped to the exemplars compared to the annotated populations (Figure 2A, Table S1). For example, 90,503 sRNA species of the leaf library mapped to the exemplars of all six families combined, compared to 310,548 that mapped to the annotated elements. This result indicates that a substantial fraction of sRNA information is lost when using exemplars. Nonetheless, the majority of sRNAs that mapped to exemplars do relate to the six families, because 89% (162,493 of 181,188 for all libraries and families combined) also mapped to the annotated TE populations.

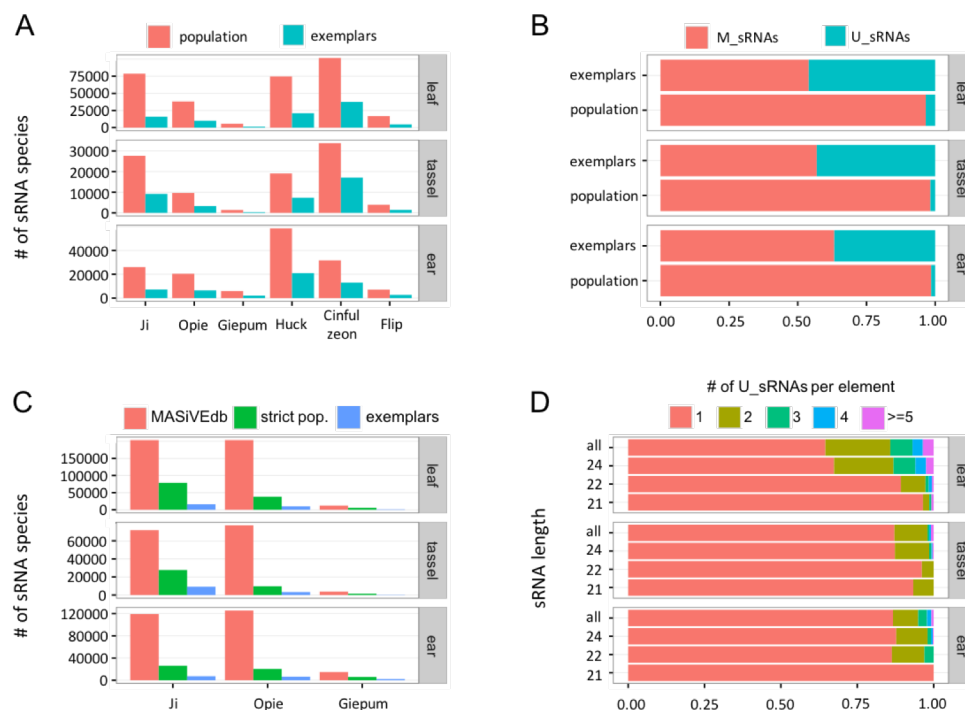


Figure 2. sRNA metrics on TE exemplars and annotated TE populations. (A) Total number of sRNA species that mapped to each family. (B) Proportion of U_sRNAs and M_sRNAs for all families combined. (C) Total number of sRNA species that mapped to different datasets of the three *Copia* families. (D) Proportion of the number of U_sRNAs that mapped per TE.

U_sRNA and M_sRNA ratios differ between datasets

Previous research has suggested that U_sRNAs may exert a stronger effect on TE silencing compared to M_sRNAs (25,34). Accordingly, several studies have used only U_sRNAs as the basis for inference, derived either from mapping to genomes or to exemplars (29,30,35,36). Our analysis showed that there is a massive difference in the U:M sRNA ratio as a function of the reference dataset: a much higher proportion of

sRNAs map uniquely to exemplars (43% of all sRNAs for all libraries and families combined) compared to annotated TE populations (2.6%) (Figure 2B, Table S2). In fact, the vast majority of U_sRNAs that map to exemplars become M_sRNAs when mapped to the genome. This finding raises two important points. First, U_sRNAs appear to be overrepresented in the analysis of exemplar datasets and hence their use over M_sRNAs (e.g., (29,30)) should be carefully considered. Second, U_sRNAs may convey little information at the genome level for TE studies, because there may be only a few mapping to each element, at least in TE-rich and large genomes like that of maize.

Low levels of sRNA ‘cross-talk’ between families are consistent between datasets

The extent to which sRNAs can map to elements of multiple TE families is largely unknown. Yet, this epigenetic ‘cross-talk’ may be important for host defenses, by spreading TE silencing through homology-based mechanisms (37,38). Our analysis of both TE exemplars and annotated TE populations was consistent in showing that sRNAs are generally family specific. For all libraries combined, only 0.2% of 181,188 exemplar-mapping and 1.7% of 550,755 population-mapping sRNA species cross-talked between families. For completeness, we also investigated the degree of sRNA cross-talk between TEs and exons from the maize Filtered Gene Set. The low levels of cross-mapping (<0.2% for each library) imply that exon-mapping sRNAs (545,601 sRNAs for all libraries combined) serve biological roles unrelated to TE silencing.

sRNA patterns along TE sequences differ between datasets

We then examined the mapping characteristics of each TE dataset by averaging sRNA targeting per nucleotide of each full-length element, so as to allow comparisons among TEs. For this analysis, we used sRNA species but also the metric of “sRNA expression” that indicates the number of reads in a library for each sRNA species. In addition, we did not normalize M_sRNAs by their number of mapping locations. (We compare weighted and un-weighted data further below). Our analysis showed that, despite the low number of sRNAs that mapped to exemplars, there was agreement across datasets (Figure S1): 24nt sRNAs, measured either as species or expression, generally targeted all families more intensely than 21-22nt sRNAs, although this was not as evident in tassel and for *Ji* that was previously categorized as a “22nt” family (39).

Next, we examined if this consistency persists during in-depth analysis of sRNA mapping along the length of elements. For this comparison we focused on the three *Copia* families, because of the preexisting annotation of their genomes, including information about complex palindrome motifs in the *cis*-regulatory region of the LTRs that are sRNA targeting hotspots (26). We found that both datasets produced highly similar patterns, with one intriguing exception: the exemplars were not mapped by sRNAs in the palindrome-rich regions (Figure 3A). Closer investigation of the exemplar sequences revealed that they contain long runs of N nucleotides in these regions (Figure 3B). We presume that the process of generating consensus sequences does not perform

‘Contamination’ of annotated TE populations affects sRNA mapping profiles

Our reference dataset representing the three *Copia* families is a strictly curated subset of the complete population of maize Sireviruses available from MASiVEDb (<http://bat.infospire.org/databases/masivedb/>). This larger dataset contains 6,283 *Ji*, 6,881 *Opie* and 221 *Giepum* full-length elements (Figure 1) that have been identified as *bona fide* Sireviruses (44). However, unlike our reference dataset, these TEs have not been carefully examined for the presence of ‘contaminating’ DNA from insertions of other elements or even from captured genes. As a result, an unknown number of elements may harbor foreign DNA, which in turn may affect data interpretation.

To examine this issue, we compared the mapping characteristics of the reference dataset against the complete MASiVEDb population. The number of sRNA species that mapped to each TE family increased substantially for MASiVEDb. Collectively, 626,836 sRNAs from the three sRNA libraries mapped to the 13,385 TEs of MASiVEDb, but only a third (206,589) of that total mapped to our reference dataset (Figure 2C, Table S1). Hence, the MASiVEDb population represents far more sRNA variation, but the extent to which TE contamination contributes to this increase is unclear. It is also hard to assess, because even very small fragments may map several sRNAs that could collectively account for a large fraction of the sRNA pool.

One indication may be provided by comparing the level of sRNA cross-talk between families of the two datasets. Our conjecture is that higher levels of cross-talk in MASiVEDb will reflect the presence of fragments of one family within elements of another family, thereby artificially increasing their ‘common’ sRNAs. Our analysis showed that indeed this was the case. For example, 30.8% (188,926 of 612,225 for all libraries combined) of the *Ji* and *Opie* sRNAs mapped to elements of both families using MASiVEDb compared to 3.1% (6,033 of 194,582) using the reference dataset. Likewise, cross-talk increased with the *Gypsy* families, for example from 0.2% to 5.3% between *Ji* and *Huck*, and from 0.2% to 10% between *Opie* and *Cinful-zeon*. To further test this conjecture, we screened for foreign TE fragments within the two datasets using non-Sirevirus maize TE exemplars as queries (BLASTN, max *E*-value 1×10^{-20}). We detected only two elements of the reference dataset with foreign TEs, compared to 1,158 elements of MASiVEDb that contained fragments (of 189nt median length) from 451 non-Sirevirus families.

Using MASiVEDb as an example, this analysis implies that a substantial fraction of the full-length elements of TE databases contain foreign TE fragments. Yet, that is not to say that these databases are of low quality; they all employ sophisticated algorithms based on various structural and homology criteria to identify high-quality full-length elements. To support this argument, we note that MASiVEDb generated similar sRNA mapping patterns as the reference dataset along the TE sequence, including the palindrome-rich region (Figure S2). It is likely, however, that other types of epigenetic

analyses might be negatively affected by TE contamination. Hence, it is advisable that the sequence quality of TEs is considered prior to mapping of sRNA data.

Normalization: Complexities regarding the use of M_sRNAs

Inclusion of M_sRNAs is necessary in TE studies

The handling of sRNAs with multiple mapping locations is an issue that has long troubled scientists. Often, in an effort to avoid methodological complications, M_sRNAs are excluded from analyses (29,30,35,36). However, even though U_sRNAs correlate more consistently with TE silencing than M_sRNAs (25), a significant proportion of RNA-directed DNA methylation (RdDM) is thought to be mediated by M_sRNAs (34). Moreover, our data in Figure 2B suggest that there may not be enough U_sRNAs to make meaningful inferences about TEs in hosts with large genomes.

To examine potential U_sRNA differences among plant species, we calculated the median density of 24nt U_sRNAs per nucleotide of maize TEs (for all libraries and families combined) and compared it to those of *Arabidopsis thaliana* and *lyrata* TEs previously reported by Hollister et al. (2011). While the median densities were only twofold different between *thaliana* and *lyrata* (0.11 vs. 0.06), these two species had a 69-fold and 37-fold difference with maize respectively (0.0016 24nt U_sRNAs per nucleotide of maize TEs). Comparative data were not available for 21-22nt U_sRNAs from the Hollister et al. study, but given that only 3,522 21-22nt U_sRNAs from all libraries mapped to the 15,532 full-length elements of the *Copia* and *Gypsy* datasets combined, it is clear that most elements did not map U_sRNAs in maize. These findings i) suggest that the effect of U_sRNAs on TE silencing may be less prominent in maize and other large genomes, and ii) decisively argue in favor of including M_sRNAs in TE analyses.

Normalization of M_sRNAs varies across genomic regions and between datasets

Besides excluding M_sRNAs from analyses as discussed previously or sometimes even allocating them randomly to single loci (45-47), the most common approaches for handling M_sRNAs is either to count all mapping locations so that each location has a value of 1.0, or to weight for multiple mapping so that each location is assigned a value of $1/x$, where x is the total number of locations for a given M_sRNA. This normalization can be applied to both “sRNA species” and “sRNA expression”. Nonetheless, it is unclear if and how these normalization strategies affect downstream research. One parameter that may provide valuable insights is the extent of multiple mapping (i.e., the number of mapping locations) for M_sRNAs that target various parts of a genome or different reference datasets (Figure 1). The reasoning is that the smaller the x , the weaker the differences between strategies will be and *vice versa*. To our knowledge, such a systematic examination has not been conducted so far. We therefore compared the mapping locations of M_sRNAs that target our *Copia* and *Gypsy* families i) across the

genome, ii) within their annotated full-length populations, and iii) across the TE exemplar database, so as to keep in line with the various strategies of previous studies.

Focusing first on the entire genome, we find that M_sRNAs have an exceptionally high number of mapping locations. For example, the median number of locations for all families combined was up to 513 among the three libraries, while the average often exceeded 1,500 (Table 1). Interestingly, separate investigation of the *Copia* and *Gypsy* families revealed that M_sRNAs targeting *Copia* elements had twice as many mapping locations compared to *Gypsy* elements (Table S3), even though each group of families comprises ~20-25% of the maize genome (32,33). This discrepancy may reflect different levels of sequence heterogeneity, i.e. families with many highly similar elements provide substantially more mapping loci for a given M_sRNA, compared to families with more divergent elements. Second, in contrast to the numerous mapping locations across the entire genome, there was a marked decrease in the number of locations within the annotated full-length populations (Table 1). We found that, on average, only a fifth of the genomic locations correspond to full-length elements, indicating that most M_sRNAs map to other types of sequences related to the six families, presumably unidentified full-length elements, degraded copies or solo LTRs. Third, the decrease was even more dramatic within the TE exemplar dataset, where the M_sRNAs of the six families only had three to five mapping locations each (Table 1).

Table 1. Number of locations for M_sRNAs that mapped to different parts of the maize genome

library	sRNA length	# of loci for sRNAs of the six families ^a			# of genomic loci for exon sRNAs ^a	# of genomic loci for other sRNAs ^a
		genome	TE population	exemplar db		
leaf	21	283 - 1,397	66 - 298	3 - 5	4 - 12	5 - 37
	22	262 - 1,261	70 - 284	3 - 5	4 - 11	5 - 42
	24	82 - 613	11 - 121	3 - 4	4 - 12	4 - 21
	all	127 - 854	18 - 179	3 - 4	4 - 11	4 - 26
tassel	21	425 - 2,033	114 - 419	3 - 5	4 - 18	6 - 57
	22	380 - 1,615	118 - 369	3 - 5	4 - 15	7 - 60
	24	199 - 1,017	26 - 194	3 - 4	5 - 17	4 - 25
	all	277 - 1,353	60 - 281	3 - 5	5 - 17	4 - 34
ear	21	513 - 2,130	86 - 411	4 - 5	4 - 14	6 - 55
	22	454 - 1,748	83 - 359	4 - 5	4 - 15	7 - 56
	24	147 - 897	19 - 170	3 - 5	4 - 17	5 - 26
	all	219 - 1,231	31 - 241	3 - 5	4 - 16	5 - 32

^aThe median (left) and average (right) number of mapping locations are shown for each category.

The above findings were derived from the most abundant TE families in maize and hence represent the most repetitive parts of a large genome. To contrast them with lower copy regions, we calculated the genomic locations of two additional sets of

M_sRNAs: exon-mapping M_sRNAs and all other M_sRNAs that did not map to either exons or the six TE families. We assume that a substantial proportion of the last category corresponds to less abundant TE families. (Note that the libraries were filtered for tRNAs, rRNAs, snoRNAs and miRNAs). Our analysis showed that the mapping locations of both categories did not exceed a handful of sites (Table 1); nonetheless, the average number of locations of the ‘other’ M_sRNAs was three-fold higher than the exon-mapping M_sRNAs, implying that a large proportion of the former type may indeed target low copy TEs.

Altogether, these data show that there can be up to a hundred-fold variation in the number of locations for M_sRNAs of maize TEs, depending on the reference dataset that is employed. This huge variation particularly affects the most abundant TE families, because mapping to TE exemplar datasets turns extreme multi-mappers to essentially semi-U_sRNAs (Table 1). Moreover, this effect likely holds true for the majority of plants, as most have genomes larger than maize and with concomitant TE content (24). Based on this evidence, and given the rapid accumulation of sequence data for several organisms, exemplars should be avoided when possible, and whole genomes or (at least) annotated TE populations should be preferred for the normalization of M_sRNAs.

Impact of normalization on data inference

To gain further insights into how sRNA metrics can change as a function of methodology, we compared the two extremes of a theoretical ‘normalization spectrum’, i.e. un-weighted vs. genome-weighted sRNA data, by investigating two questions of increasing complexity. The first was the general mapping characteristics of TE families, i.e. levels of sRNA targeting per nucleotide of full-length elements. We found that this metric was not altered substantially by the two extremes, because 24nt sRNA targeting remained stronger than 21-22nt sRNAs across most tissues for all families (Figure S3).

The second question was how sRNA mapping correlates with other TE variables. For this question, we studied sRNA targeting as a function of the insertion age of TEs. Age was calculated based on the sequence divergence of each LTR pair and profiled at the family level (Figure 4A). Use of un-weighted data generated strong negative correlations between age and both sRNA species and sRNA expression for all combinations of tissue, family and sRNA length (average Spearman $r = -0.67$, $P < 10^{-20}$; Figures 4B, S4). Critically, use of genome-weighted data retained this pattern only for 21-22nt sRNAs (average Spearman $r = -0.35$, $P < 10^{-20}$ in most cases), while for 24nt sRNAs there was discordance both between sRNA metrics and among families. We detected a positive correlation for *Ji*, *Opie* and *Huck* using sRNA species, which was often reversed or not statistically supported using sRNA expression (Figures 4B, S4). In contrast, there was a negative correlation for *Cinful-zeon*, *Flip* and *Giepum* across most tissues and for both sRNA metrics.

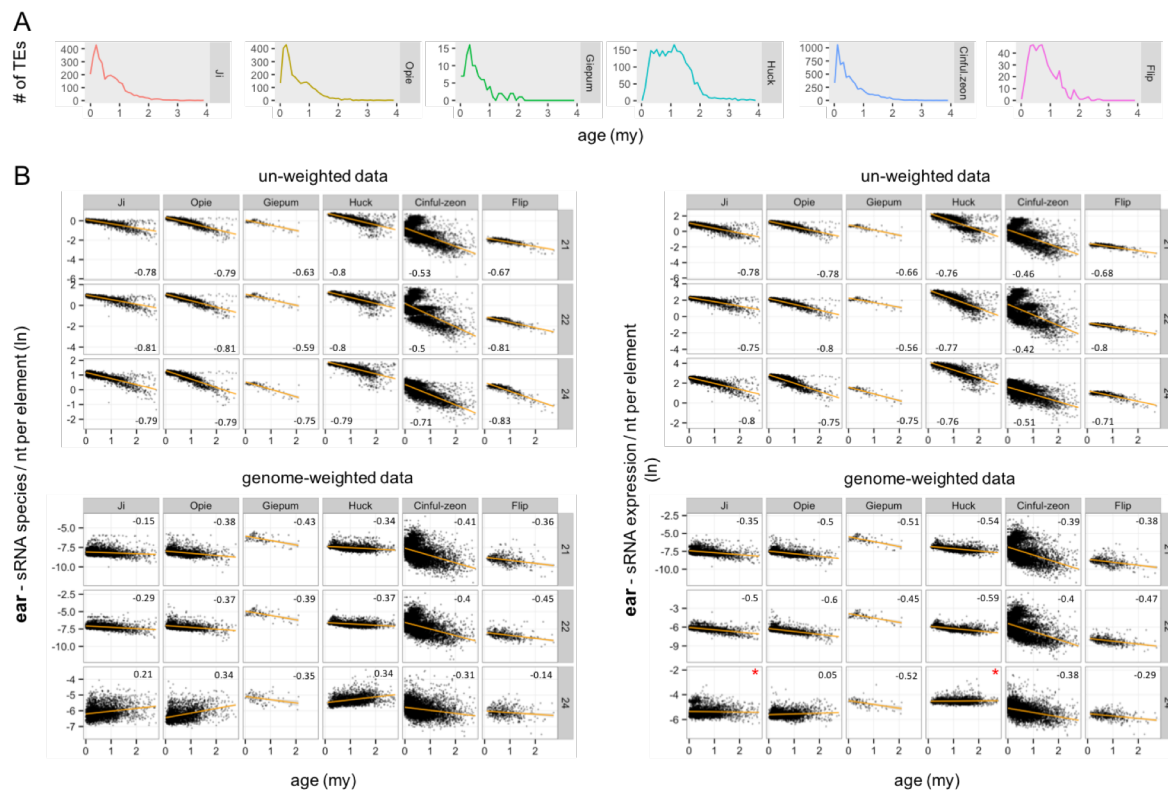


Figure 4. Relationship between TE age and sRNA mapping using un-weighted and genome-weighted approaches. (A) Age distribution in million years (my) of TE families. (B) Mapping of sRNA species (left panels) or expression (right panels) from ear tissue was calculated per nucleotide of full-length elements for each family. Age is cutoff at 3my to allow sufficient visualization of the x -axis. The Spearman r coefficient is shown for each plot, calculated for all elements and not only for those <3my. P values were <0.01, except those indicated by an asterisk.

Finally, we contrasted the above analysis with how normalization affects another type of commonly used HTS libraries, that is of mRNA expression data. It is known that mRNA libraries have a considerably smaller amount of multiple mapping reads (10% against the genome, in comparison to 40-90% of sRNA libraries), which can be attributed to the longer read lengths and mostly single-copy transcription loci (23). We therefore retrieved mRNA data from three biological leaf replicates and examined (as we did with sRNAs) i) their general mapping characteristics to the genome, ii) the expression patterns of TE families, and iii) the relationship between expression and TE age. First, our analyses confirmed the aforementioned ~10% of M_mRNAs at the genome levels, but we additionally discovered that the vast majority TE-mapping reads were M_mRNAs (Table S4). Furthermore, the median number of locations for these TE-mapping M_mRNAs across the genome or within the annotated full-length elements (Table S4) was approximately only two-fold lower to those of the TE-mapping M_sRNAs (Table 1).

Second, both normalization approaches generated the same relative expression levels among families despite their widely different sizes (Figure S5A). Finally, both types of data produced strong negative correlations between mRNA expression and age for all possible combinations (average Spearman $r = -0.61$, $P < 10^{-20}$; Figure S5B).

Taken together, these findings suggest that the choice of treatment of HTS data can affect biological inference, clearly evident in the inconsistent relationship between 24nt sRNAs and age. And although the conclusions of many of the analyses were unchanged, we note that the strength of the correlations with age were substantially weaker for genome-weighted than un-weighted data (average r of 0.32 vs. 0.67 for sRNAs and 0.72 vs. 0.52 for mRNAs, using absolute values). This is counterintuitive, because, as we showed earlier (Table 1), weighting-by-location is expected to have a stronger impact on high-copy than low-copy sequences. Yet, 21-22nt sRNA and mRNA profiles did not change as a function of age within each family, whereby the numerous young and highly similar elements were mapped by more sRNAs (Figures 4B, S4) or mRNAs (Figure S5B) than their few, old and divergent relatives in both normalization approaches. Also, the relative mRNA expression levels of families of varying sizes did not change by using weighted or un-weighted data (Figure S5A). Based on these insights and the unpredictability of how normalization may further impact other research questions, we argue that multiple approaches should be used in parallel to validate results.

U_sRNA-guided mapping of M_sRNAs may be problematic for TE studies

An alternative approach for mapping M_sRNAs assigns reads to single loci using as guide the local densities of U_sRNAs (23). This method, which is at the core of the ShortStack tool (12), aims to find the true generating locus of each read, instead of allocating M_sRNAs across their targets or even excluding them altogether. Historically, this concept was initially tested with mRNA data where it significantly improved placement of M_mRNAs (48). For sRNAs, recent analysis of simulated libraries by Johnson et al. (2016) showed that the U_sRNA-guided mode outperforms other methodologies in selecting the correct locus from which an M_sRNA may have originated.

However, our data suggest that two properties of TEs may pose a real challenge to this process. First, there is a very small number of U_sRNAs that align to TEs. For example, only 2,166 of 147,034 sRNA species of the ear library that collectively mapped to our *Copia* and *Gypsy* elements are U_sRNAs (Figure 2B, Table S2); furthermore, the vast majority of these U_sRNAs mapped to different TEs (Figure 2D). As a result, and given that the length of our TEs ranges between 7-15kb and that ShortStack examines 250nt windows (23), it is expected that most windows will not have a U_sRNA score and hence vast amounts of M_sRNAs will be discarded. The second issue concerns the numerous genomic locations for M_sRNAs targeting TEs (Table 1). These are far above the 50-target cutoff that Johnson et al. suggest leads to a high rate of misplacement.

Finally, ShortStack can also guide M_sRNA allocation by calculating the densities of both U_sRNAs and weighted M_sRNAs; however, this option did not perform as well as the U_sRNA-only option at the genome level in Arabidopsis, rice and maize (23) and, hence, it is likely that its performance will be further compromised in TE-focused analyses.

A last important point is that a distinction should be made between sRNA-generating vs. sRNA-targeting loci. ShortStack appears to work beautifully for allocating M_sRNAs to their single locus of origin, and future developments may make this approach more efficient for TE data as well. Nonetheless, studies that investigate sRNA targeting patterns may benefit more by methods that allow multiple mapping. This may be especially important for TEs, where it is possible that a given sRNA mediates silencing of more than one locus. Although not empirically proven yet, this conjecture is supported by evidence for the importance of M_sRNAs in RdDM (34), the homology-based *trans* silencing pathway among TEs (38), and the cytoplasmic step of Argonaute loading that dissociates sRNAs from their generating loci (49).

sRNA metrics: unexpected differences between sRNA species and sRNA expression

So far, our analysis has indicated that sRNA species and sRNA expression generally produce similar results. However, this is not always true. When we examined the relationship between sRNAs and age separately for the LTRs and the internal (INT) domain of TEs using un-weighted data, we observed that the plots of the *Opie* family were markedly different in one case. The expression levels of 24nt sRNAs from leaf on the LTRs split the *Opie* elements in two distinct groups, whereby the ‘upper zone’ was mapped by approximately twice as many reads compared to the ‘lower zone’ (Figure 5A). Species of 24nt sRNAs did not generate the same pattern, nor did other combinations of sRNA lengths and metrics in *Opie* (Figure 5A), or in other families or tissues (not shown).

To our surprise, closer investigation revealed that this ‘zoning’ was triggered by sRNAs that mapped to a narrow region on the sense strand of the LTRs (Figure 5B). This region was targeted by ~115x more reads in the elements of the upper zone compared to the elements of the lower zone (median coverage of 1,610 and 14 reads/nt respectively), while there was only a three-fold difference (6.1 vs. 2.1 reads/nt) along the rest of the LTR. This finding implies that highly expressed sRNA species mapping to this region of the elements of the upper zone may drive the *Opie* split. We retrieved 836 24nt sRNA species from all *Opie* elements and, surprisingly, only one appeared to be responsible for the zoning. This sRNA species combined very high expression (1,976 reads) and number of target LTRs (3,228 LTRs that comprise the upper zone), ranked 1st and 7th respectively among the 836 sRNAs. In contrast, most other highly-mapped sRNAs of the same region had expression levels of <10 reads, and therefore did not contribute to the zoning.

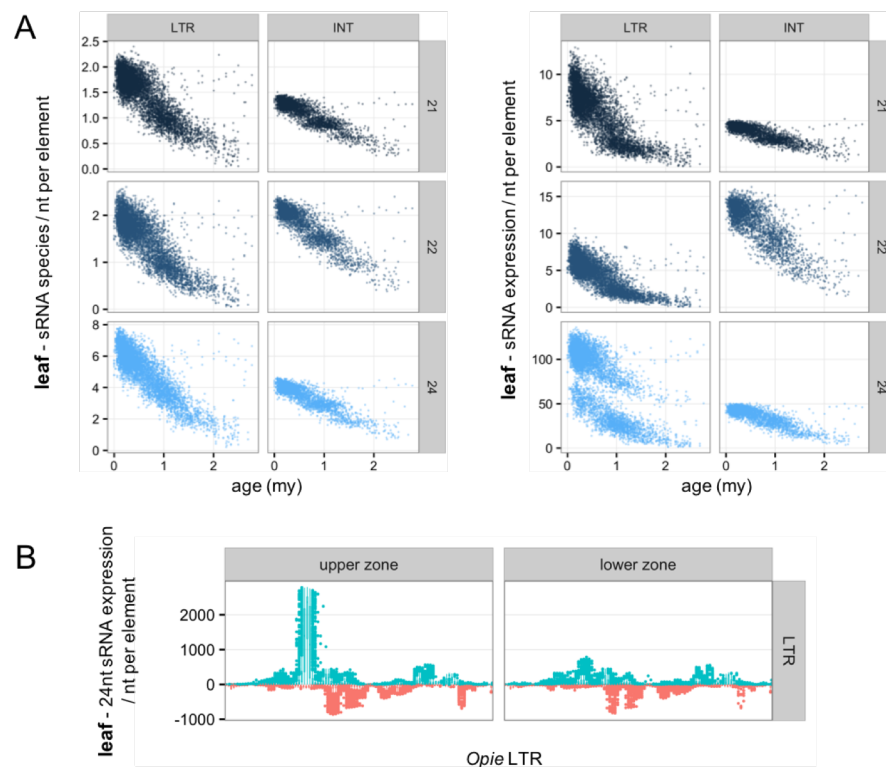


Figure 5. *Opie* population split based on sRNA expression data from leaf tissue. (A) Relationship between TE age and number of sRNA species (left) or expression (right) calculated per nucleotide of the *Opie* LTRs and INT domain. Age is cutoff at 3my to allow sufficient visualization of the *x*-axis. **(B)** Mapping patterns (calculated as in Figure 3A) of 24nt expression data along the LTRs of the two distinct *Opie* subpopulations. sRNA data in **A** and **B** were not weighted by their number of genomic loci.

The fact that a single sRNA generates this spectacular pattern raises several methodological concerns. First, it is likely that such very high expression levels may be the outcome of biases during library construction (15). Second, our data imply that the use of sRNA species is more robust than sRNA expression, because it appears to be less sensitive to errors that can occur, e.g., during PCR amplification. Finally, and perhaps most importantly, these findings denote the need for the confirmation of such observations. This can be achieved by cross-examining results from different normalization approaches. We checked if genome-weighted sRNA expression data reproduce the zoning, but this does not seem to be the case (not shown). However, given the inconsistencies of normalization approaches as discussed previously, the most appropriate way is the inclusion in the experimental design of technical and/or biological replicates. In previous years the lack of sRNA studies could be attributed to the high costs of sequencing. These costs are now much lower and, hence, replicates should be typically included in epigenetic studies to help identify aberrancies.

CONCLUDING REMARKS

In this work we attempted to address the complex issue of mapping and analyzing sRNAs in the context of TEs, which comprise the vast majority of most plant genomes. With the goal of providing insights that might help guide future studies, we compared mapping to TE exemplars vs. annotated TE populations; explored the extent of multiple mapping across different genomic regions or datasets; examined how various strategies for mapping M_sRNAs (or M_mRNAs) affect biological inference or are applicable to TEs; and finally presented an unexpected inconsistency between the two sRNA metrics of species and expression. Summarizing each section into take-home messages:

1) TE exemplars have been widely popular thus far for various reasons, including the absence of sufficient sequence information or, indeed, because research would not truly benefit from the burdensome analysis of annotated TE populations. However, their usage comes with several limitations that scientists need to be aware of, such as the presence of unresolved/masked regions in their sequences, the overrepresentation of U_sRNAs, and the large volumes of unmapped data that may severely bias analyses.

2) Annotated TE populations appear to be more informative than exemplars for mapping epigenetic data. Given that sequence data are now available for a large number of genomes, the use of TE populations – additionally filtered for foreign DNA if needed – should be preferred over exemplars when possible.

3) Our analyses also suggest that the inclusion of M_sRNAs in TE studies is necessary, and we strongly advocate against a focus solely on U_sRNAs. We favor using comparative un-weighted and weighted (i.e. normalized) mapping approaches in parallel to validate biological inferences. Crucially, whole, or even partially, sequenced genomes should be preferred over exemplars for weighting M_sRNAs. Furthermore, approaches that assign M_sRNAs to single loci based on U_sRNAs density are very promising but are not yet applicable for TE studies.

4) Finally, the metric of sRNA expression (and to a lesser extent sRNA species) may be prone to errors during HTS library construction. However, sRNA expression is often a crucial measurement, for example during differential expression analysis. Aided by the rapidly falling costs of sequencing, the inclusion of technical and/or biological replicates in sRNA studies should now be standard.

METHODS

TE datasets

For the *Ji*, *Opie* and *Giepum* Sirevirus families of the *Copia* superfamily we used the strictly curated set of full-length elements that were previously analyzed in Bousios et al. 2016. We also retrieved the complete annotated populations from the MASiVEDb website (<http://bat.infospire.org/databases/masivedb/>) (31).

For the *Huck*, *Cinful-zeon* and *Flip* families of the *Gypsy* superfamily we first retrieved the repeat annotation file of the maize TE consortium (ZmB73_5a_MTEC+LTR_repeats.gff) from ftp.gramene.org. This file, however, does not specify whether an annotated region represents full-length or fragmented TEs. Hence, we plotted the frequency distribution of the lengths of the annotated regions to identify peaks for each family that would correspond to the size of full-length elements as calculated by Baucom et al. (32) (Figure S6A). This approach identified a single peak for *Huck* that nearly overlapped with the Baucom full-length average (13.4kb), two peaks for *Cinful-zeon* that flanked the Baucom average (8.2kb), and two peaks for *Flip* – one nearly overlapping with the Baucom average (14.8kb) and one residing in close proximity (Figure S6A). Based on these results, we selected regions between 13.3-14.1kb for *Huck*, 7.1-7.5kb and 9.2-9.7kb for *Cinful-Zeon*, and 14.8-15.6kb for *Flip* as candidates for full-length elements, retrieving 2,614, 6,965 and 607 sequences respectively. We then ran LTRharvest (50) with parameters *xdrop* 25, *mindistltr* 2000, *maxdistltr* 20000, *ins* -3, *del* -3, *similar* 50, *motif* TGCA, *motifmis* 1, *minlenltr* 100, and *maxlenltr* 5000 in order to identify the borders between the LTRs and the INT domain, and to also calculate the canonical LTR length of each family. Based on our approach, we selected LTR lengths between 1-1.8kb for *Huck*, 450-750nt for *Cinful-zeon*, and 4.1-4.5kb for *Flip*, finally yielding 2,460, 6,276 and 483 full-length elements for each family respectively (Figure S6B).

The insertion age of each *Copia* and *Gypsy* TE was calculated by first aligning the LTRs using MAFFT with default parameters (51) and then applying the LTR retrotransposon age formula with a substitution rate of 1.3×10^{-8} mutations per site per year (52).

Finally, all maize TE exemplars were downloaded from <http://maizetadb.org>. Note that we removed one *Ji* (RLC_ji_AC186528-1508) and two *Giepum* (RLC_giepum_AC197531-5634; RLC_giepum_AC211155-11010) exemplars from our analysis, based on evidence from Bousios et al. (2012) that indicated that the specific exemplars are not true representatives of these families.

Mapping sRNA and mRNA libraries

We used published sRNA data from leaf (GSM1342517), tassel (GSM448857), and ear (GSM306487) tissue, and mRNA data from three technical replicates from leaf (SRR531869, SRR531870, SRR531871) tissue. Adapters and low quality nucleotides were removed using Trimmomatic and the FASTX toolkit respectively, until every read had three or more consecutive nucleotides with a Phred quality score of >20 at the 3'-end. The libraries were filtered for miRNAs (<http://www.mirbase.org/>), tRNAs (<http://gtrnadb.ucsc.edu/>), and rRNAs and snoRNAs (<http://rfam.sanger.ac.uk/>); sRNA reads of 21nt, 22nt and 24nt length and mRNA reads longer than 25nt were mapped to the maize B73 genome (RefGen_V2) and the maize TE database using BWA with default

settings and no mismatches (22). Following previous work (26), each distinct sRNA or mRNA sequence was termed “species”, and the number of its reads was its “expression”.

After mapping, each species was tagged as either U_sRNA/U_mRNA or M_sRNA/M_mRNA separately for the genome and the exemplar database. M_sRNAs/M_mRNAs were either normalized by their number of mapping locations or not normalized, depending on the analysis. Finally, we calculated the total number of sRNA species that mapped to a TE ‘locus’ (i.e. the full-length sequence, LTRs or INT domain), but also the number of sRNA species and sRNA expression (weighted or un-weighted) per nucleotide of each locus. The per nucleotide measures allow comparisons of averages among TEs and also analysis along the length of the TE sequence.

SUPPLEMENTAL MATERIAL

The sequences and other information of the annotated TE populations are available as Supplemental Material in bat.infospire.org/sireviruses/Submission_Data/.

ACKNOWLEDGMENTS

AB is supported by the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement no: PIEF-GA-2012-329033; BSG by NSF grant IOS-1542703 and a fellowship from the Borchard Foundation; ND by research grant AZV-MZ-CR 16-34272A-4/2016 and project CEITEC 2020 (LQ1601).

REFERENCES

1. Castel, S.E. and Martienssen, R.A. (2013) RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. *Nature Reviews Genetics*, **14**, 100-112.
2. Axtell, M.J. (2013) Classification and Comparison of Small RNAs from Plants. *Annual Review of Plant Biology*, Vol 64, **64**, 137-159.
3. Borges, F. and Martienssen, R.A. (2015) The expanding world of small RNAs in plants. *Nature Reviews Molecular Cell Biology*, **16**, 727-741.
4. Matzke, M.A. and Mosher, R.A. (2014) RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nature Reviews Genetics*, **15**, 394-408.
5. An, J.Y., Lai, J., Lehman, M.L. and Nelson, C.C. (2013) miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. *Nucleic Acids Research*, **41**, 727-737.
6. Garmire, L.X. and Subramaniam, S. (2012) Evaluation of normalization methods in mammalian microRNA-Seq data. *Rna-a Publication of the Rna Society*, **18**, 1279-1288.
7. Hackenberg, M., Rodriguez-Ezpeleta, N. and Aransay, A.M. (2011) miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Research*, **39**, W132-W138.

8. Li, Y., Zhang, Z., Liu, F., Vongsangnak, W., Jing, Q. and Shen, B. (2012) Performance comparison and evaluation of software tools for microRNA deep-sequencing data analysis. *Nucleic Acids Research*, **40**, 4298-4305.
9. Sablok, G., Milev, I., Minkov, G., Minkov, I., Varotto, C., Yahubyan, G. and Baev, V. (2013) isomiRex: Web-based identification of microRNAs, isomiR variations and differential expression using next-generation sequencing datasets. *Febs Letters*, **587**, 2629-2634.
10. Srivastava, P.K., Moturu, T.R., Pandey, P., Baldwin, I.T. and Pandey, S.P. (2014) A comparison of performance of plant miRNA target prediction tools and the characterization of features for genome-wide target prediction. *Bmc Genomics*, **15**.
11. Zhu, E.L., Zhao, F.Q., Xu, G., Hou, H.B., Zhou, L.L., Li, X.K., Sun, Z.S. and Wu, J.Y. (2010) mirTools: microRNA profiling and discovery based on high-throughput sequencing. *Nucleic Acids Research*, **38**, W392-W397.
12. Axtell, M.J. (2013) ShortStack: Comprehensive annotation and quantification of small RNA genes. *Rna-a Publication of the Rna Society*, **19**, 740-751.
13. Hardcastle, T.J., Kelly, K.A. and Baulcombe, D.C. (2012) Identifying small interfering RNA loci from high-throughput sequencing data. *Bioinformatics*, **28**, 457-463.
14. Luo, G.-Z., Yang, W., Ma, Y.-K. and Wang, X.-J. (2014) ISRNA: an integrative online toolkit for short reads from high-throughput sequencing data. *Bioinformatics*, **30**, 434-436.
15. McCormick, K.P., Willmann, M.R. and Meyers, B.C. (2011) Experimental design, preprocessing, normalization and differential expression analysis of small RNA sequencing experiments. *Silence*, **2**, 2-2.
16. Rueda, A., Barturen, G., Lebron, R., Gomez-Martin, C., Alganza, A., Oliver, J.L. and Hackenberg, M. (2015) sRNAtoolbox: an integrated collection of small RNA research tools. *Nucleic Acids Research*, **43**, W467-W473.
17. Stocks, M.B., Moxon, S., Mapleson, D., Woolfenden, H.C., Mohorianu, I., Folkes, L., Schwach, F., Dalmay, T. and Moulton, V. (2012) The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. *Bioinformatics*, **28**, 2059-2061.
18. Sun, J., Nishiyama, T., Shimizu, K. and Kadota, K. (2013) TCC: an R package for comparing tag count data with robust normalization strategies. *Bmc Bioinformatics*, **14**.
19. MacLean, D., Moulton, V. and Studholme, D.J. (2010) Finding sRNA generative locales from high-throughput sequencing data with NiBLS. *Bmc Bioinformatics*, **11**.
20. Moxon, S., Schwach, F., Dalmay, T., MacLean, D., Studholme, D.J. and Moulton, V. (2008) A toolkit for analysing large-scale plant small RNA datasets. *Bioinformatics*, **24**, 2252-2253.
21. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**.
22. Li, H. and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589-595.

23. Johnson, N.R., Yeoh, J.M., Coruh, C. and Axtell, M.J. (2016) Improved Placement of Multi-mapping Small RNAs. *G3-Genes Genomes Genetics*, **6**, 2103-2111.
24. Tenaillon, M.I., Hollister, J.D. and Gaut, B.S. (2010) A triptych of the evolution of plant transposable elements. *Trends in Plant Science*, **15**, 471-478.
25. Hollister, J.D., Smith, L.M., Guo, Y.-L., Ott, F., Weigel, D. and Gaut, B.S. (2011) Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 2322-2327.
26. Bousios, A., Diez, C.M., Takuno, S., Bystry, V., Darzentas, N. and Gaut, B.S. (2016) A role for palindromic structures in the cis-region of maize Sirevirus LTRs in transposable element evolution and host epigenetic response. *Genome Research*, **26**(2):226-37.
27. Ragupathy, R., You, F.M. and Cloutier, S. (2013) Arguments for standardizing transposable element annotation in plant genomes. *Trends in Plant Science*, **18**, 367-376.
28. Diez, C.M., Meca, E., Tenaillon, M.I. and Gaut, B.S. (2014) Three Groups of Transposable Elements with Contrasting Copy Number Dynamics and Host Responses in the Maize (*Zea mays ssp mays*) Genome. *Plos Genetics*, **10**.
29. Gent, J.I., Ellis, N.A., Guo, L., Harkess, A.E., Yao, Y., Zhang, X. and Dawe, R.K. (2013) CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize. *Genome Research*, **23**, 628-637.
30. Regulski, M., Lu, Z., Kendall, J., Donoghue, M.T.A., Reinders, J., Llaca, V., Deschamps, S., Smith, A., Levy, D., McCombie, W.R. *et al.* (2013) The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA. *Genome Research*, **23**, 1651-1662.
31. Bousios, A., Minga, E., Kalitsou, N., Pantermali, M., Tsaaballa, A. and Darzentas, N. (2012) MASiVEDb: the Sirevirus Plant Retrotransposon Database. *Bmc Genomics*, **13**.
32. 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. (2009) Exceptional Diversity, Non-Random Distribution, and Rapid Evolution of Retroelements in the B73 Maize Genome. *Plos Genetics*, **5**.
33. Bousios, A., Kourmpetis, Y.A.I., Pavlidis, P., Minga, E., Tsaftaris, A. and Darzentas, N. (2012) The turbulent life of Sirevirus retrotransposons and the evolution of the maize genome: more than ten thousand elements tell the story. *Plant Journal*, **69**, 475-488.
34. Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H. and Ecker, J.R. (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell*, **133**, 523-536.
35. Law, J.A., Du, J.M., Hale, C.J., Feng, S.H., Krajewski, K., Palanca, A.M.S., Strahl, B.D., Patel, D.J. and Jacobsen, S.E. (2013) Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1. *Nature*, **498**, 385-+.
36. Zhai, J.X., Bischof, S., Wang, H.F., Feng, S.H., Lee, T.F., Teng, C., Chen, X.Y., Park, S.Y., Liu, L.S., Gallego-Bartolome, J. *et al.* (2015) A One Precursor One siRNA Model for Pol IV-Dependent siRNA Biogenesis. *Cell*, **163**, 445-455.

37. Bousios, A. and Gaut, B.S. (2016) Mechanistic and evolutionary questions about epigenetic conflicts between transposable elements and their plant hosts. *Current Opinion in Plant Biology*, **30**, 123-133.
38. Fultz, D., Choudury, S.G. and Slotkin, R.K. (2015) Silencing of active transposable elements in plants. *Current opinion in plant biology*, **27**, 67-76.
39. Barber, W.T., Zhang, W., Win, H., Varala, K.K., Dorweiler, J.E., Hudson, M.E. and Moose, S.P. (2012) Repeat associated small RNAs vary among parents and following hybridization in maize. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 10444-10449.
40. Bennetzen, J.L., Schmutz, J., Wang, H., Percifield, R., Hawkins, J., Pontaroli, A.C., Estep, M., Feng, L., Vaughn, J.N., Grimwood, J. *et al.* (2012) Reference genome sequence of the model plant *Setaria*. *Nature Biotechnology*, **30**, 555-+.
41. Shulaev, V., Sargent, D.J., Crowhurst, R.N., Mockler, T.C., Folkerts, O., Delcher, A.L., Jaiswal, P., Mockaitis, K., Liston, A., Mane, S.P. *et al.* (2011) The genome of woodland strawberry (*Fragaria vesca*). *Nature Genetics*, **43**, 109-116.
42. Diez, C.M., Vitte, C., Ross-Ibarra, J., Gaut, B.S. and Tenaillon, M.I. (2012) Using Nextgen Sequencing to Investigate Genome Size Variation and Transposable Element Content. *Book: Plant Transposable Elements. Topics in Current Genetics (eds. Grandbastien MA and Casacuberta JM)*, 41-58.
43. Tenaillon, M.I., Hufford, M.B., Gaut, B.S. and Ross-Ibarra, J. (2011) Genome Size and Transposable Element Content as Determined by High-Throughput Sequencing in Maize and *Zea luxurians*. *Genome Biology and Evolution*, **3**, 219-229.
44. Darzentas, N., Bousios, A., Apostolidou, V. and Tsaftaris, A.S. (2010) MASiVE: Mapping and Analysis of SireVirus Elements in plant genome sequences. *Bioinformatics*, **26**, 2452-2454.
45. He, G., Chen, B., Wang, X., Li, X., Li, J., He, H., Yang, M., Lu, L., Qi, Y., Wang, X. *et al.* (2013) Conservation and divergence of transcriptomic and epigenomic variation in maize hybrids. *Genome Biology*, **14**.
46. McCue, A.D., Nuthikattu, S. and Slotkin, R.K. (2013) Genome-wide identification of genes regulated in trans by transposable element small interfering RNAs. *Rna Biology*, **10**, 1379-1395.
47. Wang, X., Elling, A.A., Li, X., Li, N., Peng, Z., He, G., Sun, H., Qi, Y., Liu, X.S. and Deng, X.W. (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.
48. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, **5**, 621-628.
49. Ye, R.Q., Wang, W., Iki, T., Liu, C., Wu, Y., Ishikawa, M., Zhou, X.P. and Qi, Y.J. (2012) Cytoplasmic Assembly and Selective Nuclear Import of Arabidopsis ARGONAUTE4/siRNA Complexes. *Molecular Cell*, **46**, 859-870.
50. Ellinghaus, D., Kurtz, S. and Willhoeft, U. (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *Bmc Bioinformatics*, **9**.

51. Katoh, K. and Standley, D.M. (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, **30**, 772-780.
52. Ma, J.X. and Bennetzen, J.L. (2004) Rapid recent growth and divergence of rice nuclear genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 12404-12410.