the plant journal

SEB Society for Experimental Biology

The Plant Journal (2012) **72**, 142–153 doi: 10.1111/j.1365-313X.2012.05072.x

The sunflower (*Helianthus annuus* L.) genome reflects a recent history of biased accumulation of transposable elements

S. Evan Staton[1], Bradley H. Bakken[2], Benjamin K. Blackman[3,†], Mark A. Chapman[4,‡], Nolan C. Kane[5], Shunxue Tang[6,§], Mark C. Ungerer[2], Steven J. Knapp[6,¶], Loren H. Rieseberg[5] and John M. Burke[4,*]

[1]*Department of Genetics, University of Georgia, Athens, GA 30602, USA,*

[2]*Division of Biology, 426 Ackert Hall, Kansas State University, Manhattan, KS 66506, USA,*

[3]*Department of Biology, Jordan Hall, 1001 East Third Street, Indiana University, Bloomington, IN 47405, USA,*

[4]*Department of Plant Biology, University of Georgia, Athens, GA 30602, USA,*

[5]*The Biodiversity Research Centre and Department of Botany, 3529-6270 University Blvd, University of British Columbia, Vancouver, BC, Canada V6T 1Z4, and*

[6]*Institute for Plant Breeding, Genetics, and Genomics, University of Georgia, Athens, GA 30602, USA*

Received 5 August 2011; revised 25 April 2012; accepted 1 June 2012; published online 30 July 2012.

*For correspondence (e-mail jmburke@uga.edu).

[†]Present address: Department of Biology, Duke University, Durham, NC 27708, USA.

[‡]Present address: Department of Plant Sciences, University of Oxford, South Parks Road, Oxford, OX1 3RB, UK.

[§]Present address: Trait Genetics and Technologies, Dow AgroSciences LLC, 9330 Zionsville Road, Indianapolis, IN 46268, USA.

[¶]Present address: Monsanto Vegetable Seeds, 37437 California Highway 16, Woodland, CA 95695, USA.

SUMMARY

Aside from polyploidy, transposable elements are the major drivers of genome size increases in plants. Thus, understanding the diversity and evolutionary dynamics of transposable elements in sunflower (*Helianthus annuus* L.), especially given its large genome size (~3.5 Gb) and the well-documented cases of amplification of certain transposons within the genus, is of considerable importance for understanding the evolutionary history of this emerging model species. By analyzing approximately 25% of the sunflower genome from random sequence reads and assembled bacterial artificial chromosome (BAC) clones, we show that it is composed of over 81% transposable elements, 77% of which are long terminal repeat (LTR) retrotransposons. Moreover, the LTR retrotransposon fraction in BAC clones harboring genes is disproportionately composed of chromodomain-containing *Gypsy* LTR retrotransposons ('chromoviruses'), and the majority of the intact chromoviruses contain tandem chromodomain duplications. We show that there is a bias in the efficacy of homologous recombination in removing LTR retrotransposon DNA, thereby providing insight into the mechanisms associated with transposable element (TE) composition in the sunflower genome. We also show that the vast majority of observed LTR retrotransposon insertions have likely occurred since the origin of this species, providing further evidence that biased LTR retrotransposon activity has played a major role in shaping the chromatin and DNA landscape of the sunflower genome. Although our findings on LTR retrotransposon age and structure could be influenced by the selection of the BAC clones analyzed, a global analysis of random sequence reads indicates that the evolutionary patterns described herein apply to the sunflower genome as a whole.

Keywords: transposable elements, chromodomain, *Helianthus annuus*, Asteraceae, LTR retrotransposon, genome evolution.

INTRODUCTION

Transposable elements (TEs) are mobile DNA sequences that are present in the nuclear genomes of virtually all eukaryotes. A common feature of TEs is the potential to replicate faster than the host, thereby allowing them to increase in abundance, sometimes drastically (e.g. Naito *et al.*, 2009; Belyayev *et al.*, 2010), from one generation to the next. Variation in TE amplification rates can thus generate enormous variation in TE content within and

between the genomes of even closely related species (e.g. Piegu *et al.*, 2006; Ungerer *et al.*, 2006; Wicker *et al.*, 2009). Differences in TE abundance amongst genomes may be explained by differences in host-encoded mechanisms that limit transposition, modes of TE replication or specific properties that limit TE removal from the genome (Lippman *et al.*, 2004; Du *et al.*, 2010).

Class-I TEs (i.e. retrotransposons) replicate through an RNA intermediate that is reverse transcribed into a DNA copy that can insert elsewhere in the genome (Kumar and Bennetzen, 1999). These elements can be classified into five taxonomic orders (Wicker *et al.*, 2007). The most abundant and diverse order in plants, the long terminal repeat retrotransposons (LTR-RTs), is primarily composed of two superfamilies, *Ty1*/*copia* and *Ty3*/*gypsy* (referred to hereafter as *Copia* and *Gypsy*, respectively; Wicker *et al.*, 2007), which can be distinguished based on the order of their coding domains as well as the similarity of their reverse transcriptase sequences (Xiong and Eickbush, 1990; Kumar and Bennetzen, 1999). Certain *Gypsy* clades exhibit an extra coding domain known as the 'chromodomain', which is thought to confer insertion site specificity (Gao *et al.*, 2008). Although *Copia* and *Gypsy* elements are present in all plant genomes (Voytas *et al.*, 1992; Suoniemi *et al.*, 1998), their relative proportions vary between species (Hua-Van *et al.*, 2011). This variation may result from different insertion site preferences (Peterson-Burch *et al.*, 2004; Gao *et al.*, 2008), but could also be driven by variation in the efficacy of illegitimate recombination and/or unequal homologous recombination in removing LTR-RTs from the genome (Devos *et al.*, 2002; Ma *et al.,* 2004). In contrast, class-II TEs (i.e. DNA transposons) use a DNA-based enzymatic method for excision and transposition of the parent copy without creating a new copy (Wicker *et al.*, 2007). Consequently, class-II TEs are generally less abundant than retrotransposons.

Despite their differences in genomic abundance, both retrotransposons and DNA transposons are potent sources of genetic variation (e.g. McClintock, 1984; Hilbrict *et al.*, 2008; Zeh *et al.*, 2009). Transposable elements also have a large impact on, and appear to be integral components of, the chromatin landscape of the host genome (Biemont, 2009). In *Arabidopsis thaliana*, for example, epigenetic regulation of TEs and tandem repeats contributes to genome organization and the regulation of neighboring genes (e.g. Lippman *et al.*, 2004; Hollister and Gaut, 2009), and TEs also contribute to expression divergence between Arabidopsis species (Pereira *et al.*, 2009; Warenfors *et al.*, 2010; Hollister *et al.*, 2011). Given the potential influence of TEs on the structure and function of plant genomes, we investigated their contribution to *Helianthus annuus* L. (sunflower) genome evolution.

Sunflower is a diploid ($2n = 34$) species with an estimated genome size of ~3.5 Gb (Baack *et al.*, 2005). Because the total number of retrotransposon copies in the genome of a plant species typically correlates with genome size (Bennetzen, 2000, 2007; Devos, 2010), we expected the sunflower genome to contain billions of base pairs of retrotransposon DNA. Indeed, previous studies have suggested that the sunflower genome may be 62–78% repetitive (Cavallini *et al.*, 2010; Kane *et al.*, 2011), and a few studies have also investigated the genomic organization of retrotransposons in sunflower. Retrotransposons are known to be transcriptionally active in this species (Vukich *et al.*, 2009; Cavallini *et al.*, 2010; Kawakami *et al.*, 2011), and fluorescent *in situ* hybridization studies have indicated that the *Gypsy* and *Copia* superfamilies are enriched in the heterochromatic regions of the pericentromeres and telomeres, respectively (Santini *et al.*, 2002; Natali *et al.*, 2006; Staton *et al.*, 2009). This genomic organization of *Gypsy* elements has been conserved in hybrid sunflower species derived from the common sunflower, despite massive amplifications of these elements in the hybrid species' genomes (Ungerer *et al.*, 2006, 2009; Staton *et al.*, 2009).

Many basic questions about the contributions of transposons to sunflower genome evolution remain unanswered, however, because previous studies have relied on *in situ* hybridization techniques that only offered chromosome-level resolution (Natali *et al.*, 2006; Staton *et al.*, 2009; Cavallini *et al.*, 2010). For example, what has been the evolutionary time scale over which these sequences have been active? Were these, and the majority of other LTR-RT sequences, present in the common ancestor of sunflower and related species, or did they arise following the origin of the sunflower lineage (0.74–1.67 Ma; Heesacker *et al.*, 2009)? Also, given that the sunflower genome is ~1 Gb larger than the *Zea mays* (maize) genome, what type of TE diversity resides in the sunflower genome? And what is the relative importance of selective removal versus selective amplification of TEs in shaping sunflower genome composition?

Here, we address these questions through a global survey of sequence composition and a fine-scale analysis of genomic structure. Specifically, we interrogated a large set of whole-genome shotgun (WGS) sequence reads representing approximately 25% of the sunflower genome, as well as the sequences of 21 unique bacterial artificial chromosome (BAC) clones. The random sequence reads allowed us to generate an unbiased and accurate estimate of sunflower genome composition, whereas the BAC sequences allowed for a detailed analysis of full-length TEs. We show that the sunflower genome is highly biased towards one superfamily of LTR-RTs, discuss the diversity of LTR-RT families identified in this study, and investigate the evolutionary time scales over which all types of LTR-RTs in this species appear to have been active. The sunflower-specific repeats identified in this study will aid in efforts to assemble the sunflower genome, which is currently being sequenced (Kane *et al.*,

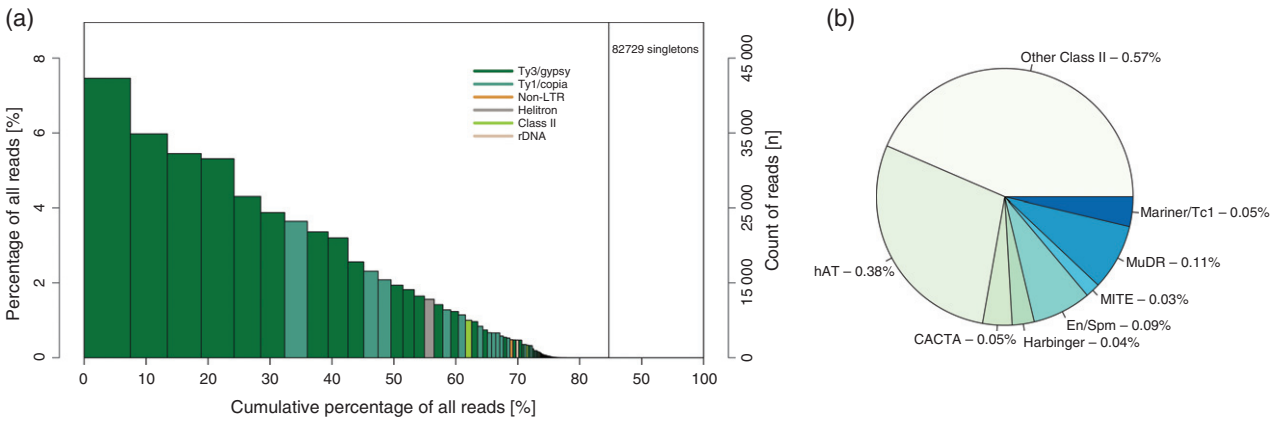

Figure 1. Repeat abundance based on 540 574 reads [a subset of all the whole-genome shotgun (WGS) reads; see Experimental procedures].
(a) Each bar in the histogram shows the individual size (height) of each cluster and the size relative to the total (width). The composition of each cluster is indicated by color, and single-copy, unclustered sequences are reflected to the right of the vertical bar.
(b) The genomic composition of subclass I of class-II transposable elements (TEs). The genome-wide abundance of each superfamily, based on the same subset of WGS reads as in (a), is shown because their low abundance made them difficult to visualize in (a).

2011), and will greatly improve future repeat-masking and gene annotation efforts in the Asteraceae.

RESULTS

Sunflower genome composition

We investigated repeat content and abundance in a collection of WGS reads corresponding to 0.23X coverage of the sunflower genome. Through our analyses we estimated that the sunflower genome is at least 81.1 ± 1.1% (mean ± SD) TEs and ribosomal repeats, with 77.7 ± 1.8% being composed of LTR-RTs, 57.9 ± 1.4% of which belong to the *Gypsy* superfamily (see Experimental procedures; Figure 1a). Conversely, subclass I (comprising all terminal inverted-repeat transposons) of class-II TEs and *Helitrons* (which are the only class-II, subclass-II TEs found in plants) accounted for just 1.3 ± 0.4% and 0.7 ± 1.6% of the genome, respectively (Figure 1a). Non-LTR retrotransposons appeared to occupy even less genomic space than class-II TEs, accounting for only 0.6 ± 0.4% of the sunflower genome, and were almost entirely composed of LINE-like lineages. Our graph-based analyses found that ~15% of the genome was single-copy, as represented by singletons, and an additional 4% of the genome was described as multi-copy genic sequences or low-copy transposable element families. The most abundant class-II, subclass-I TEs were the hAT and *Mutator* superfamilies, comprising 0.38 ± 0.04 and 0.11 ± 0.06% of the genome, respectively (Figure 1b).

In addition to analyzing the WGS data for repeat composition and abundance, we also analyzed the repeat composition of 21 BAC clones (~2.5 Mb), 20 of which were selected for sequencing because they carry genes of interest (see Experimental procedures). To characterize the diversity and demography of LTR retrotransposons in these BACs, we used both model-based and structure-based methods. All

Table 1 Statistics for long terminal repeat (LTR) retrotransposon superfamilies derived from bacterial artificial chromosome (BAC) clone sequences (top) and whole-genome shotgun (WGS) reads (bottom)

Super-family	Count	Overall length[a]	LTR length[a]	Percentage of BACs[b]	Solo:FL:TR[c]
Copia	28	9061	775	9.86 ± 10.6	0.53:1:0.03
Gypsy	79	9918	1551	30.47 ± 26.7	0.15:1:0.07
Total	107	9693	1346	40.33 ± 24.0	0.14:1:0.06

Super-family	Percentage of WGS reads[b]	LTR:RVT[d]
Copia	19.83 ± 2.8	2.27:1
Gypsy	57.93 ± 1.4	1.53:1
Total	77.75 ± 1.84	1.90:1

[a]Lengths are presented as the average (in bp).
[b]Percentage composition of BAC clones and WGS reads along with the standard deviation for each superfamily.
[c]Ratio of solo LTRs (Solo) to full-length (FL) to truncated (TR) LTR retrotransposon copies.
[d]The ratio of BLAST hits for LTR sequences (LTR) to reverse transcriptase (RVT) sequences from the WGS reads (see Experimental procedures).

BAC clones were composed of, on average, 40.3% intact LTR-RTs, with *Gypsy* families alone accounting for over 30% of the BAC clone sequences (Tables 1 and S2). The lower frequency of TEs in the BAC data was likely linked to the fact that the majority of these clones were selected for sequencing because they contained genes of interest, as noted above. We identified 16 families of LTR-RTs based on coding domain and terminal repeat similarity from intact and fragmented elements. The largest family, RLG-iketas,

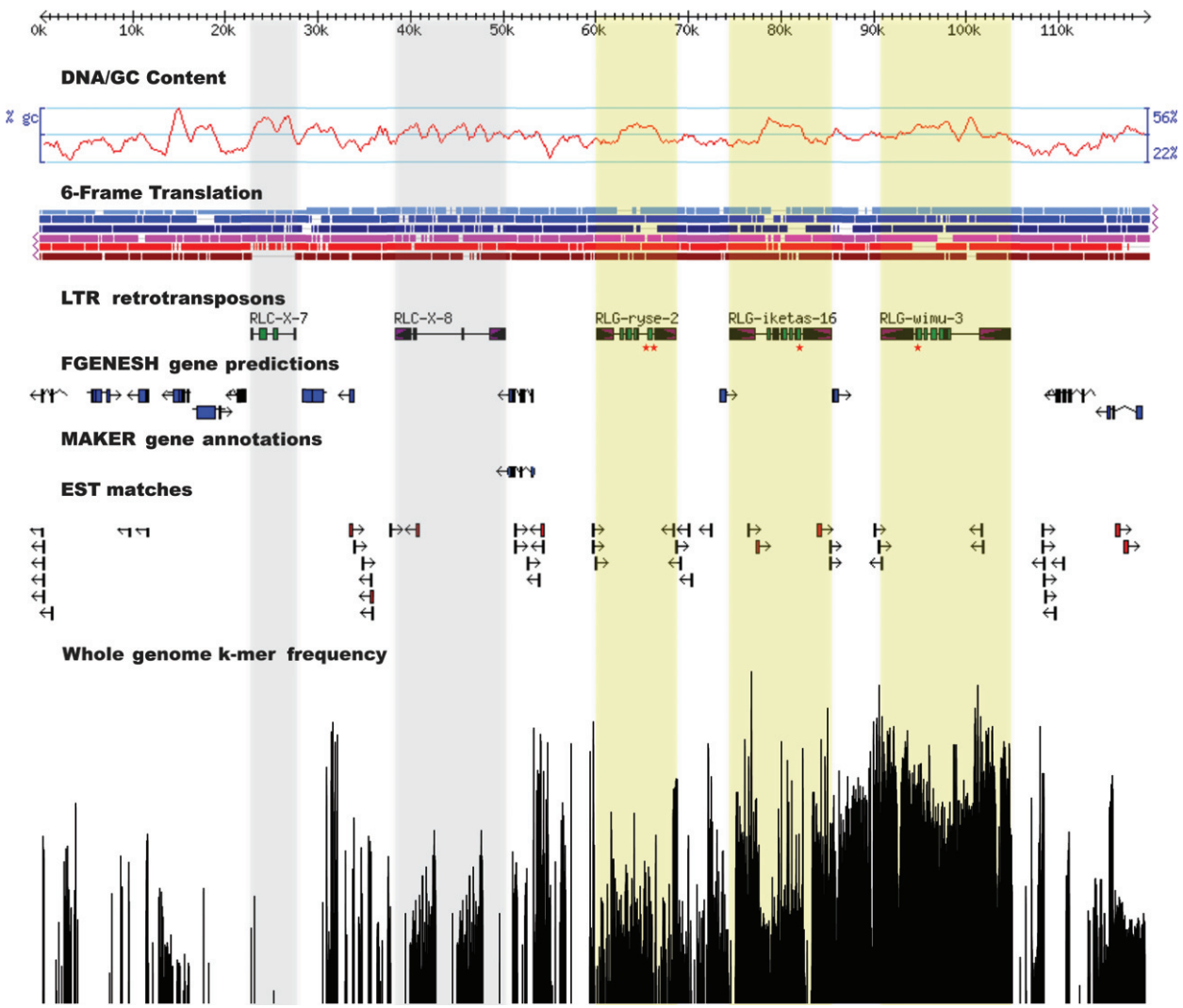

Figure 2. Fine-scale structure of bacterial artificial chromosome (BAC) clone 254L24 (see Table S1). The track displaying long terminal repeat (LTR) retrotransposons demonstrates the characteristic lack of coding domains (green) for *Copia* elements (columns shaded in gray), as compared with the prevalence of coding domains found in *Gypsy* (columns shaded in yellow) chromovirus sequences (location of chromodomains indicated with a star). The name above each element denotes its family designation (see Table S2). The bias in the EST matches to *Gypsy* elements, and the biased genomic abundance of these sequences are shown in the tracks below the predicted genes. Gene predictions were made using FGENESH (http://www.softberry.com) and MAKER (http://gmod.org/wiki/MAKER). The relative genome-wide frequency (plotted on a log scale) of genomic elements in this region is shown in the bottom track.

accounted for 19% of the LTR-RTs contained in the BAC clones analyzed. Consistent with the much lower frequency of the class-II transposable elements observed in the WGS data set, the BAC sequences contained only a single *Mutator* element, four putative *Helitrons* families of between two and four copies per family, and four putative MITE families of between five and eight copies per family. In total, *Helitrons* and MITEs accounted for just 0.09 and 0.12% of the total BAC sequences, respectively. To further investigate the genomic abundance of specific LTR-RT families identified in the BAC clones, we compared an index of *k*–mers from the WGS reads to the BAC clones (see Experimental procedures). In agreement with our estimates of family-level abundance based on the BAC clones, the WGS data have a high frequency of sequences matching the coding domains of *Gypsy* elements relative to *Copia* elements (Figures 2 and S1).

Demography of LTR retrotransposons in the sunflower genome

To better understand the dynamics of LTR-RTs during sunflower genome evolution, we analyzed the structure and age of all elements from the BAC clones analyzed, including those not belonging to any of the 16 families described here. The *Copia* superfamily had a higher percentage of solo LTRs compared with *Gypsy* elements (Tables 1 and S2). Although this result could potentially be an artifact of the non-random sample of BAC clones analyzed, Cavallini *et al.* (2010) also

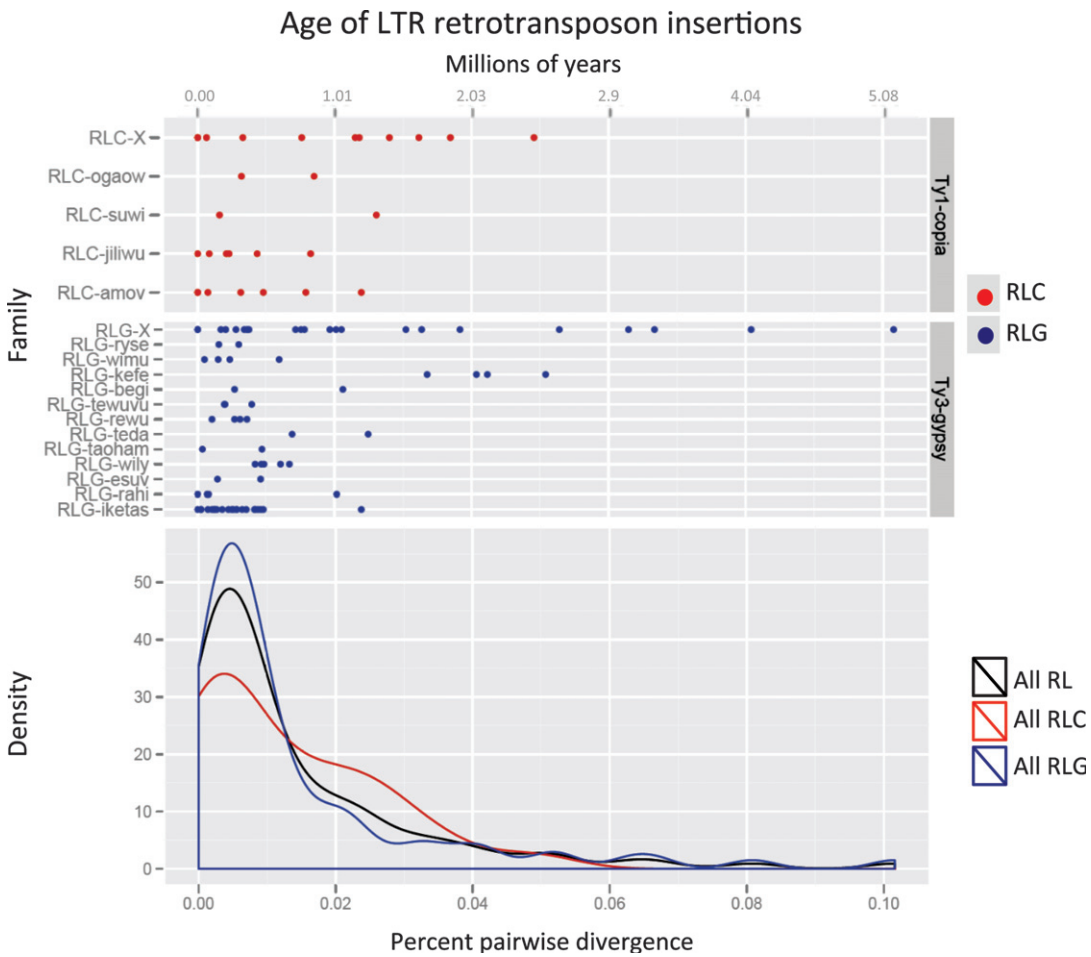

Figure 3. Age distribution of (LTR) retrotransposon insertions. The top panel shows the divergence between the LTRs of each individual retrotransposon insertion by family, and the bottom panel shows the same for each superfamily. The values along the lower *x*-axis represent the level of nucleotide divergence between the LTRs of each LTR-RT, whereas the values along the top *x*-axis represent the corresponding age of each element.

reported a similar finding using a hybridization-based approach. In addition, an analysis of solo LTRs on a genome-wide scale revealed that *Copia* solo LTRs and truncated elements appear to be more abundant than those from *Gypsy* elements, compared with intact elements (Table 1). The average length of the solo LTRs was just 200 bp, whereas the average length of all LTRs was 1346 bp (Table S2). All truncated LTR-RTs and solo LTRs appeared to have arisen within the past 1.4 Myr (0–1.4 Myr for solo LTRs and 0.28–1.18 Myr for truncated copies; as determined by the method described by Vitte *et al.*, 2007). In addition, an analysis of the age distribution of all LTR-RTs found that the majority of copies identified in this study arose within the past 1 Myr (Figure 3). Although many LTR-RT families were quite young (mean = 0.70 Myr), the mean age of individual families was >2 Myr in some cases (e.g. RLG-kefe; Figure 3; Table S2).

The chromodomain-containing *Gypsy* families accounted for over 55% of all *Gypsy* elements, and these particular *Gypsy* families were characterized by an absence of solo LTRs in our data set. Moreover, all but one family (RLG-ryse; Table S2) contained all of the coding domains necessary for activity. Although the BAC clones analyzed represent a non-random sample of the genome, this finding is unlikely to be artifactual, as a comparison with the WGS reads revealed a high frequency of sequences matching to the chromoviruses, including the chromovirus coding domains identified in this study (Figures 2 and S1). We infer that these retrotransposons are likely to be autonomous, based on the presence of multiple intact domains and translated open reading frames (ORFs) longer than 500 amino acids in 81.8% of the elements (22.7% contained translated ORFs longer than 1000 amino acids; see also Bachlava *et al.*, 2011), as well as evidence of transcriptional activity. Indeed, all chromoviruses also had at least eight, and as many as 26, unique matches to sunflower expressed sequence tags (ESTs), giving a total of 574 unique ESTs matching the chromovirus sequences identified in this study

(e.g. Figures 2 and S1), indicating that these sequences are expressed. This is in contrast to the *Copia* domain organization where only the reverse transcriptase and integrase were detectable. This latter finding may be related to the fact that the average age of *Copia* retrotransposons identified in this study was approximately twice the average age of the *Gypsy* superfamily described here (963 000 years versus 552 000 years).

Phylogenetic diversity and structure of chromoviruses in sunflower

Because over half of the ~3.5 Gb sunflower genome is likely to be composed of LTR retrotransposons belonging to a phylogenetic clade referred to as the chromoviruses, we asked whether there were yet unknown novel clades of chromoviruses in sunflower. We also pursued this question because previous studies of chromovirus diversity have focused on a biased sample of plant genomes, limited mainly to cereal crops and a few model dicot species (Gorinsek *et al.*, 2004; Novikova *et al.*, 2008). The phylogenetic placement of sunflower chromovirus sequences indicates that all sequences fall into known clades, with nearly all sequences belonging to the Tekay clade, whereas a single sequence falls in the Reina clade (see Supporting information).

The two recognized groups of chromodomains – groups I and II – are defined by the presence of three aromatic residues (Gao *et al.*, 2008). All plant chromodomains appear to lack the first of these residues, and some plant species also lack the third aromatic residue (Gorinsek *et al.*, 2004; Gao *et al.*, 2008; Novikova *et al.*, 2008). As in other plant chromoviruses, sunflower chromodomains lack the first aromatic residue (position 6; Figure 4) but contain the second aromatic residue, which is characteristic of group-II

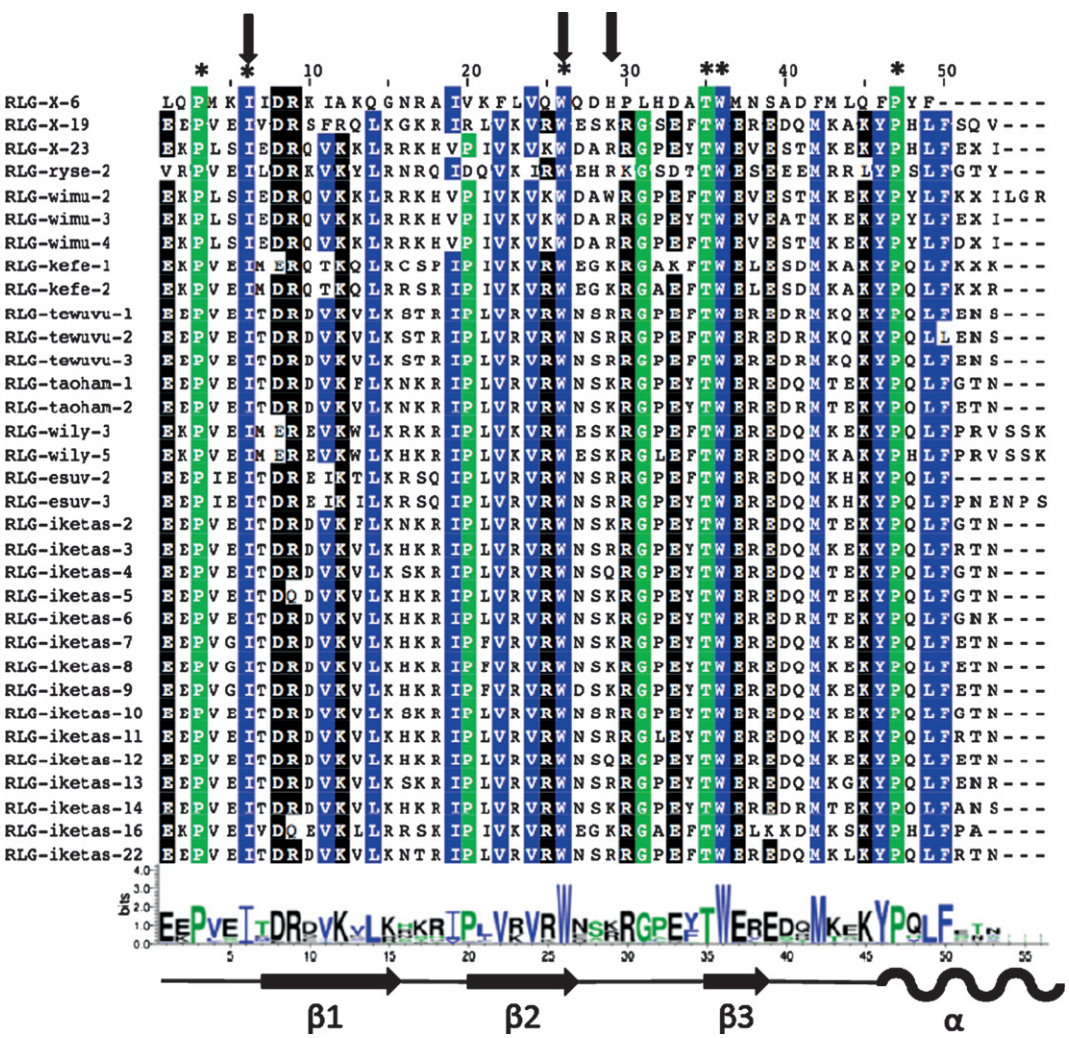

Figure 4. Alignment of sunflower chromodomain sequences. Only a single domain for each chromovirus was used in the alignment. Residues with conservation levels above 80% are highlighted, and the composition of each position is indicated in the sequence logo below the alignment. Aromatic residues characteristic of chromodomains are indicated with arrows (top), invariant sites are indicated with asterisks (top) and the predicted secondary structure is shown below the alignment.

chromodomains. One chromodomain (RLG-wimu-2; Figure 4) does contain a tryptophan at the third site, although this is not uncharacteristic of group-II chromodomains (Gao *et al.*, 2008). By aligning the chromodomains from sunflower with predicted chromodomain secondary structures, we inferred the structure of these domains (Ball *et al.*, 1997; Figure 4). This alignment of chromodomains revealed the presence of duplications of entire chromodomains within individual retrotransposons in the sunflower genome. Nearly 85% (28/33) of the chromoviruses contained a single duplication of the chromodomain, varying in length from 49 to 56 amino acids. Additionally, two chromoviruses from different BAC clones contained three perfect tandem duplications of a 53-amino-acid chromodomain: the amino acid sequence of the chromodomain for these two retrotransposons varied by a single residue at position 51. In contrast, only 9% (3/33) of the chromoviruses contained just one chromodomain (52 or 53 amino acids). This pattern is also evident when looking at the whole-genome level. For example, of the 4318 unique WGS reads with homology to a chromodomain, 74.4% were derived from a duplicated chromodomain [23.43% (1012/4318) with homology to a tandem chromodomain; 50.97% (2201/4318) with homology to more than two tandem chromodomains], as compared with 25.6% (1105/4318) being derived from a solo chromodomain. A phylogenetic analysis of duplications for all chromoviruses in sunflower revealed no evidence for multiple origins of tandem chromodomains (data not shown).

DISCUSSION

It is evident that the sunflower genome contains many thousands of retrotransposon copies (this study; Santini *et al.*, 2002; Natali *et al.*, 2006; Ungerer *et al.*, 2006), and numerous retrotransposon families are transcriptionally active in both cultivated (Vukich *et al.*, 2009) and wild populations (Kawakami *et al.*, 2011). However, there is a paucity of information regarding TE diversity and the mechanisms influencing the abundance of individual TE families in the sunflower genome. Thus, it seems clear that a comprehensive analysis of the diversity and dynamics of TEs would yield valuable insights into the role of TEs in the evolution of this important species.

Sunflower genome composition: pattern and process

Sunflower is distantly related to any plant species for which there is a curated set of genomic repeats (e.g. the estimated divergence time from *A. thaliana* is ~120 Myr, i.e. the divergence time between Asterids and Rosids; Cenci *et al.*, 2010). Therefore, to create a library of repeats for sunflower, we relied on a de novo repeat-finding method, rather than strictly homology-based methods (Novak *et al.*, 2010). To assess the composition of the sunflower genome we analyzed over 811 Mb of WGS reads (~0.23X genome coverage see Experimental procedures), using the method of Novak *et al.* (2010). LTR-RTs were the most abundant form of DNA in the sunflower genome, with the *Gypsy* superfamily alone accounting for ~58% of the genome (see also Cavallini *et al.*, 2010). Interestingly, analysis of intact LTR-RTs in BAC clone sequences revealed that the largest density of all LTR-RT insertions occurred within the last 1 Myr. That is, they arose since, or concomitantly with, the origin of sunflower as a species (Figure 3; Heesacker *et al.*, 2009). Although this dating procedure is an approximation, and may not reflect the true time since insertion, the finding of recent insertions is concordant with a previous study demonstrating that LTR-RTs are transcriptionally active in multiple wild populations of *H. annuus* and other annual sunflower species (Kawakami *et al.*, 2011). Although many insertions are likely to predate the origin of the *H. annuus* lineage (Figure 3), all insertions are within the age estimates for the origin of the genus *Helianthus* (i.e. the extant lineages arose 1.7–8.2 Ma; Schilling, 1997). Thus, the diversity and dynamics of LTR-RTs presented here are likely to reflect properties unique to the sunflower lineage, a finding consistent with those of Buti *et al.* (2011), where LTR-RT age was analyzed in three gene-harboring BAC clones. Biases towards recent (i.e. <5 Ma) LTR-RT insertions have also been noted in other plant genomes (Ma and Bennetzen, 2004; Vitte *et al.*, 2007; Wang and Liu, 2008; Du *et al.*, 2010), and this pattern likely reflects an ongoing struggle (i.e. 'genomic turnover') between the addition and removal of repetitive elements (Ma and Bennetzen, 2004).

We investigated how this process may have shaped the sunflower genome by analyzing the structure of LTR-RTs in order to assess the relative efficacy of unequal homologous recombination and illegitimate recombination in counteracting expansion of the sunflower genome. The formation of solo LTRs and truncated elements results from unequal homologous recombination between LTRs of a single LTR-RT or between elements at different genomic locations, respectively (Devos *et al.*, 2002; Bennetzen *et al.*, 2005), and this process appears to have been an effective DNA removal mechanism in the *Oryza sativa* (rice) and *Hordeum vulgare* (barley) genomes (Shirasu *et al.*, 2000; Vitte and Panaud, 2003). However, the process of illegitimate recombination, which involves microhomology, and occurs independently of the normal recombinational machinery, may have a greater impact on counteracting genome expansion through the formation of truncated elements (Chantret *et al.*, 2005), as appears to be the case in *A. thaliana* (Devos *et al.*, 2002) and *Medicago truncatula* (Wang and Liu, 2008).

In sunflower, solo LTRs and truncated LTR-RTs appeared to be in lower abundance than full-length elements (0.14:1.0:0.6 ratios of solo LTR:intact LRT-RT:truncated LTR-RT for all sunflower LTR-RTs; Table S2), as has been observed in maize (0.2:1.0 ratio of solo LTR:intact LTR-RT; SanMiguel *et al.*, 1996; Devos *et al.*, 2002). Solo LTRs were

also biased towards the *Copia* superfamily, and the majority of *Copia* solo LTRs analyzed (10/15) showed no divergence, suggesting a recent origin in our data set. In addition, a ratio of greater than 2:1 for LTR:reverse transcriptase sequences on a whole genome scale could indicate that: (i) *Copia* solo LTRs are more abundant that intact elements; (ii) there is a paucity of coding domains for *Copia* elements in the genome; or (iii) both of these factors contribute to the observed patterns, and the latter possibility is supported by our results from the analysis of 21 BAC clones (Tables 1 and S2). These differences in solo LTR formation between superfamilies may be driven by insertion preferences and LTR length (e.g. elements containing longer LTRs may be biased towards solo LTR formation; Vitte and Panaud, 2003; Du *et al.*, 2010), although *Copia* LTRs are half the length of *Gypsy* LTRs on average. In addition, the solo LTR fragments detected in this study averaged only 200 bp in length, which may reflect selection against the removal of larger stretches of DNA in genic regions (Tian *et al.*, 2009). Despite finding a paucity of solo LTRs, however, we did find a large number of deletions (278 in total, ranging from 10 to 17 bp each) flanked by short (4–9 bp) direct repeats (Figure S2; Table S3).

Although results from analyses of genomic structure can vary depending on the genomic regions being analyzed (e.g. Ma and Bennetzen, 2004, 2006), the foregoing findings highlight important processes that may be contributing to sunflower genome evolution. First, the observed bias in sunflower genome composition appears to have been driven, at least in part, by the selective removal of *Copia* LTR-RTs, as opposed to solely resulting from the amplification of *Gypsy* elements (Table S2). This result is supported by hybridization-based studies using *Gypsy* and *Copia* LTR sequences in sunflower (Cavallini *et al.*, 2010), and may have a significant impact on TE composition because solo LTR formation may remove more LTR-RT DNA than illegitimate recombination alone over short evolutionary time scales (Devos *et al.*, 2002). However, the frequency of putative illegitimate recombination events analyzed for the *Gypsy* and *Copia* superfamilies was proportional to their abundance (Tables S2 and S3). Second, our observation that solo LTRs were rare in regions harboring genes, where they might be expected to be more abundant (Tian *et al.*, 2009; Du *et al.*, 2010), suggests that illegitimate recombination may play an important role in regulating the DNA content in the sunflower genome. The high percentage of small deletions associated with sunflower LTR-RTs was also strongly suggestive of illegitimate recombination (Figure S2; Table S3). Even so, the relative importance of unequal homologous recombination and illegitimate recombination is likely to vary over evolutionary time (Tian *et al.*, 2009), and further investigation of the nature of recombination in sunflower will be required to determine the absolute genomic impact of these processes.

We also found a disproportional abundance of LINE-like lineages of non-LTR retrotransposons, as compared with the abundance of SINE-like lineages in our WGS data. In contrast, despite a slight bias towards the hAT superfamily, all types of class-II (subclass-I) TEs appear in nearly equal abundance (Figure 1b). This variation in proportionality may indicate differences in insertion preferences and host control between class-I and class-II TEs in sunflower.

Chromovirus structures and their potential impact on the sunflower genome

Chromoviruses appear to be the most abundant lineage of *Gypsy* LTR-RTs among flowering plants (Gorinsek *et al.*, 2004; Kordis, 2005); this pattern was concordant with our observations in sunflower, where over 55% of intact *Gypsy* elements identified in the BAC sequences contained a chromodomain. Based on work in *Schizosaccharomyces pombe*, it has been shown that chromodomains mediate the integration of chromovirus sequences by interacting with dimethyl and trimethylated lysine-9 residues on histone H3, an epigenetic mark of heterochromatin (Gao *et al.*, 2008). Notably, the most highly conserved residues of chromodomains in sunflower chromoviruses, four of which are invariant, reside within the regions predicted to mediate interactions with methylated lysine residues on histone H3 (Figure 4; Jacobs and Khorasanizadeh, 2002; Nielsen *et al.*, 2002).

Interestingly, nearly 85% of the chromovirus sequences identified in the BAC sequences contain at least one tandem duplication of the chromodomain, and nearly 75% of the chromodomain-derived sequences identified in the WGS reads appear to have been derived from tandem arrays of chromodomains. Given that tandem chromodomains recognize methylated lysine-4 on histone H3 in *Drosophila* and humans, which is a mark of transcriptionally active euchromatin (Flanagan *et al.*, 2005, 2007), and that the abundance of elements with duplicated chromodomains is marginally higher in gene-containing BACs versus the genome as a whole, it is tempting to infer that a similar function could be employed by certain sunflower chromovirus sequences. Analyses of randomly selected BAC clones could provide insight into the genome-wide co-occurrence of chromoviruses and genes. This finding also raises the possibility that chromatin remodeling factors associated with sunflower chromoviruses could potentially lend to their stability in the genome (Lippman *et al.*, 2004), and help to explain the biased composition of TEs in the sunflower genome. Whether these findings represent yet unknown active targeting mechanisms for chromoviruses or are the result of aberrant integration arising from mutations (i.e. duplication of the chromodomain), it is evident that these sequences have played an active and presumably continuing role in shaping the sunflower genome.

EXPERIMENTAL PROCEDURES

WGS and BAC clone sequencing

In order to obtain an unbiased estimate of the sunflower genome composition, 2 325 196 random genomic sequences (i.e. WGS sequences; mean length 403 bp, GC 39.05%; ~811 Mb in total) were generated via Roche 454 GS FLX (Roche, http://www.roche.com) sequencing of a highly inbred line derived from sunflower cultivar HA412-HO (PI 642777) using XLR (Titanium) chemistry. With the exception of sequences showing similarity to rDNA genes and organellar genomes (see below), all of these sequences were used in the analysis of genome composition.

Twenty-one BAC clones from sunflower cultivar HA383 (PI 578872) were selected for sequencing based on the presence of genes of evolutionary and/or agronomic importance (Table S1). BAC clones were prepared using standard protocols (Bachlava *et al.*, 2011; Blackman *et al.*, 2011). Sixteen of these BAC clones were sequenced using a Sanger shotgun approach at either Washington University or the Joint Genome Institute, with automatic and manual finishing. Assembly and editing were carried out with PHRAP and CONSED, respectively (Ewing and Green, 1998; Ewing *et al.*, 1998; Gordon *et al.*, 1998). Four additional clones were sequenced in the Georgia Genomics Facility using a Roche 454 GS FLX sequencer with XLR (Titanium) sequencing chemistry. Final assemblies were generated with MIRA 3.0.3 (Chevreux *et al.*, 1999; see Supporting information for details). The final BAC clone was selected by probing the same sunflower BAC library (filter Ha_HBa_A) with a *Gypsy* integrase sequence fragment and selecting a clone address exhibiting a strong hybridization signal. Sequencing, assembly, and editing of this BAC clone were performed at the Clemson University Genomics Institute (CUGI). The WGS and BAC clone sequences described above are available for download at http://www.sunflower.uga.edu/data.

Repeat identification from WGS and BAC clone sequences

All sequences containing chloroplast, mitochondrial or ribosomal fragments were removed using BLAST similarity searches and custom PERL scripts (Altschul *et al.*, 1990); low-complexity sequences were removed with the DUST algorithm (Hancock and Armstrong, 1994). First, to identify putative repeat families, a graph-based clustering method was applied to the cleaned, reduced set of genomic sequences (2 088 836 in total; Novak *et al.*, 2010). Despite having removed ribosomal and low complexity sequences, clustering was not feasible on the full data set because of computational requirements, so the data were split into four subsets containing ~500 000 sequences each. Briefly, clustering was performed by first using an all-by-all search with MGBLAST with the following parameters: –F 'm D' –D 4 –p 85 –W18 –UT –X40 –KT –JF –v90000000 –b90000000 –C80 –H 320 –a 8 (Pertea *et al.*, 2003; Novak *et al.*, 2010). Next, a custom script was used to select read pairs that had at least 90% identity and covered at least 15% of the length of the matching sequences. The bitscore for read pairs that passed these thresholds was used for clustering with the methods and software described by Novak *et al.* (2010). Lastly, all clusters containing at least 500 reads were assembled using GSASSEMBLER 2.5.3 (Roche), and contigs were searched for coding domains with HMMSCAN 2.3.2 (Eddy, 1998) using the translated nucleotide sequences as a query against the Pfam database (release 24.0; Finn *et al.*, 2010). We also performed nucleotide searches (BLASTN searches with an e-value of $1e^{-5}$) with the contigs using a custom repeat database, comprising Repbase 15.06 (Jurka *et al.*, 2005), mips-REdat 4.3 (Spannagl *et al.*, 2007) and the JCVI maize characterized repeats V4.0 (http://maize.jcvi.org/repeat_db.shtml), as the target. The size and composition of clusters for each of the four subsets showed very little variation with respect to abundance; thus, we have reported the abundance of each transposable element type as an average of the subsets, as well as the standard deviation for each estimate.

The program LTR_FINDER (Xu and Wang, 2007) was used with default settings, and executed with the batch_ltrfinder.pl script from DAWGPAWS (Estill and Bennetzen, 2009), in order to discover intact LTR retrotransposons from the BAC clones. In addition, LTRHARVEST 1.3.4 (Ellinghaus *et al.*, 2008) was used to discover LTR-RTs using the default settings, except for the following parameter changes: –mintsd 4 –mindistltr 4000 –maxlenltr 4000. Given that Ellinghaus *et al.* (2008) demonstrated a higher rate of true positive recovery with LTRHARVEST when combined with a clustering step, as compared with other LTR-RT prediction methods, and that LTR_Finder recovered a low percentage of elements with TSDs, the output of LTRHARVEST was used to search for binding sites and coding domains. To identify coding regions within the predicted retrotransposons, the program LTRDIGEST (Steinbiss *et al.*, 2009) was run on the LTR-RTs predicted by LTRHARVEST. Complete, or intact, LTR-RTs were defined as having at minimum of two flanking TSDs, two nearly intact LTRs, a primer binding site and a poly purine tract (see Ma *et al.*, 2004). Solo LTRs and truncated LTR-RTs were identified by searching the BAC clone sequences with the full-length LTR-RTs (see Supporting information). Putative sites of illegitimate recombination were identified by first aligning all full-length members of an LTR-RT family (see below), and then comparing (with the BLAST program BL2SEQ) the 20 bp of sequence upstream and downstream of gap sites for direct repeats. To eliminate artifacts, we only analyzed gap sites of >10 bp that were flanked by direct repeats of >4 bp, which had no more than two non-matching bases intervening the matching repeats and a gap (see also Devos *et al.*, 2002; Ma *et al.*, 2004). Deletions shared by more than one element were assumed to represent an ancestral event, and were counted once (Ma *et al.*, 2004).

The LTR-RT superfamilies (e.g. *Gypsy* and *Copia*) were constructed using evidence from matches to Hidden Markov Models (HMMs) for the Reverse Transcriptase (RVT) domain and matches to the custom repeat database described above. LTR-RT families were identified by clustering separately the primer binding site, the 5′ LTR sequence and internal coding domains (i.e. gag, reverse transcriptase, integrase, RNase H and chromodomain) with VMATCH (http://vmatch.de) following the methods described in Steinbiss *et al.* (2009). All LTR-RT families were named according to Wicker *et al.* (2007). Each LTR-RT copy that could not be unambiguously assigned to a family but could be assigned to a superfamily (see Wicker *et al.*, 2007) was classified as RLG-X or RLC-X for *Gypsy* unclassified or *Copia* unclassified, respectively. The procedure for dating each LTR-RT family was adapted from Vitte *et al.* (2007) and Baucom *et al.* (2009), but also see SanMiguel *et al.* (1996). Briefly, the K80 model (Kimura, 1980) within the BaseML module of PAML 4.2a (Yang, 2007) was used to obtain a likelihood divergence estimate for each LTR-RT based on the similarity of the two LTRs. This divergence value (which we will refer to as *d*) was used to determine age with the formula $T = d/2r$, where $r = 1.0 \times 10^{-8}$, as determined for host-encoded genes (Strasburg and Rieseberg, 2008), and the multiplier of two accounts for the elevated rates of evolution of TEs, as compared with genes (Baucom *et al.*, 2009). Putative class-II transposons and *Helitrons* were identified using MITEHunter as well as through similarity searches using HMMER and INTERPROSCAN (Eddy, 1998; Zdobnov and Apweiler, 2001), and HELSEARCH (Yang and Bennetzen, 2009), respectively.

To compare the frequency of intact repeats identified from BAC clones with their frequency in the whole genome, we generated 20-mers for each BAC clone and compared those sequences with an index of 20-mers from all of the WGS reads using TALLYMER

(Kurtz *et al.*, 2008). Plotting the relationship between the length of *k*-mers and the uniqueness ratio for each value of *k* from 1 to 100 revealed a natural inflection at *k = 20*, similar to the maize genome (Kurtz *et al.*, 2008), representing a value that would maximize the information and resolution in the *k*-mers being compared (Kurtz *et al.*, 2008). Custom PERL scripts were then used to format matches between the WGS index and BAC clone 20-mers for viewing in GBROWSE 2.40 (Figure 2; Stein *et al.*, 2002). The genome-wide frequency of solo LTRs was estimated with similarity searches using BLAST, where the WGS read set was the subject and the LTR and reverse transcriptase sequences (from intact LTR-RTs identified in the BAC clones) were used as the query (see Supporting information). This same procedure was used for determining the relative frequency of chromodomain duplications in the genome wherein the sequences of single and tandemly duplicated chromodomains (identified in the BAC clone sequences) were used to interrogate the WGS reads. A unique match in the WGS reads was scored as single if it had only a single matching region up to the length of a chromodomain, and tandem matches were scored by the presence of two (or more) regions where one match begins at the end site of the previous match. All scripts described herein are available upon request.

ACKNOWLEDGEMENTS

We kindly thank Dr Dusan Kordis for sharing plant chromovirus sequences with us, as well as Navdeep Gill and members of the Burke Laboratory for comments on an earlier version of the article. This work was supported by grants from the National Science Foundation (DBI-0820451 to J.M.B., S.J.K. and L.H.R.; DEB-0742993 to M.C.U.) as well as the USDA National Institute of Food and Agriculture (2008-35300-19263 to J.M.B.).

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Figure S1. Fine-scale structure of BACs a) P102A12, a) P189P24, and a) Contig36_P245O15 (see Table S1).

Figure S2. Depiction of shared and non-shared direct repeats flanking deletions, representing putative cases of illegitimate recombination.

Table S1. Length statistics for BAC clone sequences.

Table S2. Demography statistics for LTR retrotransposon families derived from BAC clone sequences.

Table S3. Statistics for putative events of illegitimate recombination for each LTR-RT family.

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

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Baack, E.J., Whitney, K.D. and Rieseberg, L.H. (2005) Hybridization and genome size evolution: timing and magnitude of nuclear DNA content increases in *Helianthus* homoploid hybrid species. *New Phytol.* **167**, 623–630.

Bachlava, E., Radwan, O.E., Abratti, G., Tang, S., Gao, W., Heesacker, A.F., Bazzalo, M.E., Zambelli, E., Leon, A.J. and Knapp, S.J. (2011) Downy mildew (PI8 and PI14) and rust (RAdv) resistance genes reside in close proximity to tandemly duplicated clusters of non-TIR-like NBS-LRR-encoding genes on sunflower chromosomes 1 and 13. *Theor. Appl. Genet.* **122**, 1211–1221.

Ball, L.J., Murzina, N.V., Broadhurst, R.W., Raine, A.R., Archer, S.J., Stott, F.J., Murzin, A.G., Singh, P.B., Domaille, P.J. and Laue, E.D. (1997) Structure of the chromatin binding (chromo) domain from mouse modifier protein 1. *EMBO J.* **16**, 2473–2481.

Baucom, R.S., Estill, J.C., Leebens-Mack, J. and Bennetzen, J.L. (2009) Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome. *Genome Res.* **19**, 243–254.

Belyayev, A., Kalendar, R., Brodsky, L., Nevo, E., Schulman, A.H. and Raskina, O. (2010) Transposable elements in a marginal plant population: temporal fluctuations provide new insights into genome evolution of wild diploid wheat. *Mob. DNA*, **6**, 1.

Bennetzen, J.L. (2000) Transposable element contributions to plant gene and genome evolution. *Plant Mol. Biol.* **42**, 251–269.

Bennetzen, J.L. (2007) Patterns in grass genome evolution. *Curr. Opin. Plant Biol.* **10**, 176–181.

Bennetzen, J.L., Ma, J. and Devos, K.M. (2005) Mechanisms of recent genome size variation in flowering plants. *Plant Mol. Biol.* **95**, 127–132.

Biemont, C. (2009) Are transposable elements simply silenced or are they under house arrest? *Trends Genet.* **25**, 333–334.

Blackman, B.K., Rasmussen, D.A., Strasburg, J.L., Raduski, A.R., Burke, J.M., Knapp, S.J., Michaels, S.D. and Rieseberg, L.H. (2011) Contributions of flowering time genes to sunflower domestication and improvement. *Genetics*, **187**, 271–287.

Buti, M., Giordani, T., Cattonaro, F., Cossu, R.M., Pistelli, L., Vukich, M., Morgante, M., Cavallini, A. and Natali, L. (2011) Temporal dynamics in the evolution of the sunflower genome as revealed by sequencing and annotation of three large genomic regions. *Theor. Appl. Genet.* **5**, 779–791.

Cavallini, A., Natali, L., Zuccolo, A., Giordani, T., Jurman, I., Ferrillo, V. *et al.* (2010) Analysis of transposons and repeat composition of the sunflower (*Helianthus annuus* L.) genome. *Theor. Appl. Genet.* **120**, 491–508.

Cenci, A., Combes, M.C. and Lashermes, P. (2010) Comparative sequence analysis reveals that Coffea (Asterids) and Vitis (Rosids) derive from the same paleo-hexaploid ancestral genome. *Mol. Genet. Genomics*, **283**, 493–501.

Chantret, N., Salse, J., Sabot, F., *et al.* (2005) Molecular basis of evolutionary events that shaped the *Hardness* locus in diploid and polyploidy wheat speces (Triticum and Aegilops). *Plant Cell*, **17**, 1033–1045.

Chevreux, B., Wetter, T. and Suhai, S. (1999) Genome sequence assembly using trace signals and additional sequence information. *Comput. Sci. Biol. Proc. German Conf. Bioinf.* **99**, 45–56.

Devos, K.M. (2010) Grass genome organization and evolution. *Curr. Opin. Plant Biol.* **13**, 139–145.

Devos, K.M., Brown, J.K.M. and Bennetzen, J.L. (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res.* **12**, 1075–1079.

Du, J., Tian, Z., Hans, C.S., Laten, H.M., Cannon, S.B., Jackson, S.A., Shoemaker, R.C. and Ma, J. (2010) Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison. *Plant J.* **63**, 584–598.

Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Ellinghaus, D., Kurtz, S. and Willhoeft, U. (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, **9**, 18.

Estill, J.C. and Bennetzen, J.L. (2009) The DAWGPAWS pipeline for the annotation of genes and transposable elements in plant genomes. *Plant Methods*, **5**, 8.

Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.* **8**, 186–194.

Ewing, B., Hillier, L., Wendl, M.C. and Green, P. (1998) Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.* **38**, D211–D222.

Flanagan, J.F., Mi, L., Chruszcz, M., Cymborowski, M., Clines, K.L., Kim, Y., Minor, W., Rastinejad, F. and Khorasanizadeh, S. (2005) Double chromodomains cooperate to recognize the methylated histone H3 tail. *Nature*, **438**, 1181–1185.

Flanagan, J.F., Blus, B.J., Kim, D., Clines, K.L., Rastinejad, F. and Khorasanizadeh, S. (2007) Molecular implications of evolutionary differences in CHD double chromodomains. *J. Mol. Biol.* **369**, 334–342.

Gao, X., Hou, Y., Ebina, H., Levin, H.L. and Voytas, D.F. (2008) Chromodomains direct integration of retrotransposons to heterochromatin. *Genome Res.* **18**, 359–369.

Gordon, D., Abajian, C. and Green, P. (1998) Consed: a graphical tool for sequence finishing. *Genome Res.* **8**, 195–202.

Gorinsek, B., Gubensek, F. and Kordis, D. (2004) Evolutionary genomics of chromoviruses in eukaryotes. *Mol. Biol. Evol.* **21**, 781–798.

Hancock, J.M. and Armstrong, J.S. (1994) SIMPLE34: an improved and enhanced implementation for VAX and Sun computers of the SIMPLE algorithm for analysis of clustered repetitive motifs in nucleotide sequences. *Comput. Appl. Biosci.* **10**, 67–70.

Heesacker, A.F., Bachlava, E., Brunick, R.L., Burke, J.M., Rieseberg, L.H. and Knapp, S.J. (2009) Karyotypic evolution of the common and silverleaf sunflower genomes. *Plant Genome*, **2**, 233–246.

Hilbrict, T., Varotto, S., Sgaramella, V., Bartels, D., Salamini, F. and Furini, A. (2008) Retrotransposons and siRNA have a role in the evolution of desiccation tolerance leading to resurrection of the plant *Craterostigma plantagineum*. *New Phytol.* **179**, 877–887.

Hollister, J.D. and Gaut, B.S. (2009) Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res.* **19**, 1419–1428.

Hollister, J.D., Smith, L.M., Guo, Y., 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. *Proc. Natl Acad. Sci. USA*, **108**, 2322–2327.

Hua-Van, A., Le Rouzic, A., Boutin, T.S., Filee, J. and Capy, P. (2011) The struggle for life of the genome's selfish architects. *Biol. Direct*, **6**, 19.

Jacobs, S.A. and Khorasanizadeh, S. (2002) Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail. *Science*, **295**, 2080–2083.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467.

Kane, M.C., Gill, N., King, M.E., Bowers, J.E., Berges, H., Gouzy, J., Bachlava, E. *et al.* (2011) Progress towards a reference genome for sunflower. *Botany*, **89**, 429–437.

Kawakami, T., Dhakal, P., Katterhenry, A.N., Heatherington, C.A. and Ungerer, M.C. (2011) Transposable element proliferation and genome expansion are rare in contemporary sunflower hybrid populations despite widespread transcriptional activity of LTR retrotransposons. *Genome Biol. Evol.* **3**, 156–167.

Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120.

Kordis, D. (2005) A genomic perspective on the chromodomain-containing retrotransposons: Chromoviruses. *Gene*, **347**, 161–173.

Kumar, A. and Bennetzen, J.L. (1999) Plant Retrotransposons. *Annu. Rev. Genet.* **33**, 479–532.

Kurtz, S., Narechania, A., Stein, J.C. and Ware, D. (2008) A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics*, **9**, 517.

Lippman, Z., Gendrel, A., Black, M., Vaughn, M.W., Dedhia, N., McCombie, W.R. *et al.* (2004) Role of transposable elements in heterochromatin and epigenetic control. *Science*, **430**, 471–476.

Ma, J. and Bennetzen, J.L. (2004) Recent growth and divergence of rice nuclear genomes. *Proc. Natl Acad. Sci. USA*, **101**, 12404–12410.

Ma, J. 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., 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.

McClintock, B. (1984) The significance of responses of the genome to challenge. *Science*, **226**, 792–801.

Naito, K., Zhang, F., Tsukiyama, T., Saito, H., Hancock, C.N., Richardson, A.O., Okumoto, Y., Tanisaka, T. and Wessler, S.R. (2009) Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature*, **461**, 1130–1134.

Natali, L., Santini, S., Giordani, T., Minelli, S., Maestrini, P., Cionini, P.G. and Cavallini, A. (2006) Distribution of Ty3-*gypsy*- and Ty1-*copia*-like DNA sequences in the genus *Helianthus* and other Asteraceae. *Genome*, **49**, 64–72.

Nielsen, P.R., Nietlispach, D., Mott, H.R., Callaghan, J., Bannister, A., Kouzarides, T., Murzin, A.G., Murzina, N.V. and Laue, E.D. (2002) Structure of the HP1 chromodomain bound to histone H3 methylated at lysine 9. *Nature*, **416**, 103–107.

Novak, P., Neumann, P. and Macas, J. (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics*, **11**, 378.

Novikova, O., Mayorov, V., Smyshlyaev, G., Fursov, M., Adkison, L., Pisarenko, O. and Blinov, A. (2008) Novel clades of chromodomain-containing Gypsy LTR retrotransposons from mosses (Bryophyta). *Plant J.* **56**, 562–574.

Pereira, V., Enard, D. and Eyre-Walker, A. (2009) The effect of transposable element insertions on gene expression evolution in rodents. *PLoS One*, **4**, e4321.

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y. *et al.* (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, **19**, 651–652.

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.

Piegu, B., Guyot, R., Picault, N., Roulin, A., Saniyal, A., Kim, H., Collura, H. *et al.* (2006) Doubling genome size without polyploidization: dynamics of retrotransposon-mediated genome expansions in *Oryza australensis*, a wild relative of rice. *Genome Res.* **16**, 1262–1269.

SanMiguel, P., Tikhonov, A., Jin, Y., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A. *et al.* (1996) Nested retrotransposons in the intergenic regions of the maize genome. *Science*, **274**, 765–768.

Santini, S., Cavallini, A., Natali, L., Minelli, S., Maggini, F. and Cioini, P.G. (2002) *Ty1*- and *Ty3/gypsy*-like retrotransposon sequences in *Helianthus* species. *Chromosoma*, **111**, 192–200.

Schilling, E.E. (1997) Phylogenetic analysis of *Helianthus* (Asteraceae) based on chloroplast restriction-site data. *Theor. Appl. Genet.* **94**, 925–933.

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.

Spannagl, M., Noubibou, O., Haase, D., Yang, L., Gundlach, H., Hindemitt, T., Klee, K., Haberer, G., Schoof, H. and Mayer, K.F.X. (2007) MIPSPlantsDB—plant database resource for integrative and comparative plant genome research. *Nucleic Acids Res.* **35**, D834–D840.

Staton, S.E., Ungerer, M.C. and Moore, R.C. (2009) The genomic organization of Ty3/gypsy-like retrotransposons in Helianthus (Asteraceae) homoploid hybrid species. *Am. J. Bot.* **96**, 1646–1655.

Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.* **10**, 1599–1610.

Steinbiss, S., Willhoeft, U., Gremme, G. and Kurtz, S. (2009) Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Res.* **37**, 7002–7013.

Strasburg, J. and Rieseberg, L.H. (2008) Molecular demographic history of the annual sunflowers Helianthus annuus and H. petiolaris - large effective population sizes and rates of long-term gene flow. *Evolution*, **62**, 1936–1950.

Suoniemi, A., Tanskanen, J. and Schulman, A.H. (1998) Gypsy-like retrotransposons are widespread in the plant kingdom. *Plant J.* **13**, 699–705.

Tian, Z., Rizzon, C., Du, J., Zhu, L., Bennetzen, J.L., Jackson, S.A., Gaut, B.S. and Ma, J. (2009) Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? *Genome Res.* **19**, 2221–2230.

Ungerer, M.C., Strakosh, S.C. and Zhen, Y. (2006) Genome expansion in three hybrid sunflower species is associated with retrotransposon proliferation. *Curr. Biol.* **16**, R872–R873.

Ungerer, M.C., Strakosh, S.C. and Stimpson, K.M. (2009) Proliferation of Ty3/gypsy-like retrotransposons in hybrid sunflower taxa inferred from phylogenic data. *BMC Biol.* **7**, 40.

Vitte, C. and Panaud, O. (2003) Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice Oryza sativa L. *Mol. Biol. Evol.* **20**, 528–540.

Vitte, C., Panaud, O. and Quesneville, H. (2007) LTR retrotransposons in rice (*Oryza sativa*, L.): recent burst amplifications followed by rapid DNA loss. *BMC Genomics*, **8**, 218.

Voytas, D.F., Cummings, M.P., Konieczny, A., Ausubel, F.M. and Rodermel, S.R. (1992) Copia-like retrotransoposons are ubiquitous among plants. *Proc. Natl Acad. Sci. USA*, **89**, 7124–7128.

Vukich, M., Giordani, T., Natali, L. and Cavallini, A. (2009) *Copia* and *Gypsy* retrotransposons activity in sunflower (*Helianthus annuus* L.). *BMC Plant Biol.* **9**, 150.

Wang, H. and Liu, J. (2008) LTR retrotransposon landscape in *Medicago truncatula*: more rapid removal than in rice. *BMC Genomics*, **9**, 382.

Warenfors, M., Pereira, V. and Eyre-Walker, A. (2010) Transposable elements: insertion pattern and impact on gene expression evolution in Hominids. *Mol. Biol. Evol.* **27**, 1955–1962.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A. *et al.* (2007) A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **12**, 973–982.

Wicker, T., Taudien, S., Houben, A., Keller, B., Graner, A., Platzer, M. and Stein, N. (2009) A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. *Plant J.* **59**, 712–722.

Xiong, Y. and Eickbush, T.H. (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.* **9**, 3353–3362.

Xu, Z. and Wang, H. (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268.

Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591.

Yang, L. and Bennetzen, J.L. (2009) Distribution, diversity, evolution and survival of Helitrons in the maize genome. *Proc. Natl Acad. Sci. USA*, **106**, 19922–19927.

Zdobnov, E.M. and Apweiler, R. (2001) InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.

Zeh, D.W., Zeh, J.A. and Ishida, Y. (2009) Transposable elements and an epigenetic basis for punctuated equilibria. *BioEssays*, **31**, 715–726.

