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

S. Evan Staton*, Bradley H. Bakken†, Benjamin K. Blackman², Mark A. Chapman³, Nolan C. Kane⁵, Mark C. Ungerer², Steven J. Knapp⁶*, Loren H. Rieseberg⁵ and John M. Burke⁴

1Department of Genetics, University of Georgia, Athens, GA 30602, USA,
2Division of Biology, 426 Ackert Hall, Kansas State University, Manhattan, KS 66506, USA,
3Department of Biology, Jordan Hall, 1001 East Third Street, Indiana University, Bloomington, IN 47405, USA,
4Department of Plant Biology, University of Georgia, Athens, GA 30602, USA,
5The Biodiversity Research Centre and Department of Botany, 3529-6270 University Blvd, University of British Columbia, Vancouver, BC, Canada V6T 1Z4, and
6Institute 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 46288, 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
but 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.,
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\% (mean SD) TEs and ribosomal repeats, with 77.7\% 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\% and 0.7\% 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\% 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\% and 0.11\% of the genome, respectively (Figure 1a).

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 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,
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
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 by 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 TE 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 39.65%; 8,111 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, mitochondrion or ribosomal fragments were removed using BLAST similarity searches and custom PERL scripts (Altitsch 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 sequences. The K80 model (Kimura, 1980) within the BaseML module of PAML 4.2a (Yang, 2007) was used to obtain a likelihood divergence dating each LTR-RT family was adapted from Vitte et al., 2007) 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 LTR sequence and internal coding domains (i.e. gag, reverse transcriptase, integrase, RNase H and chromodom) 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 RLK-X or PLC-X for Gypsy unclassified or Copia unclassified, respectively. The procedure for dating each LTR-RT family was adapted from Vitte et al. (2007) and Baucoum 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 (Baucoum 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 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_ltfinder.pl script from Juneau (Estill and Bennetzen, 2009), in order to discover intact LTR retrotransposons from the BAC clones. In addition, LTRHAVEST 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 the Blat 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 BLASTX) 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 LTR sequence and internal coding domains (i.e. gag, reverse transcriptase, integrase, RNase H and chromodom) 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 RLK-X or PLC-X for Gypsy unclassified or Copia unclassified, respectively. The procedure for dating each LTR-RT family was adapted from Vitte et al. (2007) and Baucoum 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 TE's, as compared with genes (Baucoum et al., 2009). Putative class-II transposons and Helitrons were identified using the 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.
Figure S1. Table S3. Table S2. Demography statistics for LTR retrotransposon families
Figure S2. Please note: As a service to our authors and readers, this journal
et al.

k

viewing in GBROWSE 2.40 (Figure 2; Stein

The Plant Journal

ª


Sunflower transposon abundance and diversity


© 2012 The Authors

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 chromodomains duplications in the genome wherein the sequences of single and tandemly duplicated chromodomains (identified in the BAC clone sequences) where 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_P245015 (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


