Heredity (2013) 110, 194–204

www.nature.com/hdy

Heredity

ORIGINAL ARTICLE

The dynamics of LTR retrotransposon accumulation across 25 million years of panicoid grass evolution

MC Estep[1], JD DeBarry[1] and JL Bennetzen

Sample sequence analysis was employed to investigate the repetitive DNAs that were most responsible for the evolved variation in genome content across seven panicoid grasses with >5-fold variation in genome size and different histories of polyploidy. In all cases, the most abundant repeats were LTR retrotransposons, but the particular families that had become dominant were found to be different in the *Pennisetum*, *Saccharum*, *Sorghum* and *Zea* lineages. One element family, *Huck*, has been very active in all of the studied species over the last few million years. This suggests the transmittal of an active or quiescent autonomous set of *Huck* elements to this lineage at the founding of the panicoids. Similarly, independent recent activity of *Ji* and *Opie* elements in *Zea* and of *Leviathan* elements in *Sorghum* and *Saccharum* species suggests that members of these families with exceptional activation potential were present in the genome(s) of the founders of these lineages. In a detailed analysis of the *Zea* lineage, the combined action of several families of LTR retrotransposons were observed to have approximately doubled the genome size of *Zea luxurians* relative to *Zea mays* and *Zea diploperennis* in just the last few million years. One of the LTR retrotransposon amplification bursts in *Zea* may have been initiated by polyploidy, but the great majority of transposable element activations are not. Instead, the results suggest random activation of a few or many LTR retrotransposons families in particular lineages over evolutionary time, with some families especially prone to future activation and hyper-amplification.
Heredity (2013) **110,** 194–204; doi:10.1038/hdy.2012.99

Keywords: DNA amplification; genome size variation; polyploidy; repetitive DNA; sample sequence analysis; transposable elements

INTRODUCTION

Flowering plant (angiosperm) genomes are enormously unstable at the levels of chromosome number, genome size and repetitive DNA content. In maize (*Zea mays*), barley (*Hordeum vulgare*) and other grasses with genomes >2000 Mb, most genes exist as single-gene islands that are surrounded by seas of nested transposable elements (TEs) (SanMiguel and Bennetzen, 1998). Where haplotype variation has been investigated in maize, any two alleles of the same gene that have diverged for >2 million years differ by >50% in their contents of flanking TEs (Wang and Dooner, 2006). Gene content and organization are more stable, but still vary substantially, especially in copy number and gene order (Bennetzen, 2007; Springer *et al.*, 2009).

Over the last 15 years, the primary mechanisms of genome rearrangement have been discovered (reviewed in Bennetzen, 2007). Polyploidy is a frequent and dramatic contributor to genome variation. Although some lineages can undergo fixed (that is, successful) polyploid events several times in just a few million years, other lineages escape this process for tens of millions of years. For example, the last polyploidy in the sorghum (*Sorghum bicolor*) lineage was about 70 million years ago (mya), many millions of years before the origin and broad diversification of the grass family (Paterson *et al.*, 2004). Beyond doubling genome size, polyploidy has been observed to serve as a 'genomic shock' that activates TE amplification and resultant genome rearrangement, possibly through altering the balance in their epigenetic silencing (O'Neill *et al.*, 1998; Ozkan *et al.*, 2001; Madlung *et al.*, 2005; Parisod *et al.*, 2009; Petit *et al.*, 2010). After polyploidy, an eventual diploidization process occurs that leads to exclusive disomic inheritance and the loss of a subset of the genes that were, for instance, doubled in nuclear copy number by a diploid to tetraploid polyploid event. This gene loss is not random, involving a 'fractionation' where genes are lost more frequently (in any given chromosomal domain) from one parent of the tetraploid rather than the other and also involving preferential loss of genes that encode dosage-sensitive proteins (Thomas *et al.*, 2006; Schnable *et al.*, 2011).

However, the major determinants of genome structure in angiosperms have been shown to act on a more rapid time scale than even recurrent polyploidy. TE amplification and removal are the major determinants of genome size in grass lineages, for instance (reviewed in Bennetzen *et al.*, 2005), and this correlation appears to hold generally across the flowering plants. In most angiosperm genomes, the LTR retrotransposons are the most significant contributor to genome size, contributing over 75% of the nuclear DNA to even moderate-sized genomes like maize (Schnable *et al.*, 2009). Most LTR retrotransposons families exist in low copy numbers (SanMiguel and Bennetzen, 1998; Baucom *et al.*, 2009), but the amplification of a few families that individually contribute >100 Mb of DNA to a genome are the major causes of 'genomic obesity' (Bennetzen and Kellogg, 1997) in plants. LTR retrotransposon amplifications by only one or a

Department of Genetics, University of Georgia, Athens, GA, USA
[1]These authors contributed equally to this work.
Correspondence: Dr JL Bennetzen, Department of Genetics, University of Georgia, Davison Hall, Athens, GA 30602, USA.
E-mail: maize@uga.edu
Received 18 June 2012; revised 22 October 2012; accepted 23 October 2012

few families are sufficient to more than double genome size in a just a few million years, as shown for *Oryza australiensis* (Piegu *et al.*, 2006).

Genomic DNA removal can also be exceedingly rapid. LTR retrotransposons commonly mutate to solo LTRs by unequal recombination, especially in regions (for example, near genes) where homologous recombination is a frequent process (Ma and Bennetzen, 2006). However, the major mechanism for DNA removal involves small deletions associated with illegitimate recombination (Devos *et al.*, 2002), which has been shown to remove hundreds of megabases of LTR retrotransposon DNA in as little as 2 million years (Ma *et al.*, 2004). This process acts across the entire genome (including in those genes lost in the fractionation process; Ilic *et al.*, 2003), leaving highly degenerate legacies of earlier genome constituents that are peppered with small deletions and thus become unrecognizable within a few million years. Hence, any TEs observed in a grass genome must have been active in the last 5–10 million years, or much more recently, or they would no longer be detectable. It seems likely that regions composed mostly of degenerated fragments of LTR retrotransposons and other TEs, are responsible for most of the 'unannotated' DNA in a genome, although many TE fragments have evolved host-beneficial roles (Hudson *et al.*, 2003; Bundock and Hooykaas, 2005), especially in gene regulation (White *et al.*, 1994; Michaels *et al.*, 2003; Lisch and Bennetzen, 2011).

The combination of very active DNA removal and very active TE amplification creates an exceptionally dynamic genome balance. In many lineages, genomes seem to be tending primarily toward growth, while others appear to be shrinking (Leitch *et al.*, 1998; Kellogg and Bennetzen, 2004; Hawkins *et al.*, 2009). We do not know why any particular plant lineage is trending in one direction or another, whether these trends are caused by novel patterns in TE amplification or removal, or the degree to which selection on genome size (Bennett, 1972) plays a role in this process. As a prerequisite to understanding the processes that differentially regulate genome composition, studies are needed to investigate the details of genome dynamics in a set of closely related and genetically tractable species.

The panicoid grass lineage is about 26 million years old (Bennetzen *et al.*, 2012), and includes such important crops as maize (*Zea mays*), sorghum (*Sorghum bicolor*), sugarcane (*Saccharum* spp.) and pearl millet (*Pennisetum glaucum*). The maize and sorghum genomes have been sequenced and extensively annotated (Paterson *et al.*, 2009; Schnable *et al.*, 2009), so they provide foundations for genetic analyses within the panicoids. For any full-genome to full-genome comparison, a single species provides one data point. Hence, any full-genome analysis in the panicoids requires analyses of multiple species, but this is expensive at the full-genome level with current genome sequencing technologies. Sample sequence analysis (SSA) (Brenner *et al.*, 1993; Devos *et al.*, 2005; Liu *et al.*, 2007) uses statistical analysis of a small and randomly chosen set of DNA molecules to provide an alternative to full-genome studies. Because full-genome analysis usually under-represents repetitive DNAs (due to challenges in their assembly), a randomly chosen set of DNAs contributing to an SSA can provide a more accurate description of the repetitive DNAs than even a 'completed' genome sequence (Liu and Bennetzen, 2008).

This manuscript reports an SSA of the content and evolution of the major repetitive DNAs in seven panicoid grasses, with primary concentration on the Andropogoneae tribe that includes maize, sorghum and sugarcane. The primary questions investigated are (1) the nature of the repetitive DNA content of these genomes, (2) how these repeat contents differ qualitatively and quantitatively and (3) the timing and molecular mechanisms responsible for the lineage specificity of repeat content. The results provide a window on the dynamics of genome size variation, indicating TE amplification tendencies that are established at documentable times in specific lineages, combined with a dominant trend towards amplifications of LTR retrotransposon families that are not highly amplified in closely related lineages.

MATERIALS AND METHODS

Sample preparation and sequence acquisition

Plant materials for DNA preparation were derived from seed provided by K Devos of the University of Georgia (pearl millet inbred TIFT23DB), HJ Price of Texas A&M University (*Sorghum propinquum* designated TAMU) and John Doebley of the University Wisconsin-Madison for the same teosinte lines used to estimate nuclear DNA content for *Zea luxurians* (Iltis G-5 and G-42) and *Zea diploperennis* (Iltis 1190) (Laurie and Bennett, 1985). Nuclear DNA from sugarcane (*Saccharum officinarum*) was provided by A D'Hont from CIRAD in Montpellier, France (variety Black Cheribon). Like maize and sorghum, all of these samples are from diploid species, except sugarcane, which is an octoploid with a high frequency of aneuploid derivatives. All three *Zea* species share an allotetraploidy that occurred <12 mya (Figure 1) (Swigonova *et al.*, 2004), but now behave as true diploids

For nuclear DNA preparation, ~ 100 g of leaf tissue from greenhouse-grown plants were harvested and used for isolation of nuclei following a standard protocol (Peterson *et al.*, 2000). DNA was then isolated from the nuclei and randomly sheared using a GeneMachine Hydroshear (Genomic solutions, Bath, UK) set on speed code 14 for 20 cycles to obtain 3–5 kb DNA fragments. The sheared fragments were treated with Mung bean nuclease (New England Biolabs, Ipswich, MA, USA), size selected on a 1% agarose gel, dephosphorylated using shrimp alkaline phosphatase (Roche, Bradford, CT, USA), and A-tailed using Taq DNA polymerase (Roche) and dATP (Roche). The modified fragments were ligated into a Topo-4 cloning vector (Invitrogen, Carlsbad, CA, USA) and transformed into ElectroMAX DH10B cells

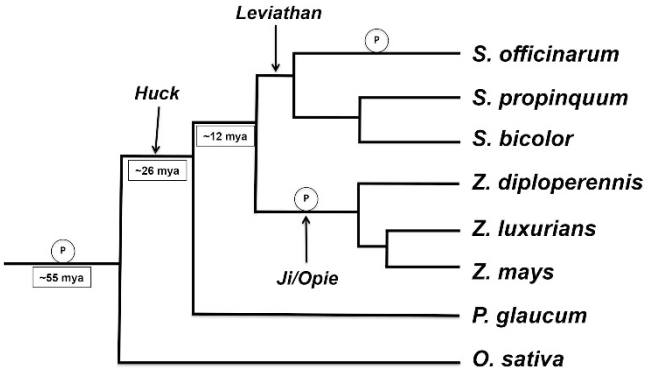

Figure 1 A cladogram indicating the relatedness of selected grass species, including points of TE family 'activation' and polyploidy. The dates indicate times of divergence for some important lineages. Circled '**P**' indicates approximate timings and precise lineages of polyploid events. Arrows indicate the approximate timing of LTR retrotransposon family transmission in a form with high potential for subsequent activation. The arrows do not refer to when the indicated TE was transposing, because such activity could only be detected with discernible elements if it occurred in the last few million years (Ma *et al*., 2004). Whether *Z. luxurians* is more closely related to *Z. diploperennis* or to *Z. mays* is controversial, but this drawing reflects our bias of a closer *Z. mays* relationship that is supported by the amplification of a specific *Huck* subfamily that is shared by *Z. luxurians* and *Z. mays*, but not by *Z. diploperennis*. Aside from this minor point regarding the relative relatedness of *Z. diploperennis*, *Z. luxurians* and *Z. mays*, the cladogram structure and the dates of polyploid events are all from previous publications (Swigonova *et al.*, 2004; Vicentini *et al.*, 2008). Arrows indicating the timing of the passage of an activatable TE family are derived exclusively from the data and analyses in this manuscript.

(Invitrogen). Three 384-well plates of clones were randomly chosen and sequenced from one direction using the T7 primer and BigDye terminator v3.1 (Applied Biosystems, Foster City, CA, USA) for the *Zea* analysis. For the other panicoid grasses, two 384-well plates were chosen and clones were sequenced from both directions, except for sugarcane where four plates were sequenced because of small insert sizes for many clones. The electropherograms obtained from the ABI3730 sequencing machine (Applied Biosystems) were analyzed with Phred (Ewing *et al.*, 1998) for base calling. Low-quality, vector, chloroplast and mitochondrial sequences were identified with Phred and Cross_match and removed from the data sets before submission to the GSS division of GenBank (accession numbers JY127741—JY133169 and JY133584—JY136902).

Whole genome sequences were also obtained for *Zea mays* (inbred B73) by downloading data from the 3–4 kb unfiltered genomic shotgun data set (NCBI accession # 33825241–34849215) from GenBank (Schnable *et al.*, 2009). A custom PERL script (available upon request) was used to randomly extract 1152 sequences from the data set without replacement.

Repeat discovery and assembly

An all-versus-all BLASTn (Altschul *et al.*, 1990) of each data set was used to assess the copy number of sequences within the samples. BLAST hits were required to show at least 70% identity over at least 100 bp to be counted (hits of a sequence to itself were ignored). Sequences were grouped arbitrarily into four categories: one copy, 2–3 copies, 4–9 copies and $\geq$10 copies. This grouping does not directly determine the copy number of sequences in the target genome. Because of the small sample size, even sequences that are present only once in the sample data may have multiple copies within the genome.

For *Zea*, where repeat sequences have been well-characterized, the repeat groups were then annotated using Cross_Match and five repeat databases (MAGI v3.1, TIGR v2.0 (Ouyang and Buell, 2004), TREP release 7, and the Maize Transposable Element database (Baucom *et al.*, 2009) (http://maizetedb.org/~maize/)). Sequences with an e-value of 1E–05 or better were annotated with the best hit for each repeat database. The percent of identified sequences in each library was calculated by dividing the number of sequences identified by the total number of sequences in that library (both repetitive and single copy). The TIGR Plant Repeat Database was able to identify most sequences from uncollapsed repeat data.

For the repeats in the panicoid grasses other than *Zea*, where excellent species-specific repeat databases were not available, all-versus-all BLASTn groupings were used to assemble repeats with the program AAARF (DeBarry *et al.*, 2008). Testing was initiated with default parameters. Final AAARF parameters were chosen after multiple tests were run for each species. Briefly, these optimal parameters were a requirement for BLAST hits to be at least 150 nt long, have a minimum sequence identity of 85%, and a maximum e-value of 1E–05. The MCS for each stage of extension was required to be at least 150 nt long with a coverage depth of at least 2. Due to the sparse data, only a single sequence was required for build extension, with a minimum length of 150 nt. Individual build steps were allowed to extend no more than 200 nt. For pairwise sequence BLASTs, the minimum hit size was 90 and the maximum e-value allowed was 1E–05. Each sequence was available for five rounds of extension in both directions. This parameter set was found to produce optimal builds among those tested for use with a sparse sample data set. The effectiveness of different sets of AAARF parameters was evaluated based on the amount of the data utilized in different tests and how well AAARF-produced builds represented known repeats.

AAARF builds belonging to known repeat families in the target species or in related species were identified via BLAST searches against several databases. BLAST reports were parsed using custom PERL scripts, and at least 60% identity over a minimum of 100 bp with a maximum e-value of 1E–05 was required to identify a hit. At least 80% continuous coverage along the build by a single, distinct family of repeats was required to identify a build as belonging to a particular repeat family. To minimize false-positive identifications, BLAST reports for all builds were also manually inspected based on the above criteria. All of the most abundant builds exhibited high homology with either no previously known family (rarely) or, most commonly, with a single previously characterized family of grass repeats.

Because LTRs are present on either end of a full-length LTR retrotransposon, LTR sequences will typically be at least twice as abundant as the internal regions within SSA data. Further, sequences from both LTRs of an element will be collapsed in a single LTR region in an AAARF build (DeBarry *et al.*, 2008). Because of this abundance of LTR retrotransposon sequence relative to the internal regions, an increase in coverage along a build is one indicator of an LTR region. Build coverage was inspected by BLASTing the builds to the SSA data used to create them and using a modified version of the AAARF algorithm to create a coverage matrix representing the number of hits to each nucleotide position along the build. A combination of sequence similarity to the 3′ end of a tRNA sequence (including the 'CCA' tRNA cap), the proximal presence of the 'CA' motif at the 3′ LTR end and at least a twofold coverage increase (relative to the putative primer binding site) along the build beginning immediately at the 'CA' motif was used to classify builds as fully intact LTR retrotransposons.

Evaluating repeat abundances across genomes

Estimates of the amount of Mb per genome for each of the highly repetitive elements were calculated using the annotated sample sequences. Hits were required to have a minimum length of 50 bp, a minimum identity of 85%, and a maximum e-value of 1E–05. In previous studies (Baucom *et al.*, 2009), it has been shown that 85% identity over 50 bp leads to absolute separation of all LTR retrotransposon families, with no intermingling of the results from separate families.

For the *Zea* comparisons, where quantitation was important, the annotation information was transformed into a repeat percentage for each sequence by dividing the repeat length in that sequence by the total sequence read length. The transformed data were then bootstrapped using SAS with 1000 permutations. The values produced in the bootstrap statistic were multiplied by genome size for each library. The 1C/1N genome size values utilized were all from the Kew C value database (http://data.kew.org/cvalues): *P. glaucum* (2620 Mb), *S. bicolor* (730 Mb), *S. propinquum* (740 Mb), *S. officianarum* (3960 Mb for the octoploid 1C/4N genome), *Z. diploperennis* (2590 Mb); *Z. mays* (2365 Mb) and *Z. luxurians* (4470 Mb). The mean and a 95% confidence interval for repeat quantities in each species or genotype were then graphed to display the genome comparisons and test the null hypothesis that the two samples being compared have equal amounts (Mb) of the TE family. If the 95% confidence interval in any pairwise comparison did not overlap, we rejected the null hypothesis and argue that the samples are significantly different in the amount of the TE family being compared.

Because sorghum and maize have excellent repeat databases, masking of SSA data was employed to find and quantify repeats using the prototypic repeat representatives from the Repbase Update data (AFA Smit, R Hubley and P Green RepeatMasker at http://repeatmasker.org). A custom PERL script (R Hubley, pers. comm.) returned the percent sample masked by each repeat.

Retroelement phylogenetic analysis

Annotated sequences of retroelements were translated into all six reading frames and then searched by BLASTp with a translated copy of the reverse transcriptase or the integrase genes to find the sequences in the data set that could be used to reconstruct a phylogeny for each high copy retroelement. The BLAST results with the highest number of sequence hits were aligned in clustalX and trimmed to incorporate the largest number of taxa with the longest alignment. Neighbor joining trees were constructed in PAUP using the default settings (using an uncorrected '*p*' distance matrix with ties broken systematically) and 1000 bootstrap replicates (Swofford, 2002). TE sequences from *Sorghum* (the closest related species with sequence data) were used as the out group for the resulting trees. In the same manner, a nucleotide alignment and NJ tree were also produced for the 180-bp knob repeat.

For comparison with *S. officinarum*, *S. propinquum* and *P. glaucom* repeats assembled with AAARF, SSA data were produced *in silico* for *Z. mays* and *S. bicolor.* Ten thousand random unfiltered shotgun sequence reads from maize (average read length 782 bp, accession numbers EI697885.1—EI684889.2) were downloaded from TIGR. For sorghum, 10 000 random sequences (average read length 975 bp) were downloaded from the NCBI GSS database (http://www.ncbi.nlm.nih.gov/projects/dbGSS/).

A custom database was used to identify LTR retrotransposon sequences from the five panicoid grasses' SSA data. The database was assembled from Panicoid-specific LTR retrotransposons from Repbase Update (Jarka *et al.*, 2005), *Zea* and *Sorghum* retrotransposons from the TIGR Plant Repeat Databases (Ouyang and Buell, 2004) (from http://plantrepeats.plantbiology.msu.edu/), and the full MIPS-REdat database (v. 4.3) (http://mips.helmholtz-muenchen.de/plant).

RESULTS

The four most abundant repeats in five panicoid grass species

AAARF assemblies (for *P. glaucum, S. officinarum* and *S. propinquum*) and genome sequence inspection (for *S. bicolor* and *Z. mays*) provided the results shown in Table 1. In even the smallest genomes, those of the two *Sorghum* species, several repeats were found to account for >1% each of the total genome. In each case, the largest contribution was from an LTR retrotransposon family, although this was a different family in each genus.

The *Huck* family was found to be an abundant element in most of the panicoids investigated, including among the top four in maize and pearl millet, and the sixth most abundant LTR retrotransposon in *S. propinquum*. However, the *Huck* element is only a middle-repetitive DNA in *S. bicolor* (Peterson *et al.*, 2002) and was not seen at all in our sugarcane data set (data not shown). This element is absent from the rice (*Oryza sativa*) genome, even at an e-value of 1E–01. *Leviathan* is a shared most abundant family among the *S. propinquum* and *Saccharum* species, but has a much lower copy number in *S. bicolor* (Peterson *et al.*, 2002). *Leviathan* is a middle-repetitive DNA in B73 maize, but none of the element is intact (that is, with two alignable LTRs), so this TE has not been active for a very long time, and thus was missed with the intact element discovery pipeline applied by Baucom *et al.* (2009). BLASTn analysis with the *Leviathan* LTR in rice yielded 14 candidate homologs, with the lowest e-value homology observed at 4.3E–10 (data not shown). *Ji* and *Opie* homologs are found in both *S. bicolor* and *O. sativa*, but there copy numbers do not exceed 50 in any of these lineages, and they are usually found as highly degenerate TEs without intact structures (data not shown). Figure 1 shows a cladogram with approximate divergence dates developed in earlier studies by Kellogg and coworkers (Vicentini *et al.*, 2008), with all of the investigated panicoid species, and indicates apparent timing for the potentiation (propensity for future activation) of specific LTR retrotransposon families. Most of these major activations do not correlate with the history of polyploidy in these lineages, with the possible exception of the *Ji* and *Opie* activation that appears to be basal to the *Zea* lineage (see below).

The most abundant repeats in the genomes of maize and two teosinte species

Although all three of the *Zea* species investigated in this study are current diploids, and shared a last polyploidization a few million years ago (Swigonova *et al.*, 2004), their genomes now vary greatly in size. *Z. luxurians* accessions G-5 and G-42 were measured as 4481 and 4525 Mb, compared with 2589 Mb and 2365 Mb for *Z. diploperennis* accession 1190 and B73 maize, respectively (Laurie and Bennett, 1985). The reason for this genome size variation is not known. Hence, we sequenced 1112 randomly chosen plasmid clones from G-5 and 1122 randomly chosen plasmid clones from G-42 for *Z. luxurians*. The sequenced *Z. diploperennis* and *Z. mays* clones numbered 1085 and 1152, respectively. Average read lengths for these four genomes sampled were a respective 722, 744, 771 and 672 bp.

The all-versus-all BLASTn indicated that 29–35% of the sequences are not highly repetitive, 9–10% of the sequences were found in two copies, 31–41% of the sequences were in 3–10 copies, and 15–29% of the sequences were in very high copy number (11 or more sequences within each library). All of the repetitive sequences were annotated for the nature of the repeat.

Although four repeat databases were used to annotate the sequences, the TIGR Plant Repeat Database (Ouyang and Buell, 2004) was able to identify most of the sequences in the library (63.1–69.9%). The remaining three repeat databases primarily confirmed the annotation that was derived from the TIGR database. MAGI was the only database that provided annotations not available within the TIGR repeat database. All of the sequences that were uniquely annotated with MAGI were statistically defined repeats that have not been experimentally verified. Moreover, they were found only as a subset of the 2–3 copy repeat class, and hence were not used in downstream analysis. Using the TIGR annotation and the information from the all-versus-all results, each sequence was grouped into a specific class (Table 2). The non-highly repetitive sequences composed ~22–29% of the four sampled genomes. LTR retrotransposons contributed the largest percentage of the samples, ranging from ~51–61%, with *Z. mays* having the largest percentage of the four sampled. The percentage of knob sequence in the samples ranged from 1.1% in *Z. mays* to almost 17% in *Z. luxurians* (G-42), indicating a very high copy number (~150 000 copies in B73 maize

Table 1 The four most abundant LTR retrotransposon families in five panicoid genomes

Species (line)	*P. glaucom*		*Z. mays (B73)*		*S. bicolor*		*S. propinquum*		*S. officianarum*	
Genome size in mb	*~2800*		*~2365*		*~730*		*~740*		*~990 (1N)*	
	Ofovin	~4.5%[a]	*Huck*	~24.8%	*Evum*	~11.0%	*Omor*	~4.7%	*Leviathan*	~3.3%
	copia	~126.0[b]	*gypsy*	~586.5	*gypsy*	~80.3	*Unknown*	~34.8	*gypsy*	~32.3
	Adyrog	~4.1%	*Ji*	~8.2%	*Omor*	~9.7%	*Leviathan*	~2.7%	*Giepum*	~1.6%
	Unknown	~114.8	*copia*	~193.9	*Unknown*	~70.8	*gypsy*	~20.0	*copia*	~15.8
	Juriah	~3.1%	*Grande*	~5.7%	*Onap*	~3.8%	*Evum*	~2.2%	*Gypsor1*	~1.1%
	gypsy	~86.8	*gypsy*	~134.8	*gypsy*	~27.7	*gypsy*	~16.3	*gypsy*	~10.9
	Huck	~2.6%	*Opie*	~3.8%	*Gypsy*-136_SBi-I	~3.7%	*Onap*	~1.4%	*Angela*	~0.7%
	gypsy	~72.8	*copia*	~89.9	*gypsy*	~27.0	*gypsy*	~10.4	*copia*	~6.9

Abbreviations: SSA, sample sequence analysis; TE, transposable element.
[a]% of SSA data set for each TE family.
[b]Estimated mb of genome (from % of SSA) for each TE family.

Table 2 Comparisons of major retrotransposon groups and specific families among the four *Zea* samples by sequence count and % of total data

Seq types	*Z. diploperennis*		*Z. luxurians (G-5)*		*Z. luxurians (G-42)*		*Z. mays (B73)*	
	Sequence #	*% Sample*	*Sequence #*	*% Sample*	*Sequence #*	*% Sample*	*Sequence #*	*% Sample*
gypsy	345	31.8	353	31.7	289	25.8	450	39
copia	240	22.1	218	19.6	265	23.6	236	20.5
LINE	0	0	0	0	0	0	1	0.1
Total	585	53.9	571	51.3	554	49.4	687	59.6
copia elements								
Dagaf	13	1.2	2	0.2	5	0.5	5	0.4
Eninu	0	0	2	0.2	1	0.1	1	0.1
Fourf	3	0.3	6	0.5	4	0.4	3	0.3
Giepum	3	0.3	2	0.2	5	0.5	10	0.1
Ji	108	10	84	7.6	91	8.1	109	9.5
Opie	73	6.7	61	5.5	68	6.1	59	5.1
PREM	32	3	52	4.7	74	6.6	40	3.5
Rire1	0	0	0	0	3	0.3	2	0.2
Ruda	6	0.6	5	0.5	11	1	4	0.4
Sto	0	0	1	0.1	1	0.1	1	0.1
Victim	2	0.2	3	0.3	0	0	0	0
Total	240	22.1	218	19.6	263	23.6	234	20.5
gypsy elements								
Bogu	0	0	1	0.1	0	0	0	0
CentA	2	0.2	2	0.2	1	0.1	0	0
Cinful	50	4.6	42	3.8	36	3.2	37	3.2
Diguus	4	0.4	5	0.5	6	0.5	5	0.4
Grande	39	3.6	32	2.9	24	2.1	49	4.3
Gyma	54	5	24	2.2	24	2.1	18	1.6
Huck	74	6.8	139	12.5	105	9.4	241	20.9
Kake	0	0	0	0	0	0	2	0.2
Milt	1	0.1	9	0.8	8	0.7	8	0.7
Rire1	12	1.1	1	0.1	0	0	0	0
Shadowspawn	5	0.5	5	0.5	2	0.2	2	0.2
Tekay	8	0.7	8	0.7	7	0.6	11	1
Xilon	33	3	39	3.5	33	2.9	38	3.3
Zeon	63	5.8	46	4.1	41	3.7	39	3.4
Total	345	31.8	353	31.7	287	25.8	450	39

to ~4 320 000 copies in G-42 *Z. luxurians*) for this 180 bp tandem repeat. Each of the four samples also contained unknown repeats ranging from ~8–10% of the sample. The remaining three groups, ribosomal repeats, centromere-specific repeats and DNA transposons, were of relatively minor abundance and showed no statistically significant variation across the samples.

Although the above analysis indicates that repetitive DNAs, especially TEs, are the major determinants of genome size variation in these *Zea* species, it does not indicate which of the many hundreds of TE families have been the primary contributors to this variation. In order to better understand the relationship between these repeats and genome size variation in the genus *Zea*, the most abundant repetitive elements were used to estimate their total quantitative contributions to each genome. The *copia* elements were found to have contributed ~425–485 Mb to the *Z. diploperennis* and *Z. mays* genomes, in contrast to ~750–920 Mb to the two *Z. luxurians* genotypes. The difference in relative abundance of these elements between the smaller genome group and the larger genome group is statistically significant (based on a permutation test with a 95% cutoff value). The *gypsy* elements account for ~670 Mb of the *Z. diploperennis* genome, ~830 Mb of the *Z. mays* genome, and ~985–1200 Mb of the *Z. luxurians* genotypes. In this analysis, differences between both of the smaller genomes and between the smaller and larger genome groups are also statistically significant. The 180-bp knob repeat contributes only ~27 Mb to the B73 maize genome, compared to ~117 Mb for the *Z. diploperennis* genome and ~588–778 Mb for the two *Z. luxurians* genomes. The ribosomal repeat group showed no statistically significant differences between any of the genomes, with estimates of 15–47 Mb.

Specific repeat family and subfamily contributions to genome size variation in *Zea*

Eleven families of *copia* LTR retrotransposons and 14 *gypsy* families were identified in the highly repetitive SSA category. Table 3 shows the four most abundant repeats in each of these genomes. As can be seen, individual families show major quantitative differences across the *Zea*

Table 3 The four most abundant LTR retrotransposon families in four different *Zea* genomes

Species (line)	*Z. diploperennis*		*Z. luxurians (G-5)*		*Z. luxurians (G-42)*		*Z. mays (B73)*	
Genome size in mb	*~2589*		*~4481*		*~4525*		*~2365*	
	Ji	~10.0%[a]	*Huck*	~12.5%	*Huck*	~9.4%	*Huck*	~20.9%
	copia	~258.9[b]	*gypsy*	~560.1	*gypsy*	~425.4	*gypsy*	~494.3
	Huck	~6.8%	*Ji*	~7.6%	*Ji*	~8.1%	*Ji*	~9.5%
	gypsy	~176.0	*copia*	~340.6	*copia*	~366.5	*copia*	~224.7
	Opie	~6.7%	*Opie*	~5.5%	*Prem*	~6.6%	*Opie*	~5.1%
	copia	~173.5	*copia*	~246.5	*copia*	~298.7	*copia*	~120.6
	Zeon	~5.8%	*Prem*	~4.7%	*Opie*	~6.1%	*Grande*	~4.3%
	gypsy	~150.1	*copia*	~210.6	*copia*	~276.0	*gypsy*	~101.7

Abbreviations: SSA, sample sequence analysis; TE, transposable element.
[a]% of SSA data set for each TE family.
[b]Estimated mb of genome (from % of SSA) for each TE family.

species, but are largely congruent for the families that are the most abundant. Moreover, the data for B73 maize by this analysis largely agrees with the *in silico* analysis presented in Table 1. The same procedure used to estimate Mb for the major groups of repeats were followed individually for the most abundant *copia* and *gypsy* families (Figure 2). The *Ji* family is estimated at ~200–300 Mb in the four *Zea* genomes, with statistically significant increases in the *Z. luxurians* (G-42) genotype compared with the two smaller genomes. The *Opie* family is estimated at ~110–260 Mb of the four genomes, with statistically significant increases in abundance in the larger genomes compared with the smaller genomes. This is also true for the *Prem* family, with estimates ranging from 55–230 Mb. For *gypsy* elements, the five most abundant families were used to estimate their Mb contributions to each genome. Statistically significant decreases were observed in *Z. mays* relative to the other three genomes for *Cinful* (~63–135 Mb), *Gyma* (~24–93 Mb) and *Zeon* (~68–150 Mb), but no significant differences were observed between the smaller *Z. diploperennis* and the larger *Z. luxurians* genomes. The *Huck* family showed the opposite pattern, with *Z. diploperennis* (~159 Mb) having the smallest estimate and being statistically different from the remaining genomes (~390–522 Mb), while no significant differences were observed between the larger *Z. luxurians* and smaller *Z. mays* genomes for this LTR retrotransposon. Finally, the *Xilon* family showed no significant differences among the four genomes, with estimates of 66–125 Mb.

We used neighbor joining trees to find evidence for sequence differences between elements in the same class or family that might help identify any possible relatedness of amplification events that were indicated by our SSA. The three high copy *copia* elements (*Ji*, *Opie* and *Prem*) were aligned using the integrase gene (Figure 3). A total of 851 sequences were annotated as one of these *copia* families: of these sequences, only 39 (4.5%) shared sufficient homology with the integrase gene. A single sequence for the integrase gene from each family was identified in *S. bicolor* and used as out-group in the phylogenetic reconstruction. The *copia* phylogram shows three major clusters, one each for the families investigated. The *Prem* clade is split into two subclusters with strong bootstrap support.

A second tree was constructed using the reverse transcriptase sequence in the *Huck* element, the only *gypsy* family with enough sequences to build a tree (Figure 4). Of the 559 sequences that were annotated as *Huck*, only 19 (3.5%) shared sufficient homology with the reverse transcriptase gene. A copy of the *Huck* element in *S. bicolor* was used as an out-group. The *Huck* phylogenetic tree shows two distinct clusters with high bootstrap support. One cluster contains sequences from all four genotypes, but the second cluster includes only *Z. mays* and *Z. luxurians* sequences.

A third tree was constructed from the nucleotide alignment of the 180-bp knob repeat (data not shown). Of the 396 sequences annotated as knob repeats, 380 (96%) were easily aligned. A copy of a similar ~180 bp tandem repeat from sorghum was used as an out-group for the phylogenetic tree. This tree contained a single derived cluster with high bootstrap support, and knob sequences from the four taxa appear to be randomly dispersed among all branches of the tree. These *Zea* repeats all exhibit a 6-bp deletion, an insertion of 4 bp and an insertion of 5 bp relative to the similar *S. bicolor* repeat.

DISCUSSION

Efficient and detailed analysis of repetitive DNA content

SSA allows analysis of the major repetitive components of a genome without the huge cost of deeply sequencing the entire genome (Brenner *et al.*, 1993; Devos *et al.*, 2005; Liu *et al.*, 2007). Because we are interested in the precise subfamilies of any repeats that were found, the short reads associated with 'next generation' sequencing (Mardis, 2008) were judged to be inappropriate. Longer reads provide extensive coverage of *cis*-linked variation that allows one element subfamily to be distinguished from another (Baucom *et al.*, 2009), while very short reads lead to an assembly that homogenizes all subfamilies into a single polymorphic assembly. With the small data sets that were generated with the longer Sanger reads, we were limited to only the highest copy number repeats in each genome analyzed, but targeting these elements was the purpose of this project. In these species, where we routinely analyzed <0.1% of the genome, repeats need to be present at copy numbers of at least a few thousand in order to be seen as repetitive in an all-versus-all BLAST analysis but, even within this group, our analysis concentrated on the 4–5 most abundant repeats so that the results would have sufficient depth to justify quantitative comparisons.

Repeat content in seven panicoid grass species

In this study, we have shown that using a very small sample of sequences from a plant genome allows discovery and description of the most abundant repeats, and their dynamics, in higher plant genomes. The similar results for two *Z. luxurians* accessions (G-5 and G-42) both confirm the rigor of this SSA approach and show that the TE dynamics observed are distinctive to a taxon, and not just a single sampled plant. A much larger data set of paired-end Illumina sequences from *Z. luxurians* generated by Ross-Ibarra and colleagues

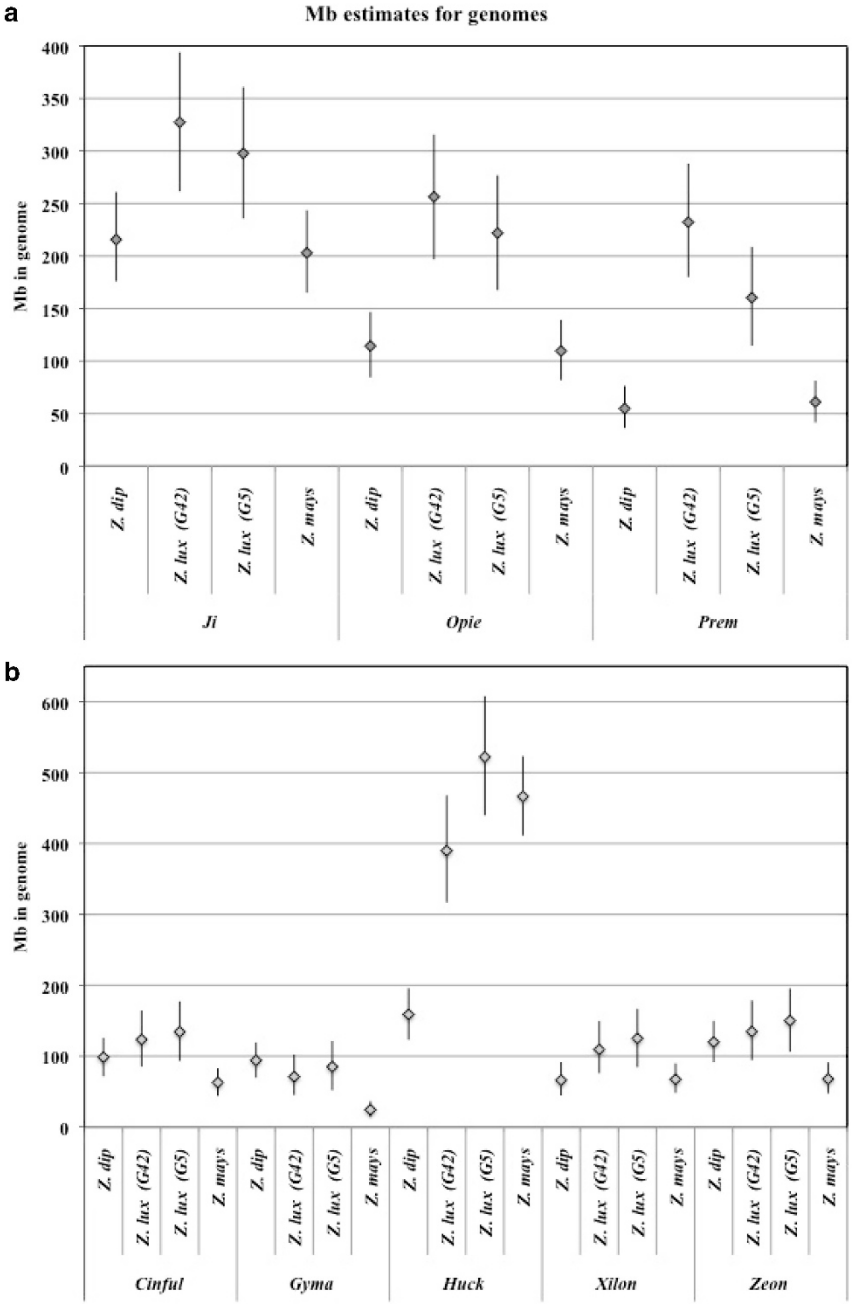

Figure 2 Megabases of DNA contributed by LTR retrotransposons in four different *Zea* genomes: (**a**) three *copia* LTR retrotransposon families; (**b**) five *gypsy* LTR retrotransposon families. Mean values of 1000 bootstrap replicates (95% CI) are indicated.

(Tenaillon *et al.*, 2011) permitted a highly quantifiable analysis of TE content variation in this species relative to maize. Tenaillon *et al.* (2011) found very similar TE properties to the ones reported for *Z. luxurians* in our study (for example, increased *Ji* and *Opie* abundance), but the shortness of the Illumina reads did not allow phylogenetic analysis of the specific LTR retrotransposon subfamilies that we found to be responsible for recent genome expansions in *Z. luxurians, Z. diploperennis* and *Z. mays*.

Grouping randomly sequenced clones using an all-versus-all BLASTn approach identifies the overall repetitive nature of the sample and helps to ensure that repetitive sequences were not missed during the annotation procedure. In the *Zea* component of this study, the TIGR repeat database was able to categorize most of the sequences identified as repetitive in our sample ($>60\%$). In an earlier study using a similar approach to investigate genomes in the genus *Gossypium*, only $\sim 3.5\%$ of the SSA data were able to be annotated using the highly conserved coding genes found in existing repeat databases for *Arabidopsis* and other *Brassica* species (Hawkins *et al.*, 2006). This illustrates the value of a high-quality repeat database from a close relative of the species targeted for genome analysis.

Identifying the major repeat classes produces information about the overall composition of the genome. LTR retrotransposons were found to be the most abundant component in all of the genomes investigated, with *gypsy* elements usually providing the most Mb of

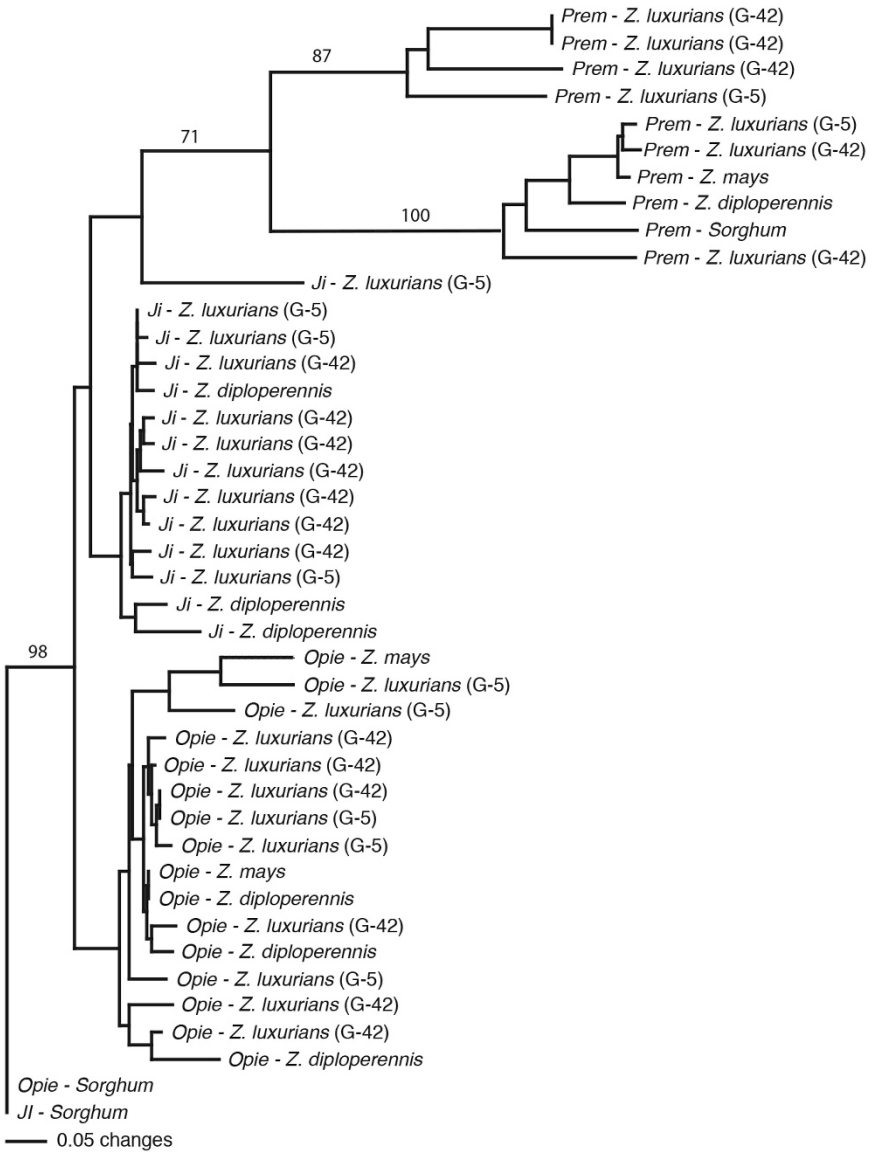

Figure 3 Neighbor joining tree of three highly abundant *copia* families from four *Zea* genomes, generated from an amino-acid alignment of the integrase gene sequences. Bootstrap values >50 are reported.

repetitive DNA. However, the most abundant LTR retrotransposon in pearl millet was found to be a member of the *copia* superfamily, so the overall predominance of *gypsy* elements is not absolute. In fact, in smaller plant genomes, it is fairly common for *copia* elements to provide as much or more DNA than the *gypsy* superfamily (Peterson-Burch *et al.*, 2004; Zuccolo *et al.*, 2007), indicating that it is variation in *gypsy* activity that is the most significant TE phenomenon affecting genome size.

Patterns in repeat accumulation across seven grass genomes

Now that it is known that the balance between TE (mostly LTR retrotransposon) amplification and DNA removal processes is responsible for genome size variability in angiosperms (reviewed in Bennetzen *et al.*, 2005), it becomes important to investigate why these factors are so variable in different plant lineages. Are there some TE families that are particularly likely to be hyper-abundant, and is it the presence (either vertically or horizontally transmitted) of these families in an active form that conditions a lineage for genome expansion? Or do all TEs have the possibility to amplify to exceedingly high numbers given the correct environmentally and genetic circumstances?

Independent activation of specific TE families in specific lineages

Analysis of the genomes in seven grass species indicates that many different LTR retrotransposons can become the major contributors to genome size. In some cases, like the *Huck* elements that have been very active in all of these seven species, an apparent tendency for hyper-amplification seems to be shared over tens of millions of years, but is more strongly manifested in some sublineages than in others. Because DNA removal processes erase most evidence of any TE insertion after only a few hundred thousand to a few million years in grass genomes (Devos *et al.*, 2002; Wicker *et al.*, 2003; Ma *et al.*, 2004; Wang and Dooner, 2006), the shared *Huck* amplifications in *Pennisetum, Saccharum, Sorghum* and *Zea* must have occurred independently within the last 2–5 million years, long after these lineages separated. Similar, but more recent, activations shared by

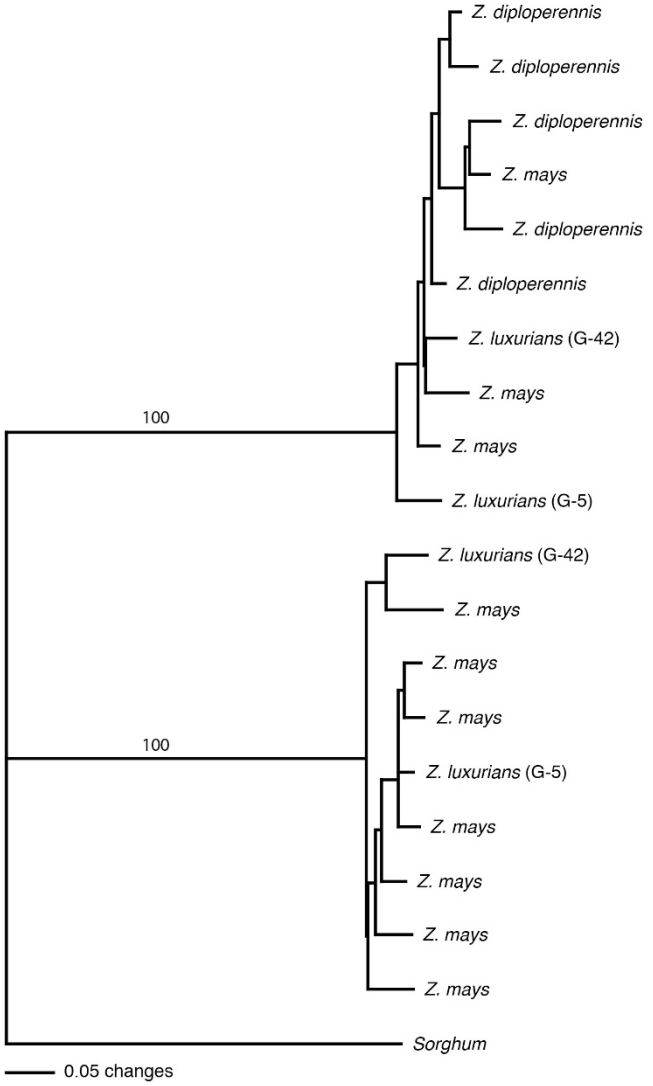

Figure 4 Neighbor joining tree of one *gypsy* element family, *Huck,* from four *Zea* genomes, generated from an amino-acid alignment of the reverse transcriptase gene sequences. Bootstrap values >50 are reported.

the *Zea* for *Ji* and *Opie* and by *Sorghum* and *Saccharuum* lineages for *Leviathan* indicate an apparently routine phenomenon, and one that can be mapped to a particular time and lineage on a phylogenetic tree.

This study was not designed to investigate horizontal transfer for any of the identified TEs. Such investigations require comprehensive analysis of multiple intermediate species across a precisely chosen set of lineages, with demonstration of more conserved sequences for a TE between two distant relatives than for those TEs in close relatives (Diao *et al.*, 2005; Roulin *et al.*, 2009). Even when observed, such data trends can also be explained by extinction of some TE subfamilies in some lineages. Although our *Ji* and *Opie* data are compatible with horizontal transmission associated with activation of hyper-amplification of a particular subfamily, further analyses across more Andropogoneae will be needed to substantiate this possibility. As shown in Figure 1, genomic shock associated with polyploidy cannot have been a factor in the timing of most of these TE activation events, but that does not mean that polyploidy might not be an activator of TEs in some linages. A more comprehensive analysis of TE behavior in a broader set of closely related lineages that differ in ploidy are needed to address this point. However, it is known that a large number of stress (Grandbastien, 1998) or genetic (Tsukahara *et al.*, 2009) states can lead to a pulse of TE activation, which might take thousands or even millions of years to be fully suppressed by the plant host.

One possible LTR retrotransposon activation by polyploidy was observed in our study, the hyper-activation of *Ji* and *Opie* in the *Zea* lineage. This activation was not shared by *Sorghum* or *Saccharum*, which had no detected copies of these elements. It will be enormously interesting to use SSA analysis on closer *Zea* relatives, like *Coix* (which did not share the polyploidy event seen in *Zea*) and *Tripsacum* (which did share the event) (Mathews *et al.*, 2002), to see how tightly the *Ji* and *Opie* hyper-accumulation correlates with the timing of the polyploidy. The presence of *Ji* and *Opie* in *Zea*, but not in *Sorghum* or *Saccharum*, could indicate horizontal transfer of these elements into *Zea* from a yet-undiscovered source, but they could also be due to extinction of inactive (and hence low copy number) *Ji* and *Opie* families by sequence decay or by segregation.

The highly amplified LTR retrotransposons in one lineage are absent or present at low copy numbers in more distant relatives (for example, in rice for *Huck*, in *Sorghum* for *Ji* and *Opie* and in *Zea* for *Leviathan*), so hyper-amplification is not a dependable family trait. Hence, the simplest model suggests that which LTR retrotransposon families become the most abundant in a genome is a stochastic outcome. Although many LTR retrotransposons, of both the *gypsy* and *copia* superfamilies, can become the major genome size determinants in plants, it is not clear that all can do so. Nor is it clear what conditions allow a particular element family in a particular lineage (for instance, *Huck* within the panicoid grasses) to be passed on in a form that greatly increases its chance of subsequent activity, even tens of millions of years after this potential was determined.

In order to investigate genome dynamics in the most detail, dramatic events over short evolutionary time frames provide the optimal opportunities. The near doubling in the last 1–2 million years of genome size in *Z. luxurians*, without polyploidy, as compared to *Z. diploperennis* and *Z. mays*, provides an excellent study system (Laurie and Bennett, 1985). Genome size is also known to be quite variable (>40%) even within *Z. mays*, but this is mostly associated with very different quantities of B chromosomes and/or knob repeats (reviewed in Poggio *et al.*, 1998), which can build up by random or selected segregation processes. However, TEs that are scattered about the genome cannot be easily concentrated by simple segregation, so we felt an investigation of repeat content in maize, *Z. luxurians* and *Z. diploperennis* would be informative.

Using the annotation of both the major repeat classes and the retroelement family diversity, we were able to estimate the amount of highly repetitive sequences and compare them across four samples using a standard bootstrap statistic to support our observations with 95% confidence intervals. In the major repeat classes, we found numerous statistically significant differences between abundances of various repeats within the larger *Z. luxurians* genomes and the two smaller genome species. The lack of significant difference for the ribosomal RNA repeats served as a baseline to measure other repeat dynamics. We found an almost twofold difference in the Mb estimates for the *copia* elements between the smaller genomes (*Z. mays* and *Z. diploperennis*) and the two *Z. luxurians* genomes. A similar twofold difference in estimates was also detected for the *gypsy* TEs, suggesting that a simple broad amplification of LTR retrotransposons from both these superfamilies was responsible for the dramatic growth of the *Zea luxurians* genome. However, investigated at the individual family level, these changes were much less uniform than expected.

Repeat dynamics and the evolution of grass genome content

Several different LTR retrotransposon families (for example, *Ji*, *Opie* and *Prem*) were dramatically more abundant in *Z. luxurians* compared with the two smaller genomes. However, the *Cinful*, *Gyma* and *Zeon* families appear to have amplified more actively in both the smallish *Z. diploperennis* genome and the large *Z. luxurians* genotypes in comparison to *Z. mays*. In contrast, the *Huck* family amplification was most dramatic in the *Z. mays* and *Z. luxurians* genotypes in comparison to the *Z. diploperennis* genome. Although phylogenetic studies have not been clear as to whether *Z. luxurians* is more closely related to *Z. mays* and or *Z. diploperennis*, the shared amplification of *Huck* in *Z. luxurians* and *Z. mays* (Figure 4) that was not shared with *Z. diploperennis* supports a more recent shared lineage for *luxurians/mays*, and hence justified the relatedness of these taxa depicted in Figure 1. From this perspective, the very high recent level of *Cinful, Gyma* and *Zeon* amplifications that were in common to *Z. luxurians* and *Z. diploperrenis* were not shared events, but similar outcomes of TEs that were independently activated, and this interpretation is supported by the phylogenetic trees for these *copia* elements (Figure 3), which have many clusters composed of elements from only one species.

The last major repetitive DNA investigated was the knob repeat. For this tandem satellite repeat, a fivefold difference was detected between the two smaller genomes, which accounts for ~20% of the genome size difference between these two genomes. A 25- to 30-fold increase was also detected in the *Z. luxurians* genotypes in comparison to the *Z. mays* genome. This difference accounts for ~15–17% of the variation in genome size seen between the smallest and larger genomes.

Taking the three primary classes of highly repetitive sequences (*copia*, *gypsy* and knob repeats) into account, we can explain ~45–50% of the variation between the two smaller genomes and the two larger *Z. luxurians* genomes. Given the large number of LTR retrotransposon families that have changed dramatically in their abundance, it is likely that differential amplification/abundance of the lower copy number LTR retrotransposon families that make up >15% of the maize genome (SanMiguel and Bennetzen, 1998; Baucom *et al.*, 2009) provides some of the additional genome size variation. Active *gypsy* elements greatly affecting genome size also have been shown in *Oryza* and *Gossypium* species (Hawkins *et al.*, 2006; Piegu *et al.*, 2006). However, unlike these earlier studies, it is not a family or two of LTR retrotransposons that has largely determined recent genome size change in *Zea*. Rather, a combined outcome of many different family activities, some increased greatly, some less active, and some not amplifying at all, has been responsible for the great differences in *Zea* genome sizes. A similar story seems to hold true in *Arabidopsis* where, even under the influence of mutations that decrease the epigenetic silencing that keeps most TEs transcriptionally and transpositionally quiescent, activation of each of several families shows unique patterns of timing and amplification intensity (Tsukahara *et al.*, 2009).

With the detailed SSA analysis of *Zea* genome dynamics, it becomes clear that simple models of genome growth due to the hyper-activity of a single or small number of families of LTR retrotransposons are not adequate to explain all dramatic genome size variation in plants. In comparison to *Z. mays*, the nearly twofold larger *Z. luxurians* genome shows higher Mb contributions from many different TE families of both *copia* and *gypsy* elements (for example, *Cinful*, *Gyma*, *Ji*, *Opie*, *Prem*, *Zeon*), but no obvious change for others (for example, *Xilon*) and less amplification of a particularly abundant family, *Huck*, compared with *Z. mays*. Hence, a mixture of multiple TE families with very different activity levels can lead to relative genome size expansion if the amplifiers are predominant. As shown for the comparisons across the panicoids, we do not yet know why or have any tools to predict which of these TEs will become active in any given lineage, for how long they will continue to be active, or how heavily they will amplify. Further searches are needed for the TE transmission with high activation potential indicated in this study. These studies should be pursued across many more plant lineages, with appropriate phylogenetic selection and depth of pursuit, to help tease out patterns in TE activity and evolution that have been responsible for the great range in genome variation that we now observe.

DATA ARCHIVING

Data generated in this study were deposited into the GSS division of GenBank with accession numbers JY127741—JY133169 and JY133584—JY136902.

ACKNOWLEDGEMENTS

We thank NSF (Grant #0607123), NIH (training Grant T32 GM07103-31) and the Georgia Research Alliance for funds that supported this research. We also thank K Devos, J Doebley, A D'Hont and HJ Price for providing biological materials employed in this project. The *Zea* data in this manuscript were released in a poster and oral presentation at the 48th Annual Maize Genetics Conference in Pacific Grove, CA, USA.

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

Baucom R, Estill J, Chaparro C, Upshaw N, Jogi A, Deragon J *et al.* (2009). Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *Plos Genet* **5**: e1000732.

Bennett MD (1972). Nuclear DNA content and minimum generation time in herbaceous plants. *Proc Roy Soc Lond Ser B Biol Sci* **181**: 109–135.

Bennetzen JL (2007). Patterns in grass genome evolution. *Curr Opin Plant Biol* **10**: 176–181.

Bennetzen JL, Kellogg EA (1997). Do plants have a one-way ticket to genomic obesity? *Plant Cell* **9**: 1509–1514.

Bennetzen JL, Ma J, Devos KM (2005). Mechanisms of recent genome size variation in flowering plants. *Ann Bot* **95**: 127–132.

Bennetzen JL, Schmutz J, Wang H, Percifield R, Hawkins J, Pontaroli AC *et al.* (2012). Reference genome sequence of the model plant Setaria. *Nat Biotech* **30**: 555–561.

Brenner S, Elgar G, Sandford R, Macrae A, Venkatesh B, Aparicio S (1993). Characterization of the pufferfish (fugu) genome as a compact model vertebrate genome. *Nature* **366**: 265–268.

Bundock P, Hooykaas P (2005). An arabidopsis hat-like transposase is essential for plant development. *Nature* **436**: 282–284.

DeBarry JD, Liu R, Bennetzen JL (2008). Discovery and assembly of repeat family pseudomolecules from sparse genomics sequence data using the assisted automated assembler of repeat families (AAARF) algorithm. *BMC Bioinformatics* **13**: 235.

Devos KM, Ma J, Pontaroli AC, Pratt LH, Bennetzen JL (2005). Analysis and mapping of randomly chosen bacterial artificial chromosome clones from hexaploid bread wheat. *Proc Natl Acad Sci USA* **102**: 19243–19248.

Devos KM, Brown JKM, Bennetzen JL (2002). Genome size reduction through illegitimate recombination counteracts genome expansion in arabidopsis. *Genome Res* **12**: 1075–1079.

Diao X, Freeling M, Lisch D (2005). Horizontal transfer of a plant transposon. *PLoS Biol* **4**: e5.

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

Grandbastien M-A (1998). Activation of plant retrotransposons under stress conditions. *Trends Plant Sci* **3**: 181–187.

Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF (2006). Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *gossypium*. *Genome Res* **16**: 1252–1261.

Hawkins JS, Proulx SR, Rapp RA, Wendel JF (2009). Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. *Proc Natl Acad Sci* **106**: 17811–17816.

Hudson ME, Lisch DR, Quail PH (2003). The FHY3 and FAR1 genes encode transposase-related proteins involved in regulation of gene expression by the phytochrome a-signaling pathway. *Plant J* **34**: 453–471.

Ilic K, SanMiguel PJ, Bennetzen JL (2003). A complex history of rearrangement in an orthologous region of the maize, sorghum, and rice genomes. *Proc Natl Acad Sci USA* **100**: 12265–12270.

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

Kellogg EA, Bennetzen JL (2004). The evolution of nuclear genome structure in seed plants. *Am J Bot* **91**: 1709–1725.

Laurie DA, Bennett MD (1985). Nuclear DNA content in the genera Zea and Sorghum. Intergeneric, interspecific and intraspecific variation. *Heredity* **55**: 307–313.

Leitch IJ, Chase MW, Bennett MD (1998). Phylogenetic analysis of DNA c-values provides evidence for a small ancestral genome size in flowering plants. *Ann Bot* **82**(Suppl 1): 85–94.

Lisch D, Bennetzen JL (2011). Transposable element origins of epigenetic gene regulation. *Curr Opin Plant Biol* **14**: 156–161.

Liu R, Bennetzen JL (2008). Enchilada redux: how complete is your genome sequence? *New Phytol* **179**: 249–250.

Liu R, Vitte C, Ma J, Mahama AA, Dhliwayo T, Lee M *et al.* (2007). A genetrek analysis of the maize genome. *Proc Natl Acad Sci USA* **104**: 11844–11849.

Ma J, Bennetzen JL (2006). Recombination, rearrangement, reshuffling and divergence in a centromeric region of rice. *Proc Natl Acad Sci USA* **103**: 383–388.

Ma J, Devos KM, Bennetzen JL (2004). Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res* **14**: 860–869.

Madlung A, Tyagi AP, Watson B, Jiang H, Kagochi T, Doerge RW *et al.* (2005). Genomic changes in synthetic arabidopsis polyploids. *Plant J* **41**: 221–230.

Mardis ER (2008). Next-generation DNA sequenceing methods. *Annu Rev Genomics Hum Genet* **9**: 387–402.

Mathews S, Spangler RE, Mason-Gamer RJ, Kellogg EA (2002). Phylogeny of andropogoneae inferred from phytochrome b, gbssi, and ndhf. *Int J Plant Sci* **163**: 441–450.

Michaels SD, He Y, Scortecci KC, Amasino RM (2003). Attenuation of flowering locus c activity as a mechanism for the evolution of summer-annual flowering behavior in arabidopsis. *Proc Natl Acad Sci* **100**: 10102–10107.

O'Neill RJW, O'Neill MJ, Graves JAM (1998). Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. *Nature* **393**: 68–72.

Ouyang S, Buell CR (2004). The TIGR plant repeat databases: a collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Res* **32**(Suppl 1): D360–D363.

Ozkan H, Levy AA, Feldman M (2001). Allopolyploidy-induced rapid genome evolution in the wheat (aegilops–triticum) group. *Plant Cell Online* **13**: 1735–1747.

Parisod C, Salmon A, Zerjal T, Tenaillon M, Grandbastien M-A, Ainouche M (2009). Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in spartina. *New Phytol* **184**: 1003–1015.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H *et al.* (2009). The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**: 551–556.

Paterson AH, Bowers JE, Chapman BA (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *PNAS* **101**: 9903–9908.

Peterson DG, Schulze SR, Sciara EB, Lee SA, Bowers JE, Nagel A *et al.* (2002). Integration of cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. *Genome Res* **12**: 795–807.

Peterson DG, Tomkins JP, Frisch DA, Wing RA, Paterson AH (2000). Construction of plant bacterial artificial chromosome (BAC) libraries: an illustrated guide. *J Agricultur Genomics* **5**, http://wheat.pw.usda.gov/jag/papers00/paper300/indexpage300.html.

Peterson-Burch B, Nettleton D, Voytas D (2004). Genomic neighborhoods for arabidopsis retrotransposons: a role for targeted integration in the distribution of the metaviridae. *Genome Biol* **5**: R78.

Petit M, Guidat C, Daniel J, Denis E, Montoriol E, Bui QT *et al.* (2010). Mobilization of retrotransposons in synthetic allotetraploid tobacco. *New Phytol* **186**: 135–147.

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

Poggio L, Rosato M, Chiavarino AM, Naranjo CA (1998). Genome size and environmental correlations in maize (*Zea mays ssp. Mays*, poaceae). *Ann Bot* **82**(Suppl 1): 107–115.

Roulin A, Piegu B, Fortune P, Sabot F, D'Hont A, Manicacci D *et al.* (2009). Whole genome surveys of rice, maize and Sorghum reveal multiple horizontal transfers of the LTR-retrotransposon *route66* in poaceae. *BMC Evol Biol* **9**: 58.

SanMiguel P, Bennetzen JL (1998). Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. *Ann Bot* **82**: 37–44.

Schnable JC, Springer NM, Freeling M (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc Natl Acad Sci USA* **108**: 4069–4074.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S *et al.* (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**: 1112–1115.

Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia Y *et al.* (2009). Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet* **5**: e1000734.

Swigonova Z, Lai J, Ma J, Ramakrishna W, Llaca V, Bennetzen JL *et al.* (2004). Close split of Sorghum and maize genome progenitors. *Genome Res* **14**: 1916–1923.

Swofford D (2002). *PAUP 4.0*. Sinauer Associates: Sunderland, MA.

Tenaillon MI, Hufford MB, Gaut BS, Ross-Ibarra J (2011). Genome size and transposable element content as determined by high-throughput sequencing in maize and *Zea luxurians*. *Genome Biol Evol* **3**: 219–229.

Thomas BC, Pedersen B, Freeling M (2006). Following tetraploidy in an arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res* **16**: 934–946.

Tsukahara S, Kobayashi A, Kawabe A, Mathieu O, Miura A, Kakutani T (2009). Bursts of retrotransposition reproduced in arabidopsis. *Nature* **461**: 423–426.

Vicentini A, Barber J, Aliscioni A, Giussani L, Kellogg E (2008). The age of the grasses and clusters of orgins of C4 photosynthesis. *Global Change Biol* **14**: 2963–2977.

Wang Q, Dooner HK (2006). Remarkable variation in maize genome structure inferred from haplotype diversity at the bz locus. *Proc Natl Acad Sci USA* **103**: 17644–17649.

White SE, Habera LF, Wessler SR (1994). Retrotransposons in the flanking regions of normal plant genes: a role for *copia*-like elements in the evolution of gene structure and expression. *Proc Natl Acad Sci USA* **91**: 11792–11796.

Wicker T, Yahiaoui N, Guyot R, Schlagenhauf E, Liu Z-D, Dubcovsky J *et al.* (2003). Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and AM genomes of wheat. *Plant Cell Online* **15**: 1186–1197.

Zuccolo A, Sebastian A, Talag J, Yu Y, Kim H, Collura K *et al.* (2007). Transposable element distribution, abundance and role in genome size variation in the genus *Oryza*. *BMC Evol Biol* **7**: 152.