Young, Intact And Nested Retrotransposons Are Abundant In The Onion And Asparagus Genomes

Authors:
C. Vitte1**, M. C. Estep,** J. Leebens-Mack and J. L. Bennetzen

Abstract

Background and Aims Although monocotyledonous plants comprise one of the two major groups of angiosperms and include .65 000 species, comprehensive genome analysis has been focused mainly on the Poaceae (grass) family. Due to this bias, most of the conclusions that have been drawn for monocot genome evolution are based on grasses. It is not known whether these conclusions apply to many other monocots.

Methods To extend our understanding of genome evolution in the monocots, Asparagales genomic sequence data were acquired and the structural properties of asparagus and onion genomes were analysed. Specifically, several available onion and asparagus bacterial artificial chromosomes (BACs) with contig sizes .35 kb were annotated and analysed, with a particular focus on the characterization of long terminal repeat (LTR) retrotransposons.

Key Results The results reveal that LTR retrotransposons are the major components of the onion and garden aspara-gus genomes. These elements are mostly intact (i.e. with two LTRs), have mainly inserted within the past 6 million years and are piled up into nested structures. Analysis of shotgun genomic sequence data and the observation of two copies for some transposable elements (TEs) in annotated BACs indicates that some families have become particu-larly abundant, as high as 4–5 % (asparagus) or 3–4 % (onion) of the genome for the most abundant families, as also seen in large grass genomes such as wheat and maize.

Conclusions Although previous annotations of contiguous genomic sequences have suggested that LTR retrotransposons were highly fragmented in these two Asparagales genomes, the results presented here show that this was largely due to the methodology used. In contrast, this current work indicates an ensemble of genomic features similar to those observed in the Poaceae.

C. Vitte1, **M. C. Estep**, J. Leebens-Mack and J. L. Bennetzen (2013) "Young, Intact And Nested Retrotransposons Are Abundant In The Onion And Asparagus Genomes" *Annals of Botany* Version of Record Available At www.academic.oup.com [doi:10.1093/aob/mct155]

INTRODUCTION

Monocotyledonous plants are divided into 11 orders: Acorales, Alismatales, Petrosaviales, Dioscoreales, Pandanales, Liliales, Asparagales, Arecales, Poales, Commelinales and Zingiberales, the last four being grouped into the Commelinid clade (Stevens *et al.*, 2001). Most of these orders include multiple agriculturally and ornamentally important groups. Among these, the Asparagales may be the second most important for agriculture and horticulture after the Poales. The Asparagales includes the hyperdiverse Orchidaceae with .30 000 species and nearly 5000 additional species distributed across the rest of the order. These include important crops such as aloe (*Aloe vera*), agave (*Agave tequilana*), asparagus (*Asparagus officinalis*), garlic (*Allium sativum*), leek (*Allium ampeloprasum*), onion (*Allium cepa*) and vanilla (*Vanilla planifolia*), as well as ornamental plants such as yuccas, amarylids, daffodils, irises and orchids. With a world production of .95 Mt, it is the third most cultivated group for vegetable production in the world after the Solanales (including potato, tomato, pepper and aubergine) and the Cucurbitales (including melons, cucumbers and gourds) (world production, http://faostat.fao.org/).

Onion (*A. cepa*) is the most economically important member of the Asparagales. As a biological model, onion has been extensively studied at the cytological and biochemical levels. Onion is a diploid (2*n* ¼ 2*x* ¼ 16), but has a genome size of approx. 16 400 Mb/1C, one of the largest among all cultivated diploid species and similar in size to that of hexaploid wheat (http://data.kew.org/cvalues/). Although a large genome can facilitate cytogenetic analyses, it has hampered the development of genomic resources, thus impeding both molecular breeding and characterization of the molecular origins of its large genome. Transcriptome analyses suggest that onion transcript characteristics are quite distinct from those of Poaceae (Kuhl *et al.*, 2004), but little is known about the transposable element (TE) composition in the onion genome. In particular, while long terminal repeat (LTR) retrotransposons are thought to be a major component of the onion genome (Pearce *et al.*, 1996), the role that TE activity has played in shaping the current genome is poorly understood. Most of the available data on onion genome composition come from re-association kinetics (Cot) and fluorescence *in situ* hybridization (FISH) analyses that revealed an abundance of middle-repetitive sequences interspersed with low copy number regions (Stack and Comings, 1979; Pearce *et al.*, 1996; Suzuki *et al.*, 2001). Analyses of bacterial artificial chromosome (BAC)-end sequences revealed a high frequency of TEs, consistent with results obtained for other plants with large genomes

(Jakše *et al.*, 2008). Sequencing of a few onion BACs (Do *et al.*, 2003; Jakše *et al.*, 2008) revealed the presence of repeated sequences, including a large quantity (approx. 50 %) of TEs, but also microsatellites and direct tandem repeats. Most of the TEs found were interpreted as highly degraded, a feature that is quite different from the pattern observed in grasses, where intact and young (,5 million year old) elements are found. However, the scarcity of genomic data and the absence of any TE database from a closely related species made the detection of TEs in Asparagales particularly challenging.

Garden asparagus (*A. officinalis*) is the third most economically important plant in the Asparagales, after onion and garlic. It is believed to be native to Europe, northern Africa and Asia, and is even more widely cultivated as a vegetable crop. Asparagus is diploid (2*n* ¼ 2*x* ¼ 20) and has one of the smallest genomes of the core Asparagales (approx. 1300 Mb/1C, http://data.kew.org/cvalues/) For this reason, garden asparagus has been proposed as a genomic model for the core Asparagales that would help investigate the gene content of other Asparagales with larger genome sizes, such as onion (Kuhl *et al.*, 2005; Telgmann-Rauber *et al.*, 2007). Synteny between asparagus and onion genomes has received preliminary investigation by mapping several onion expressed sequence tags (ESTs) on both onion chromosomes and a few studied asparagus BACs (Jakše *et al.*, 2006). This analysis posited that many genes encoding physically linked ESTs in asparagus were located on distinct chromosomes in onion, suggesting a lack of collinearity between the two genomes. Since the ESTs analysed were not single copy [e.g. from gene duplication, perhaps via TE mobilization of genes or gene fragments (Jin and Bennetzen, 1994; Jiang *et al.*, 2004; Lai *et al.*, 2005; Yang and Bennetzen, 2009)], it is possible that the sequences compared were not orthologous. However, given that the divergence of the lineages leading to these species occurred approx. 87 million years ago (Mya) (Jansson and Bremer, 2005), frequent interruptions in microcollinearity are expected (Bennetzen, 2000). Hence, more detailed analyses, as well as comparison of other species within the Asparagales, are needed to investigate and extend these results.

The recent sequencing of asparagus BACs has revealed the presence of sequences showing similarities to known LTR retrotransposons from other species (Jakše *et al.*, 2006; Telgmann-Rauber *et al.*, 2007). However, the structure of the TEs found and their precise characterization were not assessed. A comprehensive characterization of TE sequences in the garden asparagus genome is needed, especially to investigate the origin of the large differences in genome size observed between hermaphroditic African asparagus species and dioecious species from Europe and Africa (Štajner *et al.*, 2002).

Herewe describe the acquisition of new shotgun sequence data and further annotation of available onion and asparagus BAC sequences, with a particular focus on the detailed characterization of TE sequences, to unravel the structural properties of these genomes and compare them with those of several well-studied grass genomes. Our results reveal that LTR retrotransposons are the major components of the onion and garden asparagus genomes. These elements are mostly intact (i.e. with two LTRs), have inserted ,6 Mya and are piled up into nested structures, a suite of features identical to those observed in the Poaceae and in another recently analysed monocot, banana (*Musa acuminata*) (D’Hont *et al.*, 2012).

MATERIALS AND METHODS

Sequence data

Contiguous sequences (contigs) from four asparagus (***Asparagus officinalis***) BACs (available as GenBank accessions AC183410, DQ273271 and the overlapping AC183409 and AC183411) that represent three genomic regions selected to contain a sulfite reductase gene, a sucrose transporter gene and the AOB272 cDNA (described in Jakše *et al.*, 2006), and three full-length onion (***Allium cepa***) BACs (AB111058, DQ273272 and DQ273270) were retrieved from NCBI. In addition, 3235 unpaired Sanger shotgun sequences from onion were provided by C. Town, and we generated an additional 3560 unpaired Sanger sequences from garden asparagus (deposited in GenBank).

Annotation of asparagus and onion BAC sequences

Bacterial artificial chromosome sequences were used as the subject for Repeatmasker and BLASTX searches. The Repeatmasker search was performed with version 3·0 (Smit *et al.*, 1996; http://www.repeatmasker.org) using repbase (April 2012) and redat (MIPS, May 2008) as query databases. For the BLASTX search (Altschul *et al.*, 1990), version blastall 2·2·25 was used with the nr database (March 2012) as query and a threshold e-value $\leq 10^{25}$, and without filtering out low complexity regions. In parallel, asparagus and onion shotgun reads were mapped onto the asparagus and onion BACs, respectively, in order to spot repeated regions that may have been missed by the BLASTX and Repeatmasker searches due to the absence of closely related TEs in the available repeat databases. For this, BLASTN searches using BAC sequences as subjects and shotgun sequences as queries were performed.

For each contig, results of these searches were visualized using Argo v10·0·31 (http://www.broadinstitute.org/annotation/argo) and manually evaluated to provide a first annotation. Genes were predicted using the BLASTX results. Only matches to a predicted protein from a dicotyledonous species with an e-value $\leq 10^{25}$ were used, because, as onion and asparagus are monocots, genes annotated from matches to other monocots proteins with unknown function have a reasonable chance to be TE genes (Bennetzen *et al.*, 2004). Because TE genes are much less conserved than are standard plant genes when comparing distantly related species (such as monocots with dicots), a match between a monocot genome and an unknown protein from a dicot species is more likely to be a real gene. Transposable element-coding domains were predicted using BLASTX and Repeatmasker results, and TE regions were predicted from Repeatmasker and BLASTN search results with survey sequence data. To improve the annotation of TEs, LTRs were sought in the regions flanking BLASTX-defined TE-coding domains by BLASTN2 dotplot comparison (www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi), and by running LTR_STRUC (McCarthy and McDonald, 2003).

Characterization of LTR retrotransposon types and clustering of the discovered elements into families

For all LTR retrotransposons for which a polyprotein could be detected, we looked for the order of the reverse transcriptase (RT), integrase (Int) and ribonuclease H (RNaseH) domains, a feature that distinguishes elements of the ***Gypsy***-like and ***Copia***-like superfamilies. When elements were lacking some of these domains, their sub-class was determined by comparing their sequence with a database of maize LTR retrotransposons (Baucom ***et al.***, 2009) of known superfamily using tBLASTX (from blastall version 2·2·25).

For intact elements, we sought structural features such as the primer-binding site (PBS), polypurine tract (PPT) and target site duplications (TSDs). To analyse PBSs, a tRNA database (http://lowelab.ucsc.edu/GtRNAdb/) was reverse complemented and compared with the downstream region of the 5′ LTR from each element. Intact LTR retrotransposons were then clustered into families based on the overall structure of the elements (in particular the sharing of similar LTRs and PBS), and on sequence identity between copies obtained from an all vs. all retrotransposons BLASTN search with a required e-value $\leq 10^{25}$.

Estimation of intact LTR retrotransposon insertion dates

The insertion date of each LTR retrotransposon with two LTRs and a TSD was estimated following the method described in SanMiguel ***et al.*** (1998). Divergence between LTRs was computed by MEGA version 4·0 (Tamura ***et al.***, 2007), using the Kimura 2 parameter distance (Kimura, 1980) that corrects for both homoplasy and differences in the rates of transition and transversion. Rough estimation of insertion dates was obtained using the substitution rate of $1{\cdot}3 \times 10^{28}$ substitution per site per year that has been described for rice LTR retrotransposons (Ma and Bennetzen, 2004).

Estimation of TE copy divergence and TE genome coverage using genome survey sequences

For each intact TE characterized, nucleotide divergence between each sequence and the intact copy was estimated by running a BLASTN search using sequences of TEs as subject and shotgun sequences as query, with an e-value threshold of 10^{23} and a minimum read coverage of 20 % (i.e. approx. 140 bp). For each element for which a minimum of 50 cases of homology was found, identity scores between the detected regions and the corresponding reference sequence were computed to build identity histograms. Results of this search were also used to compute the length occupied by each LTR retrotransposon, leading to an estimate of genome coverage for each element family. For each element, lengths of all BLAST high scoring pairs (HSPs) found were extracted and computed to estimate the amount of shotgun sequence occupied by this element and the corresponding percentage of total shotgun sequence. This percentage was then converted into a fraction of the genome using the corresponding genome size. Confidence intervals (CIs) were estimated using a boostrap approach in which the shotgun sequence data were regenerated 100 times by resampling the same number of sequences with replacement in the original data set and performing the same analysis. Corresponding distributions were used to estimate 95 % CIs of the genome fraction occupied by each element.

RESULTS

Genome composition and structure from BAC sequences

Although a few BAC sequences from asparagus and onion have been produced over the past several years, their annotation has been hampered by the lack of TE sequences from closely related species. To better evaluate the genome structure of asparagus and onion genomes, we carefully annotated the contig sequences with size .35 000 bp from publicly available BACs of these two species, using a methodology combining searches for homology to known TEs, and structural characteristics of TEs. In addition, random shotgun reads were mapped onto the sequences of the TEs that we discovered, to assess their diversity, relative abundance and genome coverage (see the Materials and Methods). As presented in Table 1, LTR retrotransposons were found to account for the majority of both onion and asparagus contigs, while other types of TEs such as DNA transposons and LINEs (long interspersed nuclear elements) are much less abundant, contributing only a few per cent of the contig DNA. Mapping of the shotgun reads onto the contigs revealed the presence of additional repeats that could not be fully deciphered. Although these repeats could not be proven to be TEs, they may correspond to TE fragments that were missing one or both ends of the element because it was outside of the contig or had lost some segments by progressive deletion (Ma and Bennetzen, 2004) or other rearrangement. For asparagus, two

Table 1. *BAC composition*

			% BAC coverage						
Species	BAC name	Size of largest* contig (bp)	Gene	LTR retrotransposon	DNA transposon	LINE	Uncharacterized repeat	Microsatellite	Unknown
Asparagus	AC183410	39 875	0·0	80·7	0·0	0·0	0·9	0·1	18·3
Asparagus	DQ273271	42 943	0·5	36·9	23·6	12·3	1·6	0·1	25·1
Asparagus	AC183411	51 593	0·0	90·6	0·0	0·0	5·5	0·0	3·9
Onion	AB111058	35 243	6·1	4·1	0·0	8·5	33·2	0·1	48·0
Onion	DQ273272	84 316	0·0	58·9	0·0	0·0	14·2	0·0	27·0
Onion	DQ273270	108 232	0·0	52·4	6·8	0·0	4·0	0·0	36·7

LINE, long interspersed nuclear element.

* When several contigs were available, only those with size .35 000 bp were analysed. In all cases, this led to the analysis of the largest contigs only, as the second largest contig was too short to reach this threshold.

contigs (from BACs AC183410 and AC183411) are composed almost entirely of LTR retrotransposon sequences, representing approx. 81 % and approx. 91 % of the sequence analysed, respectively. On the other hand, the two onion contigs with the highest LTR retrotransposon content (from BACs DQ273272 and DQ273270) were found to include, respectively, approx. 52 % and approx. 59 % LTR retrotransposon DNA. Analysis of contig structures (Fig. 1) revealed the presence of nested LTR retrotransposons in both species. The largest set of nested elements was observed in onion BAC DQ273272, and consists of three LTR retrotransposons. The presence of traces of other LTR retrotransposons in the flanking regions of the nests suggests that the degree of nesting could be greater than observable in these short contigs.

Type, structure and copy number of asparagus and onion LTR retrotransposons

As presented in Table 2, our annotation of large contigs enabled us to retrieve 12 and eight intact LTR retrotransposon copies in asparagus and onion, respectively. Similarity clustering of the elements revealed that some of them are members of the same family. The asparagus ART2 element and the onion ORT5 element were found at two copies each.

Of the 11 distinct LTR retrotransposons found in asparagus, five could be characterized as ***Copia***-like (Table 2) and the rest could not be classified as either ***Copia***-like or ***Gypsy***-like. In onion, of the seven distinct elements found, five could be assigned to the ***Gypsy*** clade, one to the ***Copia*** clade and one could not be classified (Table 2).

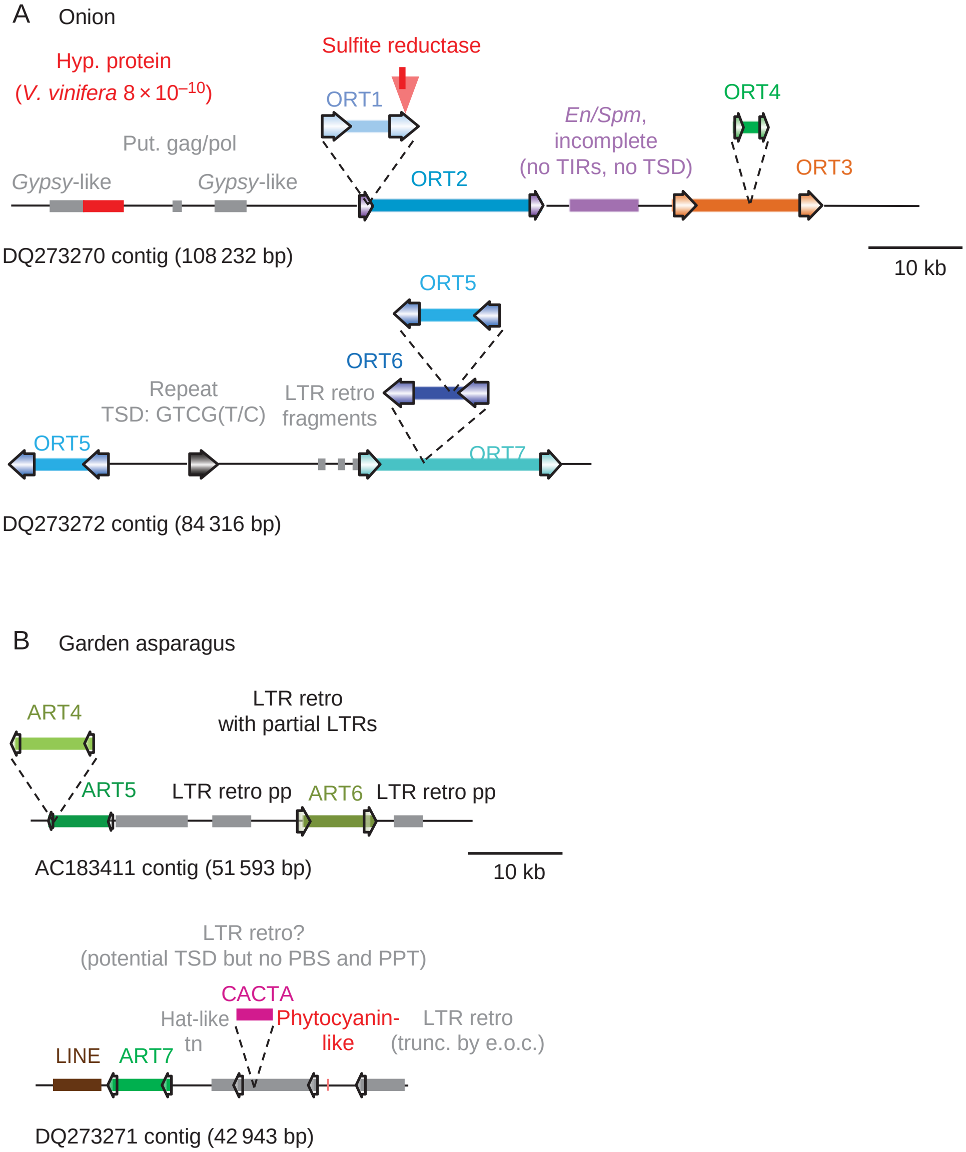

FIG. 1. Annotation of large contigs. Names of the BACs are highlighted below each drawing. Genes are shown in red. Regions with other colours correspond to fully characterized TEs. Grey bars show regions with homology to TEs for which the complete structure could not be detected or characterized. LTR retrotransposons are represented by lines with arrows highlighting the position of the LTRs. A 10 kb scale is given on the right of (A) onion BACs and (B) garden asparagus BACs. Abbreviations: put., putative; hyp., hypothetical; TIRs, terminal inverted repeats; TSD, target site duplication; retro, retrotransposon; tn, transposon; pp, polyprotein; trunc., truncated; e.o.c., end of contig.

Table 2. ***Properties of the LTR retrotransposons discovered***

Species	BAC	Name	Type	Total size	LTR size	TSD	LTR: start . . . end	PBS	PPT	Ts:Tv	LTR divergence J&C	Age J&C
Asparagus	AC183410	ART1	*Copia*	5207	656/505	CATCC	TGT . . . ACA	Arg	Yes	2·8:1	0·055	2·1
Asparagus	AC183410	ART2	*Copia*	5591	588	AATTT	TGT . . . ACA	Met	Yes	5·0:1	0·04	1·5
Asparagus	AC183410	ART3	*Copia*	5227	129	ATTTT	TGT . . . ACA	Met	Yes	0:0	0	0·0
Asparagus	AC183411	ART4	Unknown	7979	1075/1046	GTGGC	TGT . . . ACA	Arg	Yes	3·8:1	0·043	1·7
Asparagus	AC183411	ART5	Unknown	7368	624/696	ATTTT	TGGT . . . ACA	Arg	Yes	4·4:1	0·118	4·5
Asparagus	AC183411	ART6	Unknown	8216	1328/1260	TATAC	TGT . . . ACA	Arg	Yes	2·6:1	0·009	0·3
Asparagus	DQ273271	ART7	Unknown	6756	960/958	ATCGT	TGT . . . ACA	Arg	Yes	3·0:1	0·025	1·0
Asparagus	DQ273271	ART8	Unknown	6837	2122/2094	TCATC	TGT . . . ATA	Met	Yes	1·2:1	0·016	0·6
Asparagus	AC183409	ART9	Unknown	2265	430/433	TTATT	TGA . . . T(C/T)A	Met	Yes	5·5:1	0·031	1·2
Asparagus	AC183409	ART2	*Copia*	5109	693/688	TT(T/C)CT	T(G/A)T . . . ACA	Met	Yes	5·4:1	0·099	3·8
Asparagus	AC183409	ART10	Unknown	7602	1185/1156	AGGAC	TGT . . . ACA	Arg	Yes	1·7:1	0·017	0·7
Asparagus	AC183409	ART11	*Copia*	5309	617	TTGAC	TGT . . . ACA	Met	Gx3	3:1	0·02	0·8
Onion	DQ273270	ORT1	*Gypsy*	11 347	3091/3092	TCTGT	TGT . . . ACA	Met	Yes	1·7:1	0·155	6·0
Onion	DQ273270	ORT2	*Gypsy*	10 909	1463/1460	A(C/A)AAG	TGT. . . A(− /C)A	NF	Yes	1·6:1	0·217	8·3
Onion	DQ273270	ORT3	*Copia*	15 510	2377	GTTTA	TGT . . . ACA	Met	Yes	1·3:1	0·008	0·3
Onion	DQ273270	ORT4	Unknown	3659	962/963	No TSD	TCT . . . ACA	Met	Yes	NC	NC	NC
Onion	DQ273272	ORT5	*Gypsy*	10 168	2720/2718	A(C/A)TAG	T(T/G)T . . . ACA	Met	Yes	1·7:1	0·032	1·2
Onion	DQ273272	ORT6	*Gypsy*	11 194	3217/3211	CATTC	TGA . . . ACA	Met	Yes	1·2:1	0·032	1·2
Onion	DQ273272	ORT5	*Gypsy*	10 275	2770/2773	AATCT	TGT . . . ACA	Met	Yes	0:1	0·001	0·0
Onion	DQ273272	ORT7	*Gypsy*	43 875*	2225/2229	ATAT(C/T)	TGT . . . ACA	Met	Yes	1·8:1	0·062	2·4

TSD, target site duplication; PBS, primer-binding site; PPT, polypurine tract; Ts, transition; Tv, transversion; J&C, Jukes and Cantor; NF, not found; NC, not characterized.

Arg and Met correspond to tRNA types matching the PBS.

* The size of this element may be smaller due to the possible insertion of another retrotransposon within it. A duplicated structure that could correspond to LTRs has been found, but no other LTR retrotransposon feature could be detected. The total size would be 12 566 bp if only the two repeats are removed (without what could be the internal part), or 10 633 bp if the whole structure was removed.

Analysis of the frequency of homology found for each discovered element to reads in the shotgun data set revealed great variation between element families, including some with no identified homologues identical to the analysed intact elements (i.e. ART3) and hundreds with identical sequences for ART4, ART5, ART6, ART7 and ART8 (Fig. 2, Table 3). For onion LTR retrotransposons, the number of cases of identity to the intact elements ranged from a few for ORT1, ORT2, ORT6 and ORT7 to up to 250 for ORT3 (Fig. 2, Table 3).

The total length occupied by each element was also computed using the output of the BLASTN analysis of shotgun reads compared with each TE sequence of intact elements. This led to the estimation that some elements (ART5, ART6, ART7, ART8 and ART10) represent around 10 – 50 Mb of the asparagus genome, while ART4 represents .50 Mb of the genome (Table 3). On the other hand, several elements (ART1, ART2, ART3, ART9 and ART11) had too few cases of homology to estimate accurately the genomic fraction they occupy (Table 3). For onion, element ORT3 was found to be very abundant, providing 560 – 730 Mb of the onion genome. Two other elements, ORT4 and ORT5, were estimated to contribute .80 Mb to the onion genome each, while the other three identified elements were observed to represent a very small fraction of this genome (Table 3). Altogether, these results revealed that ART4, ART7 and ART8 combine to make up 8 – 12 % of the asparagus genome, while ORT3, ORT4 and ORT5 together account for .5 % to 7 % of the onion genome.

Insertion timing of asparagus and onion LTR retrotransposons

For intact copies (i.e. copies with two LTRs), estimation of insertion dates (see the Materials and Methods) revealed that all elements have inserted within the past 6 million years, most having insertion dates ,4 Mya (Fig. 3). For elements present in nested structures, insertion dates of the different elements involved in the same nest were compared (Table 2, Fig. 1), and revealed, as expected, that younger elements are always inserted into older ones: 6·0 Mya/8·3 Mya for nest ORT1/ORT2, 0·04 Mya/1·2 Mya/2·4 Mya for nest ORT5/ORT6/ORT7 and 1·6 Mya/4·5 Mya for nest ART4/ART5. Interestingly, the two oldest elements found for onion (i.e. ORT1, 6·0 Mya; and ORT2, 8·3 Mya), far older than any of the other elements found in onion (all younger than 2·5 Mya), are inserted within each other. Analysis of nucleotide pair frequencies in LTRs reveals that transitions outnumber transversions in both species, with average transition (Ts):transversion (Tv) ratios of 3·2:1 and 1·3:1 for asparagus and onion, respectively.

To get an overall idea of the timing of amplification events, we computed the distribution of nucleotide identities from pairwise alignments between each intact copy and matching random

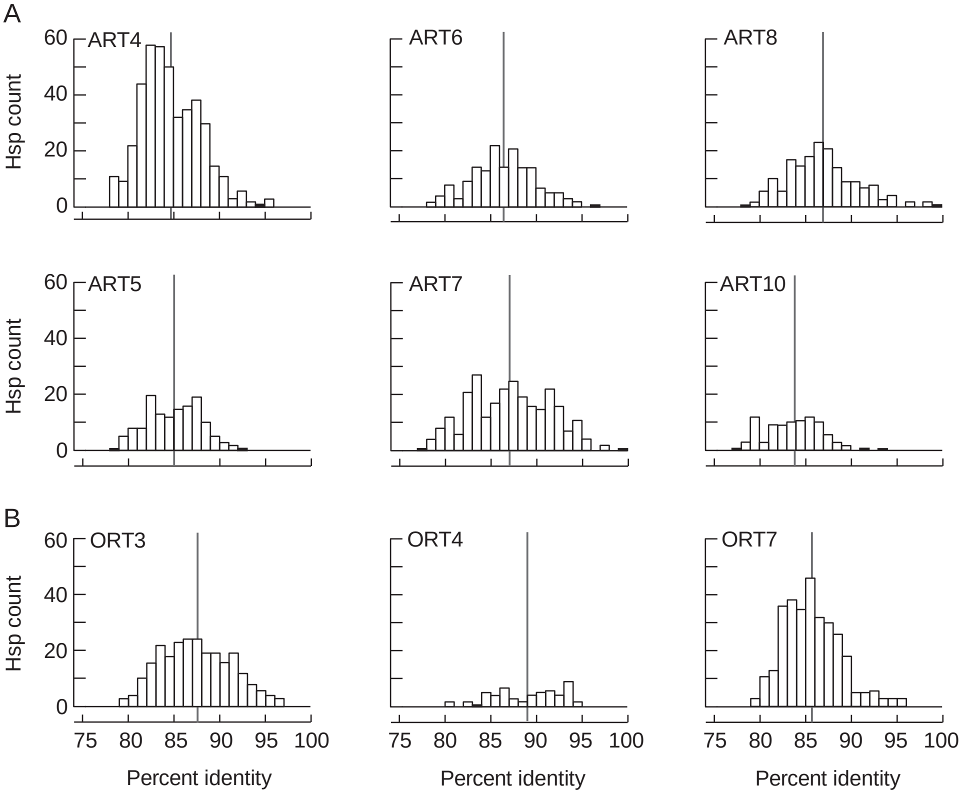

FIG. 2. Global identity scores among copies of the ART and ORT LTR retrotransposons. Pairwise identity score between each LTR retrotransposon and the random shotgun sequences with a significant BLAST homology. Onlyelements for which .50 high scoring pairs (HSPs) were found are shown. Vertical lines indicate average identity scores. For ORT7, the graph presented corresponds to that obtained with the full-length element as reference, but almost identical results were obtained when removing potential inserted elements. (A) Asparagus LTR retrotransposons (ART type); (B) onion LTR retrotransposons (ORT type).

TABLE 3. *Genomic fraction occupied by each LTR retrotransposon discovered*

		Genomic shotgun sequence occupied		Portion of genome occupied (Mb)	
Name	Class	Length (bp)	Percentage	From original data set	95 % CI from resampling
ART1	*Copia*	2802	0·1	2	NE
ART2	*Copia*	10 931	0·5	6	NE
ART3	*Copia*	0	0·0	0	NE
ART4	Unknown	109 918	4·7	61	51 – 70
ART5	Unknown	27 391	1·2	15	12 – 19
ART6	Unknown	36 565	1·6	20	15 – 25
ART7	Unknown	67 872	2·9	37	28 – 50
ART8	Unknown	49 748	2·1	27	23 – 32
ART9	Unknown	1028	0·0	1	NE
ART10	Unknown	22 434	1·0	12	10 – 16
ART11	*Copia*	8488	0·4	5	NE
ORT1	*Gypsy*	5755	0·2	33	NE
ORT2	*Gypsy*	1950	0·1	11	NE
ORT3	*Copia*	111 687	3·9	642	560 – 730
ORT4	Unknown	23 758	0·8	137	100 – 185
ORT5	*Gypsy*	20 798	0·7	120	80 – 160
ORT6	*Gypsy*	1378	0·0	8	NE

NE, not estimated.

Note that ORT7 was not analysed due to the possible presence of other elements inserted within it, thus creating the potential for an erroneous prediction of genome coverage.

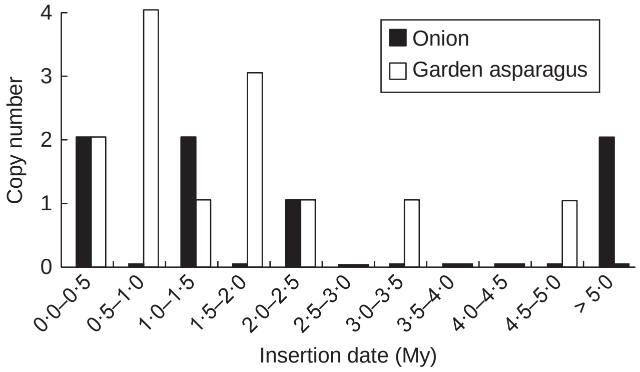

FIG. 3. Distribution of intact element insertion dates. Insertion dates were estimated using pairwise identity scores between the two LTRs for each copy (see Material and Methods for details).

shotgun reads for all elements showing .50 cases of high homology. For both onion and asparagus, all elements have an average identity .84 %. Using the same conversion rate as that used for estimating individual insertion dates, this indicates that the elements analysed were active within the last 6 million years (Fig. 3). Although these data only give a first estimate, because they are impacted by the age of the reference copy used, they are in accordance with the insertion dates found for individual insertions (Fig. 3, Table 2). Moreover, the detection of hits showing .95 % identity with our reference intact copy (for elements ART4, ART6, ART7, ART8, ORT3 and ORT7) reveals that, for a few elements, very recent amplification events have occurred in both genomes. ORT2, whose full-length copy was estimated to have inserted around 8 Mya (Table 2), had a copy number too low to make this pairwise alignments-based estimation.

Gene annotation

Several genes were predicted in these BACs in previous annotations (Jakše *et al.*, 2006, 2008). In our annotation, genes were validated only when a BLASTX hit to a gene from a eudicot species could be found, as matches to proteins of other monocots with unknown function have a reasonable chance to be TE genes, while a match between a monocot genome and an unknown protein from a dicot genome is more likely to be a real gene (Bennetzen *et al.*, 2004).

Moreover, when our extended characterization of TEs indicated overlap with a previously characterized gene, we made a concerted effort to discover whether the gene had been misannotated due to the fact that the element had not been detected, or if it corresponds to an insertion of the gene or gene fragment within the newly annotated element. Interestingly, on onion BAC DQ273270 originally selected for containing a sulfite reductase gene, we found that the LTR of the ORT1 element carries an insertion of unknown origin, within which is included part of a sulfite reductase gene with identity to both maize protein NP_001105302·1 and *Arabidopsis thaliana* protein NP_196079·1 (data not shown). Hence, the sulfite reductase homology on this BAC appears to be exclusively caused by the presence of a sulfite reductase fragment, perhaps acquired by a Pack-MULE (Jiang *et al.*, 2004), by another transposon with gene acquisition propensities (Yang and Bennetzen, 2009) or by LTR retrotransposon ORT1 itself.

DISCUSSION

LTR retrotransposons in onion and asparagus are young, diverse and organized in nested structures

Annotation of several contigs with size .35 000 bp for onion and garden asparagus revealed the presence of many distinct LTR retrotransposons. These elements make up .50 % of the regions analysed, while only a few genes could be characterized. This structure is similar to what is observed for large grass genomes, in which small gene-rich islands are interspersed in a sea of TEs, in particular LTR retrotransposons (San Miguel *et al.*, 1996), a structure also observed in banana, a non-grass monocot (D'Hont *et al.*, 2012). Because the BACs analysed were originally selected for containing agronomically important genes, the regions studied are expected to be gene rich, and this trend towards small gene islands is likely to be even more pronounced in the gene-poor regions of the genome (e.g. paracentromeric heterochromatin).

The observation of two copies for some elements in this limited BAC data set suggests that some families have become very abundant, as is seen in maize, wheat and barley (Vitte and Bennetzen, 2006). This may be particularly true for onion, as the contigs analysed correspond to only 0·001 % of the total genome size, as compared with 0·017 % for the contigs analysed from garden asparagus. This feature is reinforced by analyses of the shotgun sequence data indicating that some elements represent at least 2 % of these genomes (for instance, ART4, ART6, ART7, ART8 and ORT3 represent 3·6 – 5·5, 1·1 – 2·1, 1·8 – 4·1, 1·7 – 2·6 and 3·4 – 4·5 % of the data, respectively). Because the onion genome is much larger (approx. 16 000 Mb) than that of asparagus (approx. 1300 Mb), similar proportional genome occupancy leads to much higher sequence coverage for onion. For instance, the ORT3 element is estimated to make up .550 Mb within the onion nuclear genome (a TE abundance greater than the size of the rice genome, for instance), while the ART4 element is estimated to make up only 50 – 70 Mb of the asparagus genome sequence. Resampling the data with replacement indicated that the accuracy of these estimations is approx. 13 – 35 % for the most abundant elements (Table 3). Hence, this rough comparison of genome coverage indicates .20-fold variation in abundance of different TEs in each species, indicating that some element families have amplified much more actively in recent times than have others, a feature observed for the Poaceae, but also for other more recently analysed monocot species such as banana (D'Hont *et al.*, 2012) and date palm (***Phoenix dactylifera***) (Al-Dous *et al.*, 2011).

Both LTR comparison of intact elements and pairwise analysis of shotgun data show that the different LTR retrotransposon families found have been active within the past 6 million years. This period of activity seems to be more recent for some families such as ART4, ART6, ART7, ART8, ORT3 and ORT7. This suggests that both asparagus and onion genomes have undergone amplification of LTR retrotransposon in their recent history.

The observation of Ts:Tv ratios of 3·2:1 and 1·3:1 in asparagus and onion LTR sequences, respectively, is in accordance with previous results showing that Ts:Tv rates in LTRs differ dramatically among species, from 1·6 (barley) to 3·9 (maize) (Vitte and Bennetzen, 2006). Genic regions (including introns) typically exhibit Ts:Tv ratios of 1:1. Therefore, it has been argued that a higher Ts:Tv ratio is evidence of extensive cytosine 5-methylation, because this epigenetic DNA modification increases the C to T transition rate (SanMiguel *et al.*, 1998). Therefore, as for other plant genomes that were previously analysed, most LTR retrotransposons from onion and asparagus are probably epigenetically silenced with extensive cytosine 5-methylation, at CG, CHG and CHH sites (Feng and Jacobsen, 2011).

Impact of the TE detection process on the biological features depicted

In the past few years, several contig sequences have been generated from BAC clones for onion and asparagus. Although global gene and TE content has been estimated from these BACs, detection of TEs had been made using homology searches on existing databases, as well as pairwise comparison of BAC sequences (Jakše *et al.*, 2006, 2008; Telgmann-Rauber *et al.*, 2007). Although this methodology gives a first approximation of TE content, it is not adequate to characterize TE abundance or organization within genomes, in particular for species with scarce genomic data and which are not closely related to a model species, as is the case for onion and asparagus. First, the absence of known TEs from a closely related species limits homology-based detection to the most conserved part of TEs. Therefore, with regular annotation processes, one is likely to miss segments within elements and conclude mistakenly that the elements are highly truncated and have degenerated. Secondly, the scarcity of genomic data limits within-genome comparisons and therefore hampers the detection of TEs based on their repetitiveness. For these reasons, we felt that the previous analyses concluding that TEs from onion and asparagus genomes were highly fragmented (Jakše *et al.*, 2006, 2008) required some further investigation. We used a combination of searches based on both homology and the detection of TE structural features to detect the presence of TEs in these two Asparagales species. Our analyses led to the discovery of several intact LTR retrotransposons from both species. The LTR retrotransposon database that we constructed is the largest available for these two species, but needs to expand dramatically as more data are generated and similar analyses to ours are undertaken. Subsequent comparison of random shotgun reads with the newly discovered intact elements allowed characterization of the age of amplification events and provided crude approximations of relative abundance for different LTR retrotransposon families.

Our results also highlight the importance of annotation quality to provide a clear description of TEs in plant genomes. We propose that combining analysis of large contigs to discover new intact TEs and random shotgun reads mapping onto these new TEs is an efficient method to characterize TE dynamics for any newly investigated and/or large genome species. Although this study was performed using data that had been generated using the older Sanger technology, current and future high-throughput methodologies (Mardis *et al.*, 2008) will clearly render such procedures so powerful and affordable that TE properties can be investigated in a vast number of species, therefore allowing analysis of TE dynamics across broader evolutionary times.

Finally, our analysis reveals that some of the sequences that were previously annotated as genes were erroneously predicted. This is mainly because TEs were not accurately detected, leading to false prediction of 'host genes' in the coding regions of these elements. The finding of a fragment of the sulfite reductase gene within the LTR of one LTR retrotransposon highlights the importance of precisely delimitating TE positions to better characterize genome structure and evolution. This is particularly important when comparative genome analyses are performed, as insertions of gene fragments within TEs may lead to misleading predictions of a lack of synteny between species.

ACKNOWLEDGEMENTS

We thank Chris Town for sharing onion shotgun data, Rod Wing and Dave Kurdna for performing asparagus shotgun sequencing, Eugene McCarthy for discussions regarding LTR-STRUC results, and Johann Joets for bioinformatic support and discussion on the shotgun data analysis. This work was supported by NSF grant #0607123 (to J.L.B.) and NSF grant #0841988 (to J.L.M.).

LITERATURE CITED

Al-Dous EK, George B, Al-Mahmoud ME, *et al.* 2011. *De novo* genome sequencing and comparative genomics of the date palm (*Phoenix dactylifera*). *Nature Biotechnology* 29: 521 – 527.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215: 403 – 410.

Baucom RS, Estill JC, Chaparro C, *et al.* 2009. Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLoS Genetics* 5: pe1000732.

Bennetzen JL. 2000. Comparative sequence analysis of plant nuclear genomes: microcollinearity and its many exceptions. *The Plant Cell* 12: 1021 – 1029.

Bennetzen JL, Coleman C, Liu R, Ma J, Ramakrishna W. 2004. Consistent over-estimation of gene number in complex plant genomes. *Current Opinion in Plant Biology* 7: 732 – 736.

D' Hont A, Denoeud F, Aury JM, *et al.* 2012. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* 488: 213 – 217.

Do S, Suzuki G, Mukai Y. 2003. Genomic organization of a novel root alliinase gene, *ALL1*, in onion. *Gene* 325: 17 – 24.

Feng S, Jacobsen SE. 2011. Epigenetic modifications in plants: an evolutionary perspective. *Current Opinion in Plant Biology* 14: 179 – 186.

Jaks˘ e J, Telgmann A, Jung C, *et al.* 2006. Comparative sequence and genetic analyses of asparagus BACs reveal no microsynteny with onion or rice. *Theoretical and Applied Genetics* 114: 31 – 39.

Jaks˘ e J, Meyer JD, Suzuki G, *et al.* 2008. Pilot sequencing of onion genomic DNA reveals fragments of transposable elements, low gene densities, and significant gene enrichment after methyl filtration. *Molecular Genetics and Genomics* 280: 287 – 292.

Jansson T, Bremer K. 2005. The age of major monocot groups in- ferred from 800+rbcL sequences. *Botanical Journal of the Linnean Society* 146: 395 – 398.

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. 2004. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* 431: 569 – 573.

Jin YK, Bennetzen JL. 1994. Integration and nonrandom mutation of a plasma membrane proton ATPase gene fragment within the Bs1 retroelement of maize. *The Plant Cell* 6: 1177 – 1186.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 16: 111 – 120.

Kuhl JC, Cheung F, Yuan Q, *et al.* 2004. A unique set of 11,008 onion expressed sequence tags reveals expressed sequence and genomic differences between the monocot orders Asparagales and Poales. *The Plant Cell* 16: 114 – 125.

Kuhl JC, Havey MJ, Martin WJ, *et al.* 2005. Comparative genomic analyses in Asparagus. *Genome* 48: 1052 – 1060.

Lai J, Li Y, Messing J, Dooner HK. 2005. Gene movement by Helitron transposons contributes to the haplotype variability of maize. *Proceedings of the National Academy of Sciences, USA* 102: 9068 – 9073.

Ma J, Bennetzen JL. 2004. Rapid recent growth and divergence of rice nuclear genomes. *Proceedings of the National Academy of Sciences, USA* 101: 12404 – 12410.

Mardis ER. 2008. Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics* 9: 387 – 402.

McCarthy EM, McDonald JF. 2003. LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics* 19: 362 – 367.

Pearce SR, Pich U, Harrison G, *et al.* 1996. The Ty1-copia group retrotransposons of *Allium cepa* are distributed throughout the chromosomes but are enriched in the terminal heterochromatin. *Chromosome Research* 4: 357 – 364.

SanMiguel P, Tikhonov A, Jin YK, *et al.* 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science* 274: 765 – 768.

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. 1998. The paleontology of intergene retrotransposons of maize. *Nature Genetics* 20: 43 – 45.

Smit AFA, Hubley R, Green P. 1996 – 2010. *RepeatMasker Open-3·0*. http://www.repeatmasker.org.

Stack SM, Comings DE. 1979. The chromosomes and DNA of *Allium cepa*. *Chromosoma* 70: 161 – 181.

Stajner N, Bohanec B, Javornik B. 2002. Genetic variability of economically important *Asparagus* species as revealed by genome size analysis and rDNA ITS polymorphisms. *Plant Science* 162: 931 – 937.

Stevens PF. 2001 onwards*. Angiosperm Phylogeny Website. Version 9, June 2008* [and more or less continuously updated since]. http://www.mobot.org/MOBOT/research/APweb/

Suzuki G, Ura A, Saito N, *et al.* 2001. BAC FISH analysis in Allium cepa. *Genes and Genetic Systems* 76: 251 – 255.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4·0. *Molecular Biology and Evolution* 24: 1596 – 1599.

Telgmann-Rauber A, Jamsari A, Kinney MS, Pires JC, Jung C. 2007. Genetic and physical maps around the sex-determining M-locus of the dioecious plant asparagus. *Molecular Genetics and Genomics* 278: 221 – 234.

Vitte C, Bennetzen JL. 2006. Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. *Proceedings of the National Academy of Sciences, USA* 103: 17638 – 17643.

Yang L, Bennetzen JL. 2009. Distribution, diversity, evolution, and survival of *Helitrons* in the maize genome. *Proceedings of the National Academy of Sciences, USA* 106: 19922 – 19927.