Genomic Characterization for Parasitic Weeds of the Genus *Striga* by Sample Sequence Analysis

Matt C. Estep,* Bhavani S. Gowda, Kan Huang, Michael P. Timko, and Jeffrey L. Bennetzen

Abstract

Generation of ~2200 Sanger sequence reads or ~10,000 454 reads for seven *Striga* Lour. DNA samples (five species) allowed identification of the highly repetitive DNA content in these genomes. The 14 most abundant repeats in these *Striga* species were identified and partially assembled. Annotation indicated that they represent nine long terminal repeat (LTR) retrotransposon families, three tandem satellite repeats, one long interspersed element (LINE) retroelement, and one DNA transposon. All of these repeats are most closely related to repetitive elements in other closely related plants and are not products of horizontal transfer from their host species. These repeats were differentially abundant in each species, with the LTR retrotransposons and satellite repeats most responsible for variation in genome size. Each species had some repetitive elements that were more abundant and some less abundant than the other *Striga* species examined, indicating that no single element or any unilateral growth or decrease trend in genome behavior was responsible for variation in genome size and composition. Genome sizes were determined by flow sorting, and the values of 615 Mb [*S. asiatica* (L.) Kuntze], 1330 Mb [*S. gesnerioides* (Willd.) Vatke], 1425 Mb [*S. hermonthica* (Delile) Benth.] and 2460 Mb [*S. forbesii* Benth.] suggest a ploidy series, a prediction supported by repetitive DNA sequence analysis. Phylogenetic analysis using six chloroplast loci indicated the ancestral relationships of the five most agriculturally important *Striga* species, with the unexpected result that the one parasite of dicotyledonous plants (*S. gesnerioides*) was found to be more closely related to some of the grass parasites than many of the grass parasites are to each other.

*Striga* Lour., commonly known as witchweed, is a genus comprising 28 described species of parasitic plants with an old world distribution. Natural low-density populations infecting a diverse group of grass species can be found in tropical grasslands across much of Africa and parts of Asia (Mohamed et al., 2001). Most research has focused on the five African species that are agronomically important: *S. hermonthica* (Delile) Benth., *S. asiatica* (L.) Kuntze, *S. forbesii* Benth., *S. aspera* (Willd.) Benth., and *S. gesnerioides* (Willd.) Vatke. The latter parasites dicotyledonous plants such as cowpea (*Vigna unguiculata* (L.) Walp.) and tobacco (*Nicotiana tabacum* L.) while all other species within the genus are grass parasites. In agricultural environments, these five species are found in high-density populations parasitizing staple crops, including maize (*Zea mays* L.), sorghum (*Sorghum bicolor* (L.) Moench), and pearl millet (*Pennisetum glaucum* (L.) R. Br.).

In sub-Saharan Africa, *Striga* is the single most important biological limitation to food production (Ejeta, 2007). Despite its devastating effects on crop yields and subsistence farmer livelihoods, relatively little is known about the genetics, development, biochemistry, or physiology of this parasitic genus. Most studies have

**Abbreviations:** BAC, bacterial artificial chromosome; emPCR, emulsion polymerase chain reaction; GC, guanine plus cytosine; GS, Genome Sequencer; HI, homoplasy index; LTR, long terminal repeat; M, grown on maize; PAUP, phylogenetic analysis using parsimony; PTP, PicoTiterPlate; RVT, reverse transcriptase; S, grown on sorghum; SEB, sucrose extraction buffer; SRC, *Striga* retrotransposon copia; SSR, *Striga* retrotransposon gypsy; SS, *Striga* satellite; SSA, sample sequence analysis; ssDNA, single-stranded template DNA; TE, transposable element; TIGR, The Institute for Genomic Research; rRNA, transfer ribonucleic acid.
focused on Striga ecology (rhizosphere communication, seed bank dynamics, and haustorium formation) or control (host resistance and chemical, abiotic, or biotic control methods). Among the prominent gaps in Striga characterization is in understanding the phylogenetic relationships of the 28 described African species and the less-well-defined Asian species. The base ploidy level and the evolutionary role of polyploidyization, if any, within the genus are not clear. A genetic map has not been developed for any Striga species. One study has initiated rigorous investigations of genetic population structure in S. gesnerioides (Botanga and Timko, 2006). Another recent set of studies has explored the coding potential of S. hermonthica with expressed sequence tag sequencing and identified a single horizontal transfer of a nuclear encoded gene with unknown function between a grass host and Striga (Yoshida et al., 2010a, b).

Genetic characterization of any wild plant is a challenging exercise but particularly so with species that are obligate parasites, have no genetic maps, have few identified sequences, and have limited potential for cytogenetic analysis due to the absence of extensive root tissues. Hence, innovative technologies for efficient genome characterization need to be developed for Striga. One approach for genome characterization and genetic tool development is by sample sequence analysis (SSA) of randomly selected genomic DNA (Brenner et al., 1993). This technique has been used to explore genome composition in hexaploid wheat using the sequences of randomly selected bacterial artificial chromosome (BAC) clones (Devos et al., 2005), in the genus Oryza L. using BAC end sequences (Piegu et al., 2006), and within Gossypium L. and Pisum L. using genomic shotgun sequence data (Hawkins et al., 2006; Macas et al., 2007). All of these studies have concluded that transposable elements (TEs) constitute a large portion of plant nuclear genomes and can greatly influence genome size. Satellite (tandem) repeats have also been shown to occur in large and variable numbers that affect genome size.

In this study, we have used SSA to investigate and compare the genomes of several agronomically important Striga species. Our results indicate general similarities and some dramatic differences in genome composition among the species analyzed. We have also uncovered sequence variability that will be useful for future characterization of Striga genomes and studies of inheritance within each species.

Materials and Methods

Seven isolates representing five Striga species were used in this analysis. These isolates were originally collected from multiple African countries over a period of several years (Table 1). For purposes of DNA extractions, these seven isolates were grown on the species they were originally collected from—maize (S. asiatica [grown on maize (M)]), S. aspera, S. forbesii, and S. hermonthica [M]), sorghum (S. asiatica [grown on sorghum (S)] and S. hermonthica [S]), or cowpea (S. gesnerioides)—for approximately 60 d in a quarantine facility (Animal and Plant Health Inspection Service Plant Protection and Quarantine Permit No. 70902-P) at the University of Virginia. Vegetative above-ground stems were collected into liquid N and shipped on dry ice to the University of Georgia for nucleic acid extraction.

Phylogenetic Analysis of Selected Striga Species

Approximately 0.1 gm of frozen tissue was ground to a fine powder with liquid N in a 1.5 mL Eppendorf tube. Total genomic DNA was extracted using a modified cetyltrimethylammonium bromide procedure (Doyle and Doyle, 1987). Published primer sequences were used to amplify and subsequently sequence six loci from the chloroplast genome of each Striga species (Shaw et al., 2005). American bluehearts (Buchnera americana L.) was similarly analyzed and used as an out-group. Two of these loci are protein encoding: ribulose-bisphosphate carboxylase large subunit (RbcL) and MaturaseK (MatK) (Table 2). The remaining four loci are intergenic regions (noncoding) with primers anchored in conserved coding regions: transfer ribonucleic acid (tRNA) threonine through tRNA leucine (trnT [UGU]–trnL [UAAl]), tRNA aspartic acid through tRNA threonine (trnD [GUC]–trnT [GGU]), RNA polymerase β subunit through tRNA cysteine (rpoB–trnC [GCA]), and tRNA histidine through photosystem II protein D1 (trnH [GUG]–psbA) (Table 2). Sequences were aligned with SEAVIEW (Galtier et al., 1996) and then concatenated into a single super matrix. A parsimony analysis was conducted using phylogenetic analysis using parsimony (PAUP) (Swafford, 2002) with 1000 bootstrap replicates to evaluate relationships between the studied Striga species.

Genome Size Estimates

Genome size was measured by flow cytometry. Fresh leaf tissue was shipped to the Flow Cytometry and Imaging Core laboratory at the Virginia Mason Research Center. Fresh leaf tissue was not available for S. aspera, despite several attempts at growing this species, apparently due to a loss of seed viability in stored materials. Genome size measurements were repeated a minimum of four times with rice (Oryza sativa L.) (genotype TP-109) as a size standard using standard methodologies (Arumuganathan et al., 1999; Temsch and Greilhuber, 2000).

DNA Isolation for Sample Sequencing Analysis

Approximately 20 g of frozen tissue were ground to a fine powder with a mortar and pestle and immediately suspended in a sucrose extraction buffer (SEB) following option Y as described in Peterson et al. (2000). The suspension was filtered through two layers of cheesecloth and miracloth (EMD Chemicals, Inc.) to remove particulates, and 10% Triton X-100 (Sigma-Aldrich Corp.) (v/v) in SEB solution was then added to the filtered suspension at a 1:20 volume ratio. The Triton X-100 is used to selectively lyse plastid membranes (Gurtubay et al., 1980). Nuclei were isolated from the suspension using a series
of centrifugations at 650 × g for 15 min. Nuclei were then resuspended in 5 mL of buffer API from the DNeasy plant maxi-kit (Qiagen), and manufacturer’s instructions were followed to isolate plastid-free DNA.

Library Construction for Sanger Sample Sequencing Analysis

For each sample, 10 μg of DNA were sheared using a HydroShear (GeneMachines) set on speed code 14, for 20 cycles, to obtain DNA fragmented in the size range of 3 to 5 kb. The sheared fragments were converted into blunt-ended molecules using mung bean [Vigna radiata (L.) R. Wilczek] nuclease (F. Hoffmann-La Roche Ltd.), and A-tailed using Taq DNA polymerase (F. Hoffmann-La Roche Ltd.) and A-tailed using shrimp alkaline phosphatase (F. Hoffmann-La Roche Ltd.), and A-tailed using 

<table>
<thead>
<tr>
<th>Species</th>
<th>Host</th>
<th>Country of origin and year</th>
<th>Sequences</th>
<th>Total bp</th>
<th>% GC</th>
<th>Genome size pg per 2C</th>
<th>Mb per 1C</th>
<th>% genome sampled</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Striga forbesii</em></td>
<td>maize</td>
<td>Zimbabwe in 1984</td>
<td>2280</td>
<td>2,002,164</td>
<td>39.50</td>
<td>5.02 ± 0.05</td>
<td>-2460</td>
<td>-0.08</td>
</tr>
<tr>
<td><em>Striga asiatica</em></td>
<td>sorghum</td>
<td>Tanzania in 1992</td>
<td>2251</td>
<td>1,978,367</td>
<td>36.80</td>
<td>1.21 ± 0.02</td>
<td>-590</td>
<td>-0.3</td>
</tr>
<tr>
<td><em>S. asiatica</em></td>
<td>maize</td>
<td>Zimbabwe in 2001</td>
<td>2251</td>
<td>1,943,407</td>
<td>37.00</td>
<td>1.31 ± 0.02</td>
<td>-640</td>
<td>-0.3</td>
</tr>
<tr>
<td><em>Striga aspera</em></td>
<td>maize</td>
<td>Mali in 1984</td>
<td>2271</td>
<td>1,823,629</td>
<td>39.70</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td><em>Striga gesnerioides</em></td>
<td>cowpea</td>
<td>Benin in 2008</td>
<td>9945</td>
<td>3,744,566</td>
<td>40.60</td>
<td>2.71 ± 0.07</td>
<td>-1330</td>
<td>-0.31</td>
</tr>
<tr>
<td><em>Striga hemonthica</em></td>
<td>sorghum</td>
<td>Sudan in 1995</td>
<td>2233</td>
<td>1,876,213</td>
<td>40.30</td>
<td>2.88 ± 0.03</td>
<td>-1410</td>
<td>-0.13</td>
</tr>
<tr>
<td><em>S. hemonthica</em></td>
<td>maize</td>
<td>Nigeria in 1998</td>
<td>2123</td>
<td>1,870,527</td>
<td>41.20</td>
<td>2.95 ± 0.03</td>
<td>-1440</td>
<td>-0.13</td>
</tr>
</tbody>
</table>

Table 1. Identified genome characteristics of the investigated *Striga* species.

1 GC, guanine plus cytosine.

Table 2. Sequence of polymerase chain reaction primers used for phylogenetic analysis.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Forward (5’-3’)</th>
<th>Reverse (5’-3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rbp1</td>
<td>5’- CTCAATTGATATGGTTTGGTTCAG</td>
<td>5’- GCGGCGCCGTCTTTTGGTTCAG</td>
</tr>
<tr>
<td>Mark</td>
<td>5’- CTTGCTATGATGGTTTGGTTCAG</td>
<td>5’- CTTTATGCTATGATGGTTTGGTTCAG</td>
</tr>
<tr>
<td>tmG1</td>
<td>5’- TATGGATATGGTTTGGTTCAG</td>
<td>5’- GCGG CGCCGTCTTTTGGTTCAG</td>
</tr>
<tr>
<td>tmG2</td>
<td>5’- ACCCATGATGGTTTGGTTCAG</td>
<td>5’- CTTGCGGCCTATGATGGTTTGGTTCAG</td>
</tr>
<tr>
<td>npsG</td>
<td>5’- CACGATGATGGTTTGGTTCAG</td>
<td>5’- CCGGCGCCGTCTTTTGGTTCAG</td>
</tr>
<tr>
<td>16S 454</td>
<td>5’- GTGATGATGGTTTGGTTCAG</td>
<td>5’- CCGGCGCCGTCTTTTGGTTCAG</td>
</tr>
</tbody>
</table>

454 Sequencing

Full-scale 454 sequencing was performed at the Genomics Core Facility (Department of Biology, University of Virginia). For full-scale sequencing, the populated DNA capture beads were added to the DNA Bead Incubation Mix (F. Hoffmann-La Roche Ltd.) (containing DNA polymerase) and layered with enzyme beads (F. Hoffmann-La Roche Ltd.) (containing sulfurylase and luciferase) onto the PicoTiterPlate (PTP) (F. Hoffmann-La Roche Ltd.) device with two large regions. The loaded PTP device was placed into the GS FLX instrument and run using the Roche XLR70 Titanium sequencing chemistry kit.

454 Post-Run Analysis

Flow data were assembled using the GSAssembler package (454 Life Sciences Corp., 2008b) as described in the GS Data Analysis Software Manual (454 Life Sciences Corp., 2008a).
Corp., 2008a). Total raw and keypass wells for both regions were 1,147,687 and 1,127,168. The total number of sequences that passed all quality filters for both regions of the PTP was 820,210 with an average read length of 374 nucleotides representing 307,208,175 bases.

**Sampling 454 Data**

A total of 30,000 sequence reads were extracted from the total data set by randomly choosing 10,000 reads three times with a Perl script. Chloroplast and mitochondrial sequences were identified with cross_match (Ewing et al., 1998) and removed from the data sets before further analyses. Sequence annotation procedures (described below) were implemented on all three 454 data sets (~10,000 reads) with no observed differences in repetitive content or the repeats identified; therefore, results from only one data set are reported.

**Sequence Identification and Reconstruction**

Each of the six Sanger sequence data sets and the single 454 data set was subjected to an “all versus all” BLASTN (Altschul et al., 1990) using an expect value of $1 \times 10^{-5}$ or smaller as an acceptance threshold to assess the repetitive nature of the samples. Sequences were grouped into three categories: single copy, between 2 and 10 copies, and 11 or more copies within the data set. Each of the libraries was also annotated by comparing them to known repetitive sequences in The Institute for Genomic Research’s (TIGR’s) repeat database version 8.0 (Ouyang and Buell, 2004) using programs BLASTN and TBLASTX. Extended repeat pseudomolecules were also constructed for each of the libraries using the AARF software package (DeBarry et al., 2008). The 18 largest pseudomolecules (>3 kb) built with AARF were structurally annotated using blast searches (BLASTN and TBLASTX) against several Pfam databases (Finn et al., 2008). These databases included those for retroelement domains such as GAG (PF00077), reverse transcriptase (RVT) (PF00078), integrase (PF00665), ribonuclease H (PF00075), reverse transcription primer binding sites (Wilhelm et al., 1994), and tRNA priming sites. Pfam’s transposase (PF00872) database was also used (http://pfam.sanger.ac.uk/). Known TEs were sought using the TIGR plant repeat database (Ouyang and Buell, 2004) and the MIPS REdat database (Spannagl et al., 2007). These sequences were also subjected to analysis with Tandem Repeat Finder (Benson, 1999). All of the BLAST searches for each of the pseudomolecules assembled by AARF were visualized simultaneously with the Apophlo genome annotation browser (http://apollo.berkeleybop.org/current/index.html) and hand curated to identify class and family designations for each pseudomolecule. A Striga-specific repeat database was constructed using the pseudomolecules identified in this analysis to produce a more accurate annotation of each library using BLASTN.

**Annotation and Ninety-Five Percent Confidence Intervals**

Many TEs share regions of homology (see above) and therefore must be identified or annotated by a competitive process. All of the sample sequences were annotated using BLASTN (Altschul et al., 1990) and the Striga-specific repeat database constructed from the identified pseudomolecules. An expect value of $1 \times 10^{-10}$ or lower was used as the criterion for accepting an annotation. The annotated sequences from a single sample were then transformed into a percent of total base pairs for each repeat by dividing the total length (bp) of the homology to a repeat by total sequence length. For example, 654 bp repeat length/764 bp sequence read length = 0.856% of total bases for a single sequence. This approach allowed a precise count for each repetitive base pair identified in the samples instead of counting the entire sequence as repetitive. The transformed data on percent genome occupation were then randomly inserted to a data matrix containing the same number of cells as sequence reads for each sample library and subjected to a bootstrap analysis with replacement using SAS software (SAS Institute, 1994) with 1000 permutations. The 1000 bootstrap results were sorted by value and the 25th, 975th, and mean values were graphed to display differences between the genomes as a mean with a 95% confidence interval for each species or genotype.

**Phylogenetic Analysis of Retroelements**

All of the nucleotide sequences reads were compared by TBLASTX (Altschul et al., 1990) to the Pfam database (PF00078, version 22.0; http://pfam.sanger.ac.uk/; Finn et al., 2008) representing all known RVT sequences associated with retroelements in flowering plants. Sequences with homology were translated into all six possible reading frames and BLASTX was used to identify the correct translation. These sequences and 30 RVT sequences from a diverse selection of flowering plants were then aligned with ClustalX (Thompson et al., 1997) and SEAVIEW (Galtier et al., 1996), removing unalignable sequences and removing any overhanging base pairs so that all sequence comparisons contained the same number of amino acid residues. The final alignment was then used to construct a neighbor-joining tree with bootstrap values representing 1000 replicates.

**Results**

**Phylogenetic Analysis**

Phylogenetic analysis of the agriculturally important *Striga* species was performed in PAUP (Swafford, 2002) using parsimony. An alignment of six chloroplast loci concatenated together for all seven accessions and an outgroup included 4434 nucleotides character states (Supplemental Table S1). Four thousand ninety-eight of the characters states were constant. Three hundred thirty-six character states were variable, of which 180 were parsimony-informative characters. A single most parsimonious tree was generated (Fig. 1). This tree has a consistency index of 0.9691, a homoplasy index of 0.0309, a retention index of 0.9560, and strong bootstrap support. The tree indicates that *S. forbesii* is a basal lineage within the genus
and that *S. aspera*, *S. gesnerioides*, and *S. hermonthica* form the most derived clade. This analysis also suggests that *S. aspera* and *S. gesnerioides* are more closely related to each other than either is to *S. hermonthica*.

**Genome Size Estimates**

Flow cytometric estimates of genome size were generated from fresh plant material using *O. sativa* (genotype TP-109) as a standard. Each sample was measured a minimum of four times using previously published methods (Arunagamanathan et al., 1999; Temsch and Greilhuber, 2000). The raw values were averaged to generate a mean value with standard deviation (Supplemental Table S2). Mean genome sizes ranged from 1.21 pg per 2C in *S. asiatica* (S) to 5.02 pg per 2C in *S. forbesii* (Table 1). These values were converted into megabase pair estimates by multiplying by 980 Mb per pg and dividing by 2 to obtain haploid sizes, resulting in values ranging from ~590 Mb per 1C to ~2460 Mb per 1C (Table 1). No significant differences were observed between the two *S. asiatica* or the two *S. hermonthica* genome size estimates. While each species was sampled with about the same number of DNA sequences in the subsequent analyses (except *S. gesnerioides*), the approximate four-fold range in genome size means that the percent of each genome sampled by Sanger SSA varied from ~0.08% in *S. forbesii* to ~0.3% in *S. asiatica* (Table 1).

**Characteristics of Each Sequencing Library**

The average read length for all of the Sanger sequences generated was ~850 bp, with the *S. aspera* library having the smallest average read length of ~803 bp and the *S. hermonthica* (M) library having the largest average read length of ~880 bp. The average read length for the *S. gesnerioides* 454 library was ~377 bp. After removing vector, chloroplast, mitochondrial, and bacterial sequences, ~2200 sequences were generated from each Sanger sequencing library to produce ~1.8 to ~2.0 Mb of nucleotide sequence per sample. About 10,000 sequences were sampled from the 454 library to produce ~3.7 Mb of nucleotide sequence (Table 1). The guanine plus cytosine (GC) content was calculated for each library, ranging from 36.8% in *S. asiatica* (S) to 40.6% in *S. gesnerioides*. Guanine plus cytosine content was not significantly different between the two samples of *S. asiatica* (grown on maize [M] and grown on sorghum [S]) or the two samples of *S. hermonthica* (M and S).

**Accessing Repetitive Nature of Sample Sequences**

To determine the repetitiveness of each genome sampled, Sanger sequences or 454 sequences from a given library were compared against themselves by an all-by-all BLAST (Altschul et al., 1990), and the total number of identified homologies (“hits”) for each sequence was recorded. Sequences were grouped into three categories based on the number of hits counted: single copy, repetitive (2–10 copies), and highly repetitive (>10 copies). The groupings are displayed graphically by arranging the seven data sets in phylogenetic order (Fig. 2).

By the criteria used, approximately 60% of the two *S. asiatica* samples were classified as single copy in the data set with the remainder being repetitive (~30%) or highly repetitive (~10%). The largest genome sampled, *S. forbesii*, showed equal (~33%) distribution of single copy, repetitive, and highly repetitive sequences. The remaining four samples ranged from 40 to 56% single copy sequences and 25 to 40% repetitive and 18 to 25% highly repetitive sequences. It is worth noting that the low degree of sampling for each genome means that most sequences with copy numbers of less than several hundred per genome would be found only once in the sequenced sample (and therefore called “single copy”) while all those sequences found in the repetitive and highly repetitive categories are mostly from repeats that are found in copy numbers in the thousands or tens of thousands, respectively, in the haploid nuclear genome.

**Identification and Annotation of Repetitive Sequences**

A total of 337 pseudomolecules were generated from independent investigations of each of the seven data sets (Sanger and 454) using the AAARF program (DeBarry et al., 2008) for de novo repeat assembly. The AAARF software was used multiple times on each data set, using a range of parameters that can be adjusted in the program header. The settings used were MinBlastIdentity ranging from 80 to 89%, Blast_e ranging from $10^{-10}$ to $10^{-25}$, BL2Seq_e ranging from $10^{-3}$ to $10^{-16}$, and Min_Extend_Len ranging from 0 to 10. These settings were tested to investigate reproducibility and to build the longest possible pseudomolecules. Eighteen pseudomolecules >3 kb were constructed. These 18 were further investigated and subjected to structural annotation to identify domains commonly found in TEs. Thirteen pseudomolecules were identified as probable TEs (one long interspersed element [LINE]-like retroelement, five *gypsy*-like long terminal repeat [LTR] retrotransposons, six *copia*-like LTR retrotransposons, and one DNA transposon). The remaining five pseudomolecules did not exhibit homology to common structural domains found in TEs. The Tandem Repeats Finder software (Benson, 1999) indicated that these five pseudomolecules are composed of tandemly repeated sequences. Comparing the results from each of the 18 pseudomolecule annotations revealed that several pseudomolecules built from separate data sets represented the same TE or tandem repeat, leaving a total of 14 novel repetitive sequences (Table 3). None of the 10 retroelements identified appears to be a full-length pseudomolecule due to the lack of a GAG domain in most or an integrase domain in some and the inability to identify at least one LTR for the nine *copia*-like or *gypsy*-like LTR retrotransposons.

**Annotation and Comparison of Sample Sequences**

The 14 repeats were used to investigate each of the sequences from all seven samples using BLASTN (Altschul et al., 1990). The total length (bp) of the homology to a repeat was divided by the total length of the query sequence to identify the percent of each sequence composed of a repeat. To generate a prediction of the degree of
accuracy in these estimates of total genome repeat content, transformed values for a given repeat were resampled 1000 times using bootstrap statistics to identify the mean and 95% confidence intervals for each repeat found in each library. These statistics were graphed for the different classes of TEs and tandem repeats and for representative individual repeats, to investigate the variation between samples (Fig. 3A and 3B).

The percentage of each sequence sample composed of identified LTR retrotransposons ranged from ~2.3% in the S. asiatica samples to ~11.7% in the S. gesnerioides sample. The DNA transposons ranged from <1% in the S. asiatica samples to ~2% in the S. forbesii sample and identified tandem repeats ranged from ~0.5% in the S. gesnerioides sample to ~9.4% in one S. hermonthica (M) sample (Fig. 3A). The two most repetitive copia and gypsy retrotransposons, the LINE retroelement, and the most repetitive satellite repeats were graphed to display the differences between samples (Fig. 3B). The LTR retrotransposon identified as Wico (Striga retrotransposon copia [SRC] 1) ranged from ~1.8% in the S. asiatica samples to ~9.8% of total sequences in the S. forbesii sample. The LTR retrotransposon identified as SRC4 ranged from ~0.5% in the S. aspera sample to ~3.8% of total sequences in the S. forbesii sample.

Table 3. Newly identified Striga transposable elements and satellite repeats.

<table>
<thead>
<tr>
<th>Repeat type</th>
<th>Repeat name</th>
<th>Abbreviated name</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Satellite</td>
<td>S. asiatica</td>
<td>SS1</td>
<td>155 bp</td>
</tr>
<tr>
<td>DNA transposon</td>
<td>S. asiatica</td>
<td>SS2</td>
<td>147 bp</td>
</tr>
<tr>
<td>Retroelements</td>
<td>S. asiatica</td>
<td>SS3</td>
<td>183 bp</td>
</tr>
<tr>
<td></td>
<td>S. guineensis</td>
<td>ST1</td>
<td>2690 bp</td>
</tr>
<tr>
<td></td>
<td>S. forbesii</td>
<td>SRC1</td>
<td>5759 bp</td>
</tr>
<tr>
<td></td>
<td>S. hermonthica</td>
<td>SRC2</td>
<td>4558 bp</td>
</tr>
<tr>
<td></td>
<td>S. hermonthica</td>
<td>SRC3</td>
<td>1184 bp</td>
</tr>
<tr>
<td></td>
<td>S. hermonthica</td>
<td>SRC4</td>
<td>4527 bp</td>
</tr>
</tbody>
</table>

Table 3. Newly identified Striga transposons and satellite repeats.

Figure 1. Phylogenetic tree of the agriculturally important Striga species based on six chloroplast loci (Table 2). Bootstrap values from 1000 replicates are indicated at nodes.

Figure 2. Graphical representation of repetitiveness for each Striga isolate sampled. Individual samples were compared to themselves by an all-by-all BLAST (Altschul et al., 1990) and the total numbers of homologies were counted. Sequences were grouped into three categories [single copy, repetitive [2–10 copies], and highly repetitive [>10 copies]] based on their copy number within each data set. M, grown on maize; S, grown on sorghum.

Table 3. Newly identified Striga transposable elements and satellite repeats.

<table>
<thead>
<tr>
<th>Repeat type</th>
<th>Repeat name</th>
<th>Abbreviated name</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Satellite</td>
<td>S. asiatica</td>
<td>SS1</td>
<td>155 bp</td>
</tr>
<tr>
<td>DNA transposon</td>
<td>S. asiatica</td>
<td>SS2</td>
<td>147 bp</td>
</tr>
<tr>
<td>Retroelements</td>
<td>S. asiatica</td>
<td>SS3</td>
<td>183 bp</td>
</tr>
<tr>
<td></td>
<td>S. asiatica</td>
<td>ST1</td>
<td>2690 bp</td>
</tr>
<tr>
<td></td>
<td>S. forbesii</td>
<td>SRC1</td>
<td>5759 bp</td>
</tr>
<tr>
<td></td>
<td>S. hermonthica</td>
<td>SRC2</td>
<td>4558 bp</td>
</tr>
<tr>
<td></td>
<td>S. hermonthica</td>
<td>SRC3</td>
<td>1184 bp</td>
</tr>
<tr>
<td></td>
<td>S. hermonthica</td>
<td>SRC4</td>
<td>4527 bp</td>
</tr>
</tbody>
</table>

†LINE, long interspersed element.
The LTR retrotransposon identified as Wigy (Striga retrotransposon gypsy [SRG] 1) ranged from less than 1% of the sequences in the *S. asiatica* samples to ~3% in the *S. aspera* sample. Another gypsy retrotransposon (SRG5) ranged from undetected in the *S. asiatica* samples to ~2.1% in the *S. hermonthica* (M) sample. The only repeat identified as a LINE (Striga retroelement LINE [SRL] 1) was fairly constant in genome representation, ranging from just under 1% in the *S. asiatica* and *S. forbesii* samples to ~1% in *S. hermonthica*. The satellite repeat identified as Striga satellite (SS) 1 ranged from undetected in *S. forbesii* to ~9.3% in the *S. hermonthica* (M) sample. In total, 4691 sequences were annotated as one of the 14 repeats identified using the AAARF software (DeBarry et al., 2008), and this amounted to about 20% of the total input sequence data.

Figure 3. A) Analysis of Striga samples for repetitive groups commonly found in plant genomes. Means and 95% confidence intervals were generated by bootstrap analysis of percent composition for common repeat groups for each sample. B) Comparison of Striga samples for the most repetitive copia, gypsy, long interspersed element (LINE), and tandem repeats. Means and 95% confidence intervals were generated by bootstrap analysis of percent composition for the most highly repetitive repeats for each sample. M, grown on maize; S, grown on sorghum; SRC, Striga retrotransposon copia; SRG, Striga retrotransposon gypsy; SRL, Striga retroelement LINE; SS, Striga satellite.
Phylogenic Analysis of Retroelements

In an attempt to investigate whether the LTR retrotransposons identified in our repeat analysis and pseudomolecule reconstruction actually represent distinct clades, phylogenetic trees were constructed using RVT sequences homologies identified within the total data set. Several known retroelements from other species were used to orient the phylogenetic analysis and to search for evidence of horizontal transfer for these most abundant repeats. A total of 186 RVT sequences were identified in the seven samples and translated into the correct reading frame. Two separate amino acid alignments were generated including 171 sequences. The first alignment contained 46 sequences that aligned from the start codon of RVT for ~176 amino acid residues and were all annotated as Wico (data not shown). The second alignment contained a total of 125 sequences aligned across 171 amino acids, and this represented the bulk of RVT sequences that could be aligned within our data set (Fig. 4). Each of the sequences within this alignment was annotated with the Striga repeat database generated in this project after the phylogenetic tree was constructed. The resulting annotations were mapped onto the phylogeny. Eight of the repeats identified as retrotransposons are identified on the phylogeny as monophyletic bootstrap-supported clades, except SRG5, which only contained homology to the GAG domain. It is important to note that Wico (SRC1) is underrepresented in this phylogeny because 46 sequences representing this clade could not be aligned with those used to generate this tree.

In the unrooted phylogenetic tree shown in Fig. 4, sequences (including 30 elements from a diverse set of flowering plants) identified as copia-like and gypsy-like form two distinct clades that are sister to each other while the LINE-like sequences are grouped into several clades and/or grades that all appear basal to the copia and gypsy clade. These results are in complete agreement with the general observation that LINEs are the most ancient and diverse class of retroelements and that the two superfamilies of LTR retrotransposons, gypsy-like and copia-like, were derived from a single ancestral LTR retrotransposon (Xiong and Eickbush, 1990). This phylogeny can also be used to show that Striga RVT sequences are more closely related to other dicotyledon RVT sequences, suggesting vertical inheritance of retroelements within Striga rather than horizontal inheritance from grass host species.

Discussion

The parasitic lifestyle is found in only ~1% of flowering plant species (Kuijt, 1969). While many parasitic plants are not weeds, some such as the weedy Striga species have adapted to cultivated host plants that exhibit the reduced genetic diversity that is routinely associated with domestication and agriculture. The leap from a low density population inhabiting tropical grasslands to a high density population inhabiting agricultural lands has occurred at least once and possibly several times within this single genus over the relatively short period of time since agriculture began (i.e., the last 10,000 yr). The shift in lifestyle and the rapid breakdown of host resistance observed in field trials of newly developed Striga-resistant cultivars suggest that some species within the genus are evolving rapidly. It is possible that aspects of the predicted recent and rapid evolution of Striga would be manifested in the structure and/or composition of the weeds’ nuclear genomes.

The first inspection of Striga genomes, provided here by SSA, suggests that the nuclear DNA content of several agriculturally important species is fairly typical when compared to other studied plant genomes. Repeats were found to be quite abundant, and most of these were LTR retrotransposons as in other large-genome angiosperms. The analysis uncovered distinctive properties of the genomes of each of these species because both the paired S. hermonthica samples (from very different collection locations such as Sudan and Nigeria) and the paired S. asiatica samples (from Tanzania and Zimbabwe) exhibited high species-specific similarity in all genome properties analyzed (size, GC content, repeat types, and repeat abundances).

Phylogenetic analysis reveals that S. aspera and S. gesnerioides are sister species and that S. hermonthica is closely related to both. This observation is unexpected because S. aspera and S. hermonthica are morphologically very similar and are believed to have the ability to cross hybridize while S. gesnerioides is morphologically distinct (Aigbokhan et al., 2000). While the genus contains ~28 species, only five of them are considered agricultural weeds. Our results suggest three distinct clades within the genus. The most derived clade includes S. aspera, S. gesnerioides, and S. hermonthica while the two other weedy species (S. asiatica and S. forbesii) represent independent lineages. These results are congruent with the most detailed hypothesis of relationships within the family Orobanchaceae (Morawetz et al., 2010). It was unexpected that S. gesnerioides was found to be closely related to S. hermonthica and S. aspera instead of basal within the genus, as previously hypothesized based on morphology. This suggests a rapid morphological change in the legume parasite, perhaps accompanying a recent switch in host preference. Further analysis of this important observation should include low copy nuclear loci.

A traditional approach to understanding genome evolution has been to count chromosomes in meiotic (haploid) or mitotic cells (diploid). The highly transient haustorium (attachment organ) of parasitic plants replaces the root tip, making this tissue’s use for cytogenetics impractical. Hence, other tissues have been used to estimate chromosome numbers in several Striga species (Aigbokhan et al., 2000; Iwo et al., 2008; Musselman et al., 1991). Haploid chromosome numbers were reported as 20 for S. asiatica and 30 or 40 for S. hermonthica depending on geographic location, using anther squashes from Striga flowers (Iwo et al., 2008). A second study using pollen mother cells reported haploid chromosome numbers of 27 for S. aspera and 32 for S. hermonthica (Musselman
Figure 4. Neighbor-joining analysis of *Striga* reverse transcriptase sequences. Unrooted neighbor-joining analysis of 125 *Striga* (colored circles) and 30 non-*Striga* (black circles) reverse transcriptase sequences support the monophyly of the long terminal repeat (LTR) retrotransposons identified with AAARF [DeBarry et al., 2008]. Each retroelement is boxed and shaded within the tree and identified to the right of taxon names (see Table 3 for abbreviations). Bootstrap values >50 are indicated at supported nodes. Abbreviations were used for some *Striga* species (Sherm, *S. hermonthica*; Sges, *S. gesnerioides*; Sasia, *S. asiatica*; Sforb, *S. forbesii*). LINE, long interspersed element; SRG, *Striga* retrotransposon gypsy; SRC, *Striga* retrotransposon copia; SRL, *Striga* retroelement long interspersed element (LINE).
The most recent chromosome counts using shoot tips reported haploid chromosome numbers of 18 for *S. aspera* and 19 for *S. hermonthica* (Aigbokhan et al., 2000). Chromosome counts were also reported for *S. aspera* \(n = 30\), *S. gesnerioides* \(n = 20\), and *S. forbesii* \(n = 22\) using anther squashes (Iwo et al., 2008). The wide variability (almost 2x) reported in chromosome counts for *Striga* species shared in these studies suggests that our understanding of chromosome constitution in any *Striga* species remains tenuous and is likely to be revised on further study.

An approximate four-fold difference in genome size was observed within this sampling of the genus *Striga*, suggesting that the processes of polyploidization and/or TE amplification are influencing genome evolution. The genome size values suggest a ploidy series with two rounds of polyploidization: a diploid species having a genome size of \(~600\) Mb (*S. asiatica*), a tetraploid species with a genome size of \(~1400\) Mb (*S. hermonthica*), and an octoploid species with a genome size of \(~2400\) Mb (*S. forbesii*). Several rounds of polyploidization have been observed and inferred to drive speciation in the evolution of a closely related genus, *Orobanche* L. (Schneeweiss et al., 2004).

In possible disagreement with the polyploidization hypothesis are the values calculated for percent GC content, total repetitiveness, and phylogenetic position (Table 1; Fig. 1 and 2). Both GC content and total repetitiveness values were found to increase at larger genome sizes, with an \(~3\)% variation in GC content and an \(~1.5x\) variation in percent repeat content across the species sampled. These increases would not be expected if simple polyploidy was the only cause of genome size variation, therefore indicating that other process(es) also contribute to shaping these genomes. Phylogenetic relationships do not dispute the polyploidization hypothesis but suggest independent polyploidization events within the *S. forbesii* lineage unrelated to polyploidization in the more derived clades within the genus.

Using de novo repeat assembly and annotation approaches, 14 different families of repetitive DNA were identified in the *Striga* genomes examined herein. These repeats are the most abundant sequences in these *Striga* genomes, representing approximately 10.7, 15.9, 16.1, 18.5, and 19% of the *S. asiatica*, *S. forbesii*, *S. gesnerioides*, *S. aspera*, and *S. hermonthica* nuclear genomes, respectively. Each repeat belongs to a class or group of repetitive sequences found commonly in plant genomes. All of these most abundant TEs appear related to their closest relatives in other dicot species, suggesting that none of them originated as horizontal transfers from their grass hosts. This result does not exclude the possibility of horizontal transfer of lower copy repetitive elements not identified in this study.

Comparing the percent content of individual repeats across species permits identification of the differences between the genomes sampled. As an internal control, both the *S. asiatica* and *S. hermonthica* samples represent genotypes collected from different host species and different geographic locations. No significant differences were observed between either of the paired samples. Although these results suggest that the analysis provides definitive species-specific descriptions of repeat content, the data set was not deep enough to produce full-length pseudomolecules for any of the LTR retrotransposons. Hence, the observed contributions of these repeats are a minimal estimate of their true contributions. Moreover, most or all of those repeats with copy numbers less than a few hundred are expected to be missed by this analysis. In the \(~2400\) Mb maize genome, these low-copy-number repeats, commonly LTR retrotransposons, make up about \(16\)% of the nuclear DNA (SanMiguel and Bennetzen, 1998; Baucom et al., 2009). Therefore, the overall repeat composition of all of the studied genomes is very likely to be \(>50\)% in line with comparable results for other angiosperm genomes (Bennetzen, 2000; Kazazian, 2004).

The low genomic redundancy (0.08 to 0.3% coverage) of the sample sequence data also limited the completeness and detail available for examining the phylogenetic relatedness of the discovered LTR retrotransposons. However, next generation sequencing technologies will allow more comprehensive sampling and therefore more complete discovery of repeats in any genome (Mardis, 2008). The AAARF program (DeBarry et al., 2008) for the discovery and assembly of repeats into pseudomolecules has proven competent to function with short sequence reads in this study and others, so more pseudomolecules should be available for future phylogenetic tree assembly.

In this study we used both Sanger (long reads with a cloning step) and 454 (short reads without cloning) sequences; these differences in data generation may affect comparisons between these two types of sequence data. Therefore, comparisons of *S. gesnerioides* to the other *Striga* species in this study should be treated with some caution.

Examining the percent composition differences between samples and different repeat groups (i.e., retropseudos, DNA transposons, and tandem and/or satellite repeats) revealed that each of these repeat classes are differentially influencing genome evolution in these species (Fig. 3A). Transposable elements are more abundant in the larger genomes. Satellite repeats display a different pattern, with variation that does not correlate with genome size (\(~8\)% in the small to midsized *S. asiatica*, *S. aspera*, and *S. hermonthica* genomes, \(~1\)% in *S. gesnerioides*, and \(~3\)% in the largest genome, *S. forbesii*). The *S. forbesii* genome contains about the same number of megabase pairs of tandem and/or satellite repeat as *S. asiatica*; however, the *S. forbesii* genome is about four times the size of the *S. asiatica* genome. The data indicate that LTR retrotransposons and satellite repeats are the major contributors to the variation in genome size. However, the differences in observed repeat content in the species studied cannot fully explain the observed overall genome size variation. From this analysis, the “single copy” components of *S. asiatica*, *S. hermonthica*, *S. gesnerioides*, and *S. forbesii*
are ~370, ~640, ~687, and ~810 Mb, respectively, once again suggesting a ploidy series. It is not clear whether this might be a 2x, 4x, and 6x series or only a 2x and 4x series, with species that differ in further DNA variations (e.g., segmental duplications, aneuploidies, other repeat amplifications) not identified in this study.

Further examination of LTR retrotransposon composition by subclass (i.e., copia and gypsy) indicates that gypsy-like retrotransposons are relatively more abundant in S. aspera, S. gesnerioides, and S. hermonthica samples than in the larger S. forbesii genome while copia-like retrotransposons are more abundant in the S. forbesii sample than in the midsized genome samples. Similar results have been observed in other plant species, and cases in cotton (Gossypium species) and Oryza australiensis. Domin suggests that amplification of one or a few LTR retrotransposons families have been responsible for most of the nuclear genome growth seen in particular lineages (Hawkins et al., 2006; Piegù et al., 2006). However, the situation in Striga is similar to that in the genus Zea in which LTR retrotransposons and satellite repeats of several different families are differentially amplified across several studied genomes (M.C. Estep and J.L. Bennetzen, unpublished data, 2011). As in the case of Zea, the larger Striga genomes such as S. forbesii contain more of some repeats but less of others compared to their smaller genome relatives. Overall, a greater quantity of repeats accounts for part of the increased genome size of the larger-genome Striga species, but this is not explained by a single element family or an exclusive genome growth trend in one direction. Hence, it appears that perhaps random but certainly stochastic activation of different TE and satellite repeats accounts for much of the genome size variation in the genus Striga.

Examining each individual repeat suggests that Wico (SRC1) had the largest effect on differences seen within the copia superfamily and that Wigy (SRG1) and SRG5 had the largest effect within the gypsy superfamily. Wigy comprises significantly more of the S. aspera sample (~3.1%) than the S. forbesii sample (~1.5%, ~40 Mb), but neither is significantly different from the S. hermonthica samples (~2.2%, ~30 Mb). Striga retrotransposon gypsy 5 on the other hand, constitutes significantly less of the large S. forbesii genome (~0.2%, ~5 Mb) than of the S. aspera (~1.6%), S. gesnerioides (~1.6%, ~21 Mb), or S. hermonthica samples (~1.9%, ~27 Mb).

Based on length, percent GC content, and lack of identifiable coding potential, the three tandem and/or satellite repeats identified in this study are similar to other satellite repeats found in centromeres and heterochromatin knob structures of several plant species (Plohl et al., 2008). The repeat SS1 constitutes ~8% of most of the genome samples, similar in percent sample composition in some maize lines to the amount of 180 bp satellite repeat responsible for knobs (Peacock et al., 1981). Although no SS1 was detected in S. forbesii, it has additional copies of SS2 and SS3 that make its overall satellite repeat content (~2.5% of the genome, ~62 Mb) very similar to that found in S. asiatica (~8.1% of the genome, ~50 Mb). In specific genetic backgrounds of maize, those chromosomes with the greater numbers of knob repeats were preferentially passed on to offspring but only through the female gametophyte. This process of distorted segregation was later termed meiotic drive (Rhoades and Dempsey, 1966). Meiotic drive may explain the selfish overaccumulation of satellite repeats over time and therefore their abundance in many Striga genomes.

Conclusions
In this study, SSA was applied to an important but challenging and understudied set of plant species to investigate global processes that can shape nuclear genomes. These modes of genome variation include polyploidization, TE amplification, and possibly meiotic drive. Striga species were found to have typically complex angiosperm genomes, with LTR retrotransposons and satellite repeats the major players in genomic size variation and overall diversity. Evidence of recurrent polyploidy was also uncovered.

Supplemental Information Available
Supplemental material is available at http://www.crops.org/publications/tpg.
Supplemental Table S1. Concatenated DNA sequence alignment used for phylogenetic analysis of Striga species.
Supplemental Table S2. Raw flow cytometry values used to estimate genome size.

Acknowledgments
We thank Dr. P. SanMiguel and the core facility at Purdue University for sequencing shotgun libraries, Dr. J. DeBarry (Univ. of Georgia) for help in implementing the AAARF algorithm, and Dr. E. Kellogg for lab resources used during the review process. We also thank John Chuckalovcak and Anthony Spano for technical assistance with the 454 data. This work was supported by grants from the Fulbright Foundation, the Guggenheim Foundation, the Georgia Research Alliance, and the Giles Professorship. MPT, BSG, and KH were supported by a grant from NSF (IBN-0322420) and (DBI-07-01748; CR-20006-477283).

References
454 Life Sciences Corp. 2008b. GSAssembler package version 2.0.00.22. A Roche Company, Bradford, CT.


