

Comparative Genomics of Transposable Elements in the Grasses

by

Patrick French

Biology Departmental Honors Thesis

Appalachian State University

May 2020

## **Abstract**

Transposable elements (TE's) are the most abundant genetic loci found in eukaryotic genomes and often occupy more than 70% of the genome landscape. These genetic elements were first described in maize (*Zea mays*) and have been found in all eukaryotic genomes investigated. The grass family (Poaceae), has long been used as a model system to study transposable elements. Transposable element content has been analyzed in many grass species including, Maize, Sorghum (*Sorghum bicolor*), Sugarcane (*Saccharum officinarum*), Rice (*Oryza sativa*), and many others. This project aims to explore and characterize the repetitive elements in six grass genomes that are closely related within the grass tribe Andropogoneae, with the ultimate goal of better understanding the mechanisms that have driven the diversification of this important grass clade. Bioinformatic software Galaxy was used to identify and characterize the abundance of TE's in six grass species that have not been investigated to date. The presence/absences of specific TE's were then mapped onto a phylogeny of the Andropogoneae to better understand the dynamics of TE evolution. An improved understanding of repetitive elements across the grass phylogeny may uncover the mechanism behind the explosive evolutionary radiation of the grasses.

## Introduction

Commonly known as the grasses, Poaceae is a clade of ubiquitous monocotyledonous flowering plants that cover ~ 25% of terrestrial habitats. Poaceae includes the economically important cereal grasses, bamboos, and those of natural grasslands and pastures. Grass stems, known as culms, are cylindrical and hollow except at the nodes, which makes them unique from other graminoid plants (Clayton WD and Renvoise, SA. 1986). Grasses include both annual and perennial forms and have alternate distichous leaves with parallel venation. Leaves are borne in sheaths which contain a low apical meristem. As a result, grass blades elongate at the base of the leaf, allowing for quick growth after herbivory (Cope T and Gray A. 2009).

Many species found in Poaceae tend to be ecologically dominant in temperate and tropical grasslands across the world in part due to C4 photosynthesis (Sage R and Monson R. 1999). In times of drought or high heat, C3 plants tend to photorespire wasting energy created during photosynthesis. In this process RuBisCo oxygenates RuBP rather than carboxylating it, lowering photosynthetic output by up to 25% (Sharkey T. 1988). C4 plants which refer to the four-carbon molecules synthesized during carbon fixation in plant chloroplasts, avoid this by keeping the RuBisCo in a CO<sub>2</sub> rich environment (Slack CR and Hatch MD. 1967).

In an ever-growing world, grasses are the world's main dietary energy supply. In a 2005 study, it was found that grasses made up 51% of all dietary energy consumed, with rice and wheat each contributing 20%, maize contributing 5% and other grains

contributing 6% (Fresco L. 2005). Maize alone is a multi-billion dollar industry in the US. According to the USDA, Americans planted 91.7 million acres of maize in 2019 (Capehart T and Proper S. 2019). About one-third of which was used for feeding livestock, another third used for biofuel and the rest used for human consumption (Capehart T and Proper S. 2019). For years, grasses have been planted in lawns and along roads to reduce erosion. Grass along roads has been found to reduce soil erosion by more than 3 times (Cao CS, et al. 2006). On a global scale, natural and cultivated grasslands contribute 15% of global primary production (Raven J. 2010). A global food supply that also brings services like erosion reduction and primary production, the grasses are extremely important to human life.

Poaceae was named by John Hendley Barnhart in 1895 (Barnhart JH. 1895) after the poa genus described by Carl Linnaeus in 1753 (Linnaeus C. 1753). With 768 genera and 11,506 species, Poaceae is the fifth-largest plant family behind Asteraceae, Orchidaceae, Fabaceae, and Rubiaceae (Stevens PF. 2001). The grass family can be broken down into 13 subfamilies, including the summer grasses in the subfamily panicoideae (Soreng RJ, et al. 2017). The panicoid grasses (2<sup>nd</sup> largest subfamily) can further be subdivided into 12 tribes (Soreng RJ, et al. 2017). The tribe Andropogoneae contain some of the most economically important crops on the planet including maize (*Zea mays*), sorghum (*Sorghum bicolor*) and sugarcane (*Saccharum officinarum*). The panicoid lineage is roughly 26 million years old (Bennetzen JL, et al. 2012) and although some of its species like *Zea mays* and *Sorghum bicolor* have been sequenced and heavily annotated (Paterson, et al. 2009; Schnable, et al., 2009), the reason behind this lineage's rapid radiation has yet to be revealed. Uncovering the evolutionary driver of

this powerful group will help us better understand why the poaceae are such a diverse lineage.

With the development of next-generation sequencing technologies in 2007, many large eukaryotic genomes have been studied. This research has elucidated that there is much genomic variation among the grasses with regards to size, ploidy, and transposable element content. Traditionally, genome size is measured in picograms of DNA in a haploid nucleus. Coined the C-value (where  $1C=1\text{pg}$ ) by Hewson Swift in 1950 (Swift H. 1950), it was first thought that more complex organisms would have more DNA and therefore a larger C-value. This hypothesis was soon disproved by the discovery of the extreme variation of genome size among the eukaryotes. For example, the largest eukaryotic genome is in *Amoeba dubia* at 700 pg, 200 times the human genome (Gregory TR. 2005). This realization that genome size does not correlate to gene number became known as the C-value paradox (Thomas CA. 1971). Poaceae like other angiosperm lineages show a great diversity in genome size. Among Poaceae species, *Brachypodium stacei* has the smallest genome at  $C=0.28$  (Catalán P, et al. 2012), and octoploid Triticale contained the largest genome at  $C=26.00$  (Gregory TR. 2005), and the mean C-value across all Poaceae species is 5.14 (Leitch IJ, et al. 2019).

One important contributor to genome size variation in angiosperms is polyploidy. Whole-genome duplication (WGD) events have occurred across many lineages of land plants causing the genomic contents to double in a single generation. In these scenarios, generally two diploid individuals combine to produce a tetraploid offspring (Moriyama Y and Koshiha-Takeuchi K. 2018). Whole-genome duplication/polyploidy has been proposed as an important driver of speciation (Ohno S. 1970) because in a single

instant, a new species containing twice the amount of DNA as its parent, is formed. Whole-genome duplication events have been linked with species diversification and the acquisition of novel traits in many land plants (Schranz ME, et al. 2012) and ancient polyploidy events of early plants correlate with major land-plant radiations (Jiao Y, et al. 2011). Studies have found that WGD events cause great changes in gene expression, transposable element activity and morphology (Doyle JJ, et al. 2008). This is prevalent in cases of allopolyploidy like in *Zea mays* where a WGD event occurs in a cross between two species creating an offspring with genomes from both progenitors (Doyle JJ, et al. 2008). Despite the rationale that WGDs cause diversification, evidence to the contrary (WGDs do not cause diversification) has been found (Stebbins GL. 1971). Stebbins agreed that “polyploidy has been important in the diversification of species and genera, but not in the origin of the families and orders themselves” (Stebbins GL. 1971). Among the panicoid grasses, recent evidence has suggested that polyploid lineages, in fact, have lower speciation rates and higher extinction rates than diploid lineages (Estep MC, et al. 2014).

After the discovery of DNA, scientists soon found out that genes were not the only thing found in the genetic material. Genomes are mostly made up of repetitive DNA consisting mainly of transposable elements (TEs) and simple repeats like tandem or satellite repeats (Jie Z, and Lonardi S. 2005). First described by Barbara McClintock in 1950 when studying maize, she found that TEs or “jumping genes” were responsible for altered pigmentation in maize kernels (McClintock B. 1950). This monumental discovery would later award her a Nobel Prize in 1983. Simply put, transposable elements are DNA fragments that can insert themselves to new locations throughout the genome,

often duplicating themselves (Feschotte C, Jiang N, and Wessler SR. 2002). In this process, TEs greatly influence gene regulation, expression, and function and can even make novel genes (Bennetzen JL and Wang H. 2014). In recent years, many geneticists like JL Bennetzen have shined the spotlight on TEs as the drivers of genomic novelty. In 2014 he proposed that “genome size, gene content, gene order, centromere function, and numerous other aspects of nuclear biology are driven by TE activity” (Bennetzen JL and Wang H. 2014). Gene duplications can be a direct result of retrotransposition, the process by which retrotransposons copy and paste themselves in the genome. This duplicated gene then has the ability to undergo neofunctionalization, meaning they gain new functions distinct from the ancestral gene (Conant CC and Wolfe KH. 2008). Also, gene regulatory elements like promoters, enhancers and silencers are often shifted during retrotransposition which results in new combinations of regulation (Sabot F and Schulman AH. 2006). Because these moving elements can directly influence gene regulation, expression, and formation of genes, it will be extremely important to further study them, giving us a better grasp of evolution as a whole.

Since their discovery, TEs have been classified into a hierarchical classification system based on mechanism and enzymatic criteria (Wicker T, et al. 2004). The first types of elements known as retrotransposons can self replicate and move, generating genomic plasticity. Retrotransposons are Class I TEs and have an RNA intermediate. Class I molecules use a ‘copy and paste’ mechanism where they are transcribed into RNA and reverse transcribed back into DNA at an insertion site (Finnegan DJ. 1989) These elements have been found to make up a majority of the genetic material in eukaryotes where Long Terminal Repeat retrotransposons (LTR-RTs) are the most

common in plants (Kumar A & Bennetzen JL. 1999). The LTR region, flanking the elements, can range from a few hundred base pairs to five thousand (Wicker T, et al. 2007). LTR retrotransposons have an open reading frame (ORF) containing a GAG gene, encoding structural and protective proteins and a POL gene, encoding aspartic proteinase, reverse transcriptase, RNase H and DDE integrase enzymes. The two most important super families, which greatly contribute to genome size, are *Gypsy* and *Copia* elements. Both variations include the same GAG and POL genes, but differ in the order of the reverse transcriptase and integrase genes within the POL reading frame (Wicker T, et al. 2007). Interestingly, it is hypothesized that retroviruses likely evolved from a *Gypsy* LTR that developed an envelope protein and a few additional regulatory proteins (Frankel AD and Young JA. 1998).

Class II transposable elements are known as transposons and are generally found in low to moderate amounts in almost all eukaryotes (Wicker T, et al. 2004). Characterized by a terminally inverted repeat at both ends, Class I DNA transposons use a 'cut and paste' method and do not have an RNA intermediate. Although they are not copied, their number can increase by transposing during DNA replication. Transposons that have already been replicated can cut and move to a region that has not been replicated, copying itself (Greenblatt IM and Brink RA. 1969). Class II transposons, known as *Helitron* replicate via a rolling-circle mechanism where only a single DNA is cleaved (Kapitonov V and Jurka J. 2001). These unique elements were actually the first TEs identified by modern computational analysis of whole-genome sequence reads, however, their protein domains and their retrotransposition mechanism was not discovered until 2001 (Surzycki SA, et al. 1999).

As mentioned previously, angiosperms genomes vary greatly in size (Gregory TR. 2005). While some of this variation is due to ploidy (Ohno S. 1970), a majority is attributed to TEs. In tandem with their important function to influence gene expression, TEs have been found to be the single largest component in the genomes of eukaryotes (Feschotte C, and Pritham EJ. 2007) where retrotransposons make a majority of total DNA in eukaryotes and upwards to 75% in most angiosperms (Feschotte C, Jiang N, and Wessler SR. 2002). TEs are extremely diverse and found in all kingdoms. With tens of thousands of families across plants, these mobile DNA are important players in the genome (Wicker T, et al. 2004). In plants, a majority of LTR families remain in low copy numbers (Sanmiguel P and Bennetzen JL. 1998), but activity and amplification of a few families can contribute more than >100 Mb of DNA to a genome, causing 'genome obesity' in some lineages (Bennetzen JL and Kellogg EA 1997). In a relative to rice, *Oryza australiensis*, the genome was more than doubled over a few million years due to the amplification of just a few families of LTR retrotransposons (Piegu B, et al. 2006). Interestingly, TEs have been shown to have a dynamic life cycle. Due to factors like random DNA mutations and illegitimate recombination, LTRs tend to fragment and degrade over about 4 million years (Devos KM, et al. 2002), so any visible intact retrotransposons likely were inserted in the last 4 million years.

Amplification and reduction of TEs have been found to be a major cause of genome size plasticity in the grasses (Bennetzen JL, Ma J, Devos KM. 2005). In maize, it was found that the entire genome had doubled in the last 6 million years due to LTR retrotransposons (Sanmiguel P, and Bennetzen JL. 1998). Some TE activity may be caused by ploidy shifts, but the majority of such activity is in fact not. Results have

shown that instead there is random activation of LTR retrotransposon families over evolutionary time (Estep MC, DeBarry JD, and Bennetzen JL. 2013). Retrotransposons have also been found to contribute to genomic DNA removal. This very active process occurs when there is illegitimate recombination (recombination at non-homologous sites) causing small deletions that add up over time (Devos KM, et al. 2002). When expanded out to millions of years, this mechanism can delete large chunks of DNA like in rice which had a genome reduction of a giga base over two million years (Ma J, Devos KM, and Bennetzen JL. 2004). This ability to both increase and decrease the size of the genome creates a dynamic balance that can lean toward genome expansion in some species and genome reduction in others (Hawkins JS, et al. 2009). With the certainty that TEs can create massive changes in genome size as well as influence gene expression and create novel genes, it would not be difficult to imagine transposable elements as the principal drivers of genome change and speciation, possibly responsible for the radiation of the panicoid grasses.

The panicoid grasses have experienced an explosion of species radiation over the last 23 million years, producing some of the most important crops in the world. Previously, polyploidization events like genome duplication were thought to be major molecular drivers of evolution (Soltis DE. 2008), however recent evidence has suggested that polyploid lineages in fact have lower speciation rates and higher extinction rates than diploid lineages (Estep MC, et al. 2014). In this study we examine the transposable element landscape of six panicoid species, five of which belong to the Andropogoneae tribe and the 6<sup>th</sup>, a closely related outgroup, to unearth transposable elements' relation to genome evolution. In order to better understand the processes that

drive diversification, repetitive elements were identified, annotated and compared across six previously undescribed species: *Apluda mutica*, *Arthraxon prionodes*, *Arundinella hirta* (outgroup), *Chasmapodium caudatum*, *Chrysopogon zizanioides*, and *Vossia cuspidata*. Producing an overall description of the genomic landscape across these six species, we give insight into the molecular driving forces behind the expansion of the grasses over the last 23 million years.

## **Methods**

### **Genome sequencing:**

Plants from the grass species *Apluda mutica*, *Arthraxon prionodes*, *Arundinella hirta* (outgroup), *Chasmapodium caudatum*, *Chrysopogon zizanioides*, and *Vossia cuspidata* were grown in a greenhouse and ~100g of leaf tissue was harvested and frozen with liquid Nitrogen. Whole-genome DNA was extracted following a modified CTAB method (Estep MC, et al. 2013). An Illumina sequencing library was constructed and sequenced for each species at the WVU core facility (Morgantown WV). Using the Paired Fastq Filtering tool on the Galaxy web-based platform, fastq paired-end reads files for each taxon were preprocessed, which includes trimming, quality filtering, and removal of adapter sequences (Novak P, et al. 2013). After the preprocessing, interlaced fasta files were produced for each of the six taxa.

### **Repeat Explorer**

In this study, a web-based platform for computational research known as Galaxy was utilized. The main Galaxy tool, Repeat Explorer2, uses graph-based clustering of

next-generation sequence reads to detect all repetitive DNA found in the genome (Novak P, et al. 2013). The classification of transposable elements is based on similarity to the Repeat Explorer database of transposable element protein domains (REXdb). Upon detection of repetitive elements like transposable elements, protein domains may be identified and analyzed using the Galaxy tool DANTE. The identification of protein domains allows for the verification of clustered transposable elements and the possible discovery of novel repetitive elements (Novak P, et al. 2013). In this study, the tools found on Galaxy aid in identifying and annotating the repetitive elements found in the Andropogoneae tribe.

#### CAP3 contigger

After repeats were identified with Repeat Explorer, contigs from clusters in the top 0.1% of the genome were assembled using the CAP3 assembly program (Huang X and Madan A. 1999).

#### LTR annotation

Repeats were annotated based on similarity to the (REXdb). Unknown repeats were then further annotated by a custom database consisting of Long Terminal Repeat regions of LTR-RTs ( $E < 1.0 \times 10^{-6}$  and  $> 200$  bp sequence length)

#### All vs All Blast

Unknown repeats from each species were compared using command-line nucleotide blasts (Altschul SF, et al. 1990) ( $E < 1.0 \times 10^{-6}$  and  $> 200$  bp sequence length). These

results were used to find repeats unique to species as well as repeats shared across the phylogeny.

## **Results**

Genomic DNA from: *Apluda mutica*, *Arthraxon prionodes*, *Arundinella hirta* (outgroup), *Chasmapodium caudatum*, *Chrysopogon zizanioides*, and *Vossia cuspidata* was sequenced then analyzed for repetitive content in this study. The six sequencing libraries yielded between 215,010 – 1,301,972 sequences (Table 1). Galaxy identified between 229-359 repeats for each species. Of these, between 122 (43.18%) - 156 (54.67%) of identified repeats had been previously described in other grass species. A range of 107 (45.3%) - 204 (56.82%) were novel and identified for the 1<sup>st</sup> time across each species. Between 100 - 120 high-copy repeats (0.1% > of the genome) were identified with 26 (26%) - 57 (47%) being novel. A comparison of the high-copy novel repeats among the six species shows between 1-23 repeats are shared across the examined taxa.

In *Arundinella*, 500,000 sequences were assembled into 359 groups representing all known repetitive element families. Of these, 155 were known and 204 were unknown. Out of the high-copy repeats, 64 were known and 56 were unknown with 71 being unique to the taxa. In *Arthraxon* 433,959 sequences were assembled into 229 repetitive elements. In total there were 122 known and 107 unknown; and 74 known and 26 unknown in high copies. There were 84 high-copy repeats unique to *Arthraxon*. For the

taxon, *Chrysopogon* 246,258 sequences were assembled into 289 elements. Of these, 158 were known and 131 were unknown. Among the high-copy elements, 93 were known and 20 were unknown with 92 repeats being unique to *Chrysopogon*. For *Vossia* 239,610 sequences were assembled into 303 clusters. Of these clusters, there were 150 known and 153 unknown repeats containing 67 known and 42 unknown elements in high copies. *Vossia* contained 96 high-copy repeats unique to the genus. In *Chasmopodium* 500,000 sequences were assembled into 287 clusters. Of these, 153 were known and 134 were unknown. Out of the high-copy repeats, 58 were known and 42 were unknown with 85 being unique to *Chasmopodium*. There were 215,010 sequences assembled into 337 repetitive element groups in *Apluda*. Of the 337 elements, 156 were known and 181 were unknown; and 58 known and 42 unknown elements were in high copies. In *Apluda*, 79 high-copy repeats were unique to the genus (Table 1). After repeats were identified using Repeat Explorer, each high-copy element across the six species was then annotated with the Rexdb and a custom LTR database (Appendix Table 1).

Repeats found in the top 0.01% of each taxon's respective genome were deemed high-copy and further characterized. In *Arundinella*, there were 120 high-copy repeats with 64 previously annotated. Of these, 4 were satellite repeats, 1 was a DNA transposon, 38 were LTR-RTs and 21 other high-copy genes like rRNA or mitochondrial DNA. *Arthraxon* contained 100 high-copy repeats with 74 having been previously described. Of the known repeats, 2 were satellite repeats, 3 were DNA transposons, 54 were LTR-RTs with the rest being other high-copy elements. In *Chrysopogon*, 113 high-copy

repeats were found containing 93 previously described elements. Among these, 7 were satellite repeats, 3 were DNA transposons, 70 were LTR-RTs and 13 other high-copy genes. 67 of Vossia's 109 high-copy repeats were already known. Of the high-copy known repeats in Vossia, there was 1 satellite repeat, 1 DNA transposon, 46 LTR-RTs and the rest consisted of other high-copy genes. Chasmopodium contained 100 high-copy repeats with 81 described elements. There was 1 satellite repeat, 2 DNA transposons, 60 LTR-RTs and 18 other high-copy elements. Lastly, Apluda contained 100 high-copy repeat elements. 58 elements have been described, and of these, there was 1 satellite repeat, 0 DNA transposons, 41 LTR-RTs and 16 other high-copy elements (Appendix Table 1).

Our analysis on the repetitive fraction of grass genomes in the Andropogoneae tribe using the Galaxy program, Repeat Explorer2, has allowed for the re-identification of a total of 894 known repetitive elements in Arundinella, Arthraxon, Chrysopogon, Vossia, Apluda, and Chasmopodium. Further examining the high-copy repeats (>0.01% of the genome) for each sample allowed us to describe a total of 224 novel repeats in the six species.

Table 1. Descriptive statistics and repeats identified by species data

Species	<i>Arundinella hirta</i>	<i>Arthraxon prionodes</i>	<i>Chrysopogon zizanioides</i>	<i>Vossia cuspidata</i>	<i>Chasmopodium caudatum</i>	<i>Apluda mutica</i>
Raw Sequence Input	835,648	433,959	246,258	239,610	1,301,972	215,010
Analyzed Sequences	500,000	433,959	246,258	239,610	500,000	215,010
Total Repeats Identified	359	229	289	303	287	337
Known Repeats	155	122	158	150	153	156
High-copy Known Repeats	64	74	93	67	81	58
Unknown Repeats	204	107	131	153	134	181
high-copy Unknown	56	26	29	42	29	42
high-copy Unique to Species	71	84	92	96	85	79
high-copy Shared	49	16	30	13	25	21

The highest copy repeats were then mapped onto the published phylogeny (Estep MC. et al 2014) to reveal the gain of one novel LTR element, Panicoid I, across all Andropogoneae grasses (Figure 1, point 1). A series of gains and losses of other novel elements is seen throughout the phylogeny (Fig 1, point 2-10). A second highly shared element, ArPr 52, is shared among all examined panicoid taxa but falls below 0.01% in *Apluda*. This element was gained at point 2 and lost at point 10. There were two unknown elements shared by both *Chrysopogon* and *Arthraxon*, ArPr 71 (point 3) and ArPr 84 (point 4). *Vossia* contained the most (table 1) unique repetitive landscape and also had significant elements lost (falling below detectable levels), ArHi 45 (point 5) and ArHi 46 (point 6). These same elements were also lost in *Chasmopodium*. *Chasmopodium* and *Apluda* contained two elements, ChCa 71 (point 7) and ChCa 76 (point 8) which were not found in the other taxa analyzed. Unlike the other taxa, ArHi 16 was lost in the *Chasmopodium* lineage (point 9).

Point	Description
1	Gain of Panicoid 1
2	Gain of ArPr 52
3	Gain of ArPr 71
4	Gain of ArPr 84
5	Loss of ArHi 45
6	Loss of ArHi 46
7	Gain of ChCa CL 71
8	Gain of ChCa 76
9	Loss of ArHi 16
10	Loss of ArPr 52

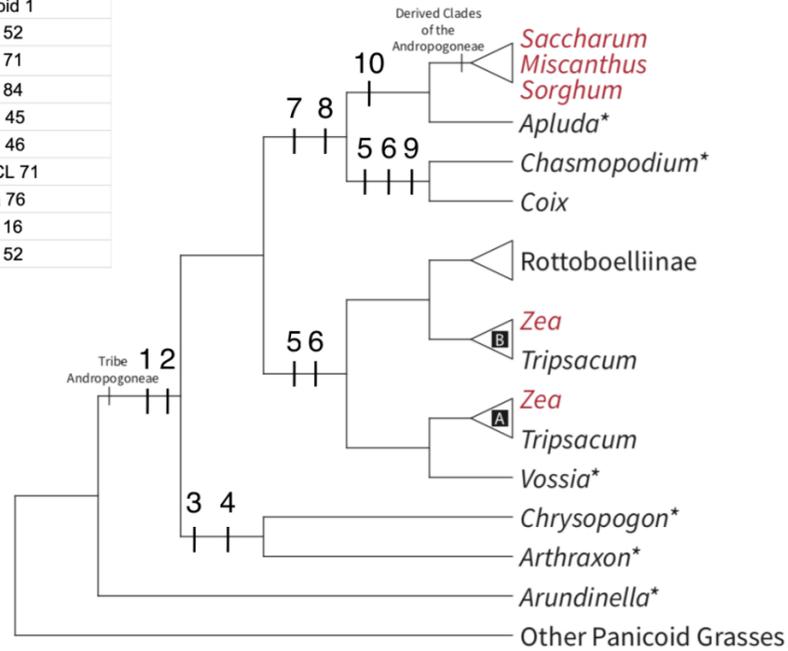


Figure 1. Phylogenetic tree (Estep, et al. 2014) of the panicoid grasses showing gains and losses of novel transposable elements.

A pair-wise comparison of high-copy repeats was carried out to show elements shared across the phylogeny. It was found that *Arundinella* shared 8 repeats with *Arthraxon*, 12 repeats with *Chrysopogon*, 2 with *Vossia*, 8 with *Chasmopodium*, and 23 with *Apluda*. The second species, *Arthraxon* shared 6 repeats with *Chrysopogon*, 2 with *Vossia*, 9 with *Chasmopodium* and 5 with *Apluda*. *Chrysopogon* shared 1 repeat with *Vossia*, 2 with *Chasmopodium*, and 5 with *Apluda*. The species with the most unique transposable element content, *Vossia* shared 2 repeats with *Chasmopodium* and *Apluda*. Lastly, the most derived taxa, *Chasmopodium* and *Apluda* shared 6 repeats (Table 2).

Table 2. Pairwise comparison showing how many unknown repeats were shared among each species.

	<i>Arundinella hirta</i>	<i>Arthraxon prionodes</i>	<i>Chrysopogon zizanioides</i>	<i>Vossia cuspidata</i>	<i>Chasmopodium caudatum</i>	<i>Apludamutica</i>
Arundinella hirta	N/A	-	-	-	-	-
Arthraxon prionodes	8	N/A	-	-	-	-
Chrysopogon zizanioides	12	6	N/A	-	-	-
Vossia cuspidata	2	2	1	N/A	-	-
Chasmopodium caudatum	8	9	2	2	N/A	-
Apludamutica	23	5	5	2	6	N/A

### **Discussion/Future Directions**

To better understand the biodiversity seen today, one must first understand the drivers of change. With modern sequencing technology, scientists are now able to investigate processes that shape current biodiversity. This study compares the transposable element content across six grass species to uncover the molecular mechanisms contributing to the group's diverse taxa.

Early hypotheses supposed that polyploidy was the main driver behind the diversification of land plants (Ohno S. 1970). Whole-genome duplication events have been linked with species diversification and the acquisition of novel traits in many land plants (Schranz ME, et al. 2012), however, among the panicoid grasses, recent evidence indicates that polyploid lineages have lower speciation rates and higher extinction rates than diploid lineages (Estep MC, et al. 2014). Transposable elements compose a majority of total DNA in eukaryotes with upwards of 75% in most angiosperms (Feschotte C, Jiang N, and Wessler SR. 2002). Along with their great abundance, TEs have been found to influence gene regulation, expression, and function as well as create novel genes (Bennetzen JL, Wang H. 2014). The transposable element content across six grass species was described and compared to uncover TE's relation to the evolution of the panicoid clade.

The repeat content across the six grass species was consistent with studies on related taxa. Focusing on the high-copy known repeats of the five species in the Andropogoneae tribe, between 68.7% - 75.3% were LTR-RTs with an average at 72.3% which is accordant to the LTR-RT content found in *Zea mays* at 75% (Schnable PS, et al. 2009). In the sister group of *Zea*, *Vossia*, 74.1% of the high-copy repeats were contributed by LTR-RTs, supporting that the retrotransposon content found here is consistent with species that have been whole-genome sequenced. Also in accordance with other studies, there were low levels of satellite repeats and DNA transposons. Satellite repeats generally make up low percentages of genome content (Garrido-Ramos MA. 2017), but vary greatly across plants ranging from 0.1% to 36%. (Garrido-

Ramos MA. 2015; Ambrožová K, et al. 2011). In all species, satellite repeats were identified in low numbers. DNA transposons were identified in all species but *Apluda*, however, there are substantial amounts of unannotated repeats in *Apluda* that certainly contain DNA transposons. In this study only LTR-RTs were further annotated, so an additional analysis of DNA transposons using transposon databases would be necessary to annotate such elements.

Using Repeatexplorer, 224 novel high-copy repeats were identified in *Arundinella*, *Arthraxon*, *Chrysopogon*, *Vossia*, *Chasmopodium* and *Apluda*. These repetitive elements were then compared across the entire clade, showing which were shared amongst taxa. In order to identify which novel TEs played a role in genome evolution, select TEs were mapped (Figure 1) on a previously published phylogeny of the group (Estep MC, et al. 2014)

One novel element, Panicoid I, was shared across all Andropogoneae grass species and not found in any outgroups. It would be interesting to explore the copy number variation of this element across the tribe. Another element, ArPr 52, was shared amongst all but the most derived of the Andropogoneae lineages, suggesting it was actively removed in derived clades. Two elements, ArPr 71 and ArPr 84, were found only in *Arthraxon* and *Chrysopogon*, which could demonstrate the birth of novel LTR-RTs ; and another two elements, ArHi 45 and ArHi 46 were lost in both *Vossia* and *Chasmopodium* which are found in distinct clades, again emphasizing the active removal within different lineages. These examples of gains and losses demonstrate the

dynamics of LTR-RTs in the genomes of the grasses. Conversely, one may suspect that speciation of taxa may be influencing the TE content in the genome. Due to the diverse ways that TEs can influence and create novel genes in the genome, and the few ways grass genomes are able to silence them, it makes intuitive sense to suspect TEs as the driver. As described, the fluctuations of high-copy novel TEs can be mapped to nodes in the Andropogoneae tribe, suggesting their involvement in the evolution process. The further classification of these novel elements may give rise to the discovery of new types of transposable elements not yet described, and will further emphasize LTR-RTs impact on the rapid diversification of this vital group of plants.

The pairwise analysis of the six species reveals that the outgroup, *Arundinella*, shared the most novel repetitive elements of any other species, this means many of the novel elements identified were already a part of the “pan-grass” genome before its diversification into the Andropogoneae tribe. *Arundinella* and *Apluda*, the two most distantly related species sampled on the phylogeny, had the most elements shared at 23. Many of the LTR-RTs identified within the six taxa were unique to that species, suggesting a rapid birth rate. *Arthraxon* and *Chasmopodium* shared high numbers of elements (9) which is consistent with their placement as sister taxa in the phylogeny. *Vossia* contained the most unique genome with regards to repetitive elements. This indicates that as *Vossia* has evolved, many elements that were present in the tribe’s common ancestor have fallen below our detection threshold (0.01%) in its genome. This assertion is consistent with the mapping of novel high-copy elements where *Vossia* and its sister taxa, *Zea*, both lost the same novel elements, ArHi 45 and ArHi 46 (Figure 1,

point 5 and 6). Also in agreement with their phylogenetic position, *Chasmopodium* and *Apluda* shared 6 high-copy repeats.

Although the panicoid phylogeny has been accurately resolved and evolutionary relationships described (Estep MC, et al. 2014), the mechanism behind the radiation of the group has yet to be uncovered. We suggest that LTR-RTs played a role in the rapid diversification of the panicoideae. Ten evolutionary significant novel LTR-RTs were mapped, showing the rise of Panicoid I at the start of the Andropogoneae tribe and a series of gains and losses of other novel LTR-RTs. These elements make up large portions of the genomes examined and likely contributed to the radiation of the group and should be further analyzed. From here, PCR sequencing assays may be constructed, allowing for a deeper analysis leading to the elements' classification. The low levels of high-copy elements shared across the Andropogoneae tribe indicates that the repetitive DNA contents across the genomes are becoming increasingly distinct. Furthermore, the high portion of repeats that the outgroup, Arundinella, share with the tribe, suggest that bursts of LTR-RTs amplification is a common pattern. To improve this study, a larger amount of analyzed sequences would provide a better resolution of the repetitive element landscape across the species. Clearly, whole-genome sequencing of these taxa would produce the most precise image of the repetitive content, but it would be very resource intensive. A deeper description of TE content in other grass genomes and then across angiosperms will demonstrate TEs' role in the diversification of the many lineages of land plants.

## References

- Altschul, S.F. et al. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.
- Ambrožová K., et al. 2011. Diverse retrotransposon families and an AT-rich satellite DNA revealed in giant genomes of *Fritillaria* lilies. Ann. Bot. 2011;107:255–268.  
doi: 10.1093/aob/mcq235.
- Barnhart, John Hendley (1895). "Family nomenclature". Bulletin of the Torrey Botanical Club. 22 (1): 1–24. doi:10.2307/2485402. JSTOR 2485402
- Bennetzen JL, Kellogg EA (1997). Do plants have a one-way ticket to genomic obesity? Plant Cell 9: 1509–1514.
- Bennetzen JL, Ma J, Devos KM (2005). Mechanisms of recent genome size variation in flowering plants. Ann Bot 95: 127–132.
- Bennetzen JL, Schmutz J, Wang H, Percifield R, Hawkins J, Pontaroli AC et al. (2012). Reference genome sequence of the model plant *Setaria*. Nat Biotech 30: 555–561
- Bennetzen, J.L., and Hao Wang (2014). "The contributions of transposable elements to the structure, function, and evolution of plant genomes." Annual Review of Plant Biology 65:505-530
- Cao, C.S. et al. (2006). "impact of planting grass on terrene roads to avoid soil erosion". Landscape and Urban Planning, ISSN: 0169-2046, Vol: 78, Issue: 3, pg: 205-216

- Capehart T.; Proper S. (2019). "Corn is America's Largest Crop in 2019". U.S. Department of Agriculture
- Catalán P, Müller J, Hasterok R, Jenkins G, Mur LAJ et al. (2012). Evolution and taxonomic split of the model grass *Brachypodium distachyon*. *Annals of Botany* 109: 385-405.
- Clayton, W.D.; Renvoise, S.A. (1986). *Genera Graminum: Grasses of the world*. London: Royal Botanic Garden, Kew
- Conant C. C., & Wolfe K. H. (2008). Turning a hobby into a job : How duplicated genes find new functions. *Nature Reviews. Genetics (Print)*, 12, 928.
- Cope, T.; Gray, A. (2009). *Grasses of the British Isles*. London, U.K.: Botanical Society of Britain and Ireland. ISBN 9780901158420.
- Devos KM, Brown JKM, Bennetzen JL (2002). Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res* 12: 1075–1079.
- Doyle JJ, et al. (2008) Evolutionary genetics of genome merger and doubling in plants. *Annu Rev Genet* 42:443–461
- Estep, M. C. et al. (2014) 'Allopolyploidy, diversification, and the Miocene grassland expansion', *Proceedings of the National Academy of Sciences of the United States of America*, 111(42), pp. 15149–15154. doi: 10.1073/pnas.1404177111.
- Estep, M. C., DeBarry, J. D., & Bennetzen, J. L. (2013). The dynamics of LTR retrotransposon accumulation across 25 million years of panicoid grass evolution. <https://doi-org.proxy006.nclive.org/10.1038/hdy.2012.99>

- Feschotte, C., Jiang, N. and Wessler, S.R. (2002). Plant transposable elements: where genetics meets genomics. *Nature Reviews Genetics*, 3(5), p.329.
- Feschotte, C., and Ellen J. Pritham (2007) "DNA transposons and the evolution of eukaryotic genomes" *Annual review of genetics* 41:331
- Finnegan, D. J.(1989) Eukaryotic transposable elements and genome evolution. *Trends Genet.* 5, 103–107
- Frankel, A. D. & Young, J. A (1998). HIV-1: fifteen proteins and an RNA. *Ann. Rev. Biochem.* 67. 1-25
- Fresco, Louise (2005). Rice is life. *Journal of Food Composition and Analysis* 18:249–253
- Garrido-Ramos M. A. (2017). Satellite DNA: An Evolving Topic. *Genes*, 8(9), 230.  
<https://doi.org/10.3390/genes8090230>
- Garrido-Ramos M.A. 2015. Satellite DNA in Plants: More than Just Rubbish. *Cytogenet. Genome Res.*;146:153–170. doi: 10.1159/000437008.
- Greenblatt, I. M. & Brink, R. A.(1962) Twin mutations in medium variegated pericarp maize. *Genetics* 47, 489–501 .
- Gregory, T.R. (2005). Animal Genome Size Database. <http://www.genomesize.com>.
- Hawkins JS, Proulx SR, Rapp RA, Wendel JF (2009). Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. *Proc Natl Acad Sci* 106: 17811–17816.
- Huang, X. and Madan, A. (1999) CAP3: A DNA sequence assembly program. *Genome Res.*, 9, 868-877.

- Jiao Y, et al. (2011) Ancestral polyploidy in seed plants and angiosperms. *Nature* 473(7345):97–100.
- Jie Zheng, & Lonardi, S. (2005). Discovery of repetitive patterns in DNA with accurate boundaries. Fifth IEEE Symposium on Bioinformatics and Bioengineering (BIBE'05), Bioinformatics and Bioengineering, 2005. BIBE 2005. Fifth IEEE Symposium on, Bioinformatics and Bioengineering, 105–112.  
<https://doi-org.proxy006.nclive.org/10.1109/BIBE.2005.23>
- Kapitonov, V. & Jurka, J. (2001) Rolling-circle transposons in eukaryotes. *Proc. Natl Acad. Sci. USA* 98,
- Kumar, A. & Bennetzen, J. (1999) Plant retrotransposons. *Annu. Rev. Genet.* 33, 479–532 .
- Leitch IJ et al. (2019) Angiosperm DNA C-values database (release 9.0, Apr 2019)  
<https://cvalues.science.kew.org/>
- Linnaeus, C. (1753) “Species Plantarum,” Vol. 1, London
- Ma J, Devos KM, Bennetzen JL (2004). Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res* 14: 860–869.
- McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences*, 36(6), 344-355
- McKain M. R. et al (2018) ‘Ancestry of the two subgenomes of maize’ *BioRxiv*, doi: 10.1038-352351
- Moriyama, Y., Koshiba-Takeuchi, K. (2018). Significance of whole-genome duplications on the emergence of evolutionary novelties. *Briefings in Functional Genomics*, 17(5), 329–338

- Novak, P. et al. (2013) - RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next generation sequence reads. *Bioinformatics*.
- Ohno S. (1970). *Evolution by Gene Duplication*. Springer-Verlag, Berlin, Heidelberg
- Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H et al. (2009). The *Sorghum bicolor* genome and the diversification of grasses. *Nature* 457: 551–556.
- Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H et al. (2006). Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res* 16: 1262–1269.
- Raven, J.; Thomas, H. (2010). "Grasses". *Current Biology*. Vol: 20, Issue 19, pg: 837-839,
- Sabot, F., and Schulman, A.H. (2006). Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. *Heredity*, 97(6), 381-388.
- Sage, R.; Monson, R. (1999). *C4 Plant Biology*. pp. 551–580. ISBN 978-0-12-614440-6.
- Sanmiguel, P., & Bennetzen, J. L. (1998). Evidence that a Recent Increase in Maize Genome Size was Caused by the Massive Amplification of Intergene Retrotransposons. *Annals of Botany*, 82(suppl\_1), 37–44.  
<https://doi-org.proxy006.nclive.org/10.1006/anbo.1998.0746>
- Schnable PS, et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science* 326: 1112–1115.

- Schranz ME, Mohammadin S, Edger PP (2012) Ancient whole genome duplications, novelty and diversification: The WGD Radiation Lag-Time Model. *Curr Opin Plant Biol* 15(2):147–153.
- Sharkey T (1988). "Estimating the rate of photorespiration in leaves". *Physiologia Plantarum*. 73 (1): 147–152. doi:10.1111/j.1399-3054.1988.tb09205.x.
- Slack, C. R.; Hatch, M. D. (1967). "Comparative studies on the activity of carboxylases and other enzymes in relation to the new pathway of photosynthetic carbon dioxide fixation in tropical grasses". *The Biochemical Journal*. 103 (3): 660–665. doi:10.1042/bj1030660. PMC 1270465. PMID 4292834.
- Soltis DE, Bell CD, Kim S, Soltis PS. 2008. The origin and early evolution of the angiosperms. *Ann NY Acad Sci.* ; 1133: 3–25. doi: 10.1196/annals.1438.005  
PMID: 18559813
- Soreng, R. J. et al. (2017) 'A worldwide phylogenetic classification of the Poaceae (Gramineae) II: An update and a comparison of two 2015 classifications.' *Journal of Systematics and Evolution*, 55(4), 259–290.  
<https://doi-org.proxy006.nclive.org/10.1111/jse.12262>
- Stebbins GL (1971) *Chromosomal Evolution in Higher Plants* (Edward Arnold, London)
- Stevens, P. F. (2001). *Angiosperm Phylogeny Website*. Version 14, July 2017
- Surzycki SA, et al. (1999). "Characterization of Repetitive DNA Elements in *Arabidopsis*". *Journal of Molecular Evolution*. 48 (6): 684–691. doi:10.1007/pl00006512. PMID 10229572.

- Swift H (1950). "The constancy of deoxyribose nucleic acid in plant nuclei".  
Proceedings of the National Academy of Sciences of the USA. 36 (11): 643–654.  
doi:10.1073/pnas.36.11.643. PMC 1063260. PMID 14808154.
- Thomas, C. A. (1971) The genetic organization of chromosomes. *Annu. Rev. Genet.* 5,  
237–256
- Wessler, S. (2006). Transposable Elements and the Evolution of Eukaryotic  
Genomes. *Proceedings of the National Academy of Sciences of the United  
States of America*, 103(47), 17600-17601. Retrieved April 1, 2020, from  
[www.jstor.org/stable/30052508](http://www.jstor.org/stable/30052508)
- Wicker, T. et al. (2007) 'A unified classification system for eukaryotic transposable  
elements', *Nature Reviews Genetics*, 8(12), pp. 973–982. doi: 10.1038/nrg2165.

## **Appendix**

Table 1. Repeat classifications based on Rexdb and a custom LTR database showing the largest 100 elements in *Arundinella*, *Arthraxon*, *Chrysopogon*, *Vossia*, *Chasmopodium* and *Apluda*.

<i>Arundinella hirta</i>	<i>Arthraxon prionodes</i>	<i>Chrysopogon zizanioides</i>	<i>Vossia cuspidata</i>	<i>Chasmopodium caudatum</i>	<i>Apluda mutica</i>
1 unknown_CL1	1 LTR Ty1/copia SIRE	1 repeat satellite	1 repeat mobile_element	1 unknown_CL1	1 LTR Ty3/gypsyTekay
2 repeat satellite	2 repeat satellite	2 LTR Ty3/gypsy Athila	2 repeat mobile_element	2 LTR Ty3/gypsyTekay	2 LTR Ty3/gypsyTekay
3 unknown_CL3	3 unknown_CL3	3 repeat satellite	3 LTR Ty3/gypsyTekay	3 LTR Ty3/gypsyTekay	3 LTR Ty3/gypsyTekay
4 unknown_CL4	4 unknown_CL4	4 unknown_CL4	4 unknown_CL4	4 unknown_CL4	4 unknown_CL4
5 unknown_CL5	5 LTR Ty3/gypsy Athila	5 LTR Ty3/gypsy Athila	5 LTR Ty3/gypsyTekay	5 unknown_CL5	5 LTR Ty3/gypsy Athila
6 unknown_CL6	6  LTR gypsy gyrna	6 repeat satellite	6 LTR X Ruda	6 LTR Ty3/gypsyTekay	6 LTR Ty3/gypsy Athila
7 LTR Ty3/gypsy Athila	7 unknown_CL7	7 unknown_CL7	7 LTR Ty1/copia SIRE	7 LTR Ty1/copia SIRE	7 LTR Ty3/gypsy Athila
8 unknown_CL8	8 unknown_CL8	8 LTR Ty3/gypsy Athila	8 LTR Ty3/gypsyTekay	8 LTR Ty1/copia SIRE	8 unknown_CL8
9 unknown_CL9	9 LTR Ty3/gypsy Athila	9 LTR Ty3/gypsy Athila	9 repeat satellite	9 LTR Ty3/gypsyCRM	9 LTR Ty3/gypsy Athila
10 unknown_CL10	10 LTR Ty1/copia SIRE	10 LTR Ty3/gypsy Athila	10 LTR Ty3/gypsyTekay	10 LTR Ty3/gypsyTekay	10 LTR Ty3/gypsyTekay
11 LTR Ty1/copia SIRE	11 unknown_CL11	11 LTR Ty3/gypsy Athila	11 LTR Ty1/copia SIRE	11 LTR Ty1/copia SIRE	11 LTR Ty3/gypsy Athila
12 unknown_CL12	12 LTR Ty3/gypsy Athila	12 LTR Ty3/gypsy Athila	12 LTR Ty3/gypsyTekay	12 LTR Ty1/copia SIRE	12 LTR Ty1/copia SIRE
13 LTR Ty1/copia SIRE	13 LTR Ty3/gypsy Athila	13 LTR Ty3/gypsyTekay	13 contamination	13 LTR Ty3/gypsyCRM	13 unknown_CL13
14 unknown_CL14	14 LTR Ty3/gypsy Athila	14 LTR Ty3/gypsyTekay	14 LTR gypsy gyrna	14 LTR Ty3/gypsy Ogre	14 LTR Ty3/gypsy Athila
15 LTR Ty3/gypsy Ogre	15 unknown_CL15	15 repeat satellite	15 LTR gypsy gyrna	15 LTR Ty3/gypsyTekay	15 LTR Ty1/copia SIRE
16 unknown_CL16	16 LTR Ty1/copia SIRE	16 LTR Ty3/gypsy Athila	16 LTR Ty3/gypsyTekay	16 LTR Ty3/gypsy Athila	16 contamination
17 unknown_CL17	17 LTR Ty3/gypsy Athila	17 LTR Ty3/gypsyTekay	17 LTR Ty1/copia SIRE	17 LTR Ty1/copia giepum	17 contamination
18 unknown_CL18	18 LTR Ty3/gypsy Athila	18 LTR Ty3/gypsy Athila	18 contamination	18 LTR Ty1/copia SIRE	18contamination
19 LTR Ty1/copia SIRE	19 unknown_CL19	19 LTR Ty3/gypsy Athila	19 contamination	19 contamination	19 LTR Ty1/copia SIRE
20 LTR Ty3/gypsy Athila	20 unknown_CL20	20 LTR Ty3/gypsy Athila	20 LTR Ty3/gypsyTekay	20 LTR Gypsy xilon-diguus	20 LTR Gypsy gyrna
21 LTR Ty1/copia SIRE	21 LTR Ty3/gypsy Athila	21 LTR Ty3/gypsy Athila	21 repeat	21 contamination	21 unknown_CL21
22 unknown_CL22	22 LTR Ty3/gypsy Athila	22 LTR Ty3/gypsy Retand	22 repeat	22 LTR Ty1/copia	22 LTR Ty1/copia SIRE
23 unknown_CL23	23 LTR Ty3/gypsy Athila	23 LTR Ty3/gypsy Retand	23 LTR Ty3/gypsyTekay	23 LTR Ty3/gypsy Ogre	22 contamination
24 unknown_CL24	24 LTR Ty3/gypsyTekay	24 LTR Ty3/gypsy Athila	24 LTR Ty3/gypsyTekay	24 LTR Ty3/gypsyTekay	23 contamination
25 unknown_CL25	25 LTR Ty3/gypsy Ogre	25 LTR Ty3/gypsyTekay	25  TIR EnSpm/CACTA	25 unknown_CL25	25 LTR Ty1/copia SIRE
26 unknown_CL26	26 LTR Ty3/gypsyTekay	26 LTR Ty3/gypsy Retand	26 contamination	26 LTR Ty3/gypsyTekay	26 LTR Ty1/copia SIRE
27 LTR Ty3/gypsy Ogre	27 LTR Ty3/gypsy Athila	27 LTR Ty3/gypsy Retand	27 unknown_CL27	27 LTR Ty1/copia SIRE	27 unknown_CL27
28 unknown_CL28	28 LTR Ty3/gypsyTekay	28 LTR Ty3/gypsyTekay	28 contamination	28 LTR Ty3/gypsy Athila	28contamination
29 LTR Ty1/copia SIRE	29 unknown_CL29	29 LTR Ty3/gypsy Athila	29 LTR Ty3/gypsyTekay	29 LTR Ty3/gypsyTekay	29 LTR Ty3/gypsy Athila
30 LTR Ty3/gypsy Ogre	30 Class_1 LTR gypsy uwum	30 LTR Ty3/gypsy Athila	30 LTR Ty3/gypsyTekay	30 LTR Ty3/gypsy Athila	30 LTR Ty3/gypsy Athila
31 LTR Ty3/gypsy Athila	31 unknown_CL31	31 contamination	31 LTR Ty1/copia SIRE	31 LTR Ty3/gypsy Athila	31 LTR Ty3/gypsyTekay
32 unknown_CL32	32 LTR Ty3/gypsy Athila	32 LTR Ty3/gypsyTekay	32 LTR Ty1/copia SIRE	32 unknown_CL32	32 unknown_CL32
33 unknown_CL33	33 unknown_CL33	33 LTR Ty3/gypsy Athila	33 LTR Ty3/gypsyTekay	33 LTR Ty3/gypsyCRM	29 contamination

34 LTR Ty1/copia SIRE	34 repeat satellite	34 LTR Ty3/gypsyTekay	34 LTR Ty1/copia SIRE	34 TIR EnSpm/CACTA	34 LTR Ty3/gypsy Ogre
35 LTR Ty3/gypsyOgre	35 contamination	35 LTR Ty3/gypsy Athila	35 LTR Copia Ji	35 contamination	35 unknown_CL35
36 LTR Ty3/gypsy Athila	36 LTR Ty3/gypsy Athila	36 LTR Ty3/gypsy Athila	36 LTR Ty3/gypsyTekay	36 unknown_CL36	36 LTR Ty3/gypsy Athila
37 organelle plastid	37 LTR Ty3/gypsy Retand	37 repeat satellite	37 repeat	37 LTR Ty3/gypsy Athila	37 LTR Ty3/gypsy Athila
38 unknown_CL38	38 pararetrovirus	38 LTR Ty3/gypsy  Retand	38 repeat	38 LTR Ty1/copia	38 LTR Ty3/gypsy Athila
39 unknown_CL39	39 LTR Ty3/gypsy Athila	39 LTR Ty3/gypsy Retand	39 LTR Ty3/gypsyTekay	39 LTR Ty3/gypsyCRM	39 LTR Ty1/copia SIRE
40 unknown_CL40	40 LTR Ty1/copia SIRE	40 repeat rDNA 45S_rDNA	40 LTR Ty1/copia SIRE	40 LTR Gypsy CRM4	40 LTR Ty1/copia SIRE
41 LTR Copia Ji	41 LTR Ty3/gypsy  Ogre	41 LTR Ty3/gypsy Retand	41 LTR gypsy xilon-diguus	41 LTR Ty1/copia SIRE	41 LTR Ty3/gypsy Athila
42 unknown_CL42	42 LTR Ty3/gypsy Athila	42 unknown_CL42	42 organelle plastid	42 LTR Ty3/gypsy Athila	42 LTR Ty3/gypsy Athila
43 LTR Ty1/copia SIRE	43 LTR Ty3/gypsy Athila	43 LTR Ty3/gypsy Retand	43 LTR Ty1/copia SIRE	43 LTR Ty3/gypsyTekay	43 LTR Ty3/gypsyTekay
44  TIR EnSpm/CACTA	44 organelle plastid	44 LTR Ty3/gypsy Athila	44 LTR Ty3/gypsyTekay	44 LTR Ty3/gypsy Athila	44LTR Ty3/gypsyTekay
45 unknown_CL45	45 TIR EnSpm/CACTA	45 repeat rDNA 45S_rDNA	45 unknown_CL45	45 LTR Ty1/copia SIRE	45 LTR Ty3/gypsy Athila
46 unknown_CL46	46 LTR Ty1/copia  keros	46 LTR Ty3/gypsy Retand	46 LTR Ty3/gypsyTekay	46 LTR Ty3/gypsyCRM	46 unknown_CL46
47 organelle plastid	47 LTR Ty3/gypsy Retand	47 LTR Ty3/gypsy Athila	47 repeat	47 LTR Ty3/gypsy Retand	47LTR Ty1/copia SIRE
48 unknown_CL48	48 LTR Ty3/gypsyTekay	48 contamination	48 organelle plastid	48LTR Ty3/gypsyCRM	48 LTR Ty3/gypsy Athila
49 organelle plastid	49 organelle plastid	49 LTR Ty3/gypsy Athila	49 unknown_CL49	49 LTR Ty1/copia SIRE	49 LTR Ty3/gypsy Athila
50 organelle plastid	50 unknown_CL99	50 LTR copia SIRE	50 unknown_CL99	50 LTR Copia Ji	50 LTR gypsy  Tekay

51 unknown_CL51	51 LTR gypsy   Athila	51 repeat satellite	51 unknown_CL51	51  LTR gypsy   Athila	51 unknown_CL51
52 unknown_CL52	52 unknown_CL52	52  LTR gypsy   Retand	52  LTR copia SIRE	52  LTR gypsy  Tekay	52 unknown_CL52
53 unknown_CL53	53  LTR gypsy   Athila	53 contamination	53 unknown_CL53	53  LTR gypsy  Tekay	53 unknown_CL53
54 unknown_CL54	54 unknown_CL54	54  LTR gypsy   Athila	54 unknown_CL54	54  LTR copia SIRE	54 unknown_CL54
55 organelle plastid	55 unknown_CL55	55 unknown_CL55	55  LTR gypsy  Tekay	55  LTR gypsy   Athila	55 unknown_CL55
56 unknown_CL56	56  LTR copia SIRE	56  LTR gypsy   Retand	56 unknown_CL56	56  LTR gypsy   Ogre	56 unknown_CL56
57  LTR copia SIRE	57  LTR gypsy   Athila	57  LTR gypsy   Athila	57  LTR gypsy  Tekay	57 repeat satellite	57  LTR gypsy   Athila
58 organelle plastid	58 organelle plastid	58  LTR copia Dijap	58 organelle plastid	58 contamination	58 organelle plastid
59 unknown_CL59	59 organelle plastid	59  LTR gypsy  Tekay	59  LTR gypsy  Tekay	59  LTR gypsy   Athila	59 organelle plastid
60 organelle plastid	60   TIR EnSpm/CACTA	60 repeat	60  LTR copia SIRE	60  LTR gypsy   Ogre	60 unknown_CL60
61 unknown_CL61	61 organelle plastid	61  LTR gypsy   Athila	61 unknown_CL61	61  LTR gypsy   Ogre	61 unknown_CL61
62 organelle plastid	62 unknown_CL62	62  LTR gypsy   Retand	62 organelle plastid	62 unknown_CL62	62 organelle plastid
63  LTR gypsy  Tekay	63  LTR copia SIRE	63  LTR gypsy   Athila	63 unknown_CL63	63  LTR gypsy   Athila	63 unknown_CL63
64  LTR gypsy   Athila	64 unknown_CL64	64  LTR gypsy   Retand	64 organelle plastid	64 organelle plastid	64  LTR copia SIRE

65  LTR copia SIRE	65  LTR gypsy   Athila	65  LTR gypsy   Athila	65  LTR gypsy  Tekay	65 unknown_CL65	65 I
66  LTR copia SIRE	66 organelle plastid	66 unknown_CL66	66 unknown_CL66	66  LTR copia SIRE	66 organelle plastid
67  LTR copia Ikeros	67 organelle plastid	67  LTR gypsy   Retand	67  LTR gypsy  Tekay	67  LTR gypsy  Tekay	67 unknown_CL67
68 repeat satellite	68  LTR gypsy   Athila	68  LTR gypsy   Ogre	68 unknown_CL68	68 organelle plastid	68 unknown_CL68
69  LTR gypsy   Athila	69 organelle plastid	69  LTR gypsy   Ogre	69  LTR copia SIRE	69 unknown_CL69	69 organelle plastid
70  LTR copia SIRE	70 unknown_CL70	70 repeat rDNA 45S_rDNA	70  LTR copia SIRE	70 unknown_CL70	70 unknown_CL70
71  LTR	71 unknown_CL71	71  LTR gypsy   Ogre	71 unknown_CL71	71 unknown_CL71	71  LTR gypsy   Athila
72 organelle plastid	72  LTR gypsy   Retand	72 repeat	72  LTR copia Angela	72   TIR EnSpm/CACTA	72 unknown_CL72
73  LTR gypsy   Athila	73 organelle plastid	73 unknown_CL73	73 unknown_CL73	73 organelle plastid	73 repeat satellite
74 organelle plastid	74 unknown_CL74	74  LTR gypsy  Tekay	74 repeat rDNA 5S_rDNA	74  LTR gypsy   Athila	74 unknown_CL74
75 unknown_CL75	75 organelle plastid	75  LTR gypsy   Athila	75 unknown_CL75	75 unknown_CL75	75 organelle plastid
76 unknown_CL76	76  LTR gypsy  Tekay	76   TIR EnSpm/CACTA	76 unknown_CL76	76 unknown_CL76	76  LTR gypsy   Athila
77 organelle plastid	77   TIR EnSpm/CACTA	77  LTR gypsy   Athila	77  LTR copia SIRE	77 unknown_CL77	77 organelle plastid
78 unknown_CL78	78 organelle plastid	78  LTR gypsy   Retand	78 unknown_CL78	78  LTR copia SIRE	78 unknown_CL78
79 unknown_CL79	79  LTR gypsy   Athila	79 repeat rDNA 5S_rDNA	79 unknown_CL79	79 organelle plastid	79 unknown_CL79
80  LTR copia giepum	80  LTR gypsy  Tekay	80 unknown_CL80	80 unknown_CL80	80  LTR copia Angela	80 unknown_CL80
81 unknown_CL81	81  LTR gypsy  Tekay	81 unknown_CL81	81  LTR gypsy  Tekay	81 unknown_CL81	81 organelle plastid
82  LTR copia SIRE	82 1 LTR gypsy Ji	82  LTR copia SIRE	82 unknown_CL82	82  LTR copia SIRE	82 unknown_CL82
83 unknown_CL83	83  LTR gypsy   Athila	83  LTR gypsy   Athila	83 unknown_CL83	83 unknown_CL83	83 organelle plastid
84 organelle plastid	84 unknown_CL84	84  LTR gypsy   Athila	84 unknown_CL84	84 unknown_CL84	84 unknown_CL84
85  LTR gypsy   Athila	85 organelle plastid	85  LTR copia Angela Wiwa	85  LTR copia Angela	85 organelle plastid	85 unknown_CL85
86 organelle plastid	86  LTR copia SIRE	86  LTR copia SIRE	86 LTR Copia xilon-diguus	86  LTR gypsy   Athila	86  LTR gypsy   Athila
87 organelle plastid	87 organelle plastid	87   TIR EnSpm/CACTA	87  LTR copia SIRE	87  xilon-diguus	87 organelle plastid
88 unknown_CL88	88  LTR gypsy   Athila	88 unknown_CL88	88  LTR copia TAR	88 contamination	88 organelle plastid
89 unknown_CL89	89 unknown_CL89	89 repeat satellite	89 unknown_CL89	89 organelle plastid	89  LTR gypsy   Athila
90 unknown_CL90	90  LTR gypsy   Athila	90  LTR gypsy   Ogre	90 unknown_CL90	90 unknown_CL90	90  LTR gypsy  Tekay
91 unknown_CL91	91  LTR copia TAR	91 repeat	91 unknown_CL91	91 unknown_CL91	91 unknown_CL91
92 unknown_CL92	92 unknown_CL92	92 unknown_CL92	92 unknown_CL92	92 unknown_CL92	92 unknown_CL92
93 repeat satellite	93 organelle plastid	93  LTR gypsy   Retand	93 unknown_CL93	93  LTR gypsy   Ogre	93 unknown_CL93
94 unknown_CL94	94 unknown_CL94	94  LTR gypsy   Ogre	94 unknown_CL94	94  LTR copia TAR Gudyeg	94 organelle plastid
95 unknown_CL95	95  LTR gypsy  CRM	95 unknown_CL95	95 organelle plastid	95  LTR copia Angela	95 organelle plastid
96 unknown_CL96	96  LTR gypsy   Athila	96  LTR gypsy  Tekay	96 unknown_CL96	96 organelle plastid	96 organelle plastid
97  LTR gypsy	97 unknown_CL97	97 unknown_CL97	97 unknown_CL97	97  LTR copia Bianca	97 unknown_CL97
98 organelle plastid	98 unknown_CL98	98  LTR copia SIRE	98 unknown_CL98	98 organelle plastid	98  LTR gypsy   Athila
99 organelle plastid	99 unknown_CL99	99  LTR copia SIRE	99 unknown_CL99	99  LTR Copia Ji	99  LTR gypsy  Tekay
100 unknown_CL100	100  LTR copia SIRE	100  LTR gypsy   Ogre	100 unknown_CL100	100 unknown_CL100	100 unknown_CL100