COMPARATIVE ANALYSIS OF REPEAT LANDSCAPES IN *AVENA* (OAT)

by

Shelvasha Burkes

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics & Computational Biology

Charlotte

2020

Approved By:

Dr. Jessica Schlueter

Dr. Robert Reid

Dr. Elizabeth Cooper

Dr. Rebekah Rogers

Dr. Adam Reitzel

ABSTRACT

SHELVASHA BURKES. Comparative Analysis of Repeat Landscapes in Avena (Oat). (Under the direction of DR. JESSICA SCHLUETER).

Avena sativa, or common oat, is a staple crop and member of the Poaceae or Grass family. Following behind wheat, maize and rice, oats account for 10.5 million hectares of the world's produced crops as of 2017. Phytocompounds such as β-glucan, avenanthramides, vanillic, syringic, ferric, and caffeic acids have shown to benefit cardiovascular health and represent potential benefactors to human health. However, further investigation into these potential factors requires research that surpasses past works in breadth and scope. Much has been done to bridge the gap in genomic resources for oats, such as the development of high throughput markers, consensus linkage maps and most recently genome sequencing efforts. However, the relative complexity of cultivated oat, an allohexaploid with highly similar subgenomes, provides additional challenges to the development of these resources. A final layer of complexity is the genome size of hexaploid oats, estimated to be approximately 12.8 gigabases, of which a significant portion is composed of complex repetitive elements. Characterization of these highly complex regions is difficult as repetitive regions contained within reads are characteristically difficult to map, thereby complicating assembly efforts and resulting in misassembly and gaps. Through investigation of repetitive elements by creating a novel pipeline capable of offering enhanced resolution, repetitive elements were further examined within well-characterized diploid Avena genomes, with concluding phylogenetic analyses examining evolutionary relationships between repetitive elements.

DEDICATION

I dedicate my doctoral thesis to those who encouraged and believed in me throughout this journey. To my parents, they were constant motivators that nurtured my curiosity, and I wouldn't be here without their moral lessons of discipline instilled in me from an early age. I also dedicate this to my high school teacher Nassim, who inspired me to pursue Biology. I dedicate this to my advisor Jess, who was the guiding light every step of the way as I researched for this dissertation, I apologize for the worry and stress I likely caused during this journey. Finally, dedicate this to my grandparents, they only wanted the best for us and taught us the value of education and the opportunities that they were not fortunate to have. I dedicate this dissertation to the memory of my grandfather, who always believed in my abilities to achieve anything I set my mind to.

ACKNOWLEDGEMENTS

I would first like to thank my thesis advisor Dr. Jessica Schlueter of the Bioinformatics and Genomics department. She always had her door open and available whenever I ran into a trouble spot or had a question about my research or writing. She consistently pushed me to question and think critically as a scientist, steering me in the right direction whenever I encountered ever-present roadblocks along the way. I would like to thank our collaborators Jeff Maughan, Tim Langdon and Robert Reid for their insight and data made available for use in my dissertation. I would also like to thank my committee members Dr. Cooper, Dr. Rogers, Dr. Reid and Dr. Reitzel, and department chair Dr. Gibas for their unwavering support when delays and real life created difficulties. Without their cooperation, support and input, I would not have been successful and I am gratefully indebted to the Department of Bioinformatics and Genomics for all of the valuable assistance throughout my thesis. Finally, I must express my very profound gratitude to my parents and to my fiancé for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

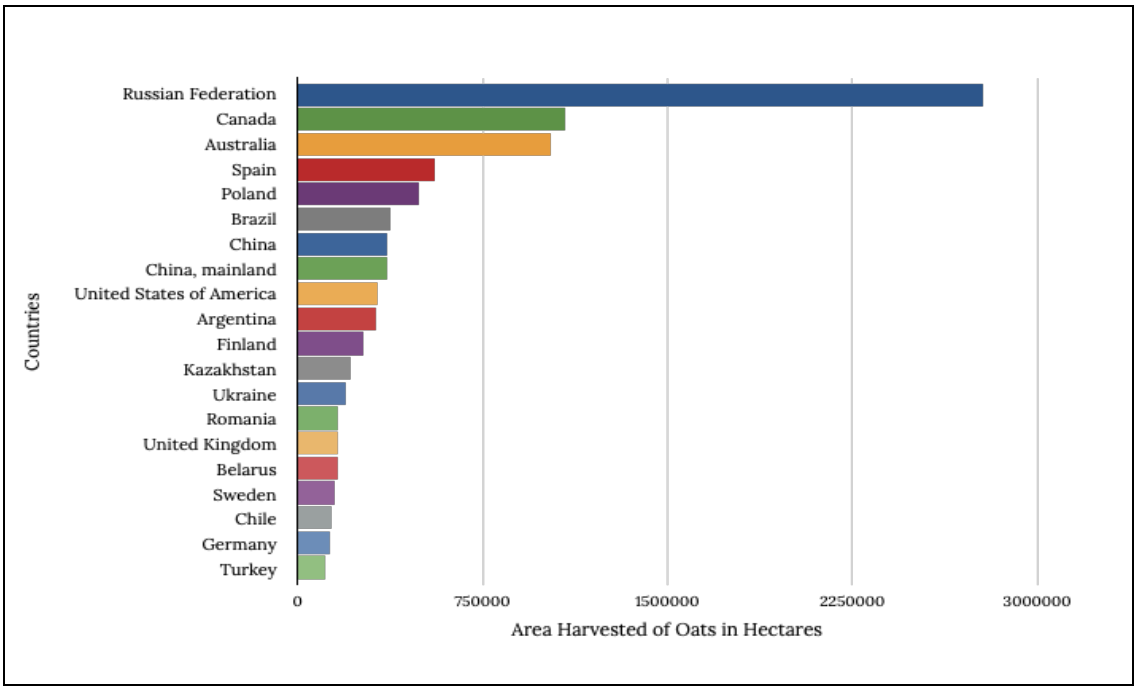
TE	Transposable Element
LTR	Long Terminal Repeat
SINE	Short Interspersed Nuclear Element
LINE	Long Interspersed Nuclear Element
DIRS	Dictyostelium Intermediate Repeat Sequence
SSR	Site-Specific Recombination
WGD	Whole Genome Duplication
DNA	Deoxyribonucleic Acid
TPRT	Target-Primed Reverse Transcription
hAT	hobo-Ac-Tam3
TIR	Terminal Inverted Repeats
MULE	Mutator and Mutator-like elements
MITE	Miniature Inverted Terminal Element
TRF	Tandem Repeat Finder
PLE	Penelope-like Element
TSD	Terminal Site Duplication
RT	Reverse Transcriptase

Chapter 1: INTRODUCTION

Overview

Throughout the course of human history, cereal crops have existed as indispensable sources of nutrition to the communities that utilize them, many of these plants eventually forming the backbone of many civilizations around the globe **[1][2]**. *Avena sativa*, better known as the common oat, is a staple food crop and member of the Poaceae or Grasses family. Following behind wheat, maize, rice, barley and sorghum, oats account for 10.5 million hectares (Figure 1) of the world's produced crops as of 2017, approximately 320,000 of which were harvested in the United States **[3]**. In comparison to related cereal crops, research into the genomics of oats has lagged, with progress on *Avena* progressing at a slower rate. We feel this is unfortunate, as many studies, namely those investigating the relationship of β-glucan to cardiovascular health **[4]**, have also indicated the myriad of potential of oats have to human health. Other compounds also include vanillic acids, syringic acids, ferric acid, caffeic acid, and avenanthramides **[5]**. Previous studies into these phenolic compounds indicate high antioxidative potential, as well as potent anti-inflammatory agents, as was observed in Sur et. al., where avenanthramides were noted to markedly decreasing dermal inflammation **[6]**. Further investigations into the potential health benefits of oats to human health requires resources and research that surpasses past works and requires an increase to the breadth and scope of studies on the oat genome. Recent works shed light on some of the unknown genetic characteristics of oat, with research revolving on understanding the phenolic compounds that constitute *Avena* species, the effect the environment has on the plant, and what additional roles those compounds can play in human health and wellness **[7][8]**.

Figure 1, Oats Harvested Internationally, in hectares



Food and Agriculture Organization of the United Nations (2017) [3]

In recent years, a significant amount of work has been done to bridge the gap in resources for oats with the development of high throughput markers, consensus linkage maps and most recently genome sequencing efforts. The relative complexity of the oat genome, namely cultivated oats being an allohexaploid with two of the subgenomes postulated to be quite similar to one another, provides additional challenges to the development of these resources. Adding to this complexity, the genome size of hexaploid oats is believed to be approximately 12.8 gigabases **[9]**, of which, a significant portion is composed of complex repetitive elements. Characterization of these highly complex repetitive regions is difficult as repetitive regions contained within reads are characteristically difficult to map, thereby complicating assembly efforts and resulting in misassembly and gaps. Despite the difficulties that accompany studying Avena, there are fortunately, diploid and tetraploid *Avena* species that are capable of alleviating

some of the aforementioned difficulties, allowing investigations to edge somewhat closer to in-depth analyses of hexaploid oat.

Objectives

Our primary objective was to investigate the repeat landscape of oats. To do this, we developed of a pipeline capable of offering enhanced resolution and detection of repetitive elements. With the advancement of repeat identification, our secondary objective was to make use of this novel information about the repeatome, or all repetitive elements in a given genome, to further examine transposable elements within two diploid *Avena* genomes. Using this analysis in conjunction with our previously developed pipeline, our final objective shifts towards beginning to understand the evolution of families of repetitive elements among *Avena* species, with the goal of gaining insight into the role transposable elements across *Avena* as a whole.

Background: History and Importance of Transposable Elements

Transposable elements were first observed by Barbara McClintock in the late 1940's with her studies in *Zea mays* **[10]**. Her discovery of transposable elements long preceded understanding of genetic processes of the time, and were largely ignored until the late 1960's, when the mechanisms of transposition were observed in bacteria, yeast, and bacteriophages **[11]**, providing evidence to conclude that this process was found in all organisms. Much later, it was found that these elements were related to genetic alterations and phenotypic expression, and these discoveries later earned McClintock the Nobel Prize of Physiology & Medicine in 1983. From McClintock's findings, transposable elements have been found in almost all living organisms. Transposable elements are defined as sequences of DNA capable of moving from one location in the genome to another location. Due to this ability, transposable elements, or

TEs, are commonly referred to as jumping genes **[12]**. The mobility exhibited by these elements has, in some cases, led to the implication of these TEs as facilitators to various processes within the genome **[13]**.

Genome Size and Rearrangement

Genome size is defined as the amount of DNA contained within the cell nucleus **[26]** and intuitively, genome sizes vary widely across all organisms, with some of the largest and more variant genomes being observed in eukaryotes. In eukaryotes, the causes of genome variance has been widely studied, and especially in plants, where genome size typically displays an even higher degree of size diversity. There are several theories as to why we see such variance. In *Pellicer et al (2018)* **[27]**, two main causes of genome size variation are discussed, the accumulation of repeats in the genome, and polyploidy. As repeats proliferate, these sequences, which can vary in size from hundreds to thousands of base pairs in length, gradually contribute more and more to the overall DNA contained within the genome. This proliferation can and is occasionally kept in check by machinery within the genome to contain uncontrolled growth, but the degree to which this proliferation is controlled can vary between organisms, and therefore is non-trivial **[27]**. In the case of *A. thaliana* and *G. nigrocaulis*, two plant species that possess small, compact genomes that are believed to have genome reduction due to double-strand break repairs inadvertently reducing DNA while undergoing repair **[28]**, and deletion bias, a process often seen in prokaryotes that favors loss of genetic material **[29]**. In cases where proliferation of transposable elements are not as regulated, transposon proliferation can also provide an alternative method by which a plant genome can quickly increase in size, or in other words, it is a potential alternative to polyploidization **[26]**. Polyploidy describes a state where the genome possesses multiple sets of homologous chromosomes. Plants in particular have experienced

many instances of polyploidization and or whole genome duplication (WGD). Where polyploidy is a duplication of a genome, TE proliferation is simply an increase of a portion of genome content, such as the TE and its genes, or sites it inserts into. Advantages of polyploidy include heterosis (hybrid vigor), increase in genetic diversity through gene redundancy, and ability to produce asexually **[30]**.

As transposable elements contribute to genome's size, alterations of the genome's structure can occur, such as in the duplication or insertion of a large retrotransposon into a loci. A consequence of this event can potentially be large-scale duplication or deletion events due to impairment in the repair of the loci, possibly resulting in segmental duplications and subsequent increase to genome content. The process of polyploidization, as illustrated in *(Figure 2),* begins with a single diploid that has diverged into separate species with distinct subgenomes. Over time, evolutionary events, such as duplications, and hybridizations, result in organisms that present with differing degrees of ploidy. A polyploid derived from the same diploid parent is what is known as an autopolyploid, and in the case that diploid parents are different, the polyploid is what is known as an allopolyploid. The degree of polyploidy, i.e. the number of copies of the genome, is designated by Greek prefix e.g. (tri- 3, tetra- 4, penta- 5, hexa- 6 and octa- 8). For example, an organism with four copies of its genome, derived from two different parents would be referred to as an allotetraploid **[31]**.

Figure 2, Overview of the Process of Polyploidization

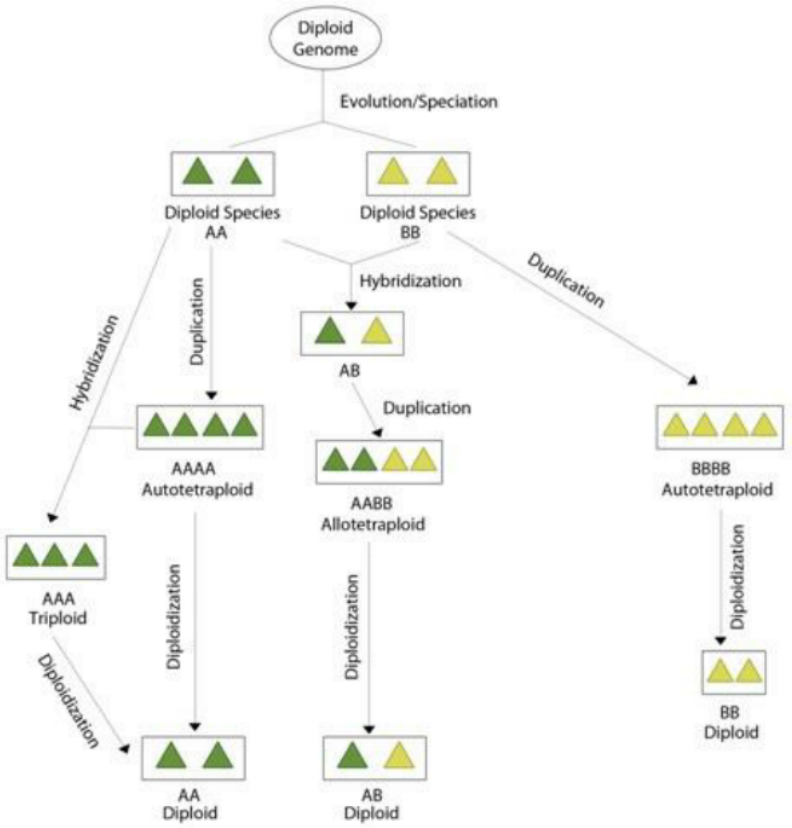


Figure 2, The methods of polyploidization. A polyploid derived from the same diploid parent is what is known as an autoploid, and in the case the diploid parents are different, the polyploid is what is known as an alloploid. The degree of polyploidy, i.e. the number of copies of the genome, is designated by Greek prefix e.g. (tri- 3, tetra- 4, penta- 5, hexa- 6 and octa- 8). For example, an organism with four copies of its genome, derived from two different parents would be referred to as an allotetraploid [31]

However, polyploidization and increases to genome content are not exclusively positive. Disadvantages of polyploidy include drastic changes in cellular architecture, epigenetic instability, and instability in chromosome pairing during mitosis and meiosis. Changes in cellular architecture are a consequence of more volume being occupied by DNA, and the cells are challenged to accommodate a sudden doubling of DNA over a single generation **[30]**. In observing the methods by which these genomes balance the pressures of increasing genome size, be it through transposon proliferation, polyploidization, or a decreasing genome via DSB repair or

deletion bias, we gain additional insight as to how we should approach our genomic studies and potential role TEs serve within them.

Perspectives on the Role of Transposable Elements

Transposon Mutagenesis

Early perspectives on the role of transposable elements were largely negative, resulting in TEs gaining the label of "junk DNA" **[14]**, and their evolutionary role assumed to be negative **[15]**. *Werren et al (2011)*, was an early publication that introduced the concept of TEs being "parasitic" or "selfish" elements, concluding they had a deleterious effect on the host genome. This detriment, as explained in *Werren et. al*, is founded on the premise of uncontrolled transposon proliferation and belief that it can potentially lead to "genetic conflict". This hypothesis is driven by the assumption that transposable elements within an individual are capable of influencing phenotypic traits by behaving "antagonistically" and preventing transcription **[15]**. The potential of transposable elements to effect genes and gene function has been observed, in a process termed transposon mutagenesis. Transposon mutagenesis is defined as an event where a transposable element relocates, with the insertion of the element into a new location disrupting gene function in the form of a mutation or frameshift mutation **[16]**. The importance of gene disruption by either insertion or removal of an element and the combined potential for an element to lose the ability to transpose, leads to the potential generation of new genes via mutation **[17]**. A well-known example of gene disruption is the initial discovery of transposable elements by McClintock and her observations of kernel color in *Zea mays*. The earliest TEs were first observed in *Zea mays*, where McClintock observed instances where the expected wild-type *Z. mays* would present with sporadic lack of kernel color. The spontaneous nature of this phenotype prompted McClintock to study the cytology of maize in greater detail,

her studies intending to understand what causes this phenomena **[10]**. It was later discovered that the gene sequence of UPD-glucose:flavonol 3-O-D glucosyltransferase, an enzyme responsible for the production of anthocyanin, was disrupted by insertion of a TE **[18]**. In some cultivars, the absence of pigmentation is fixed, creating new genes and distinct cultivars of *Zea mays* in the process. Other studies presenting additional instances of gene creation in plants include *Jiang et al (2004)* **[19]** in *Arabidopsis thaliana* and *Wang et al (2006)* **[20]** in *Oryza sativa*. In both studies, transcripts were found to contain chimeric gene fragments that were later expressed as chimeric transcripts. Incorporation of these sequences into a given genome as a new gene can directly influence the evolution of not only a given genome, but a population once the mutation is fixed. As we've come to better understand TEs and the impact they can have in interacting with genes, transposons are now being perceived as one of the many drivers of evolution **[21]**.

Genetic Markers and Linkage Mapping

With advances in sequencing technology, perspectives on what role TEs serve in a genome has evolved, and interpretations of these elements have expanded beyond "selfish genes" and or as a source of occasional mutations, currently being regarded as robust option for genetic markers. *Kumar et al (1999)* discusses the use of transposons as genetic markers, and notes that high copy number, once viewed as a negative trait of transposable elements, as markers, make them quite advantageous. Genetic linkage maps are representations of known genetic markers on chromosomes based on recombination events and segregation data in a given mapping population **[22]**. The importance of linkage maps include: (1) Localization of genes of interest (i.e. knowledge of where they are located in the genome). (2) Highly linked traits that can be selected for. This is helpful for plant breeders and any subsequent biological applications that rely on trait-association; (3) Complete and accurate survey of the genome ensures that gaps and unmarked

regions do not exist within a genome assembly. This is important, as gaps equate to an incomplete genome map, and this greatly impacts any analysis dependent on genetic markers, such as marker-assisted breeding and quantitative and qualitative analysis of traits **[23]**.

The rationale underlying use of TEs as advantageous markers is supported by the natural processes that occur during TE replication and insertion. As elements proliferate, new insertions cause changes in genomic structure and generate distinct polymorphisms within the genome. These polymorphisms can then be detected within and between species **[24]**. An example of this is seen in *Nicotiana debneyi*, a wild relative of *N. tabacum*, where transposons are used as tags to identify a virus resistant gene **[25]**.

Mechanisms of Movement in Transposable Elements

In addition to the classification of elements based on their intermediates, there are also differing mechanisms of movement for transposable elements chiefly consisting of three methods: transposition, conservative-site recombination, and target-primed reverse transcription **[32]**.

Transposition

Transposition and related retroviral integration are processes where an element is inserted into a different location within the genome **[32]**. Class II transposable elements utilize transposition as their primary mechanism of movement, and these elements will typically contain a gene encoding for transposase, as well as DNA binding sites at the end of the element having a characteristic inverted repeat sequence structure that is essential for transposition to take place **[33]**. Transposase creates a complex between the element and these regions, and transposase will then cleave the element at these ends. When inserting an element into a new position, insertion will occur almost exclusively at staggered locations on the DNA backbone **[34]**, and this results in

the characteristic target site duplication seen in transposition **[32]**. Transposition can be done in two methods: "copy-paste" or replicative transposition where a copy of the element is left behind at the original site **[12]**, and "cut-paste" transposition where the element is excised from the site. Different classes of elements perform this in a variety of ways, but the general process consists of performing a double-stranded DNA break at the end of an element following the formation of a hairpin structure **[34]**.

Autonomous and non-Autonomous Transposition

Transposition can be further broken down into two categories: (1) autonomous, defined as elements capable of independent movement and (2) non-autonomous, defined as elements dependent on other transposable elements for movement. In elements classified as autonomous, all of the required proteins are encoded for, and the element is essentially able to transpose on its own. Alternatively, elements that are classified as non-autonomous do not have the ability to move and insert on their own.

Site-Specific Recombination (SSR)

Site-Specific Recombination (SSR), also known as conservative-site recombination (CSR), is a process by which a segment of DNA moves between recombination sites using a recombinase-like enzyme **[32]**, the chief protein involved in homologous recombination in prokaryotes and eukaryotes. Movement can occur within a chromosome and or among other chromosomes **[35]**. This type of movement of mobile elements occurs primarily in bacteriophages, yeast and the integron system in bacteria **[36]** and has not been observed in plants. Despite this, the mechanism is markedly different from traditional transposition and is therefore important to note.

Target-Primed Reverse Transcription (TPRT)

Target-primed reverse transcription is a method of transposon movement that occurs in non-LTR retrotransposons (LINEs and SINEs), and found exclusively in mammalian genomes, bacteria and bacteria derived organelles **[32]**. Like SSR or CSR, it has not yet been observed in plants. The process involves an endonuclease cleaving a targeted site of DNA. The first step only cleaves one strand of the double-stranded DNA. In the second step of the process, that single strand is then used as a template for hybridization of transposon RNA. The third step follows with reverse transcriptase reverse transcribing the complement of the transposon. The process proceeds into the fourth step, where the second cleavage of the remaining strand of double-stranded DNA occurs and the reverse transcribed cDNA is integrated into the original DNA.

Structural Characteristics of Retrotransposons (Class I)

Class I transposable elements or retrotransposons are mobile elements that use RNA intermediates to facilitate propagation and movement. Retrotransposons consist of two subgroups of elements: LTR and non-LTR. LTRs or long terminal repeats are characteristic regions consisting of very long, sometimes exceeding 25kb in length as is the case with Ogre family of LTRs discovered in Pisum sativum **[37]**, at the 5’ and 3’ ends of the element. Non-LTR retrotransposons, which include Long Interspersed Nuclear Elements (LINEs), Short Interspersed Nuclear Elements (SINEs) and elements identified in *Dictyostelium*, referred to as Dictyostelium intermediate repeat sequences (DIRS), do not contain the long terminal repeat regions, but still use RNA intermediate to propagate and move.

Long Terminal Repeats (LTR) Retrotransposons

LTR retrotransposons are transposable elements that utilize RNA-intermediates to transpose or move throughout the genome and are characterized by a region of long terminal repeats on its

3’ and 5‘ ends. The general structural characteristics of LTR elements are outlined in (Figure 3). Difference within the structural components define subfamilies within the LTR classification. Within the coding region of an LTR there is a collection of functional sites and protein coding regions essential to replication and transposition. Functional and protein coding regions include: Protein Binding Site (PBS), followed by a Polypurine Tract (PPT) and genes encoding for *gag*, *pol*, and *int*, all of which are proteins essential for transposition **[24]**. The *gag* region encodes for capsid and nucleocapsid proteins, which serve the role of encapsulating the sequence while transposing, while the *pol* region encodes for protease, integrase, RNase H and reverse transcriptase, proteins responsible for replication and integration of a newly replicated retroelement **[38]**. The structural characteristics of LTRs are very similar to retroviruses. The coding regions possess a high level of similarity to retroviral coding regions, implying a similar function and mechanism of replication. This similarity was later used to cluster LTRs with significant sequence similarity for classification.

Figure 3, Structure of LTR Retrotransposons

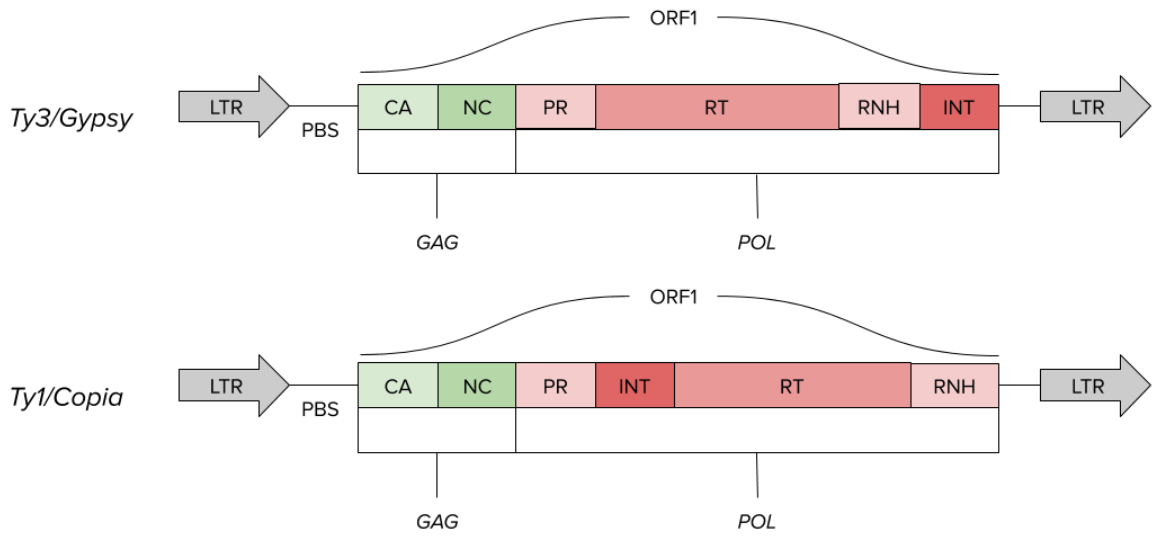


Figure 3, LTR retrotransposons consist of a single open reading frame (ORF), and the following components: LTR: Long terminal repeat, PBS: Primer Binding Site, GAG polyprotein gene consisting of CA: Capsid protein, and NC: Nucleocapsid protein. POL polyprotein gene consisting of PR: Protease, RT: Reverse Transcriptase, RNH: RNAse H, and INT: Integrase.

There are two superfamilies of LTRs observed in plants; *Ty3-Gypsy* and *Ty1-Copia* **[39]**. LTRs within these families are grouped depending on the order of *int* and *pol* domains within the retroelement **[24]** as well as similarity between amino acid sequences of the encoded reverse transcriptase, as it has been shown to be highly conserved **[40]**. Ty3-Gypsy LTR retrotransposons, also known as *Metaviridae*, are a group of LTR retrotransposons characterized by *gag*-RT-*int* ordering of its protein domains. Notable families of Ty3-Gypsy include Ale/Retrofit, Angela/Tork, Bianca, Ivana/Oryco, Maximus/Sire and TAR/Tork lineages **[41]**. Ty1-Copia, also known as Pseudoviridae, is characterized by *gag*-*int*-RT protein domain order and this includes CRM/CR, DEL/Tekay, Galadriel, Reina and TAT/Athila **[41]**. There is a third group of LTRs that share structural characteristics to *Ty3-Gypsy* and *Ty1-Copia*, and these are designated as ”-like” LTR retrotransposons and are classified as either *Ty3-Gypsy*-like or *Ty1-Copia*-like depending on their degree of homology between their protein-coding regions **[42]**.

Long Interspersed Nuclear Elements (LINEs)

Long Interspersed Nuclear Elements or LINEs, are an autonomous class of non-LTR retrotransposons that are thought to be some of the oldest retrotransposons found in eukaryotes **[43]**. LINEs were primarily observed in mammals but are also seen to a smaller extent, in plants. LINEs that are seen in plants exclusively belong to the LINE-1 (L1) clade of LINEs **[42]** and some observed LINEs include *Cin4* in *Z. mays* **[44]**, *Ta 11-1* in *A. thaliana* **[45]**, and *Karma* in *O. sativa* **[46]**. Differing from LTR retrotransposons, replication and proliferation of LINEs occurs via the previously mentioned Target Primed Reverse-Transcription or TPRT **[32]**. Illustrated in *(*

Figure 4, Structure of Long Interspersed Nuclear Elements*),* we observe a deviation of prior structural characteristics observed in LTRs, namely a lack of a long terminal region and addition of an ORF or open reading frame. The typical structure of elements in the L-1 family of LINEs consists of TSD: target site duplications, followed by two outer flanking regions, known as untranslated regions (UTRs). Within these flanking regions, there usually exist two open reading frames, ORF1 and ORF2. ORF1 consists of the 5' UTR and RBP: ribosomal promoter region. The second ORF consists of EN: endonuclease, RT: reverse transcriptase, and the 3' UTR that also contains a Poly(A) tail. These unique structural attributes not only define LINE elements, but contribute to its distinct method of transposition.

Figure 4, Structure of Long Interspersed Nuclear Elements

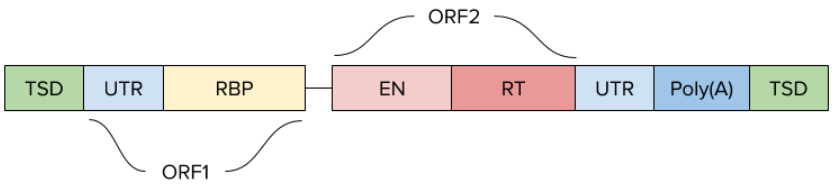


Figure 4, Long Interspersed Nuclear Elements. The typical structure of elements in the L-1 family of LINEs consists of TSD: target site duplications, followed by two outer flanking regions, known as untranslated regions (UTRs). Within these flanking regions, there usually exist two open reading frames, ORF1 and ORF2. ORF1 consists of the 5' UTR and RBP: ribosomal promoter region. The second ORF consists of EN: endonuclease, RT: reverse transcriptase, and the 3' UTR that also contains a Poly(A) tail.

Short Interspersed Nuclear Elements (SINEs)

Short Interspersed Nuclear Elements or SINEs, are non-autonomous, non-LTR retrotransposons similar to LINEs and are believed to have originated from tRNAs due to the high degree of homology between tRNA head coding regions and the head regions of SINEs. Similar to LINEs, these elements are also thought to be some of the oldest retrotransposons in eukaryotes **[47]**. SINEs are structurally distinct from LTRs, with SINEs typically containing a head region at the 5' and a A-rich tail region at the 3' end. As illustrated in Figure 5, SINEs have two ORFs (open reading frames) and flanking TSD (target site duplications) generated during insertion. ORF1 consists of a UTR (untranslated region) followed by two monomer subunits connected by an A-rich linker region. The second ORF contains RT: reverse transcriptase, and the 5' UTR region followed by a Poly(A) tail. The structural characteristics of SINEs has been used in establishing evolutionary relationships, namely those investigated include primates, plant families Gramineae, Fabaceae and Brassicaceae **[48]**, with many of these lineages of SINE elements being derived from the characteristic 5' region, as this region is notable for containing of highly conserved sequences that are traceable to its tRNA origins **[47]**.

Figure 5, Structure of Short Interspersed Nuclear Elements

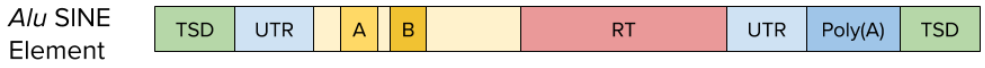


Figure 5, Short Interspersed Nuclear Elements. The general structure of SINE elements is very similar to LINEs, typically consisting of two ORFs and flanking TSD (target site duplications generated during insertion). ORF1 consists of a UTR followed by two monomer subunits connected by an A-rich linker region. The second ORF contains RT: reverse transcriptase, and the 5' UTR region followed by a Poly(A) tail

Dictyostelium Intermediate Repeat Sequences (DIRS)

Dictyostelium Intermediate Repeat Sequences (DIRS) are a relatively new family of retroelements found in the *Dictyostelium discoideum* **[49]**, a soil-dwelling slime mold. These elements possess an LTR region and multiple reading frames, representing a hybrid structure of LTR and LINE/SINE retroelements. The terminal regions also differ from the terminal regions seen in LTRs in that the regions are inverted, similar to TIRs commonly seen in DNA transposons; however, these inverted regions are not identical **[49]**. The long terminal repeats are present on the 3' and 5' ends of the element, followed by the first ORF on the 3' end. ORF1 contains GAG polyprotein, which encodes for capsid and nucleocapsid proteins. This is then followed by the second reading frame, which contains three genes: RT: reverse-transcriptase, RNH: RNAse H, and MT: methyl transferase. A final unusual feature of DIRs is that these elements encode for a unique variant of endonuclease, known as tyrosine recombinase, deviating from the normally observed DDE-type integrase or an aspartic-type protease observed in most transposase enzymes **[50]**.

Figure 6, General Structure of Dictyostelium Intermediate Repeat Sequences (DIRs)

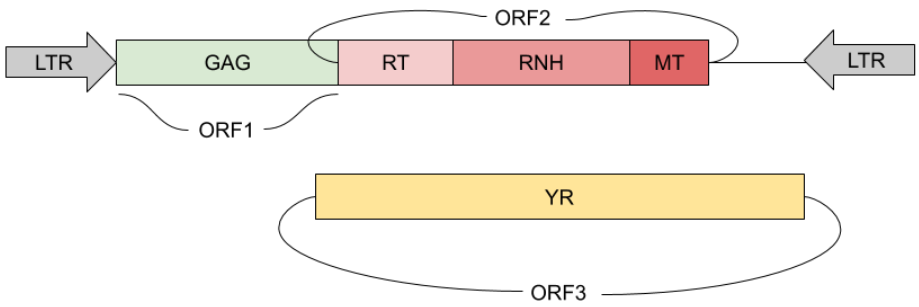


Figure 6, General gene structure of Dictyostelium Intermediate Repeat Sequences (DIRs). DIRs contain long terminal repeats on the 3' and 5' ends of the element, followed by the first ORF on the 3' end. ORF1 contains GAG polyprotein, which encodes for capsid and nucleocapsid proteins. This is then followed by the second reading frame, which contains three genes: RT: reverse-transcriptase, RNH: RNAse H, and MT: methyl transferase.

Penelope & Penelope-Like Elements (PLEs)

Penelope and Penelope-like-elements (PLEs), are unique retroelements originally found to be a causative agent of hybrid dysgenesis syndrome, a high rate of mutation that results from crosses of lineages of *Drosophila virilis* **[51][52]** that possess autonomous Penelope elements, and another lineage that lacks them. Structurally, these elements are similar to other non-LTR retroelements, containing a characteristic endonuclease domain; however, these elements also possess an ability to gain introns **[53]**. From initial discovery in Drosophila, Penelope and Penelope-like transposable elements have also been found in metazoans as well recently in conifers, specifically Loblolly pine **[54]**.

Figure 7, General Amino Acid Structure of Penelope-like Elements

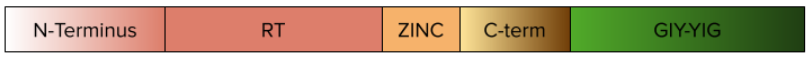


Figure 7, General amino acid structure of Penelope-like elements. N-terminus domain, followed by RT (reverse transcriptase) domain, ZINC (zinc-fingerlike domain), C-terminus, and GIY-YIG (a family of endonuclease).

Structural Characteristics of DNA Transposons (Class II)

Class II transposable elements do not use an RNA intermediate to facilitate movement, but instead encode for and/or rely on enzymes to excise and insert them into new regions. Class II

elements can be further broken down into two subclasses: (1) Terminal Inverted Repeats, and (2) Helitron/Maverick elements.

Subclass I: Terminal Inverted Repeats (TIRs)

TIRs all share a similar structure with variance in the transposase binding site that allows them to bind to specific regions of DNA. Variability is also seen in the length of TIRs in each TIR-type element.

The Ac/Ds System

The earliest example of a DNA transposon is the Ac/Ds system and its constituent *Activator* (*Ac*) and *Dissociator* (*Ds*) elements **[10]**. The Ac/Ds system was discovered by McClintock in her work in maize and it describes a system in which autonomous transposons, labelled *Activator* elements by McClintock, would activate an Ac-derived non-autonomous transposon or what McClintock called *Dissociators*, and the result of these interactions was the disruption of genes. These disruptions produced distinct phenotypic differences, which McClintock observed as differences in kernel pigmentation. Gene disruption and subsequent mutant phenotypes, occurs when either the activator or dissociator components of the AcDs system insert into the gene. Later work in *Z. mays* revealed *p1* as the gene responsible for kernel/pericarp pigmentation and also the target gene that activator and dissociator elements acted on during McClintock's initial studies. Ac and Ds are both capable of gene insertion, Ac autonomously, and Ds non-autonomously. With Ac insertions, they are approximately 4.5kb in length, and upon insertion, *p1* function is impaired for that cell resulting in spotted pigmentation. When coupled with Ds insertions, which are also approximately 4-4.5kb in size, the result ranges between spotted phenotype as in the case of Ac only insertions, or complete loss of function of *p1*, resulting in a colorless kernel. As a

consequence of these initial studies, researchers now understand the impact elements such as Ac and Ds can have on gene expression and gene functionality.

PIF-Harbinger

PIF-Harbinger is a superfamily of autonomous class II transposable elements. The families that make up the PIF-Harbinger superfamily include Harbinger, a family of elements originally observed in *A. thaliana* and P instability factor or PIF elements, a family of elements observed in *Z. mays* **[55]**. These two families were then grouped in the superfamily PIF-Harbinger due to sequence similarity, the length of their terminal inverted repeats and the characteristic 3bp target site duplication generated when transposition occurs. These elements were later observed across many species within the plant kingdom, initially just *O. sativa*, and then elements being identified in *Triticum*, *M. truncatula*, *D. carota* **[38]**.

Figure 8, Structure of PIF-Harbinger Elements

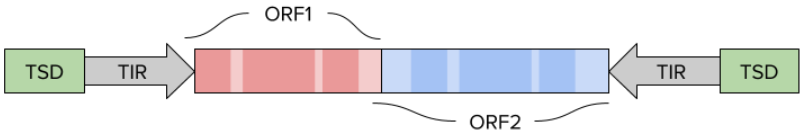


Figure 8, Structure of PIF-Harbinger Elements. TSD: Target site duplication, TIR: terminal inverted repeat, ORF1: open reading frame 1, ORF2: Open reading frame 2.

Tc1-Mariner

Tc1-Mariner is a superfamily of transposable elements that make up one of the larger families of TIR and MITE elements. The general structure of Tc1-Mariner elements consists of a transposase gene flanked by terminal inverted repeats (TIRs) region and terminal site duplications (TSDs) **[56]** . The sub-groups Tc1-Mariner are Tc1 and Mariner and they are distinguished by the catalytic domains of their transposases, as well as the length of TIRs **[56]**. Tc1 and Mariner encode for similar transposases that consist of three DNA-binding domains, designated D, D and E. These two elements are distinguished by their length with Tc1 possessing much longer

catalytic regions, and Mariner possessing shorter regions **[57]**. These domains are quite specific, and these differences directly impact where these elements insert.

Figure 9, Structure of Tc1-Mariner Elements

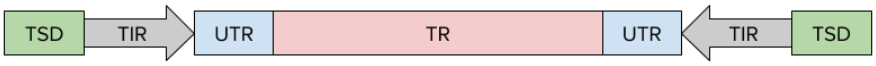


Figure 9. Structure of Tc1-Mariner Elements. TSD: target site duplication, TIR: terminal inverted repeat, UTR: untranslated region, TR: transposase.

Miniature Inverted-repeat Transposable Elements (MITEs)

Miniature Inverted repeat transposable elements are non-autonomous DNA transposons that exist in plants, animals and fungi **[58]**. MITEs were first observed in plants, and they are typically segmented into two groups: Tourist-like and Stowaway-like. This classification is based on the differences in their derivation; Tourist-like descending from PIF-Harbinger, and Stowaway descending from Tc1-Mariner **[37]**. The typical structure of a MITE consists of flanking terminal inverted repeats, and terminal site duplication sequences. Because they are non-autonomous, a MITE will not encode for transposase, and the source of transposase will be another element with similar TSD flanking regions, such as a PIF-Harbinger class element in *Zhang et al (2001)* **[55]**.

Figure 10, Structure of Miniature Inverted-repeat Transposable Elements

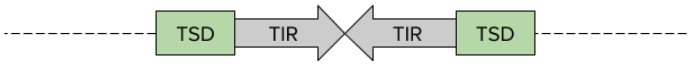


Figure 10, Miniature Inverted-repeat Transposable Elements. TSD: target site duplication, TIR: terminal inverted repeats.

hAT (hobo-Ac-Tam3)

hAT (hobo-Ac-Tam3) elements are a superfamily of autonomous DNA transposable elements **[59]** consisting of three sub-elements: Maize (*Z. mays*) element Ac (*Activator*), Drosophila element hobo and Snapdragon (*A. majus*) element Tam3 **[60]**. These seemingly unrelated elements share structural similarity to Ac Activator initially observed by McClintock **[10]**, and for this reason, all are classified as Activator-like and grouped into hAT elements. The structure of hAT elements typically consists of a target site duplication and terminal inverted repeat regions associated with class II transposable elements. Due to the fact that Ac and Activator-like elements are autonomous, there is also a region that encodes for transposase, with members of the hAT family possessing a characteristic amino acid motif unique to the family **[59]**.

Figure 11, Structure of hAT transposable elements and an Example of Sequences with Conserved Motifs

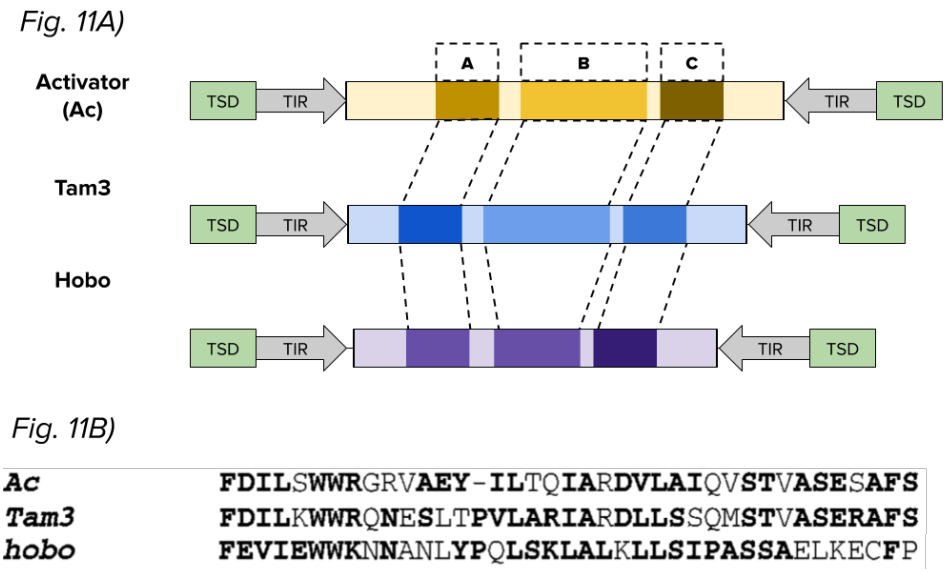


Figure 11, hAT transposable elements and sequences of conserved motifs. TSD: target site duplications, TIR: terminal inverted repeats. Regions (A), (B), and (C) correspond to highly similar motifs within each element, the coding regions of these motifs being represented in fig. 11B.

Mutator/MULE

Mutator and Mutator-like elements (MULEs) are a family of class II DNA transposable elements historically observed in maize **[16]**. Mu and Mu-like elements are distinguished by unusually long TIRs, in some cases spanning approximately 210-220 base pairs in length **[61]**. In the case of the element MUDR, the left and right TIRs were noted to also contain promoters. *Raizada et al* (2008) **[61]**, observed that mutator elements shift in transposition frequency depending on cellular development stages. Specifically, they observed that upregulation of MUDR in pollen development was correlated with higher insertion rates of mutator elements. This is of note, as transposable elements are known to be disruptive to gene function and upregulation of MUDR and increases in insertion can increase the likelihood of transposon mutagenesis occurring **[62]**.

Figure 12, Structure of Mutator and Mutator-like Elements

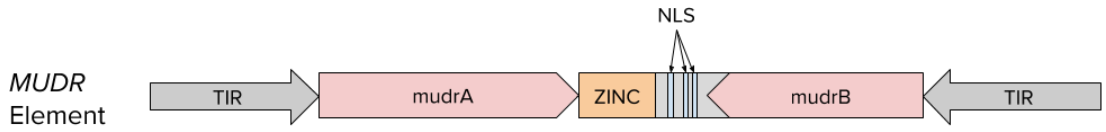


Figure 12, Mutator and Mutator-like Elements. TIR: terminal inverted repeats, mudrA: subunit A of mutator element A, mudrB: subunit B of mutator element that acts like a promoter, ZINC: encodes for a zinc-finger domain in transposase for some mutator elements. NLS: nuclear localization signals, responsible for tagging a sequence for transport.

CACTA

CACTA elements are a family of autonomous class II transposable elements characterized by a highly conserved flanking terminal inverted regions with the motif "CACTA", which serves as a recognition site for transposase protein **[63]**. The first characterized element of the CACTA family, En/Spm, was observed in Z. mays, and other CACTA elements have been observed in other grasses such as rice and sorghum. *Wicker et al., 2003)* **[12]** identified a subfamily of CACTA elements called Caspar. They observed that in Z. mays that these elements occur with high copy number implicating them as potential drivers of increased genome size when left unchecked. In addition, many of the CACTA elements identified by *Wicker et al* were observed to be non-functional or defective, possibly due to deletions and frameshifts resulting in loss of transposase **[62]**.

Figure 13, Structure of CACTA Elements and Subsets of Caspar family Elements

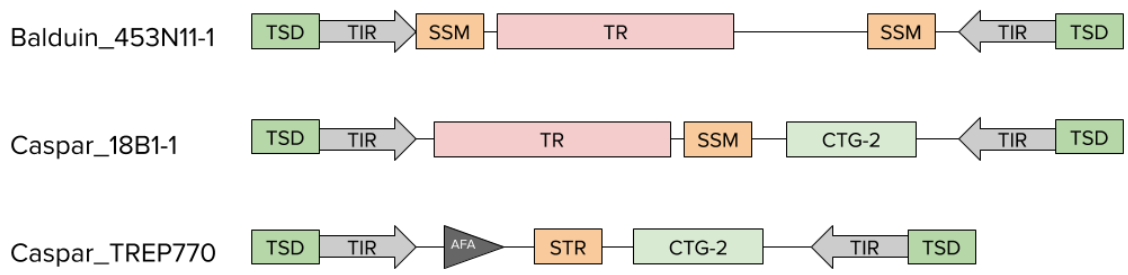


Figure 13, Subset of Caspar family of CACTA Elements. Gene structure depicted included introns. CACTA elements Balduin and Caspar all contain flanking TSD: target site duplications and TIR: Terminal inverted repeats characteristic of class II elements.

Subclass II: Helitrons & Maverick Transposons

Helitrons

Helitrons are a relatively new family of autonomous DNA transposons present in plants and animals originally discovered in Arabidopsis thaliana **[64]**. A notable differentiator of helitron elements when compared to other DNA transposons is the characteristic lack of generating a TSD **[65]** during propagation in addition to a very different method of propagation termed "rolling-circle" transposition. The general structure of a helitron consists of flanking terminal inverted regions and several protein encoding regions; a transposase, a variable combination of helicase, repA protein, or a combination of both as a single protein RepHel **[64]**. Transposase acts at the region of insertion by nicking the dsDNA and this begins the rolling circle mechanism of proliferation and insertion. This rolling-circle transposition is similar to the transposition method employed by virus and bacteria specific elements, but had not been seen in eukaryotes prior to the identification of this element **[66]**. Helitron movement consists of four primary steps: (1.) Recognition: An enzyme, typically RepA in bacteria, will recognize and bind to the 5' end of the helitron **[44]**. This enzyme is also called an initiator protein. (2.) Cleavage: Nick of the 5' end of

dsDNA by the initiator protein. (3.) Elongation: The initiator protein forms a protein complex with helicase, called RepHel in *Xiong et al (2016)* **[67]** and this protein complex will hold open the binding site while another enzyme, polymerase, binds to the strand of DNA, and elongation proceeds. (4.) Release: At this point the leading lagging strand is closed using DNA ligase, and the newly synthesized DNA is released from the RepHel. Another proposed method suggests that transposase excises the entire transposon region, facilitated by the left and right terminal regions interacting with one another to form the characteristic circular structure **[68]**.

Figure 14, General Structure of Helitron Elements

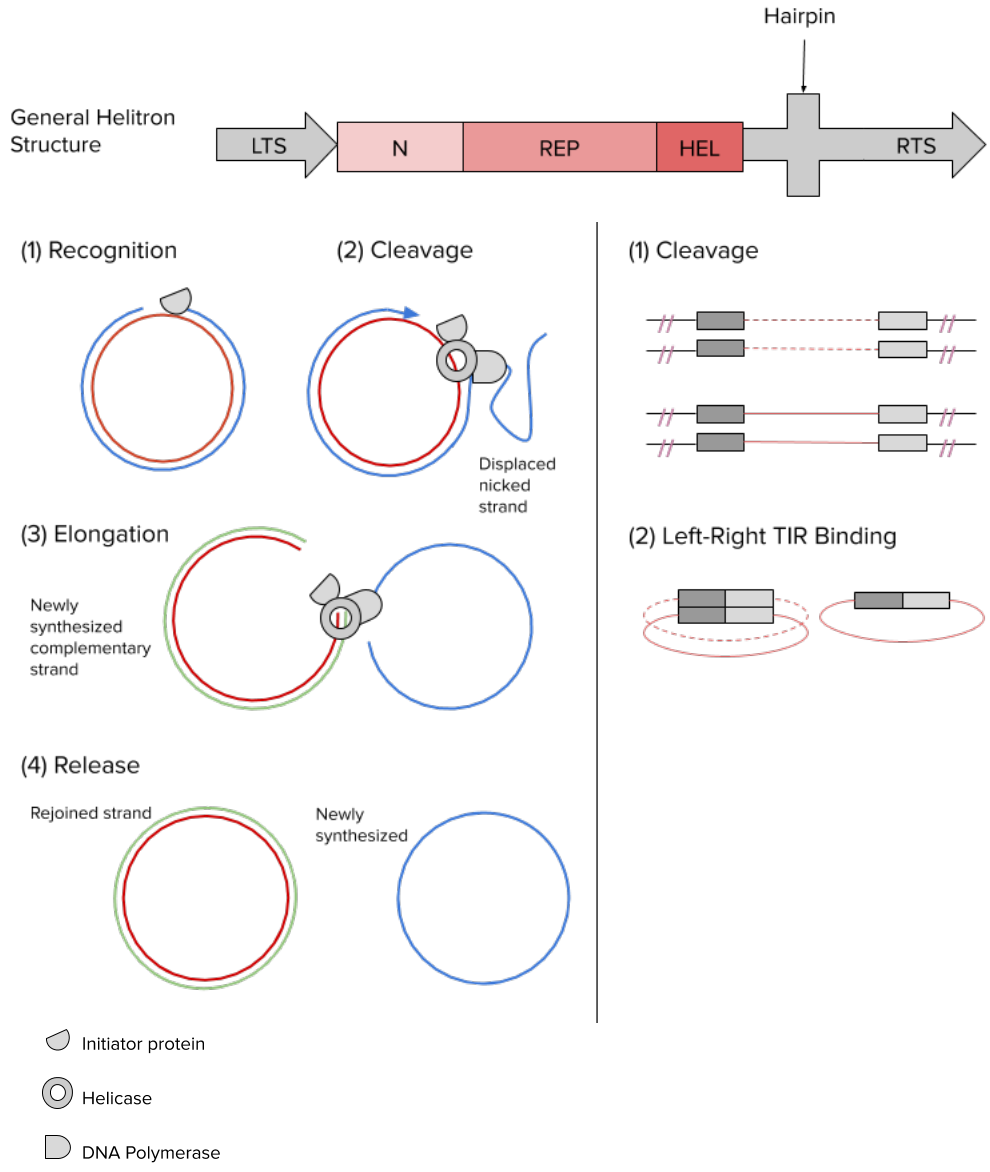


Figure 14, General structure of helitrons and mechanism of rolling-circle replication and transposition in helitrons.

Helentrons and other Helitron-like Elements

Helentrons are a subtype of non-autonomous transposable elements that also employ the rolling-circle replication method of transposition **[69]**. Once assumed to be a variant of helitrons, Helentrons were classified as their own class of transposable element by *Poulter et al (2003)* **[69]**, due to significant structural differences when compared to helitrons observed plants. The differences seen in *Poulter et al*, was the presence of an endonuclease like domain and lack of introns. Maverick (Polinton) elements are another relatively new class of transposable element that also replicates via the rolling-circle-replication method seen in helitrons and helentrons, but has not yet been observed in plants, although it has been observed in prokaryotes, fungi, vertebrates and invertebrates **[70]**.

Figure 15, General structure of Helentrons and helitron-like elements

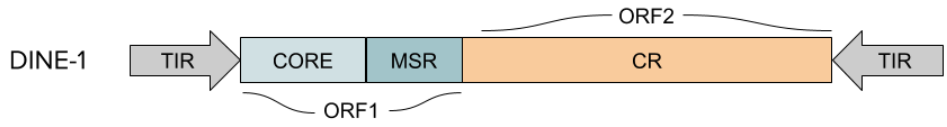


Figure 15, General structure of Helentrons and helitron-like elements. DINE-1, helentrons discovered in D. melanogaster, consists of the characteristic TIR: terminal inverted repeats, flanking the outside of the element. DINE-1 has two reading frames, ORF1 and ORF2. ORF1 contains two regions CORE, MSR. ORF2 contains CR: central repeats.

Methods of Identification of Transposable Elements

Identifying transposable elements is a highly computational task, and it is one of the most challenging steps in genome characterization due to the nature of TEs themselves **[71]**. This difficulty is due to several factors: (1) Repetitive elements are dispersed across genomes, and are subject to mutations from generation to generation, with relatively little selective pressure allowing for varying degrees of divergence making it challenging to categorize into known families **[71]**. Additionally, these elements are well known to insert themselves into other

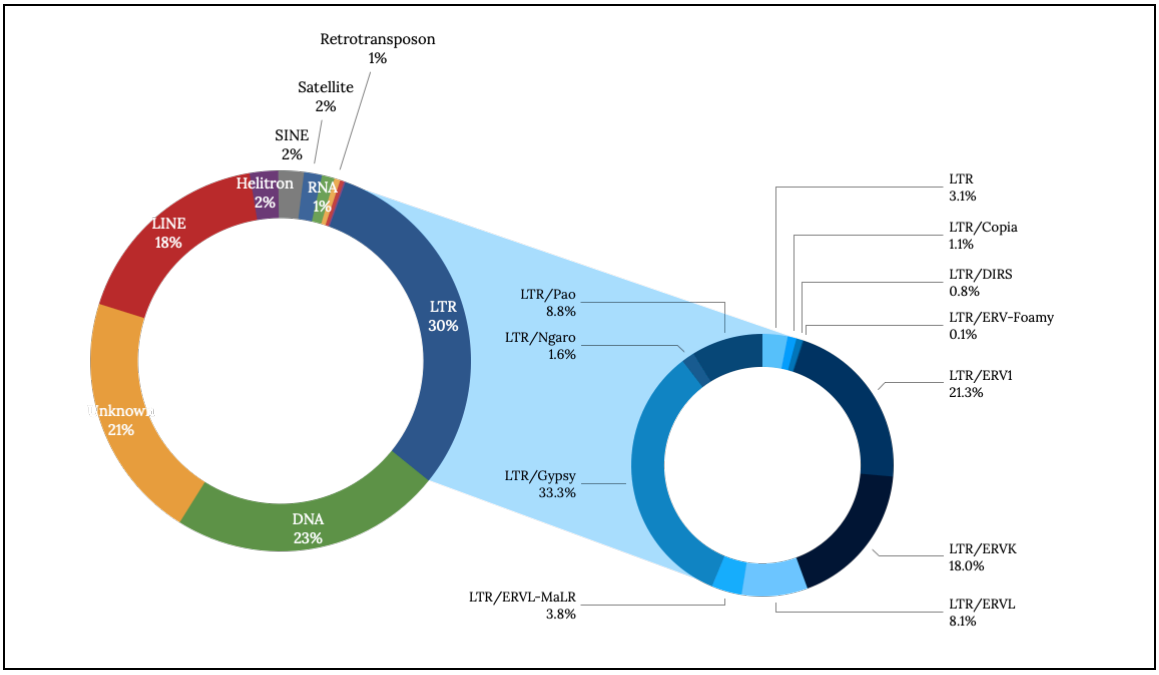
elements. Further compounding this effect, is (2) the exponential increase in the amount of data we are able to sequence and analyze. In the case of plant genomes, which tend to be larger in size and ploidy, analysis and complete identification is difficult to accomplish **[12]**. Despite these difficulties, there has been progress in terms of algorithms that can more accurately identify and categorize repetitive elements. Currently, there are four primary approaches utilized in the detection of TEs: homology-based methods, structure-based methods, comparative genomics methods **[72]** and *de novo*-based methods. No single method is considered the "best" approach to annotating repeats within a genome and most projects rely on the combination of several or all of these approaches.

Homology-Based Methods

Homology-based methods are approaches to TE detection that rely on knowledge of known TE protein-coding sequences **[72]**. These methods are advantageous to other methods as sequences that display similarities to prior confirmed sequences (prior knowledge) are more likely to be a transposable element **[72]**. The confidence in elements identified using homology allows for the detection of low-copy TEs, as they are not relying on copy-number as a filter for identification. With these benefits, homology-based methods are preferred whenever possible. There are caveats, however, to this method of TE identification, in particular when TEs are novel and/or if the potential TE lacks protein-coding regions. TEs that are underrepresented in databases or not yet characterized can lead to a bias against any potential TE too dissimilar to known sequences, resulting in further perpetuation of homology searches, as only elements that related to those known are detected. Elements that lack recognizable protein-coding regions, such as those observed in MITEs and SINEs, are also unfortunately negatively biased and commonly go undetected during TE identification. Perhaps the most popular homology-based repeat

identification software available is RepeatMasker. Repeat libraries, such as Dfam and Dfam_consensus (Version 3.2) **[73]**, RepeatPeps (version 2.0) **[74]** and Repbase (Release 20181026) **[75]**, that are distributed as RepeatMasker metadata files, consist of various TEs, with majorities belonging to LTR and DNA families *(Figure 16).* Diversity observed in these methods of TE identification is essential, as this can directly affect homology-based identification by RepeatMasker, conveying the importance of increasing diversity to expand repeat identification capabilities.

Figure 16, TE Family Representation in RepeatMasker Library



Homology-based methods identify TE elements by using one or more different types of analysis. The more common methods entail pairwise alignment and clustering techniques to assess similarity of the potential TE to elements present within a database. Perhaps the most popular pairwise alignment tool used in repeat identification is BLAST as well as being the most popular bioinformatics tool to determine sequence similarity, or homology, between a query

sequence and database of sequences **[76]**. Within the scope of RepeatMasker, it serves as an initial filter for repeats that share a high degree of homology. For elements that encode for regions with a known sequence or structure, BLAST is capable of confidently identifying those regions, and does so efficiently, saving time during increased computationally intensity. The BLAST algorithm was created to search and compare unknown sequences to known sequences within a database faster than performing a local (Smith-Waterman) alignment. There are several benefits of BLAST; namely speed, user-friendliness, statistical rigor, and increased sensitivity to query matches (Table 1).

Table 1, Outline of BLAST Algorithm as used in RepeatMasker

Consists of identification k-mers, or substrings of length k from the query. In querying DNA, the interval of matches of k-length is longer than compared to protein queries. This is due to the fact that there is a higher probability of finding a match the smaller the probability of a specific match (e.g. $\frac{1}{4}$ versus $\frac{1}{20}$). These k-mers are then put into a hash. To increase accuracy, the algorithm will also allow some level of mismatch of the k-mers to match in the database. This creates what is called a neighborhood of each k-mer. *The algorithm quickly finds sequences in the database that contain at least one of the k-mers found in the first step. This is done is to take the k-mer identified from the hash and locate it in the database if it is present.* *Upon discovery of a hit for the k-mer in the database, BLAST then looks to the left and right of the k-mer that matched to extend the seed, allowing for some degree of mismatches. If the result overall aligned sequence has a high enough score, the sequence would go on a list. This is called a high scoring pair or HSP. BLAST returns a score for each hit, and sorts them by the size of the score.* *The algorithm also returns an e-value. The e-value is the expected number of hits of score SS or better if DD is a database of the same size and composition as the real database using the same query sequence. The e-value is derived by looking at the maximum of a set of random variables. Under the assumption of the database being random, the random variable is in fact, the score. The distribution of the random variables is an extreme-value distribution or Gumbel distribution.*	**1. Generate words from sequence above threshold (e.g. T=11)** Query Sequence: >gi|16329320 (residues 412 to 594) SGANFARQLRTHKRQRIARQATTETQADRTQQAVGRIIGSIGVVTTQTTG RHQGILTSWVSQASFTPPGIMLAIPGEFDAYGLAGQNKAFVLNLLQEGRS VRRHFDHQPLPKDGDNPFSRLEHYSTQNGCLILAEALAYLECLVQSWSNI GDHVLVYATVQAGQVLQPNGITAIRHRKSGGQY Fragmentation into words: SWVSQASFTPPGIM → SWV WVS VSQ SQA QAS ASF SFT ... Selection of words scoring above threshold (for word SWV): Substitution Matrix* R G I K F S T W V R 5 0 -1 -1 -2 1 0 -3 0 G 6 -4 -2 -3 0 -2 -2 -3 I 4 -3 0 -2 -1 -3 3 K 5 -3 0 -1 -3 -2 F 6 -2 -2 1 -1 S 4 1 -3 -2 T 5 -2 0 W 11 -3 V 4 *A portion of the BLOSUM 62 matrix SWV (4+11+4 = 19) SWI (4+11+3 = 18) TWV (1+11+4 = 16) GWV (0+11+4 = 15) KWV (0+11+4 = 15) SWS (4+11-2 = 13) SFV (4+1+4 = 9) SRV (4-3+4 = 5) Synonyms above threshold 11... (others not shown) Synonyms below threshold 11... (others not shown) **2. Search the database for words matching those generated** **3. Extend matching hits in both directions** RHQGILTSWVSQASFTPPGIMLAIPGEFDAYGLAGQNK ..TAMLVSWVSQASFNPPGLTIALAKE.RAEGLDHSGD Word match from Step 1 Extension until score drops **4. Generate alignment and calculate statistics** >ref|YP_002482587.1| flavin reductase domain protein FMN-binding [Cyanothece sp. PCC 7425] gb|ACL44226.1| flavin reductase domain protein FMN-binding [Cyanothece sp. PCC 7425] Length=585 Score = 176 bits (446), Expect = 1e-42, Method: Compositional matrix adjust. Identities = 95/196 (48%), Positives = 125/196 (63%), Gaps = 16/196 (8%) Query 1 SGANFARQLRTHKRQRIARQATTETQADRTQQAVGRIIGSIGVVTTQTTGRH-------- 52 +G++FA+ L+ K+QR RQ+ E Q+DRT+QAVGRIIGS+ V+T + H Sbjct 393 AGSDFAQVLKKAKKQRSPRQSILEVQSDRTEQAVGRIIGSLCVLTAKQQQTHPHPEVEEP 452 Query 53 -----QGILTSWVSQASFTPPGIMLAIPGEFDAYGLAGQNKAFVLNLLQEGRSVRRHFDH 107 +L SWVSQASF PPG+ +A+ E A GL AFVLN+L+EG ++RRHF Sbjct 453 QLEVPTAMLVSWVSQASFNPPGLTIALAKE-RAEGLDHSGDAFVLNVLKEGMNLRRHFSK 511 Query 108 QPLPKDGDNPFSRLEHYSTQNGCLILAEALAYLECLVQSWSNIGDHVLVYATVQAGQVLQ 167 P G++ F+ L +NGC +L + LAYLEC VQS GDH L+YATV G+VLQ Sbjct 512 SFAP--GEDRFAGLNIQWAENGCPVLQDCLAYLECTVQSRMECGDHWLIYATVNNGKVLQ 569 Query 168 PNGITAIRHRKSGGQY 183 P G TA++HRKSG QY Sbjct 570 PTGTTAVQHRKSGNQY 585 ***Steps in the BLAST algorithm. Kerfeld et al (2011) [77]***

There are also modified versions of BLAST developed with much of the same underlying algorithm of BLAST but additional stringency parameters that result in decreased computation time of an already somewhat fast heuristic algorithm. One software package like this is BLAT (Blast-like Algorithm Tool), which is characterized by its **[78]** increased requirement of sequence matches, requiring an almost exact match to the database it queries. The result is speed but with

disregard to sequence variability in the process. Underlying algorithm aside, much of the appeal of RepeatMasker is speed, due in part to its use of fast pairwise alignment algorithms, and users have the choice of using cross_match or a modified BLAST, to quickly return possible homologous results from the Repbase **[75]**, a highly curated repeat database. The classification system RepeatMasker uses also adds to the appeal of the software package, as results are returned in a categorized manner based on work done in *Wicker et al., (2007)* **[12]**.

Clustering methods identify repeats by performing multiple sequence alignment of all sequences in a given cluster/group/pile. This method of analysis can be performed in Homology-based applications as well as *de novo*-based applications, so there is a bit of overlap in terms of the category of analysis that defines it. A homology-based application of this method is utilized within our pipeline is VSEARCH **[79]**. VSEARCH is a pairwise alignment clustering application that clusters sequences based on similarity derived from sequence alignment.

De novo Methods

De novo methods of repeat identification are techniques of discovering transposable elements by using only the reference genome data to identify repetitive elements and DNA fragments within the genome. The defining feature of this method that makes the approach powerful is the ability to identify elements without prior information other than the original genome **[12]**. This is extraordinarily advantageous in the situations where a genome undergoing characterization lacks a reference. This method is also heavily dependent on the quality of assembly data; therefore it is essential that the researcher obtaining the greatest quality assemblies available. The *de novo* method of transposable element identification follows a fairly straightforward strategy to identify repetitive elements, beginning with self-genome comparison. During this process, a *de novo*-

based software package scans the genome for reoccurring instances of motifs located throughout the genome.

Suffix-tree based methods are *de novo* methods that rely on suffix-tree data structures, as in the storage of subsets of strings, to efficiently perform computational processes to identify repetitive elements. A suffix-tree is defined as a compressed trie, a or a data structure that stores suffixes and positions of given sets of associate arrays that possess strings as keys **[80]**. There are many approaches used in the identification of sequences and motifs, but in the simplest case, the underlying algorithm will consist of generating suffixes, defined as derivatives of the original sequence, and placing these suffixes into a tree structure, or index, that then allows for quick and efficient processes.

K-mer-based tools are repeat identification tools that work by detecting potential repeats by analysis of overrepresented *k-mers* **[81]**. Several tools have been developed that use this method of identification, and one of the more well recognized tools is RepeatModeler. The underlying algorithm of RepeatModeler identifies potential repeat sequences by performing an initial identification of sequences using another *k-mer* analysis tool, RepeatScout **[82]**, to perform an alignment of potential repeats that are found in the genome and align these sequence to form a consensus repeat sequence, optimized using what is termed a "preferred-fit" alignment that accounts for boundaries**[82]**. This preferred-fit alignment results in a balance in the instances of over assignment or false positives, and it does so by setting a cutoff that results in a minima of sequences identified. In the case of RepeatScout's initial parameters, a minimum *k-mer* length of 15bp and a repeat frequency threshold of 3 occurrences is the threshold to which a repeat family is constructed and analyzed. Additional filtration parameters include the removal of tandem repeats (defined as repeat families with >50% of their length annotated as tandem repeats by

Tandem Repeat Finder **[83]**), low-complexity repeats, pseudogene families, partial duplications and repeat families with less that 50% of their length annotated as low-complexity by Nseg **[97]**. Other tools that use this method of *de novo* repeat identification include ReAS, REPdenovo, and RED. ReAS, is a *k-mer*-based software package that utilizes seed-and-extend methods to identify repeats **[84]**. REPdenovo is a similar tool that infers repeats via unique motifs and their k-occurrences within a given sequence or genome. One advantage of REPdenovo is that a genome does not need to be completely assembled to perform analysis **[85]**. Red is another *k-mer* based tools that identifies repeats but instead of looking for a simple count of occurrences or for known motifs, Red uses HMMs to identify predicted motifs that may not be known **[86]**. *De novo* methods allow for the discovery of completely novel TE families because there isn't a dependence on prior sequences. Possible issues of these methods is the lack of being able to differentiate between other repeat classes and distinct TE families and or degenerate repeats and, the quality of discovery is highly dependent on the quality of assembly data.

Pairwise Alignment Clustering - De novo Applications

Pairwise alignment clustering is a method of analysis that can apply previously discussed homology-based methods *(Homology-Based Methods)* and *de novo*-based methods. An example of software utilizing this method in a *de novo* context is RECON, a software package present in RepeatModeler during generation of a baseline repeat library. RECON is designed to generate *de novo* families of repeats from similar repeat alignment profiles **[87]** and families are assigned names corresponding to the order they are identified in (e.g. family-1). Another pairwise alignment tool in a *de novo* application is GROUPER, a single-linkage clustering tool that uses with overlapping constraints (utilizes the single-link clustering algorithm) **[88]** to generate sequence families. RPT is another method of *de novo* clustering but utilizes more stringent

searches for alignments, specifically requiring an overlap of 90% in ungapped alignments groups to perform single-link clustering **[89]**. There are many *de novo* applications of pairwise alignment and cluster, however the method is not without its shortcomings, in particular the time required to perform these analyses. A large downside to using pairwise similarity-based approaches, homology-based or *de novo*, is the computational intensity required to identify instances of TEs. A notable example of this is during *de novo* library generation RepeatModeler, where the first round of analysis can take upwards of days in the cases of larger genomes and can require running on a high-performance cluster to complete jobs in an adequate timeframe.

Structure-Based Methods

Structure-based methods, also called signature-based, are a strategy of transposable element detection that rely on the general structure shared by all TEs and required for proper TE function **[71]**. TE architecture is the pillar of this method, and though some knowledge of TE structural components must be known for this method to work, structure-based methods are not as biased as homology-based methods to identify TEs as structural methods are not limited by expected boundaries of a transposed elements as homologous and *de novo* methods **[72]**. Disadvantages of this method include the dependency on the structure of the TE in question. Some TEs retain their characteristic structure more so than other types and proper identification may prove more difficult if a TE does not have a defined structure, namely in cases where an element has degraded, loss of function in encoding regions via transposition error and or nesting of other TEs. Structure-based methods of TE identification also shares identification strategies with comparative genome methods, particularly when investigating larger-scale alterations by TEs, e.g. identification of larger insertions as in the case of LTRs. Both methods investigate physical similarities between sequences being compared, differing the scale of these comparisons. In

comparative genomics, whole-genome alignments are typically utilized for these methods, beginning with the search for insertion regions **[90]**. The rationale behind this method lies in capitalizing on what occurs when a TE inserts itself into a genome. Upon insertion, a transposon will insert not only genes, but also unique sequences into these insertion regions. Comparative genomic methods detect these regions and align them to identify repeating insertion regions. The repeating insertion regions are then clustered **[90]**. Tools capable of carrying out genome comparisons can be found in the software suite CoGe **[91]**, which consists of multiple tools that rely on pairwise alignment to compare genomes of organisms, LAST, a tool capable of genome scale comparisons, and PLAZA, a plant specific platform for whole genome comparisons. A potential downside of pursuing this particular method of analysis is the overwhelming dependence on the quality of whole genome alignments and the activity of TEs in a genome. TEs that have older insertion regions in relation to the species being analyzed, they will not be detected **[92]**.

Detection of LTR Elements

LTR elements are Class I TEs that possess characteristically long terminal regions often spanning kilobases in length. LTRs are often among the most abundant class of elements observed within a given genome, and detection often relies on conserved structural components, such as the aforementioned terminal regions genes that encode for highly conserved proteins. Availability of tools capable of identifying LTR transposable elements is expansive. Among one of the earliest tools available to perform this task was LTR_STRUC. LTR_STRUC used a seed-and-extend to find specific elements (TSDs, PBSs, ORFs) of a TE. Following LTR_STRUC came LTR_par, another tool that utilized a improved upon the relatively fast seed-and extend algorithm used in LTR_STRUCT. A notable improvement in the detection of LTRs came in the form of the

incorporation of enhanced suffix arrays into the analysis process. These data structures were first utilized in LTRHarvest, a *de novo* LTR prediction software package that relies on the characteristic long terminal repeat region to predict and distinguish these repeats in a given sequence **[93]**. Repbox incorporates the suffix array approach in utilizing LTRHarvest and LTR_retriever to identify LTR elements. Like LTRHarvest, LTR_retriever utilizes suffix arrays to quickly identify and predict LTRs, however LTR_retriever offers additional filtering parameters to remove false positives based upon structural characteristics of LTRs, such abnormally-sized predictions, elements lacking terminal site duplications (TSDs), removal of non-LTR proteins, and boundary determination [94]. The result of these additional filtration processes is a reported

Detection of MITE Elements

MITE elements are small Class II DNA transposons characterized by their smaller size and high abundance throughout the genome. There are several detection options available, many of which rely on structural characteristics to make predictions. Software packages considered for our novel pipeline included MITEFinder and MITE-Tracker. MITEFinder and MITETracker both use *k-mer* approaches to scan the genome for potential candidates, but differ in how motifs are discovered. MITEFinder segments the genome into 10000bp fragments and performs analysis *k-mer* searches on each fragment. Hits lacking a terminally inverted region (TIR) and tandem site duplication (TSD) are discarded and filtered from the candidates and those remaining are merged and assigned a likelihood score. The likelihood score calculated here is essential, as those that exceed the threshold are considered a true positive and classified as a MITE element. MITEFinder uses a scoring formula based upon models of positively identified MITE sequences found in Repbase and null sequences based on non-MITE sequences. The score is derived as a log ratio of

the probability of a given sequence S appearing within *M*, the true positive MITE sequences in the model dataset, divided by the probability of a given sequence *S* appearing within N, the false positive MITE sequences in the model dataset. Further consideration of the variance of MITE length is taken by MITEFinder by the assumption that the longer a given sequence, the more likely we are to observe a given fragment. Therefore, to avoid potential biases, the score is further divided by length *n*. Final selection of MITE sequences consists of final clustering based on 80% sequence identity and an all-by-all BLASTN comparison is performed at the default e-value of 1e-10.

The algorithm utilized by MITETracker is similar to MITEFinder in that it initially identifies MITE candidates by segmenting the genome and searching for sequences possessing the distinct TIR and TSD regions. These sequences are identified based on calculating a Local Composition Complexity Score (LCC) **[95]**, where sequences are scored on the basis of sequence complexity. MITETracker then performs BLAST alignments of the identified TSD and TIR regions, quickly deriving mismatches and then clusters those most similar into families. This process is performed until distinct families are formed for all identified TSDs and TIRs, and a representative sequence for each cluster or family is derived.

Detection of Helitrons

Several options for the detection of Helitrons are available, many relying on conserved structural characteristics of Helitrons to make predictions of potentially novel elements. Software packages under consideration for our pipeline were EAHelitron and HelitronScanner. EAHelitron makes predictions of potential helitron candidates by scanning for specific motifs observed in Helitron elements, namely a -TC motif on the 5’ end and -CTAG motif on the 3’ end followed by a GC-rich hairpin loop upstream of the motif **[96]**. When EAHelitron detects these motifs, the

reverse complement of these regions are also captured in order to detect other motifs of the helitron, as helitrons possess multiple hairpin loop structures that are implicated in propagation and distinct for this class of transposable element **[64]**. HelitronScanner predicts elements much in the same fashion as EAHelitron, scanning the genome for distinct -TC motifs on the 5' end and CT[A,G] at the 3' end, however it utilizes a LCV (local combinational variable) to train sets representative of the 5' and 3' ends. The derived training sets are then used to calculate threshold scores in sequences that possess characteristic motifs similar to those found in LCVs **[97]**. EAHelitron makes predictions of potential helitron candidates by scanning for specific motifs observed in Helitron elements, namely a -TC motif on the 5' end and -CTAG motif on the 3' end followed by a GC-rich hairpin loop upstream of the motif **[96]**.

Common Methods of TE Identification in Plant Genomes

Despite the great strides made in tools available for identification of TEs, annotation standards including methods or protocols of annotation, tend to lack in consistency **[98]** especially in comparison to annotation standards for genes **[99]**. In essence, many protocols fail to consider detailed characterization of TEs, as noted in *Ragupathy et al (2013)* [98]. The authors advocate for the need to improve plant-specific TE representation, as this is historically lacking in terms of genome characterization. They also discuss the need for innovation of methods taking into the consideration the role of transposable elements in plant genomes *(Chapter 1: Perspectives on the Role of Transposable Elements)* and addressing the limitations of current practices that depend so heavily on homology to repeat databases. This is not to say there is not an accepted set of tools used when looking at plant genomes, as the most popular methods currently used in repeat characterization consistently include tools such as RepeatMasker and RepeatModeler. However, the motivation behind using these tools in some cases, is to simply mask the genome, and as

Ragupathy et al (2013) note, the standardization of TE annotation in plant genomes needs improvement. This lack of standardization can result in impediments to our ability to critically understand the influence of TEs on the genomes of the plants we wish to study.

Chapter 2: DEVELOPMENT OF A BIOINFORMATICS TOOLBOX FOR REPEAT ANNOTATION

Introduction

Identification of transposable elements (TEs) is a well-known computationally intensive task. The detection of these elements has become increasingly difficult due to an increase in the size of sequenced genomes, as well as additional methods of analysis available for identifying repetitive elements **[100]**. Most of the solutions to this ever-increasing analytical load have come in the form of algorithms and software capable of efficiently analyzing the genome for TEs and providing a means of identification and classification that is relatively user-friendly, such as the well-established RepeatMasker **[101]** and RepeatModeler **[102]**.

However, the quality and diversity of available algorithms and software designed to identify repeats continues to improve and cover a wider range of specific classes of repeats, representing the potential for these newer and more innovative tools to further improve results of TE annotation. There is a plethora of software packages available that specialize in identifying specific classes of TE families. Previous works, such as *Ou et al (2019)* **[103]**, *Lerat et al (2010)* **[104]**, and *Saha et al (2008)* **[71]**, review the landscape of available software options as well as methods of software optimization in TE annotation. The consistent conclusion from these studies is that TE annotation processes need to employ some permutation of homologous, *de novo*, and structural-based methods followed up by implementation of a classifier to group identified elements into families.

Despite the progress made in this area of genomics, many of these tools are often not easily implemented by users possessing a limited bioinformatics skillset and this can prove to be a barrier for users desiring to use novel software in an effort to improve TE annotation. In addition to limitations in implementing novel software in TE analysis, the process of TE annotation is also

plagued by issues concerning the assignment of unclassified or unknown elements. The primary cause of unknown or unclassified element assignment is a lack of TE family representation, as this is a notable observation in instances of a genome that contains novel repetitive elements **[105]**.

Therefore, to address these issues, we developed Repbox, a user-friendly suite of family-specific TE detection software that incorporates all methods of identification, and is capable of identifying diverse families of repetitive elements in plant species. Our motivation for this was two-fold. First, to investigate why unclassified assignments occur by evaluating the process by which these well-used tools, RepeatModeler and RepeatMasker, assign potentially classifiable elements to an "unknown" category. Secondly, to address these large numbers of unclassified elements, we detail the development of a novel TE detection pipeline, named Repbox, that implements novel software options in an effort to improve TE annotation and thereby reduce the number of unknown elements. In this section of the dissertation, we will discuss the development process of our pipeline and results we derived in our analysis of the well characterized genomes of *Arabidopsis thaliana* and *Oryza sativa*.

Materials and Methods

Benchmarking Protocol and Repbox Algorithm

Our pipeline was developed in a three-phase process: (1) Baseline repeat annotation of our test genomes using RepeatModeler and RepeatMasker, (2) Annotation of specific classes of repeats using de novo software packages not integrated in RepeatMasker and (3) Comparison and assessment of our Repbox annotation to the baseline annotation in phase 1 as well as to published repeat annotations from the reference genomes. The genomes used for benchmarking were two well-annotated organisms, *Arabidopsis thaliana* (TAIR10, INSDC Assembly GCA_000001735.1,

Apr 2008) [and *Oryza sativa* (IRGSP-1.0, INSDC Assembly GCA_001433935.1, Oct 2015) [, both retrieved from Ensembl **[106]**.

Generation of a Control Repeat Annotation

Baseline annotations were performed with RepeatModeler (version 2.0.1) **[102]** and RepeatMasker (version 4.1.0) **[101]**. This baseline TE annotation provides a control for repeat identification for downstream comparisons to the Repbox pipeline. RepeatModeler was run using the following code to generate a *de novo* repeat library:

Code 1, Construction of a de novo Repeat Library & Genome Masking

```
# Command 1: Construction of de novo repeat library
BuildDatabase -name $DBNAME -engine ncbi $GENOME

# Command 2: Prediction of candidate repeat families
RepeatModeler -database $DBNAME -engine ncbi -pa $THREAD -LTRStruct

# Command 3: Masking of genome using de novo repeat library using RepeatMasker
RepeatMasker -pa $THREAD -e ncbi -lib $LIBRARY -gff -dir $OUTPUT -u $GENOME
```

Construction of an index or database of the genome was performed using commands in Code 1, where -name is the index name, -engine is the query engine for comparing sequences within the genome (NCBI is default), and $GENOME is the genome as an input fasta file, in this case. Following the database construction, sequences are then processed to predict repeat families using Command 2 (Code 1), where -database is the index name used in construction of the index, -engine is the engine used to query sequences, -pa is the number of CPU threads, and -LTRSruct is an optional parameter that adds ability to predict LTR candidates using *de novo* methods via LTRHarvest and LTR_Retriever. This step generates a consensi.fa.classified file containing all repeats identified by RepeatModeler. RepeatMasker is then run using Command 3 (Code 1), where -pa is the CPU threads for the analysis, -e is the engine used for querying sequences in our library, -library is the consensi.fa.classified library generated during RepeatModeler analysis, -gff

is the option to have RepeatMasker provide a gene annotation file (.gff3) as output, -dir is the output directory, -u is the original genome or reference.

MITE Benchmarking

Identification of DNA transposons for benchmarking was accomplished by comparing two software packages, MITETracker (version 1.0) **[107]** and MITEFinder(version 2.0) **[108]**. Commands for MITETracker and MITEFinder are outlined in *Code 2*, where option -m calls MITETracker's scripts, -g is the genome, -w are the number of CPUs, and -j is the name of the index generated in the analysis. Parameters used in MITEFinder are detailed in Code 1, where the -input parameter represents the genome, -output refers to a user-defined filename for the output of analysis, -pattern scoring is a default that scores motifs found during analysis, and -threshold is the minimum score allowed for a given MITE candidate (default 0.2).

Code 2: Parameters of MITE Benchmarking for MITETracker and MITEFinder

```
# Command 1: MITETracker Parameters
python3 -m MITETracker -g $GENOME -w $THREAD -j $INDEXNAME

# Command 2: MITEFinder Parameters
$miteFinder -input $GENOME -
output $INDEXNAME.mite_finder.out pattern_scoring $REPBOX_PREFIX/bin/miteFinder/profile/patter
n_scoring.txt threshold 0.2
```

Helitron Benchmarking

Annotation of Helitron elements was performed using EAHelitron (version 1.5100)**[96]** and HelitronScanner (version 1.0) **[97]**. EAHelitron ran using default parameters, which includes the length of sequences up and down stream of a predicted helitron candidate, as well as output naming conventions. Upstream and downstream base pair lengths were determined by the authors utilizing conserved flanking regions of Helitron candidates, and this was derived using sequence characteristics observed across various species known to contain these elements **[96]**. The process of how these defaults were determined are outlined in Chapter 1, but briefly, the authors of

EAHelitron utilize definitive structural motifs known to occur in helitrons, namely the 5’ TC motif paired with a 3’ CTAG motif with a 2-10nt GC-hairpin loop on the end **[67], [109], [110]**. Analysis of helitrons used commands listed below in Code 3, where option -o is in reference to the prefix for the filename of the output (default: EAHeli_out), -u is the upstream sequence length (default: 3000bp), -d is the downstream sequence length (default: 500bp), and -r is the terminal fuzzy level or degree of flexibility in the terminal regions of the Helitron candidate. A parameter sweep of the fuzziness of Helitron terminal region was performed using Code 3. The increasing fuzziness or mismatch allowance displayed in column 1 of Table 2 altered and re-ran to assess which setting would produce the greatest number of helitron elements capable of classification using BLAST. Increasing the number of mismatches, from 0-5 resulted in more Helitron candidates, however, these candidates did not share homology to known helitron elements despite increasing the allowance of mismatches. As a result, we choose the most conserved settings that identifies elements by consideration of the more conserved 3’ terminal region, as increasing the number of unknown classified sequences will obscure identification of other repeat families.

Code 3, Parameters of Helitron Benchmarking

```
# Command 1: EAHelitron Parameters
perl $EAHelitron -o EAHeli_out -u 3000 -d 500 -r 0 $GENOME

# Commands 2,3 & 4: HelitronScanner Parameters
# Consists of generation of 'head' and 'tail' predictions, Pairing of head and tail
# predictions into a candidate helitron and generation of corresponding fasta sequences
# for #helitron candidates

$HelitronScanner scanHead -lf $REPBOX_PREFIX/bin/HelitronScanner/TrainingSet/head.lcvs -g
$GENOME -bs 0 -o $INDEXNAME.head -tl $THREAD
$HelitronScanner scanTail -lf $REPBOX_PREFIX/bin/HelitronScanner/TrainingSet/tail.lcvs -g
$GENOME -bs 0 -o $INDEXNAME.tail -tl $THREAD
$HelitronScanner pairends -hs $INDEXNAME.head -ts $INDEXNAME.tail -hlr 200:20000 -o
$INDEXNAME.paired
$HelitronScanner draw -p $INDEXNAME.paired -g $GENOME -o helitronscanner_out.$INDEXNAME -
pure_helitron
```

SINE Software

Additional analysis of potential SINE elements was performed using SINE_Scan **[111]**(version 1.1.1, https://github.com/maohlzj/SINE_Scan), a structural and homology-based SINE detection software package. SINE_Scan was chosen as our method of SINE candidate analysis due to a lack of additional software options available for SINE detection, and it incorporates another SINE detection tool, SINEFinder **[112]** into its analysis. SINEFinder is a python-based script that relies on structural characteristics and motifs to identify SINE candidates, however SINE_Scan is capable of bolstering this analysis by referencing SINEBase (version 1.1), a manually curated database of known SINE elements **[113]**. This added functionality is reported by the authors to improve the detection of SINE candidates by increasing the diversity of SINE families and novel SINE elements available for homologous comparisons in elements detected by structure alone, as is the case when exclusively utilizing SINEFinder to detect candidate elements. There is a lack of additional software options available for SINE detection, and so we choose to use SINE_Scan as part of Repbox.

Code 4, Running SINE_Scan

```
perl SINE_Scan_process.pl -g $GENOME -d $DIRECTORY -o $INDEXNAME -s 123 -k $THREAD
```

SINE_Scan consists of several modules that perform identification of SINE candidates in phases, as illustrated in *Code 4.* Parameter -s corresponds to these phases of analysis with -s 1 or step one involving *ab initio* identification of SINE candidates, -s 2 or step 2 consisting of checking for SINE candidates by identifying sequence signals of TE amplification or multiple instances of SINE candidate identified in step one, and -s 3 or step three completing analysis by creating an annotation of the SINE candidates. In the code above, all steps are ran as -s 123. Other parameters, -g -d -o and -k correspond to the reference genome, any working directory, output of files from the analysis, and number of CPUs to perform the analysis, respectively.

Assessment of Novel TE Detection Software

To assess the quality and quantity of transposable elements identified by novel detection software, the reference transposon GFF files for *Oryza sativa (*IRGSP Build 5) and *Arabidopsis thaliana* (TAIR10) **[114]** were analyzed to derive information on the TE profile currently available on each genome. Additional text processing and statistical analysis was conducted using bedtools (version 2.0) **[115]**, Python (version 3.7.7) **[116]** and R (version 4.0.2), with third-party packages: dplyr (version 1.0.0) **[117]**, chromPlot (version 1.16.0), stringr (version 1.4.0) **[118]** and reticulate (version 1.16) **[119]**. Data cleanup and partitioning of all GFF files was conducted in R, followed by comparisons of the processed GFF files using coordinate data (start, end columns), the degree of overlap at those coordinates and the total number of overlaps reported in the analysis. Specific commands used in R for processing the reference GFF files is provided in the *Appendix A*. Comparisons of novel TE detection software and reference annotations are conducted by a performance assessment using a calculated equivalent to the percentage of elements identified

from each package based on the total number of corresponding elements in the reference annotation.

Equation 1, Software Assessment Score

$$Score = \frac{\#\ of\ elements\ in\ specific\ family\ identified\ -\ \#\ of\ elements\ of\ specific\ family\ in\ reference}{\#\ of\ elements\ of\ specific\ family\ identified\ in\ reference}\ x\ 100$$

Bedtools intersect was used to compare repeat feature annotations between MITE-Tracker and and MITEFinder as well as between EAHelitron and HelitronScanner by comparing each software output to the reference using coordinate data derived from the GFF files created by each package. Commands for performing the comparison were executed by bedtools in bash (version 3.2) environment (Code 5).

Code 5, Commands for Assessment of MITE & Helitron elements in Oryza sativa

```
# MITETracker to reference comparison
bedtools intersect -a Oryza_reference_annotation.gff -b mite_tracker.gff3 >
ref_MITE_MITETracker_overlap.gff3

## MITEFinder to reference Comparison
bedtools intersect -a Oryza_ref.gff -b mitefinder.gff3 > ref_MITE_MITEFinder_overlap.gff3

## EAHelitron to reference comparison
# Bash commands: bedtools intersect -a Oryza_helitron_reference.gff3 -b EAHeli_out.gff3 -f 0.8
> ref_Helitron_EAHelitron_overlap.gff3

## HelitronScanner to reference comparison
# Bash commands: bedtools intersect -a Oryza_helitron_reference.gff3 -b helitronscanner.gff3 -
f 0.8 > ref_Helitron_HelitronScanner_overlap.gff3
```

Where parameters -a corresponds to annotation ‘A’, -b corresponds to annotation ‘B’, and -f is the required sequence identity of overlap between annotation A and annotation B, where this is set at .0.8 (80%) sequence identity.

Repbox Development

The Repbox pipeline is designed to run three *de novo* software packages; EAHelitron, SINE_Scan, and MITEFinder as well as RepeatModeler (version 2.0.1) **[102]**. All corresponding output fasta files from these software packages are then clustered using VSEARCH (version 2.14

) **[79]**, to remove redundant sequences. Commands for this process are outlined in Code 7, where the -cluster option is the method of clustering used by VSEARCH, and -–id is the percentage of sequence similarity of 0.8 or 80%.

Following clustering, filtration of clustered sequences for false positives and protein-coding sequences was performed. False positives are expected when using *de novo* methods, and so to account for this, we used a protocol outlined by *Berriman et al (2017)* **[120]** as a method to remove these sequences from our consensus annotation. This protocol relies on known information, such as genes that encode for proteins and other highly conserved sequences identified across plant species, that is then used to filter out sequences corresponding to these regions but were identified as potential transposable elements. With this justification in mind, we then took our consensus fasta derived from VSEARCH and filter it for coding regions, performing a BLAST (version 2.1.0+) **[76]** analysis against any available reference coding sequences (CDS) to remove of known coding sequences. For *Oryza sativa* and *Arabidopsis thaliana*, these sequences were derived from gene annotations files available on Ensembl, and this analysis is performed using commands outlined in Code 6.

Code 6, Commands for Clustering and Filtration of Protein-coding Genes

```
# Command 1: Clustering of concatenated sequences
vsearch -cluster_fast merged-library.sorted.fa --id 0.80

# Command 2: BLAST alignment of CDS to clustered sequences with an e-value cutoff of 1e-5
blastp -query $PROT -db ~/Libraries/RepeatPeps.lib -
outfmt '6 qseqid staxids bitscore std sscinames sskingdoms stitle' -max_target_seqs 25 -
culling_limit 2 -num_threads $THREAD -evalue 1e-5 -
out proteins.fa.vs.RepeatPeps.25cul2.1e5.blastp.out

# Command 3: Removal of TE sequences from proteome
perl ./fastaqual_select.pl -f $FASTA -
e <(awk '{print $1}' proteins.fa.vs.RepeatPeps.25cul2.1e5.blastp.out | sort | uniq) > transcri
pts.no_tes.fa

# Commands 4 & 5:Creation of a database consisting of non-TE transcripts, and BLAST
# analysis of the non-TE proteome to the consensus library
makeblastdb -in transcripts.no_tes.fa -dbtype nucl
blastn -task megablast -query $FASTA -db transcripts.no_tes.fa -
outfmt '6 qseqid staxids bitscore std sscinames sskingdoms stitle' -max_target_seqs 25 -
culling_limit 2 -num_threads 48 -evalue 1e-10 -out rebox_lib.transcripts.no_tes.out

# Command 6: Removal of hits from consensus library
perl $REPBOX_PREFIX/util/fastaqual_select.pl -f $FASTA -
e <(awk '{print $1}' rebox_lib.transcripts.no_tes.out | sort | uniq) > $FASTA.filtered.fa
```

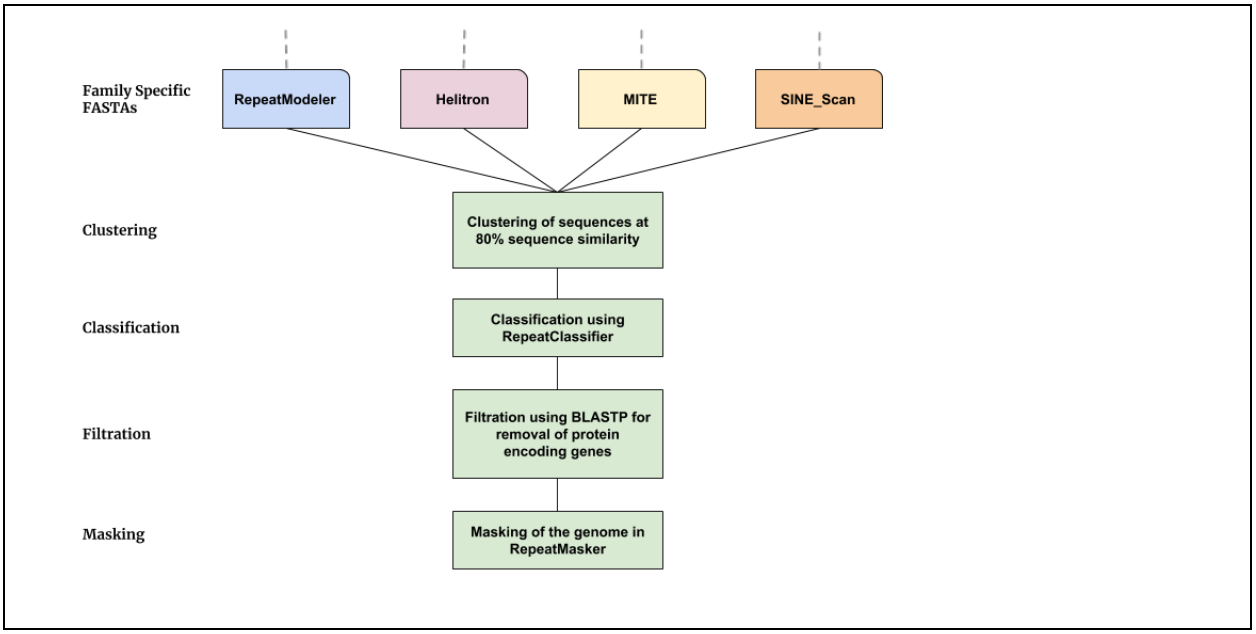
The resulting consensus fasta file, now filtered of redundant sequences and non-TE protein coding regions is then used as the custom repeat library input for RepeatMasker version 2.0.1 analysis. The commands used are detailed below in Code 8,where -pa is the CPU threads for the analysis, -e is the engine used for querying sequences in our library, -lib is the custom repeat library generated by our *de novo* tools, -gff is the output option to have RepeatMasker provide a gene annotation file (.gff3) as output, -dir is the output directory, -u is the original genome or reference.

Code 7, Commands used in Masking of Repeats

```
# RepeatMasker commands for masking of repeats
RepeatMasker -pa $THREAD -e ncbi -lib $LIBRARY -gff -dir $OUTPUT -u $GENOME
```

Requirements for implementing Repbox include (1) a genome in either .fasta, .fa, .fna, or .fas formats, (2) the Repbox repository download (available on GitHub; See *Appendix A*, and (3) Homebrew/Linuxbrew (https://docs.brew.sh), for ease of installation and required dependencies. Following the installation of all prerequisites, the user can then run the provided install script located within the Repbox GitHub repository to set up working directories and any remaining dependencies required for proper function of the pipeline. The general workflow of the pipeline is detailed below in (*Figure 17).*

Figure 17, Completion of Repbox Pipeline



Results and Discussion

Our pipeline demonstrates that the incorporation of newer family specific tools are capable of not only identifying additional TE candidates but also increase the diversity of elements that were

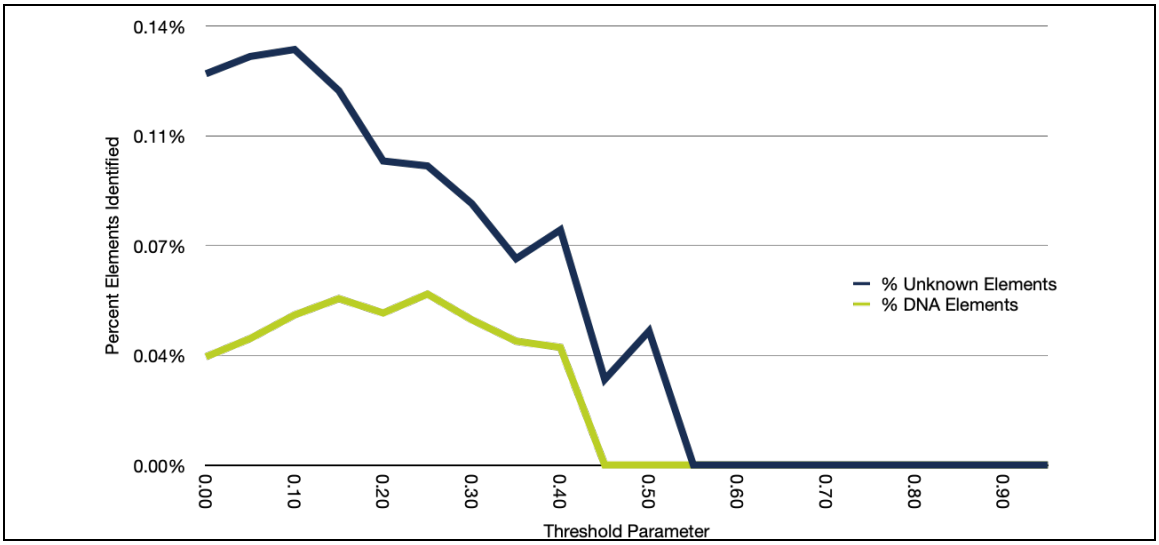
previously unobserved in our benchmarking genomes. Incorporating *de novo* identification methods for specific repeat element types also reveals a potential improvement to the overall repeat annotation process, opening the door for improved genome annotation and genome characterization. Benchmarking of the pipeline was conducted by an initial analysis of *Arabidopsis thaliana*, a very well annotated dicot genome, and expanded to include *Oryza sativa*, a well-characterized and highly repetitive monocot genome. Benchmarking was only performed with software that would improve the current repeat annotation capabilities of RepeatModeler and RepeatMasker. Recently, RepeatModeler (version 2.0.1) incorporated LTR_retriever (version 2.8.7) **[94]** and LTRHarvest (version 1.61) **[93]** as part of their code allowing users to perform additional annotation of LTR elements. These same tools were originally slated for benchmarking within our pipeline, but due to this added functionality to RepeatModeler, there was no need of further testing. As a result of these improvements, the remaining software packages, EAHelitron, HelitronScanner, MITE-Tracker, MITE-Finder, and SINE_Scan were advanced to initial benchmarking and analysis.

Parameterization of Software

In benchmarking of bioinformatics software, among the most time-intensive analysis is the parameterization of options within the software. There is also the added need to balance the efficiency or CPU time without sacrificing the accuracy of analysis by a given software package. Typically, in the initial phases of software development, authors empirically determine the appropriate default parameters for analysis by performing parameter sweeps to derive some optimized parameter setting, and for our analysis, two software packages, MITETracker and MITEFinder, were assessed with default parameter settings. For MITETracker and MITEFinder, only one (MITEFinder) provided parameter defaults recommended by the authors. MITETracker

did not possess editable parameters, and changes required substantial editing to the source code, and therefore, as suggested by the authors, recommended defaults were used in our pipeline. However for MITEFinder, a parameter sweep analysis in Arabidopsis was performed to assess the author suggested default threshold. *Hu et al* (2019) [96] determined the threshold value of MITEFinder to be 0.0, however analysis using this threshold resulted in large proportions of unclassified/unknown elements, prompting us to perform a parameter sweep to establish a more conservative threshold that balanced the count of identified MITE elements with unknown sequences (Code 8). The results of sweeping the threshold value from 0.1-0.95 are illustrated in (Figure 18). We observed large losses in MITE candidates as the threshold was increased. Based on the results of our sweep, element candidacy of MITEFinder determined a threshold of 0.2-0.3 was ideal in balancing unknown to MITE identified sequences and that is was used within our pipeline. It is important to note that if the user desires to do so, the MITEFinder parameters are mutable such as in an instance where prior knowledge about these MITE elements surpasses the defaults, and the user desires to update the threshold.

Figure 18, Parameter Sweep of MITEFinder



Code 8, UNIX commands for parameter Sweep of MITE Detection Software

```
# Parameter sweep computes MITE elements identified using a threshold ranging
# from 0.1 and 0.95.

MITEFINDER= ~/bin/miteFinder/miteFinder
for i in `seq 0.0 +0.05 1.00`; do $MITEFINDER -input $GENOME -output
$INDEXNAME.THRESHOLD_$i.mite_finder.out -pattern_scoring
$REPBOX_PREFIX/bin/miteFinder/profile/pattern_scoring.txt -threshold $i
done
```

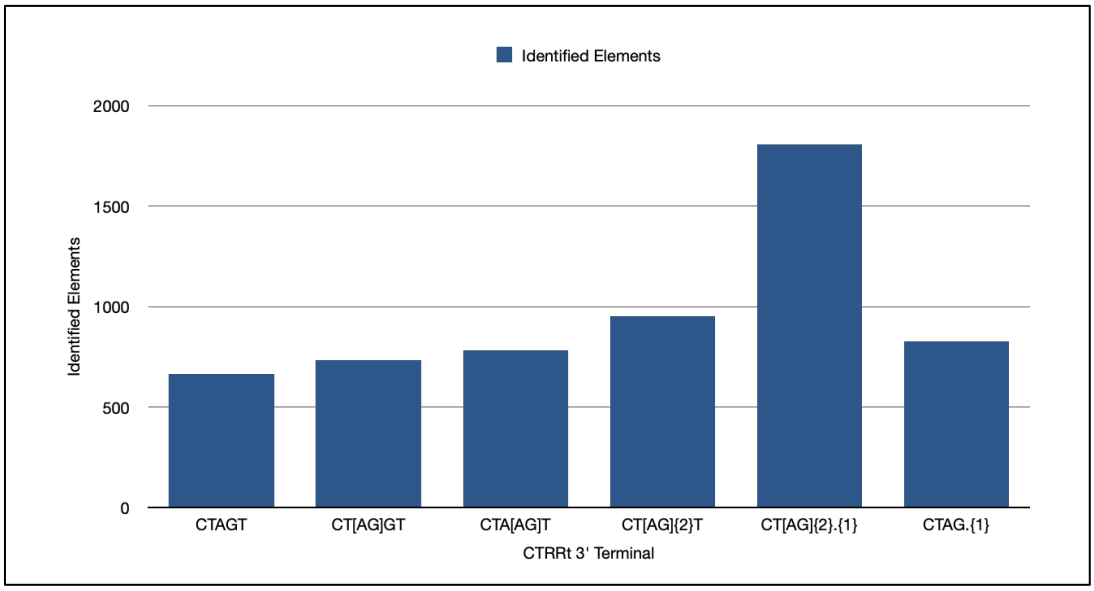
Parameterization of HelitronScanner was determined by provided training data of the head and tail (5’ and 3’) regions observed in groups of helitrons. Details of the process undertaken by the authors to derive these defaults are discussed in Chapter 1. However, to briefly review their process, the authors analyzed head and tail helitron data from various species known to possess these elements, and created a model that predicts the probability of a potential candidate being a helitron element **[97]**. Additional parameters include segmentation (-bs), which remained 0 to consider the entire chromosome and -ht or -tt parameters that defines “fuzziness” or allowed variation in the head and tail helitron regions. Parameters for benchmarking of HelitronScanner are detailed below in Code 9, where option -lf is the training set of motifs commonly found

up/downstream of a helitron, -g is the genome or fasta sequence for analysis, -bs is the number of segmentations of the genome when performing analysis (default: 0), -o is the output filename, and -tl is the number of CPU threads chosen for analysis (default: 1). The training data for the head and tail regions were derived from the author and were not altered in final iterations of Repbox, but observations of modulating the fuzziness of -ht and -tt parameter shows a tendency for HelitronScanner to derive an increase in number of predicted elements. However, increased predictions but did not improve RepeatClassifier's ability to classify element candidates, ultimately leading the decision to perform our analysis using default parameters. These parameters are below in (Table 2).

Table 2, Parameterization of HelitronScanner

Fuzziness Setting	CTRRt 3' Terminal	Identified Elements	Classified Elements
0	*CTAGT*	*665*	*0*
1	*CT[AG]GT*	*732*	*0*
2	*CTA[AG]T*	*782*	*0*
3	*CT[AG]{2}T*	*950*	*0*
4	*CT[AG]{2}.{1}*	*1808*	*0*
5	*CTAG.{1}*	*828*	*0*

Figure 19, Parameterization of HelitronScanner



Code 9, Parameterization of HelitronScanner

```
GENOME= ~/Arabidopsis_thaliana_genome.fa
INDEXNAME=$(basename $GENOME | cut -f 1 -d '.')
EAHELITRON="perl ./HelitronScanner"
THREAD=8

### Running HelitronScanner - (1)scanHead, (2)scanTail, (3) pairends, (4) draw Fasta output
HelitronScanner="java -jar $REPBOX_PREFIX/bin/HelitronScanner/HelitronScanner.jar"

## Identify upstream helitron sequences based on homology to trained helitron flanking
regions.
$HelitronScanner scanHead -lf $REPBOX_PREFIX/bin/HelitronScanner/TrainingSet/head.lcvs -g
$GENOME -bs 0 -o $INDEXNAME.head -tl $THREAD

## Identify downstream helitron sequences based on homology to trained helitron flanking
regions.
$HelitronScanner scanTail -lf $REPBOX_PREFIX/bin/HelitronScanner/TrainingSet/tail.lcvs -g
$GENOME -bs 0 -o $INDEXNAME.tail -tl $THREAD

## Pairs helitron ends
$HelitronScanner pairends -hs $INDEXNAME.head -ts $INDEXNAME.tail -hlr 200:20000 -o
$INDEXNAME.paired -lcv_filepath paired.log

## Create the fasta sequences for each helitrons
$HelitronScanner draw -p $INDEXNAME.paired -g $GENOME -o helitronscanner_out.$INDEXNAME -
pure_helitron
```

MITE Benchmarking

Analysis of the *Oryza sativa* reference repeat annotation reveals DNA class elements account for approximately 27% of the transposable elements present within the *O. sativa* genome, with

~12% or 35,813 elements generally categorized as MITEs, representing the largest proportion of DNA elements in *O. sativa*. In *Arabidopsis thaliana*, DNA class elements account for ~32% of the genome or 10,184 elements. MITE elements derived from the reference repeat annotation in *A. thaliana* are not reported directly, but the repeat annotation reports superfamilies that MITE elements derived from the broad label of MITE. The superfamilies that MITEs derive from are the following classes, *Tc1*/*Mariner/Pogo* and *Harbinger/HAT/MuDR*, and account for ~22% of the genome of *Arabidopsis thaliana* at 7,414 total elements.

Comparative analysis of the *O. sativa* reference annotation to those generated by MITETracker and MITEFinder show significant differences in counts between the packages, with 17,700 elements identified by MITETracker and 40,814 identified by MITEFinder. By direct count, MITEFinder is the most comparable to the reference MITE count, identifying 5,000 additional MITE candidates than annotated within the reference. MITETracker missed approximately 18,113 elements that were identified from the reference. Analysis in *A. thaliana* reveals similar patterns in the count of MITE element candidates observed, with MITEFinder identifying 18,576 MITE element candidates and MITETracker identifying 230 MITE element candidates. On simple count analysis alone, MITEFinder clearly outperforms MITETracker for identification of MITE candidates.

To further assess the quality of MITE predictions, the location of MITE elements identified from MITEFinder and MITETracker are overlaid with respect to the reference MITE locations. In overlaps alone, MITE elements identified by MITETracker and MITEFinder overlapped with the *O. sativa* reference annotation at 19,018 sites (53.10%) and 24,028 sites (67.09%), while overlapping with the *A. thaliana* reference at 265 sites (2.60%) and 9,796 sites (96.10%) within the reference (Tables 2 and 3). Given these results, it is clear that MITEFinder is identifying far

more potential elements within each genome in comparison to MITETracker. At 80% sequence identity between elements, i.e. at least one of the overlaps, MITETracker/MITEFinder or reference sequence, must share at least 80% sequence identity with overlapping sequences. With this restriction, MITETracker overlapped at 17,094 sites against the *O. sativa* reference, while MITEFinder overlapped at 5,918 sites against the reference. We again observed similar patterns of overlap in *A. thaliana*, with MITETracker overlapping the reference at 126 sites and MITEFinder overlapping at 2,032 sites. A visual summary using chomPlot packages in R illustrate these results by karyotypes of MITEFinder, MITETracker and reference overlaps in *O. sativa* and *A. thaliana* is shown in Figure 3. It is interesting to note that MITEFinder identified a higher number of elements on Chromosome 1 and very few on chromosomes 2 and 3 in *O. sativa.* In contrast, the distribution of MITEs by each software package is relatively consistent across the Arabidopsis genome.

Figure 20, Overlap Distribution of MITEs in O. sativa and A. thaliana

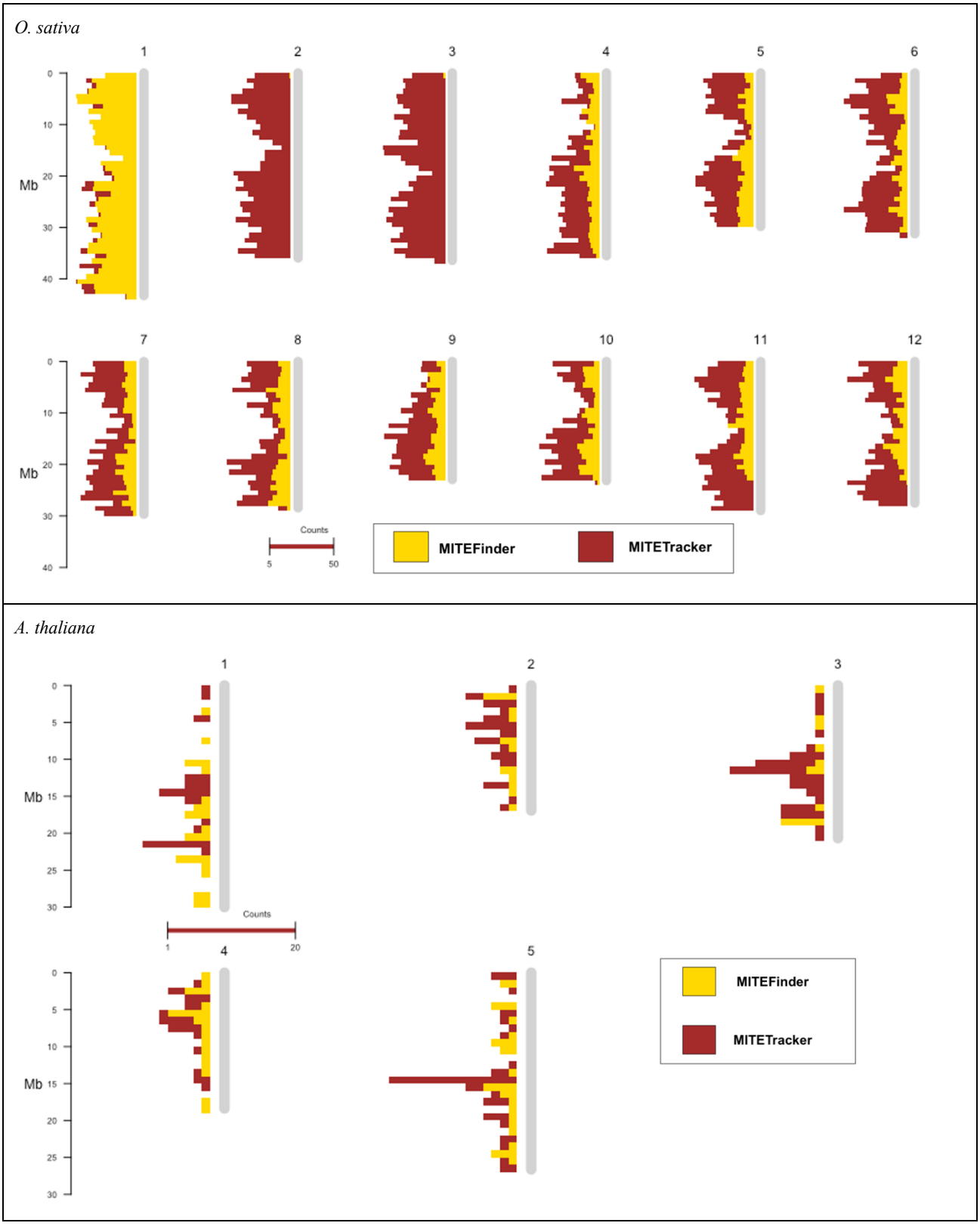


Table 3, MITE Analysis of Oryza sativa

Software	Total MITE Count	Counts of overlaps w/Reference	Overlaps >= 80% Sequence Identity	% of Sequences Overlap with Ref.	Average overlap length (bps)
MITETracker	*17,700*	*19,018*	*17,094*	*53.10%*	*249.9713*
MITEFinder	*40,814*	*24,028*	*5,918*	*67.09%*	*203.5359*

Table 4, MITE Analysis of Arabidopsis thaliana

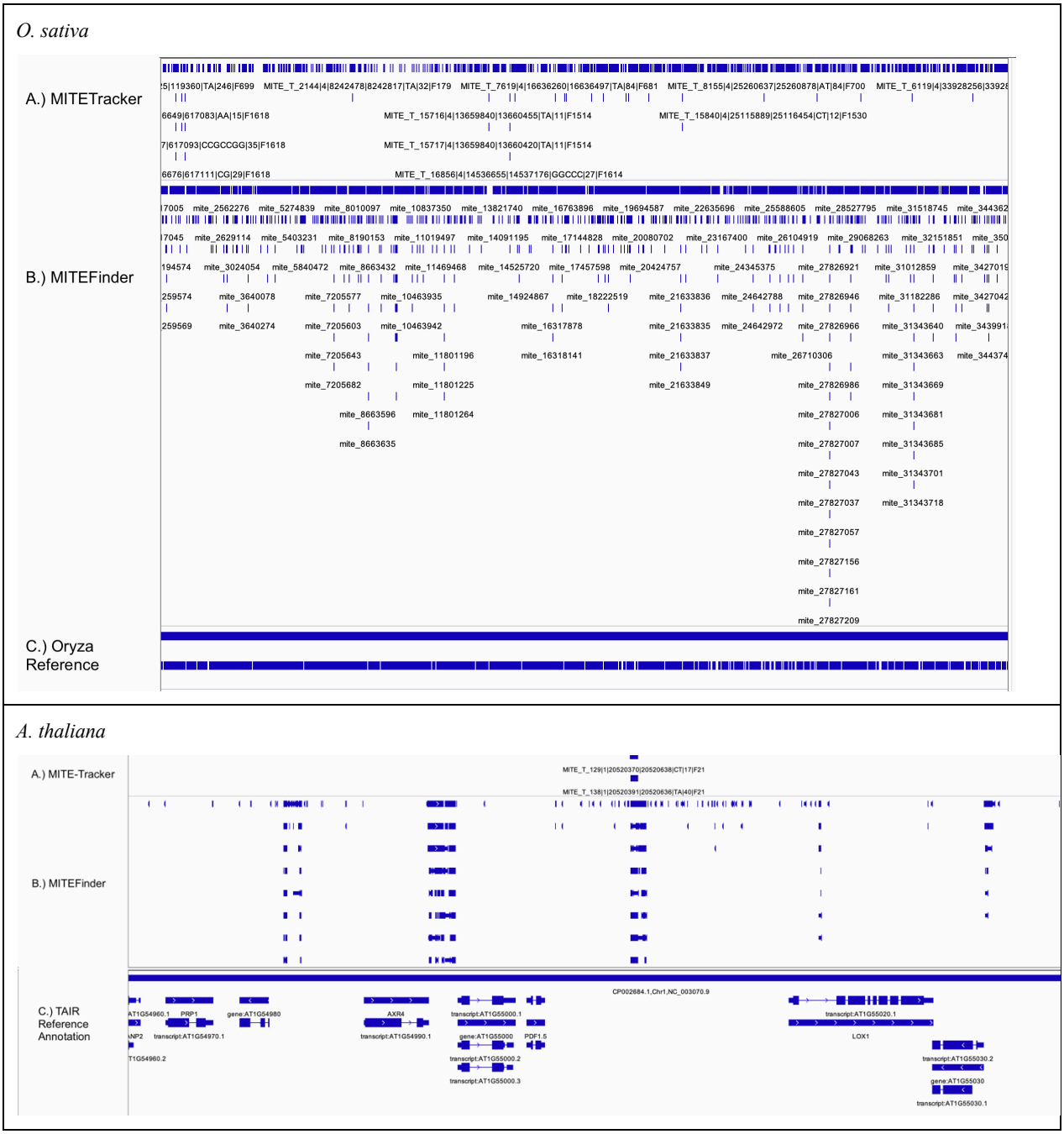
Software	Total MITE Count	Counts of overlaps w/Reference	Overlaps >= 80% Sequence Identity	% of Sequences Overlap with Ref.	Average overlap length (bps)
MITETracker	*230*	*265*	*126*	*2.60%*	*382.8889*
MITEFinder	*18,576*	*9,796*	*2,032*	*96.19%*	*253.9972*

With the 80% percent identity restriction, annotations derived from MITEFinder with an overlap to the reference genome significantly dropped, from 24,028 to 5,918 elements in *O. sativa* and from 9,796 to 2,032 elements in *A. thaliana* (Tables 2 and 3). When considering the overlap length of elements identified by each package, the average overlap length of MITEFinder is significantly smaller than that of MITETracker, a value of approximately 203bps versus 250bps in *O. sativa*, and approximately 254bps versus 383bps in *A. thaliana*. Shorter overlapping sequences will have less sequence identify to the reference, especially if the reference sequences are substantially longer. This characteristic is further compounded by the number of annotation entries generated by MITEFinder, resulting in the drastic drop observed in MITE elements with application of the 80% identity.

To visually compare the different elements found with MITEFinder and MITETracker, the coordinate locations were loaded into Integrative Genomics Viewer (IGV) browser **[121]** and are available at Appendix B. Figure 21 provides a snapshot of two genomic regions with both the

MITEFinder, MITETracker and reference annotations displayed. Clearly there are instances of overlapping regions, as well as distinct patterns and tendencies of each MITE identification method. We observed a pattern of fragmentation is more pronounced in MITETracker (A), with smaller sequence start and end coordinates for MITEs assigned. For MITEFinder, MITE annotations are significantly larger in range, with interspersed regions consistently flanking highly high coverage entries from MITEFinder's iterative assignment method.

Figure 21, IGV MITE Feature Comparison of O. sativa and A. thaliana



Helitron Benchmarking

Analysis of reference transposon annotations reveal that Helitron elements account for approximately 0.3% of the *Oryza sativa* genome and 32% of the *Arabidopsis thaliana* genome,

corresponding to a count of 2,945 and 746 respectively. This is consistent with information available in TAIR10 and RAP-DB databases where both annotations were derived. Analysis of *O. sativa* and *A. thaliana* reveal differences in the detected abundance of Helitron candidates by both software packages. From our analysis, EAHelitron detected 3,316 potential helitron elements in *O. sativa* and 665 in *A. thaliana*, whereas HelitronScanner detected 3,446 helitron candidate elements in *O. sativa* and 441 elements in *A. thaliana* (Tables 4 [1]and 5). When considering the overlap of the EAHelitron and HelitronScanner annotations with respect to the reference annotation, EAHelitron has an overlap count of 2,022 sites (271.05%) and 604 sites (4.67%) for *O. sativa* and *A. thaliana* respectively, and HelitronScanner deriving an overlap count of 22,252 sites (3206.7%) and 1,536 sites (14.23%) for *O. sativa* and *A. thaliana* respectively. Percentages over 100%, such as those observed in O. sativa, correspond to multiple instances of gene features that align at multiple sites with the reference. Similarly, large counts of overlaps from HelitronScanner are the result of overlaps between multiple entries in the HelitronScanner generated annotation and the reference annotation. When using the bedtools intersect command, A intersect B, the default is to report all intersections for all annotation entries. The annotation outputs for HelitronScanner reports multiple features for the same region, and so the result is multiple overlaps being reported for the same region. This can cause the counts of overlaps to be higher than expected. Adjustments in bedtools parameters include Increasing the overlap restriction to both -a and -b comparisons and restricting the reporting of elements that fail to meet the 80% identity with respect to the reference or -a. This parameter adjustment requires that the fraction overlap be reciprocal for A and B or in other words, if -f is 0.80 and -r is used, this requires that B overlap 80% of A and A also overlaps 80% of B. This greatly reduces the high count of Helitron elements in HelitronScanner by removal of smaller overlapping regions.

Adjusting the analysis to this restriction drastically decreases detected elements, to 207 and 60 elements in *O. sativa* and *A. thaliana* respectively.

Table 5, Helitron Analysis of Oryza sativa

Software	Total Helitron Count	Counts of overlaps w/Reference	Overlaps >= 80% Sequence Identity	% of Sequences Overlap with Ref.	Average overlap length (bps)
EAHelitron	*3,316*	*2,022*	*29*	*271.05%*	*24.93719*
HelitronScanner	*3,447*	*23,922*	*22,252*	*3206.70%*	*469.1358*

Table 6, Helitron Analysis of Arabidopsis thaliana

Software	Total Helitron Count	Counts of overlaps w/Reference	Overlaps >= 80% Sequence Identity	% of Sequences Overlap with Ref.	Average overlap length (bps)
EAHelitron	*665*	*604*	*11*	*4.67%*	*26.4404*
HelitronScanner	*441*	*1,842*	*1,536*	*14.23%*	*731.9224*

Visual comparisons of Helitron detection and overlaps using chromPlot R-package **[122]** and IGV **[121]** (Figures 22 and 23) illustrate differences in coordinate data and counts of elements between different software packages. In Figure 22, the observed distribution of helitron elements shows a pattern of dispersion of Helitron elements being fairly consistent in *O. sativa* and *A. thaliana*, as helitron elements are observed in every chromosome in both organisms. However, there are clearly more helitrons annotated across the genome with HelitronScanner. From visual comparisons in IGV, HelitronScanner is able to derive expanded annotations of helitron elements, including not only the regions corresponding to the TAIR10 Reference annotation of a transposase coding region (AT1G35470.1,AT1G35470.2 proteins), but the remaining structure of the helitron that extended beyond the coding region of transpose. This is potentially a more in-depth annotation as the transposon length reported within HelitronScanner spanning gapped regions in the reference annotation. This finding suggests that additional Helitron annotation can

potentially address fragmentation issues commonly encountered during annotation **[123]**. Greater detail into the overlapping regions, such as coordinate information, between EAHelitron and HelitronScanner are available in Appendix B.

Figure 22, Overlap Distribution of Helitrons in O. sativa and A. thaliana

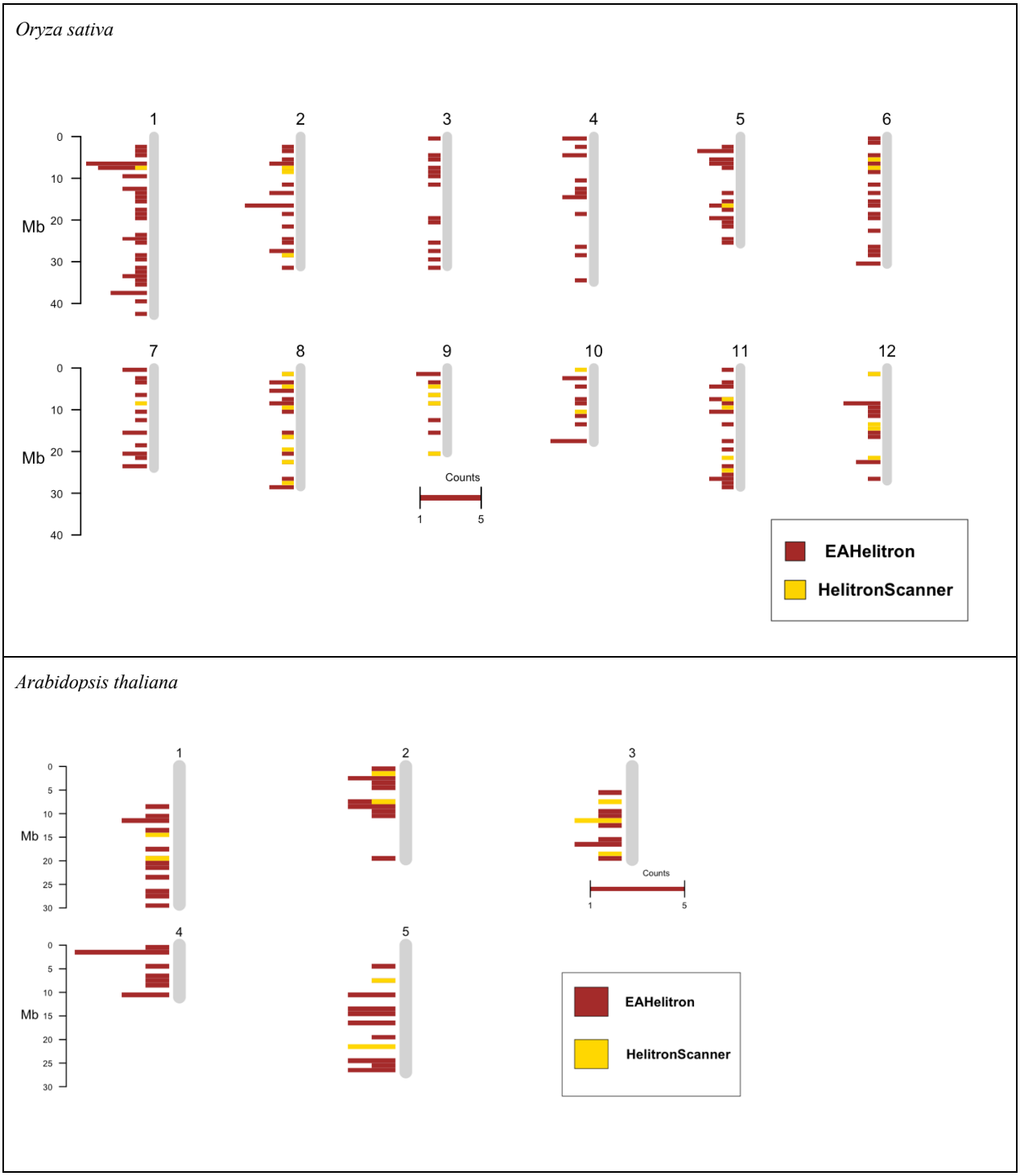


Figure 23, IGV Helitron Feature Comparison of O. sativa and A. thaliana

SINE Identification

SINE_Scan (version 1.1.1) **[111]** was tested for inclusion in Repbox to increase the diversity of SINE elements. No other SINE-specific software packages were available for comparison. Analysis of reference annotations revealed SINE elements represent approximately 2.1% (6,012 total elements) and 0.38% (131 total elements) of *Oryza sativa* and *Arabidopsis thaliana* respectively. Results of SINE_Scan analysis outlined (Table 7) reveal low counts of identified SINE elements for *O. sativa* and *A. thaliana* respectively, with most identified SINE candidates being unknown or unclassified. In total, 3,968 sequences were identified in *O. sativ*a, falling far short of the reference by ~2000 elements. However, in *A. thaliana*, 1,782 sequences were identified potentially increasing SINE candidates by ~1600 elements. Diversity in SINE families

was most apparent in *O. sativa* with 57 distinct families of SINE elements being observed, a majority of which 'SINE03_OS' family was most abundant. In *A. thaliana*, only two SINE families were observed, ATSINE2A and ATSINE4, with ATSINE4 being the most abundant.

Table 7, SINE_Scan Analysis of O. sativa and A. thaliana

O. sativa		*A. thaliana*	
Family	*Count*	*Family*	*Count*
SINE	*1*	*SINE*	*2*
SINE/ID	*2*	*SINE/5S-Deu-L2*	*14*
Non-SINE Elements	*406*	*Non-SINE Elements*	*248*
Unknown	*3559*	*Unknown*	*1518*
Total	*3968*	*Total*	*1782*

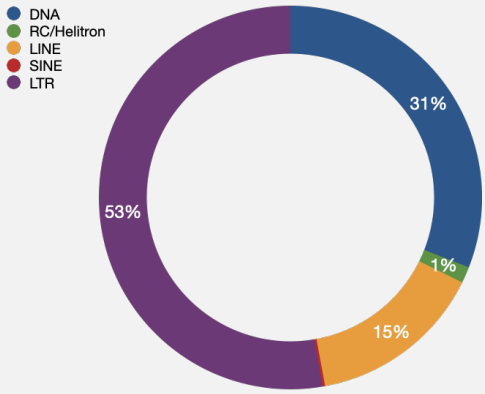
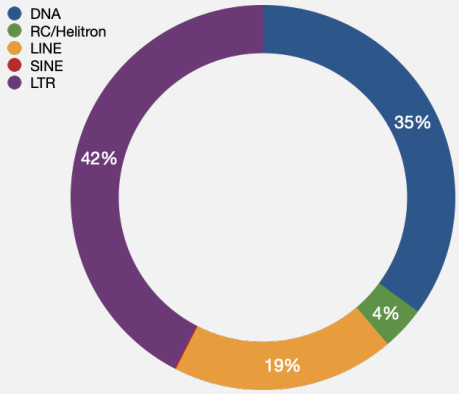
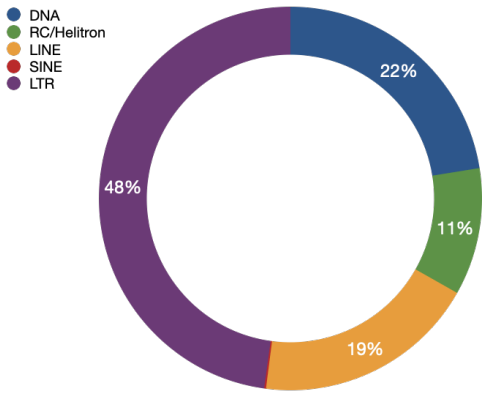
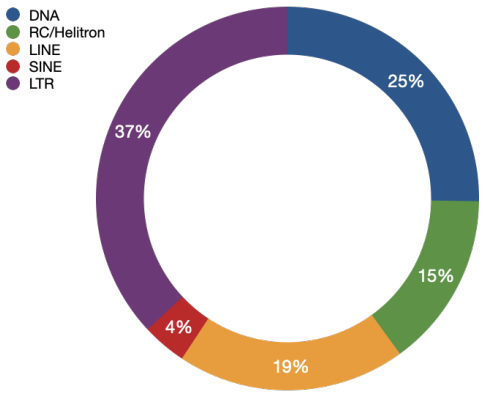
Repbox and Repeat Modeler/RepeatMasker Comparison

Benchmarking of MITE and Helitron (DNA element) software packages using *Oryza sativa* and *Arabidopsis thaliana* as references revealed an overall increase in repetitive element family diversity and interspersed repeats. As such, we chose to incorporate MITEFinder and HelitronScanner as well as SINE_Scan into the Repbox pipeline. The full Repbox pipeline begins with identification of repeat families using RepeatModeler v2.0.1, followed by de novo identification of MITE, Helitron and SINE elements by MITEFinder, HelitronScanner and SINE_Scan. This is then concluded with masking using RepeatMasker v4.1.0 . Analysis using the Repbox pipeline and traditional RepeatModeler RepeatMasker analysis was performed on both *Oryza sativa* and *Arabidopsis thaliana*. Repbox produced an increase in the raw counts of DNA, LTR, non-LTR and Helitron classes of transposable elements. In addition to increases in count, the overall diversity of repeat families is expanded. DNA transposons increased approximately 4% in *A. thaliana* and 3% in *O. sativa*. LTR elements decreased in the overall proportion of

identified repeats, not count, by approximately 9% in both organisms. This shift is illustrated below in (Figure 24).

Increases to the number of other elements, including rRNA, satellite, simple repeat, sRNA, tRNA and unknown, were also observed in both organisms from our Repbox analysis, increasing in total from 448,316 to 567,021 in *Oryza sativa* and 65,483 to 94,564 in *Arabidopsis thaliana*. This represents a 3% and 2% increase in the proportion of the genome identified as unknown for *O. sativa* and *A. thaliana* and respectively. Other notable changes to TE family composition includes increases in LINE and SINE elements in *A. thaliana* and a modest increase of SINE elements in *O. sativa*. A total of 1445 LINE elements were identified by RepeatModeler/RepeatMasker, however Repbox was able to increase this count of LINE elements to 2,844, an increase of approximately ~1400 LINE element candidates. SINE element candidates were also increased in *A. thaliana*, with an increase of 551 SINE candidates. SINE elements in *O. sativa* are modestly increased with the Repbox pipeline from 160 to 178 SINE elements identified. This modest increase by 18 additional elements falls short of the reference's count of 6,012 SINE elements and outlines the importance of manual curation **[124]**.

Figure 24, TE Family Percentage of Genome for O sativa and A. thaliana

	RepeatModeler/RepatMasker	Repbox
O. sativa		
A. thaliana		

Clustering is an essential part of Repbox and general *de novo* repeat annotation. VSEARCH clusters sequences in phases, first constructing a database of sequences similar to the query *k-mer* sequences, requiring a minimum number of consecutive nucleotides that match or overlap, followed by optimal alignments of sequences that possesses the highest number of identical *k-mers* to those within the database. Sequences that possess similarity equal or greater than the value specified by the user in the –id parameter are then accepted. Increasing the percentage of sequence identity to form a smaller but more distinct cluster is possible, however we chose to maintain 80% similarity to balance family-specific sequences and avoid excessive singleton clusters.

We also observed a shift in the composition of repeats in both *A. thaliana* and *O. sativa* in addition to the overall increase in count of elements. A majority of transposable element families are represented in both analysis pathways, but the proportions of each classes vary (Tables 8,9). Elements designated as “Other” and “Unknown” also demonstrated substantial increases in the Repbox output, however, this is likely due to a lack of homology or representation of elements in reference databases. Classification of elements relies heavily on sequences that share homology. If a TE candidate is novel or degenerate, the potential for this element to be accurately classified is much lower **[75]**. The consensus repeat library generated within the Repbox pipeline contains elements that have been identified by *de novo* detection methods, and these sequences are potentially underrepresented by TE databases such as Repbase. The result of this is many elements are classified as Unknown. Clearly manual curation of these elements would increase sequence representation within TE databases, however this is outside the scope of our analysis.

DNA elements saw the largest increase, almost doubling in count in *A. thaliana* and *O. sativa*. This is likely due to the general structure of MITE elements as well as the method by

which they are detected. MITE elements are among most common DNA elements typically found in a given genome **[107]**, and by nature of the small size of MITE elements, they are generally more abundant across the genome. As such, there is a potential bias toward DNA elements in the repeat annotation process, and this can generate an abundance of "noise" where sequences that are similar to those observed in MITEs could prevent the detection of other elements **[125]** . In comparison to the references both RepeatMasker and Repbox fall short of DNA elements identified in O. sativa and A. thaliana, which through added manual curation derives MITE element counts of 10,184 and 7,414 in each respectively. However again, this is promising as Repbox was capable of identifying a greater count of these elements, reducing the need for additional manual curation although not entirely removing it.

Table 8, RepeatModeler/Masker & Repbox Analysis of Arabidopsis thaliana

RepeatMasker	***Family***	***Count***	***Total***	***Repbox***	***Family***	***Count***	***Total***	***Reference Total***
DNA	*DNA*	*5*	*1722*	*DNA*	*DNA*	*19*	*3721*	*10184*
	DNA/CMC-Chapaev	*2*			*DNA/CMC-Chapaev*	*0*		
	DNA/CMC-EnSpm	*540*			*DNA/CMC-EnSpm*	*1323*		
	DNA/CMC-Transib	*0*			*DNA/CMC-Transib*	*15*		
	DNA/Dada	*5*			*DNA/Dada*	*9*		
	DNA/hAT	*1*			*DNA/hAT*	*0*		
	DNA/hAT-Ac	*172*			*DNA/hAT-Ac*	*113*		
	DNA/hAT-Charlie	*1*			*DNA/hAT-Charlie*	*10*		
	DNA/hAT-Tag1	*0*			*DNA/hAT-Tag1*	*8*		
	DNA/hAT-Tip100	*97*			*DNA/hAT-Tip100*	*169*		
	DNA/IS3EU	*0*			*DNA/IS3EU*	*18*		
	DNA/Kolobok-T2	*5*			*DNA/Kolobok-T2*	*0*		
	DNA/Merlin	*0*			*DNA/Merlin*	*3*		
	DNA/MULE-MuDR	*782*			*DNA/MULE-MuDR*	*1454*		
	DNA/P	*0*			*DNA/P*	*0*		
	DNA/PiggyBac	*1*			*DNA/PiggyBac*	*2*		
	DNA/PIF-Harbinger	*97*			*DNA/PIF-Harbinger*	*480*		
	DNA/TcMar-ISRm11	*0*			*DNA/TcMar-ISRm11*	*0*		
	DNA/TcMar-Pogo	*12*			*DNA/TcMar-Pogo*	*87*		
	DNA/TcMar-Stowaway	*0*			*DNA/TcMar-Stowaway*	*9*		
	DNA/TcMar-Tc1	*0*			*DNA/TcMar-Tc1*	*0*		
	DNA/TcMar-Tc2	*1*			*DNA/TcMar-Tc2*	*0*		

	DNA/Zisupton	*1*			*DNA/Zisupton*	*2*		
RC/Helitron	*RC/Helitron*	*806*	*823*	*RC/Helitron*	*RC/Helitron*	*2182*	*2182*	*12945*
	Helitron-2	*17*			*Helitron-2*	*0*		
LINE	*LINE/I*	*2*	*1445*	*LINE*	*LINE/I*	*0*	*2844*	*1447*
	LINE/I-Jockey	*1*			*LINE/I-Jockey*	*0*		
	LINE/L1	*1429*			*LINE/L1*	*2842*		
	LINE/L1-Tx1	*5*			*LINE/L1-Tx1*	*1*		
	LINE/L2	*2*			*LINE/L2*	*0*		
	LINE/Penelope	*4*			*LINE/Penelope*	*1*		
	LINE/R1	*1*			*LINE/R1*	*0*		
	LINE/RTE-X	*1*			*LINE/RTE-X*	*0*		
SINE	*SINE*	*2*	*12*	*SINE*	*SINE*	*38*	*553*	*131*
	SINE/5S-Deu-L2	*0*			*SINE/5S-Deu-L2*	*501*		
	SINE/ID	*10*			*SINE/ID*	*14*		
LTR	*LTR*	*5*	*3665*	*LTR*	*LTR*	*3*	*5453*	*5962*
	LTR/Caulimovirus	*0*			*LTR/Caulimovirus*	*2*		
	LTR/Copia	*947*			*LTR/Copia*	*1403*		
	LTR/ERV1	*8*			*LTR/ERV1*	*13*		
	LTR/ERVK	*13*			*LTR/ERVK*	*107*		
	LTR/Gypsy	*2666*			*LTR/Gypsy*	*3892*		
	LTR/Ngaro	*13*			*LTR/Ngaro*	*0*		
	LTR/Pao	*13*			*LTR/Pao*	*33*		
Other	*rRNA*	*13*	*65483*	*Other*	*rRNA*	*102*	*94564*	*NA*
	Satellite	*19*			*Satellite*	*113*		
	Simple repeat	*0*			*Simple repeat*	*29832*		
	snRNA	*14*			*snRNA*	*48*		
	tRNA	*84*			*tRNA*	*542*		
	Unknown	*65353*			*Unknown*	*63927*		
	Total	*68234*			*Total*	*117561*		

Table 9, RepeatModeler/Masker & Repbox Analysis of Oryza sativa

RepeatMasker	***Family***	***Count***	***Total***	***Repbox***	***Family***	***Count***	***Total***	***Reference Total***
DNA	*DNA/CMC-EnSpm*	*9765*	*22126*	*DNA*	*DNA/CMC-EnSpm*	*26020*	*53773*	*76131*
	DNA/DNA	*2306*			*DNA/DNA*	*2844*		
	DNA/Ginger-1	*0*			*DNA/Ginger-1*	*203*		
	DNA/hAT	*0*			*DNA/hAT*	*226*		
	DNA/hAT-Ac	*2186*			*DNA/hAT-Ac*	*7262*		
	DNA/hAT-Charlie	*0*			*DNA/hAT-Charlie*	*720*		
	DNA/hAT-Tag1	*243*			*DNA/hAT-Tag1*	*490*		
	DNA/hAT-Tip100	*1296*			*DNA/hAT-Tip100*	*4461*		
	DNA/IS3EU	*814*			*DNA/IS3EU*	*358*		
	DNA/Kolobok-T2	*133*			*DNA/Kolobok-T2*	*1030*		
	DNA/Kolobok-H	*0*			*DNA/Kolobok-H*	*131*		
	DNA/Maverick	*0*			*DNA/Maverick*	*173*		
	DNA/Merlin	*821*			*DNA/Merlin*	*161*		
	DNA/MULE-MuDR	*3546*			*DNA/MULE-MuDR*	*5412*		
	DNA/P	*0*			*DNA/P*	*292*		
	DNA/PIF-Harbinger	*741*			*DNA/PIF-Harbinger*	*2337*		
	DNA/TcMar	*0*			*DNA/TcMar*	*157*		

	DNA/TcMar-ISRm11	*0*			*DNA/TcMar-ISRm11*	*132*		
	DNA/TcMar-Stowaway	*275*			*DNA/TcMar-Stowaway*	*1364*		
RC/Helitron	*RC/Helitron*	*979*	*979*	*RC/Helitron*	*RC/Helitron*	*4975*	*5635*	*764*
	Helitron-2	*0*			*Helitron-2*	*660*		
LINE	*LINE/I-Jockey*	*393*	*10587*	*LINE*	*LINE/I-Jockey*	*0*	*28711*	*4390*
	LINE/L1	*10194*			*LINE/L1*	*26389*		
	LINE/L1-Tx1	*0*			*LINE/L1-Tx1*	*1790*		
	LINE/L2	*0*			*LINE/L2*	*65*		
	LINE/Penelope	*0*			*LINE/Penelope*	*153*		
	LINE/RTE-BovB	*0*			*LINE/RTE-BovB*	*1*		
	LINE/RTE-X	*0*			*LINE/Rex-Babar*	*313*		
SINE	*SINE/ID*	*0*	*160*	*SINE*	*SINE/ID*	*24*	*178*	*6012*
	SINE/SINE	*160*			*SINE/SINE*	*154*		
LTR	*LTR/Caulimovirus*	*252*	*37680*	*LTR*	*LTR/Caulimovirus*	*135*	*64964*	*119007*
	LTR/Copia	*9401*			*LTR/Copia*	*17586*		
	LTR/ERV1	*0*			*LTR/ERV1*	*569*		
	LTR/ERVK	*1554*			*LTR/ERVK*	*1482*		
	LTR/ERVL	*0*			*LTR/ERVL*	*374*		
	LTR/Gypsy	*25939*			*LTR/Gypsy*	*39870*		
	LTR/LTR	*269*			*LTR/LTR*	*2171*		
	LTR/Ngaro	*265*			*LTR/Ngaro*	*1702*		
	LTR/Pao	*0*			*LTR/Pao*	*1075*		
Other	*Low_complexity*	*9963*	*448316*	*Other*	*Low_complexity*	*8219*	*567021*	*77763*
	rRNA	*697*			*rRNA*	*600*		
	Satellite	*54*			*Satellite*	*409*		
	Simple_repeat	*92147*			*Simple_repeat*	*80855*		
	snRNA	*54*			*snRNA*	*54*		
	tRNA	*318*			*tRNA*	*2475*		
	Unknown	*345083*			*Unknown*	*474409*		
	Total	*519888*			*Total*	*721557*		

Conclusions

Unknown elements are a common consequence of TE annotation, with the underlying cause of which ranging from novel TEs families to a general lack of TE family representation in prior works. A consequence of unknown elements are sequences that are clustered into the label of "unclassified", which unfortunately lessens the opportunity to further characterize repetitive elements in genomes of interest. As such, we were motivated to improve the process of TE characterization by attempting to further characterize these elements by implementation of our

RepBox pipeline, with the secondary goal of capturing previously unobserved = diversity in TEs. Proper identification and classification of these elements was an important step in the process of genome annotation, and complete characterization potentially provided necessary information of not only TEs in question, but an opportunity to better overall understanding of organisms as a whole. With the development of our pipeline, much of the annotation process remained similar to traditional *de novo* methods of identification, however our developed pipeline differs as it bolsters its analyses by reliance on dependencies beyond RepeatModeler, RECON, RepeatScout and TRF (Tandem Repeat Finder), and does this by incorporating modern software tailored to identify specific repeat families. As proof of concept, we performed benchmarking to compare and contrast family-specific software, and evaluated their effectiveness of identifying family-specific repetitive elements. We observed a significant increases to DNA, LTR, non-LTR and Helitron/RC elements, potentially implicating an increase to repeat diversity in *A. thaliana* and *O. sativa*. A caveat of our findings are the observed increases to unknown elements, however we feel that further optimization of our pipeline can resolve these increases. We feel that our pipeline has the potential to enrich our understanding of repeat characteristics in a given genome, and here we have demonstrated it is possible to derive detailed information that surpasses the traditionally utilized packages when annotating repetitive elements.

Chapter 3: ANALYSIS OF REPEATS IN AVENA DIPLOID GENOMES (A. ATLANTICA AA AND A. ERIANTHA CC)

Introduction

The genus *Avena* is a member of the Poeae Tribe of the Poaceae family, a diverse family of highly nutritive food crops including wheat, rice and oats (*Avena sativa L.*) **[126]**. Studies of oat have revealed it to possess many nutritive components such as beta-glucan found in soluble fiber and anti-inflammatory agents such as avenolic acid, a derivative of linoleic and omega-6 fatty acid **[127] [1]**. Further studies investigating additive nutrition led to the characterization of the oat genome, the culmination of which has been the genome structure, described as allohexaploid (2n=6x=42) with AACCDD subgenome composition **[128]**. Current research suggests that this allohexaploid was derived from hybridization between a CCDD allotetraploid and an A_SA_S diploid **[129]**. Confounding factors to our understanding of the evolutionary history of *Avena* is the presence of several variants of A-subgenome and C-subgenome diploids, as D-subgenome diploids have not yet been identified and are even suggested to be another A-subgenome sub-variant. Another added difficulty to studying *Avena* is the rather hefty size of the hexaploid oat genome, as it is quite large at approximately 13 Gb, and sub genome diploids ranging from 3-4 Gb in size. Obstacles aside, in efforts to advance our understanding of Avena, two high quality diploid genomes, *Avena eriantha* (CC) and *Avena atlantica* (AA) have recently been assembled and annotated **[130]**. Further study into these genomes provide an excellent resource to delve deeper into the *Avena* genus **[131]**, as investigations studying repetitive elements, and the role they potentially play in *Avena* evolution as a whole, is a step toward bringing greater insight and answering some of the evolutionary questions surrounding the origins of *Avena* and *Avena sativa L.*

In-depth characterization of repetitive elements in *Avena* species is sparse given the minimal genomic sequences available prior to the work of *Maughan et al (2019)* **[130]**. Prior to this, most studies relied on either sequence generation via BAC libraries, genotype-by-sequencing or low-pass shotgun sequencing data. *Solano et al* (*1992)* **[132]** sought to identify select tandem repeat sequences such as clone pAm1 (GenBank X83958) identified as a selectively hybridizing element found in the C-subgenome of *Avena murphyi L.,* an AACC tetraploid. This pAM1 clone was identified by RepeatModeler as containing sequence matching a highly homologous (E-value 2E-82) repeat found in *A. eriantha*, but is not found in the genome of *A. atlantica* **[132]**. This suggests that *A. eriantha* is more likely related to the progenitor genome that led to tetraploid *Avena murphyi.* Similarly, *Katsiotis et al* **[133]** reported the identification of interspersed repeat *pAvKB26* (GenBank AJ297385.1) that selectively hybridized to only the A- and D-subgenomes. This *pAvKB26* repeat was found in the unknown repeat fraction of *A. atlantica* but was missing in the *A. eriantha* genome**[130]**.

In more recent studies, such as those conducted by *Liu et al*, repeats were characterized from diploid A-genome species *A. hirtula*, *A. brevis*, *A. strigosa* and hexaploid species *A. sativa L*. using whole-genome shotgun sequencing with 2×250 bp paired-end sequence data **[131]**. These genomes were characterized using RepeatExplorer followed by similarity-based clustering of raw-read subsets at 50% overlap and 90% similarity **[131]**. From this study, *Liu et al* identified subgenome specific repetitive elements within A, C and D-subgenomes and further characterized of the A-subgenome element *pAs120a*, an element originally identified by *Linares et al* from the A-subgenome of *Avena sativa* **[134]**. The *pAs120a* element displays high similarity to Ogre/Tat and Chromovirus retrotransposons, and indications that this repeat potentially originated from Ogre/Tat elements but is a degraded form of these retroelements.

Liu's work also investigates C-subgenome specific elements, however it is somewhat limited as C-specific repetitive elements are provided solely by *A. sativa* subgenome-C. Additional C-genome repeat analyses is necessary to provide greater insight into other sub-genomes types in Avena. With Liu's expansion upon prior works into A-subgenome specific repetitive elements of Avena, this work represents an area possessing the potential of uncovering additional information surrounding repetitive elements that have historically been encountered in prior works but lack current characterization

Beyond the characterization of repetitive elements, such as those outlined above, prior studies of *Avena* have explored the phylogenetic relationships among *Avena* species. Studies such as those by *Fu et al*, have investigated the potential evolutionary relationships in various *Avena* species, utilizing chloroplast and mitochondrial DNA derived from 13 diploids including *A. eriantha* and *A. atlantica*, in conjunction with 7 tetraploid, and 5 hexaploid species. This project aimed to identify phylogenetic relationships between diploid *Avena* species and gain insight as to the origins of hexaploid *A. sativa L*. In the work of *Fu. et al (2018)* **[129]**, the authors investigate the phylogeny of C-subgenome in relation to A-subgenome, hypothesizing the crown age of C-genome diploid lineage to be approximately 20 Mya, significantly older than the A-genome lineage. Other investigations into the degree of divergence with respect to *A. eriantha* and *A. atlantica*, hypothesize the evolutionary distance of these two genomes to be diverged by approximately 5.4-12.9 million years, as noted by *Maughan et al* **[130]**. Moreover, an unknown degree of drift in genomic elements including repetitive elements, has occurred between *A. atlantica* and *A. eriantha*. Analysis beyond general repeat identification and masking, such as processes typically exploited in a run of RepeatMasker **[101]** as part of genome annotation, investigation and characterization of repetitive elements of these species are largely unstudied.

With our analysis, we aimed to thoroughly characterize the repeat landscape of these two genomes by conducting additional analysis using our previously developed pipeline, Repbox. This process will provide us with insight into the elements that have accumulated, and potential sub-genome specific elements each species gained over the course of their evolution. We expect there will be particular classes of repeat elements unique to each genome as well as classes that are over or under represented, in either case novel information regarding these species is progress in our understanding of *Avena* as a whole.

Materials and Methods

Diploid Assemblies

A. atlantica (CN 7277) and *A. eriantha* (CN 19328) were sequenced using PacBio 122 RSII + Sequel and 54 SMART Sequel cells, generating a total of 31,544,396 and 28,257,346 PacBio reads and a coverage of approximately 84x and 71x coverage for *A. atlantica* and *A. eriantha* respectively **[130]**. Canu **[135]** was used in the assembly of *A. atlantica* and *A. eriantha*, with the resulting assemblies consisting of 3,914 and 8,067 contigs respectively, and an N50 of 5,544,947 and 1,385,002. Assemblies were further improved by Chicago + Dovetail Hi-C **[136]**, resulting in a scaffold N50 of 513.2Mb, and an L50 of 4, spanning a total sequence length of 3.685 Gb for *A. atlantica* and a scaffold N50 of 534.8 Mb, an L50 of 4, and spanned a total sequence length of 3.778 Gb for *A. eriantha*. The longest 8 contigs in *A. atlantica* represent 7 chromosomes (With two contigs merged into a single chromosome using linkage map information) and the longest 7 contigs in *A. eriantha* **[130]**. Quantification and classification of repetitive elements were performed using RepeatModeler (version 1.0.11) and RepeatMasker (version 4.0.7 and RepBase version v20181026) in the initial repeat analysis conducted in *Maughan et al* **[130]**.

Identification of Repeats using RepeatModeler v2 & RepeatMasker v4.1

With the implementation of an additional LTR prediction module within RepeatModeler2, analysis of *A. atlantica* and *A. eriantha* was performed to assess the effect of LTR prediction on RepeatModeler v2 derived repeat families as well as assess RepeatMasker v4.1-derived repeat profiles. RepeatModeler v2 and RepeatMasker v4.1 were ran using commands outlined in (Code 1, Commands for RepeatModeler2 & RepeatMasker v4.1).

Code 10, Commands for RepeatModeler2 & RepeatMasker v4.1

```
# RepeatModeler v2 Commands
BuildDatabase -name $DBNAME -engine ncbi $GENOME #>/dev/null
RepeatModeler -database $DBNAME -engine ncbi -pa $THREAD -LTRStruct

# RepeatMasker v4.1 Commands
RepeatMasker -pa $THREAD -e ncbi -lib $LIBRARY -gff -dir $OUTPUT -u $GENOME
```

Identification of Repeats Using Repbox Pipeline

The Repbox *de novo* pipeline and toolbox described in Chapter 2 was used for repeat identification in *A. eriantha* and *A. atlantica*. Fasta files were screened to prevent any complications due to formatting inconsistencies, such as irregular characters resulting from the of assembly process, that could potentially arise during analysis. A custom script was written to perform this task and the commands are outlined in Code 11.

Code 11, Genome Formatting Script

```
# Bash script that removes invalid characters from fasta file.
echo "Enter full path to genome..." && read GENOME
sed '/^>/ s/|/_/' $GENOME > $GENOME_clean.fa
```

Repbox *de novo* repeat detection pipeline was available for download on GitHub (https://github.com/shelvasha/repbox). All required dependencies were installed by following the installation instructions (https://github.com/shelvasha/repbox/blob/master/README.md). Running of the pipeline was executed by the *run.sh* script contained within Repbox's main directory, with function of this script being the execution of a series of bash scripts that call and

execute each sub-component of the pipeline. Corresponding *de novo* repeat libraries were generated for both *A. atlantica* and *A. eriantha*. The main script of Repbox is outlined below in (*Code 12*).

Code 12, Commands for Repbox execution in run.sh

```
# #!/bin/bash
cd $REPBOX_PREFIX
PATH=$PATH$(find $REPBOX_PREFIX/bin -type d -exec echo ":{}" \; | tr -d '\n')
$REPBOX_PREFIX/scripts/repeatmodel.sh
### Run specialized de novo TE detection software
## SINEs - sinescan.sh
## MITEs - mitefinder.sh
## Helitrons - helitronscanner.sh
$REPBOX_PREFIX/scripts/sinescan.sh
$REPBOX_PREFIX/scripts/mitefinder.sh
$REPBOX_PREFIX/scripts/helitronscanner.sh
## Classification of repeat generated from specialized de novo software packages
echo "Classification now running..."
$REPBOX_PREFIX/scripts/classify.sh
### Classify.sh process check
# Loops checking the status of classify.sh bash script.
# Loop will stop when classification is complete and the process check returns FALSE.
pid3=$(pgrep classify.sh)
while [ -d /proc/$pid3 ] ; do
    sleep 2
done
```

Identification of general repetitive elements, SINE, MITE, and Helitron elements were performed using bash scripts to call and execute each submodule of the pipeline. The following phase of analysis consisted of clustering, classification and masking of Repbox-derived *de novo* repeats from the reference genomes of *A. eriantha* and *A. atlantica*. Clustering was performed using VSEARCH **[79]**, and elements were clustered to 80% sequence identity per cluster. Classification was executed using RepeatClassifier, a submodule of RepeatModeler that classifies transposable elements by use BLAST and Repbase-derived repeat libraries **[102]**. Subsequent output from the classification were *de novo* repeat libraries classified into transposable element superfamilies consistent with RepeatMasker formatting. The final step of the Repbox pipeline was

repeat masking of the reference genomes, with this task being carried out using RepeatMasker and our classified *de novo* library.

Assessment of Generated Feature Files

Assessments between annotation methods were conducted in R (version 4.0.2) utilizing rtracklayer (version 1.48) **[137]**. Gene-feature files (GFFs) were compared using bedtools (version 2.29.2) intersect function to derive intersecting coordinates between each GFF. Corresponding repeat superfamilies/subfamilies were derived from .out files generated from RepeatMasker and incorporated into data frames to perform final comparisons in assignments. Code utilized in this analysis is outlined in Code 13. The resulting data frames for both RepeatModeler/Masker and Repbox annotations are then compared using bedtools intersect to compare assigned TE type from each respective annotation.

Code 13, R Commands for GFF Comparisons

```
# Import .out file derived from RepeatMasker
repbox_out <-
read_table2("/Users/shelvasha/Documents/Dissertation/results/eriantha/Eriantha(8_10)/Avena_eri
antha.rebox.fa.out", skip = 1)

# Import gff output from RepeatMasker
repbox_gff <- read.delim("~/repbox.out.gff3", header = FALSE)

# Rename column headers of .out and .gff imported data frames
colnames(repbox_out) <-
c("SW_Score","%_Div.","%_Del.","%_Ins.","Query_Sequence","Position_Query_Start","Position_Quer
y_End","Past_Ending Position_of_Match", "Complement_Match", "Interspersed_Repeat_Name",
"Repeat_Class", "Repeat_Match_Start", "Repeat_Match_End", "ID")

names(repbox_gff) [1:9] <-
c("Chrom","Source","Type","Start","End","Score","Strand","Phase","Attributes")

# Assign TE class to GFF coordinates from family data contained in .out
repbox_gff$Type <- repbox_out$Repeat_Class

bedtools intersect “A.gff3" -b "B.gff3" -wa -wb -f 0.8 > A_B_overlap.gff3
```

Phylogenetic Analysis of A. atlantica and A. eriantha

Gypsy Elements

Predicted *Gypsy* elements were isolated and extracted from each annotation using bedtools getfasta (version 2.29.2) **[115]** and subsequent sequences aligned in a multiple sequence alignment using MAFFT (7.453) **[138]**. These multiple sequence alignments were then used to generate phylogenetic trees using FastTree (version 2.19) **[139]** to infer phylogenetic relationships between identified *Gypsy* candidates. Code describing an overview of this analysis is outlined below (*Code 5, Phylogenetic analysis of LTR/Gypsy elements)*. Subsequent trees were output in Newick format, and were visualized using phylogram (version 2.2.4) in R.

Code 14, Phylogenetic analysis of LTR/Gypsy elements

```
# 1. Extraction of Gypsy classified elements from genome using bedtools getfasta
bedtools getfasta -fi Avena_atlantica.fa -fo Aa_Gypsy.fa -bed reclassified_gypsy.gff3 -s

# 2. Multiple Sequence alignment using MAFFT FFNS-2
mafft --retree 2 Aa_Gypsy.fa > Aa_Gypsy_output_alignment

# 3. Tree generation using FastTree from output_alignment
FastTree -gtr -nt Aa_Gypsy_output_alignment > tree
```

Results and Discussion

Initial Repeat Annotation of Avena atlantica and Avena eriantha

Maughan et. al [130] provided an initial view of the repeat landscape from *A. atlantica* and *A. eriantha* as part of the genome characterization of these species using RepeatModeler v1.0.11 and RepeatMasker v4.0.7. From their analysis, approximately 83% and 84% of *A. atlantica* and *A. eriantha* genome sequences were identified as repetitive, with long terminal repeat retrotransposons or LTRs in highest abundance. This finding is consistent with higher ploidy plant species and members of Poaceae, such as *Triticum aestivum L.*, *Oryza sativa L.*, and *Zea mays L.*, all of which are highly domesticated crops that have undergone polyploidization events over the course of their evolutions [140]. Familial repeat composition of both diploid species studied in *Maughan et al* [130] were composed of 66.7% and 62.44% of LTR class elements,

5.89% and 6.88% of DNA class elements and 1.03% and 1.09% of LINE elements in *A. atlantica* and *A. eriantha* respectively. SINE elements and Helitron elements were scarcely observed in both genomes, representing less than 0.02% of each genome. Other notable contributing elements by *Maughan et al* is the proportion of unknown repeats for each genome, with *A. atlantica* containing (8.78%) and *A. eriantha* containing (11.84%) of its genome categorized as unknown. From this result, *Maughan et al* concluded that these elements potentially represent repeat elements unique to Avena, as little is known about these sequences and they lack classification [130].

For the sake of clarity, comparisons between analysis will be referred to as V1, V2 and Repbox. V1 refers to analysis using RepeatModeler v1.0.11 and RepeatMasker 4.0.7 from *Maughan et al*, V2 refers to analysis using RepeatModeler v2 and RepeatMasker v4.1., and reference to Repbox analysis includes RepeatModeler v2, Sine_Scan, HelitronScanner, MITEFinder, and RepeatMasker v4.1.

A. atlantica Repeat Landscape

V2 analysis of *A. atlantica* demonstrated a noticeable increase in total interspersed elements; 82.59% to 85.96% between V1 and V2 analyses. Surprisingly, elements observed as highly abundant in V1 analysis, such as LTR and DNA elements, demonstrated dramatic decreases in counts and percentages: LTR displaying a decrease of 314,018; 66.7% to 52.66%; DNA elements seeing a decrease of 204,050; 5.9% to 4.05%; Changes in observations of LINE elements were mild, with the percentage of observed LINE increased by 1,836; from 0.92% to 1.23%. Finally, there was a noted decline in the observed numbers of SINE and RC/Helitrons, with these elements being unobserved in V2 analysis; decreasing by 3,080 elements; 0.01% to 0.0%. Detailed summaries detailing the losses and gains of TEs observed in *A atlantica* are outlined

below (*Table 10).* The element class with the largest change between V1 and V2 was elements reclassified to unknown, with these elements increasing from 533,080 to 1,386,441 (+853,361 elements) representing a ~20% increase to interspersed elements observed within *A. atlantica*. This result was startling as implementation of the LTR predictive module created an expectation of increased LTRs.

Table 10, Short Summary of V1 (RepeatModeler v1) and V2 (RepeatModeler v2) in A. atlantica

Repeat Class	Change in Element Count	Count of Families Lost	Count of Families Gained
DNA	*-204050*	*3*	*1*
LINE	*-1836*	*3*	*2*
LTR	*-340018*	*1*	*1*
RC	*0*	*0*	*0*
SINE	*0*	*0*	*0*
Unknown	*853361*	*	*
	Total	*7*	*4*

* indicates classes without families

Repbox derived annotations showed increases in LTRs, additional decreases to DNA elements, and overall increases to TE family diversity (defined by an increase in the number of distinct TE families). The increase of LTRs was substantial, with a total increase of ~11% or 912,780 LTR elements. Further investigation into the data indicates a primary driver of this increase is identification and classification of Gypsy elements (Table 12). This result is made more impressive by the fact that alongside the increase to Gypsy elements, there was a pronounced decrease to LTR/Copia (-250,285 elements). Although *Maughan et al* also showed that Gypsy elements are by far the largest proportion of repetitive elements detected in *A. atlantica*, the overall counts are significantly higher in Repbox's analysis indicating a greater percentage of LTR elements within the *A. atlantica* genome than previously identified. The significantly higher counts of these elements observed in Repbox's analysis was potentially indicative of underlying algorithmic/genomic differences yet to be investigated.

In addition to the substantial increase in Gypsy elements, Repbox identified previously uncharacterized families such as Ngaro (0.01% increase of total genome masked or 9,766 additional element candidates), ERV1 (254 additional element candidates), ERVK (237 additional element candidates) and Caulimovirus (105 additional element candidates). Based upon

observations of the data in conjunction to observations in prior work by *Flynn et al* (2019) **[102]** establishing RepeatModeler2, we believe the emergence of these previously unobserved elements are the result of LTR prediction implemented within RepeatModeler. ERV1, ERVK and Caulimovirus contributions to percentage of the repeat profile was less that 0.00001% of the *A. atlantica* genome. Repbox pipeline also provided increases in other superfamilies that were initially unobserved in original repeat analysis. Superfamilies include: RC/Helitrons (615); DNA families: Academ-1 (3 elements), Dada (128 elements), Ginger-1 (17 elements), TcMar-Mariner (175 elements) and hAT-Blackjack (191 elements); LINE families: LINE-I (5 elements), L1-Tx1 (71 elements), Penelope (76 elements), RTE-X (135 elements); and SINE families (120). Previous analysis of *A. atlantica* failed to detect any evidence of the aforementioned TE families.

Additional decreases of observed elements include DNA families CMC-EnSpm (167,117 count decrease in elements), MULE-MuDR (17,837 count decrease in elements), Maverick (148 decrease in elements), PIF-Harbinger (33,483 decrease in elements), TcMar-Stowaway (82,812 decrease in elements), hAT-Ac/Tag1/Tip100 (3134 count decrease in elements), and LINE families CR1 (922 count decrease in elements), Jockey (145 count decrease in elements), L1 (26,222 count decrease in elements), and R1 (1345 count decrease in elements). Despite the authors not observing significant changes to genomes utilized in benchmarking of RepeatModeler v2, we also attribute the observed decreases of DNA and LINE elements within our genomes of interest to changes implemented in RepeatModeler v2. We believe this is the case based on confirmation analysis of positional data between analysis of RepeatModeler v1 and RepeatModeler v2, which has shown to drastically change transposable element profiles by recategorization of elements to LTRs. This is discussed in greater detail below.

Table 11, Short Summary of V2 (RepeatModeler v2) and Repbox (Repbox) in A. atlantica

Repeat Class	Change in Element Count	Count of Families Lost	Count of Families Gained
DNA	*-316259*	*1*	*5*
LINE	*-28347*	*4*	*6*
LTR	*912780*	*0*	*4*
RC	*615*	*0*	*2*
SINE	*120*	*0*	*2*
Unknown	*632815*	*	*
	Total	*5*	*20*

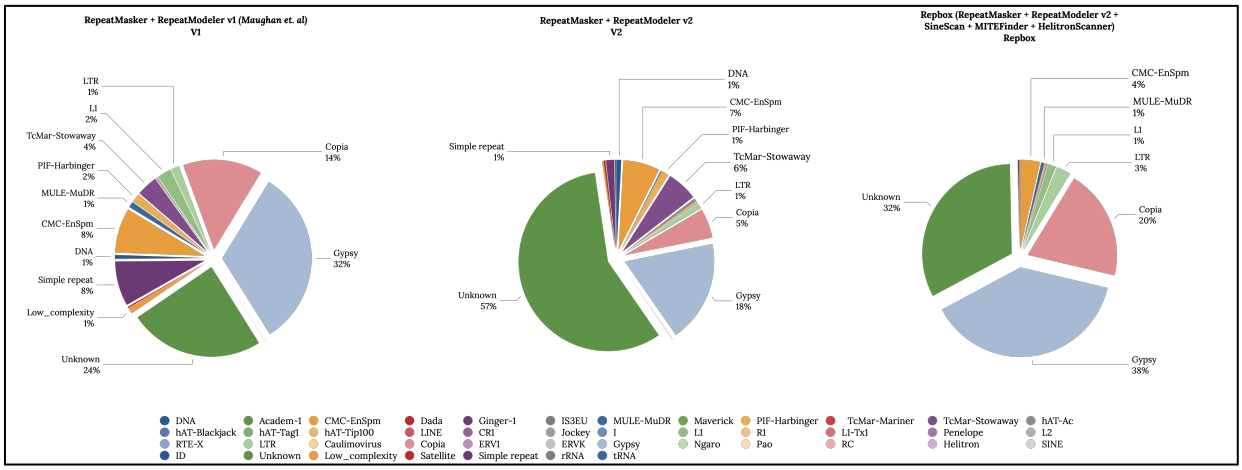
* indicates classes without families

Table 12, Summary Analysis of V1, V2 and Repbox in A. atlantica

	RepeatMasker v4.0.7 + RepeatModeler v1 (Maughan et. al)			***RepeatMasker v4.1 + RepeatModeler v2***				***Repbox (RepeatMasker v4.1 + RepeatModeler v2 + SineScan + MITEFinder + HelitronScanner)***			
	v1			*v2*				*Repbox*			
Repeat Class	*Count*	*Bases masked*	*Masked*	*Difference*	*Count*	*Bases masked*	*Masked*	*Difference*	*Count*	*Bases masked*	*Masked*
DNA											
Unclassified DNA	*12343*	*2295245*	*0.06%*	*-12343*	*0*	*0*	*0.00%*	*101*	*101*	*8588*	*0.00%*
Academ-1	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*	*3*	*3*	*571*	*0.00%*
CMC-EnSpm	*176565*	*183627714*	*5.00%*	*-98875*	*77690*	*112927752*	*3.07%*	*-68242*	*9448*	*4878007*	*0.13%*
Dada	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*	*128*	*128*	*13006*	*0.00%*
Ginger-1	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*	*17*	*17*	*517*	*0.00%*
IS3EU	*0*	*0*	*0.00%*	*1286*	*1286*	*1098582*	*0.03%*	*-1286*	*0*	*0*	*0.00%*
MULE-MuDR	*23287*	*5211297*	*0.14%*	*-1943*	*21344*	*17073516*	*0.46%*	*-15894*	*5450*	*1488970*	*0.04%*
Maverick	*148*	*18931*	*0.00%*	*-148*	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*
PIF-Harbinger	*34425*	*10378659*	*0.28%*	*-19287*	*15138*	*7564293*	*0.21%*	*-14196*	*942*	*142039*	*0.00%*
TcMar-Mariner	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*	*175*	*175*	*24236*	*0.00%*
TcMar-Stowaway	*82949*	*13125880*	*0.36%*	*-82949*	*0*	*0*	*0.00%*	*137*	*137*	*7558*	*0.00%*
hAT-Ac	*2924*	*849064*	*0.02%*	*7401*	*10325*	*7872532*	*0.21%*	*-8899*	*1426*	*185257*	*0.01%*
hAT-Blackjack	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*	*191*	*191*	*53893*	*0.00%*
hAT-Tag1	*878*	*410769*	*0.01%*	*-878*	*0*	*0*	*0.00%*	*25*	*25*	*4450*	*0.00%*
hAT-Tip100	*1561*	*651355*	*0.02%*	*3686*	*5247*	*2250652*	*0.06%*	*-4469*	*778*	*174010*	*0.00%*
LINE											
Unclassified LINE	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*
CR1	*922*	*101045*	*0.00%*	*-922*	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*
Jockey	*145*	*30396*	*0.00%*	*-145*	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*
I	*0*	*0*	*0.00%*	*288*	*288*	*61948*	*0.00%*	*-283*	*5*	*306*	*0.00%*
L1	*53523*	*33266727*	*0.91%*	*-212*	*53311*	*45110802*	*1.23%*	*-26010*	*27301*	*18093861*	*0.49%*
R1	*1345*	*438210*	*0.01%*	*-1345*	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*
L1-Tx1	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*	*71*	*71*	*10973*	*0.00%*
Penelope	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*	*76*	*76*	*13949*	*0.00%*

L2	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*
RTE-X	*0*	*0*	*0.00%*	*500*	*500*	*150083*	*0.00%*	*-365*	*135*	*43042*	*0.00%*
LTR											
Unclassified LTR	*32110*	*49218294*	*1.34%*	*14272*	*46382*	*81843883*	*2.23%*	*7204*	*53586*	*79294222*	*2.16%*
Caulimovirus	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*	*105*	*105*	*11049*	*0.00%*
Copia	*312901*	*641161159*	*17.46%*	*-79184*	*233717*	*632920041*	*17.23%*	*-171101*	*62616*	*9496065*	*0.26%*
ERV1	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*	*254*	*254*	*22747*	*0.00%*
ERVK	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*	*237*	*237*	*16231*	*0.00%*
Gypsy	*705163*	*1758990581*	*47.89%*	*-275061*	*430102*	*1219358486*	*33.20%*	*1405742*	*1835844*	*2766297667*	*75.31%*
Ngaro	*0*	*0*	*0.00%*	*474*	*474*	*77604*	*0.00%*	*9292*	*9766*	*444627*	*0.01%*
Pao	*519*	*285086*	*0.01%*	*-519*	*0*	*0*	*0.00%*	*1065*	*1065*	*244102*	*0.01%*
RC/Helitron											
Rolling Circle	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*
Helitron	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*	*615*	*615*	*105267*	*0.00%*
SINE											
Unclassified SINE	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*	*104*	*104*	*11608*	*0.00%*
ID	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*	*16*	*16*	*1089*	*0.00%*
Unknown	*533080*	*322656906*	*8.78%*	*853361*	*1386441*	*1029094595*	*28.02%*	*-220546*	*1165895*	*268572616*	*7.31%*
Total interspersed	*1977868*	*3033593943*	*82.59%*	*304377*	*2282245*	*3157404769*	*85.96%*	*894267*	*3176512*	*3149660523*	*85.75%*
Other											
Low_complexity	*22741*	*1212274*	*0.03%*	*-4487*	*18254*	*982124*	*0.03%*	*-1503*	*16751*	*868872*	*0.02%*
Satellite	*5217*	*2364614*	*0.06%*	*-4468*	*749*	*72649*	*0.00%*	*1930*	*2679*	*526363*	*0.01%*
Simple repeat	*176100*	*10467715*	*0.28%*	*-21923*	*154177*	*8180990*	*0.22%*	*-1475*	*152702*	*8082953*	*0.22%*
rRNA	*0*	*0*	*0.00%*	*1217*	*1217*	*916152*	*0.02%*	*-584*	*633*	*341532*	*0.01%*
tRNA	*3080*	*530660*	*0.01%*	*-3080*	*0*	*0*	*0.00%*	*978*	*978*	*85256*	*0.00%*
Total	*2181926*	*3047638546*	*82.97%*	*274716*	*2456642*	*3167556684*	*86.24%*	*893613*	*3350255*	*3159565499*	*86.02%*

Figure 25, TE Family Comparison of Analysis of V1, V2 and Repbox in A. atlantica



A. eriantha Repeat Landscape

V2 analysis of *A. eriantha* shows an overall increase of 82.92% in V1 analysis to 85.72% in V2. Previously more-abundant elements, such as LTR and DNA elements display noticeable decreases to overall genome percentage, with LTRs decreasing from 62.44% to 50.69% in *A. eriantha*. LINE elements between V1 and V2 in *A. eriantha* decrease slightly, from 1.09% to 1.14% (-4,836 elements). Like observations in V1, SINE and RC/Helitron were sparsely observed in V2 analysis, decreasing from 0.01% to 0.0% (-7,932 elements). The largest change between V1and is an increase in the percentage of elements reclassified to unknown, with these elements increasing from 693,112 to 1,678,345 (+985,233 elements) representing a ~19% increase to unknown elements observed within *A. eriantha*. This is quite similar to the results observed for *A. atlantica,* where large proportions of elements were reclassified into unknown with RepeatModeler v2

Table 13, Short Summary of V1 (RepeatModeler v1) and V2 (RepeatModeler v2) in A. eriantha

Repeat Class	Change in Element Count	Count of Families Lost	Count of Families Gained
DNA	*-329357*	*3*	*2*
LINE	*-4407*	*3*	*2*
LTR	*-235306*	*2*	*3*
RC	*0*	*0*	*0*
SINE	*-1695*	*0*	*0*
Unknown	*985233*	*	*
	Total	*8*	*5*

* Indicates a lack of families in class

Repbox analysis of *A. eriantha* demonstrates a decrease of total interspersed elements by ~35%, with a decrease of DNA elements (-131,074) and an overall increase in LTR elements; (increase of 449,224 elements), representing ~57% of identified elements in *A. eriantha*. There were also several previously unobserved TE families, potentially demonstrating an increase to TE

family diversity. Families including: Kolobok-T2 (35 elements), TcMar-Tc1(9 elements), TcMar-Tc2(24 elements), hAT-hAT19 (6 elements), hAT (16 elements), and Dada (2 elements). Variance of TE families noted between annotations of V2 and Repbox highlight the effect of differing underlying implementations within each pipeline, as contributions of all identified TE families strongly contrasts from prior observations (Figure 26). The more striking changes between annotations, such as observations of Gypsy elements shifting from 33% to 90.1%, MULE-MUDr elements shifting from 4.8% to 0.02%, and PIF-Harbinger shifting from 0.74% to 0.01%, are the more extreme examples of changes that occurred between pipelines. The origins and explanations as to why these substantial changes are occurring within the annotation of *A. eriantha* are discussed further below.

Table 14, Short Summary of V2 (RepeatModeler v2) and Repbox (Repbox) in A. eriantha

Repeat Class	Change in Element Count	Count of Families Lost	Count of Families Gained
DNA	*-131074*	*1*	*10*
LINE	*-42148*	*3*	*3*
LTR	*449224*	*1*	*2*
RC	*39*	*0*	*1*
SINE	*5*	*0*	*1*
Unknown	*-1567593*	*	*
	Total	*5*	*17*

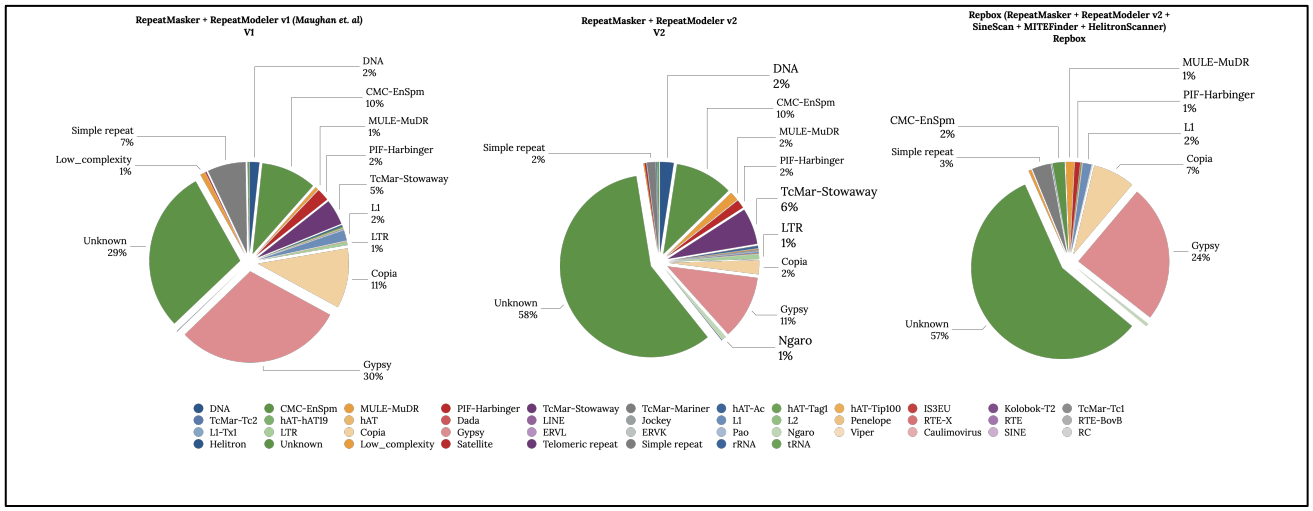
* Indicates a lack of families in class

Table 15, Summary of RepeatModeler v1, RepeatModeler v2 and Repbox in A. eriantha

			RepeatMasker v4.0.7 + RepeatModeler v1 (Maughan et. al)			***RepeatMasker + RepeatModeler v2***			***Repbox (RepeatMasker + RepeatModeler v2 + SineScan + MITEFinder + HelitronScanner)***		
			V1			*V2*			*Repbox*		
Repeat Class	*A. eriantha*	*3776743233*	*Change V1 -> V2*		*A. eriantha*	*3776743233*	*Change V2 -> Repbox*		*A. eriantha*	*3776743233*	
	Count	*Bases masked*	*Masked*	*Difference*	*Count*	*Bases masked*	*Masked*	*Difference*	*Count*	*Bases masked*	*Masked*
DNA											
Unclassified DNA	*41071*	*13892671*	*0.37%*	*-41071*	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*
CMC-EnSpm	*235692*	*181347535*	*4.80%*	*-174563*	*61129*	*75884347*	*2.01%*	*-58101*	*3028*	*2833683*	*0.08%*
MULE-MuDR	*13913*	*9147023*	*0.24%*	*27586*	*41499*	*36667126*	*0.97%*	*-40476*	*1023*	*734497*	*0.02%*
PIF-Harbinger	*53235*	*28107529*	*0.74%*	*-26168*	*27067*	*11722448*	*0.31%*	*-26266*	*801*	*352338*	*0.01%*
TcMar-Stowaway	*108832*	*20258806*	*0.54%*	*-108832*	*0*	*0*	*0.00%*	*2*	*2*	*338*	*0.00%*
TcMar-Mariner	*0*	*0*	*0.00%*	*1610*	*1610*	*543154*	*0.01%*	*-1610*	*0*	*0*	*0.00%*
hAT-Ac	*6590*	*2464996*	*0.07%*	*-6590*	*0*	*0*	*0.00%*	*26*	*26*	*4522*	*0.00%*
hAT-Tag1	*5261*	*3946889*	*0.10%*	*-149*	*5112*	*2210707*	*0.06%*	*-5109*	*3*	*804*	*0.00%*
hAT-Tip100	*1875*	*932660*	*0.02%*	*-1875*	*0*	*0*	*0.00%*	*1036*	*1036*	*578522*	*0.02%*
IS3EU	*0*	*0*	*0.00%*	*695*	*695*	*488180*	*0.01%*	*-658*	*37*	*31847*	*0.00%*
Kolobok-T2	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*	*35*	*35*	*7717*	*0.00%*
TcMar-Tc1	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*	*9*	*9*	*1151*	*0.00%*
TcMar-Tc2	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*	*14*	*14*	*1597*	*0.00%*
hAT-hAT19	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*	*6*	*6*	*1600*	*0.00%*
hAT	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*	*16*	*16*	*807*	*0.00%*
Dada	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*	*2*	*2*	*124*	*0.00%*
LINE											
Unclassified LINE	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*
Jockey	*5977*	*4017019*	*0.11%*	*-5977*	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*
L1	*44555*	*36045210*	*0.95%*	*3280*	*47835*	*42884419*	*1.14%*	*-42490*	*5345*	*2920473*	*0.08%*
L2	*573*	*326088*	*0.01%*	*-573*	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*
Penelope	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*	*820*	*820*	*1973394*	*0.05%*
RTE-X	*1738*	*877096*	*0.02%*	*-1738*	*0*	*0*	*0.00%*	*5*	*5*	*614*	*0.00%*

RTE	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*	*11*	*11*	*2460*	*0.00%*
RTE-BovB	*0*	*0*	*0.00%*	*436*	*436*	*142053*	*0.00%*	*-335*	*101*	*52315*	*0.00%*
L1-Tx1	*0*	*0*	*0.00%*	*165*	*165*	*35624*	*0.00%*	*-159*	*6*	*493*	*0.00%*
LTR											
Unclassified LTR	*14612*	*5824958*	*0.15%*	*-14612*	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*
Copia	*254114*	*522841719*	*13.84%*	*-41887*	*212227*	*471748759*	*12.49%*	*-204808*	*7419*	*3210311*	*0.09%*
Gypsy	*715788*	*1829333860*	*48.44%*	*-191638*	*524150*	*1418860862*	*37.57%*	*667043*	*1191193*	*2121284248*	*56.17%*
ERVL	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*	*267*	*267*	*199259*	*0.01%*
ERVK	*0*	*0*	*0.00%*	*147*	*147*	*487919*	*0.01%*	*-146*	*1*	*352*	*0.00%*
Pao	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*	*20*	*20*	*5130*	*0.00%*
Ngaro	*0*	*0*	*0.00%*	*12924*	*12924*	*23173341*	*0.61%*	*-12924*	*0*	*0*	*0.00%*
Viper	*469*	*281891*	*0.01%*	*-469*	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*
Caulimovirus	*0*	*0*	*0.00%*	*229*	*229*	*233384*	*0.01%*	*-228*	*1*	*271*	*0.00%*
SINE											
Unclassified SINE	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*	*5*	*5*	*533*	*0.00%*
RC											
Helitron	*1695*	*568721*	*0.02%*	*-1695*	*0*	*0*	*0.00%*	*39*	*39*	*7528*	*0.00%*
Unknown	*693112*	*447222379*	*11.84%*	*985233*	*1678345*	*1152393113*	*30.51%*	*-1567593*	*110752*	*33462426*	*0.89%*
Total interspersed	*2226957*	*3131828734*	*82.92%*	*386613*	*2613570*	*3237475436*	*85.72%*	*-1291547*	*1322023*	*1899748196*	*50.30%*
Other											
Low_complexity	*21382*	*1166133*	*0.03%*	*-3596*	*17786*	*964016*	*0.03%*	*-13660*	*4126*	*207107*	*0.01%*
Satellite	*3404*	*943623*	*0.02%*	*-3404*	*0*	*0*	*0.00%*	*115*	*115*	*89602*	*0.00%*
Telomeric repeat	*1815*	*14459837*	*0.38%*	*-1815*	*0*	*0*	*0.00%*	*0*	*0*	*0*	*0.00%*
Simple repeat	*162410*	*10363028*	*0.27%*	*-25722*	*136688*	*6095463*	*0.16%*	*-90645*	*46043*	*2019923*	*0.05%*
rRNA	*0*	*0*	*0.00%*	*2788*	*2788*	*875340*	*0.02%*	*-2788*	*0*	*0*	*0.00%*
tRNA	*6237*	*4121298*	*0.11%*	*-6237*	*0*	*0*	*0.00%*	*258*	*258*	*7651*	*0.00%*
Total	*2415968*	*3158761355*	*83.64%*	*354864*	*2770832*	*3245410255*	*85.93%*	*-1398267*	*1372565*	*2167669354*	*57.40%*

Figure 26, TE Family Comparison of Analysis of V1, V2 and Repbox in A. eriantha



Addressing Changes in Annotations

To address the changes observed between annotation versions of both *A. atlantica* and *A. eriantha*, we first looked to the differences in software versions and updates used in each analysis. *Maughan et al* **[130]** identified repeats using RepeatModeler version 1.0.11. However, since publication, RepeatModeler (version 2.0.1) **[102]**, was released, incorporating *de novo* LTR prediction with LTRHarvest **[93]** and LTR_Retriever **[94]**. Implementation of these two specific *de novo* detection tools was based upon prior work by *Jiang et al* [94], where both were noted to significantly improve the identification of LTRs in *Oryza sativa*. Each package demonstrated superior detection of LTRs, with LTRHarvest performing at a sensitivity, specificity and accuracy percentage of (92.95%, 87.70%, and 88.94%) and LTR_retriever (96.8%, 95.5% and 95.18%). *Jiang et al* concluded that these software packages provided a significant improvement in comparison to other common LTR detection methods, such as LTR_STRUCT and MGEScan-LTR **[94]**. Consistent with *Jiang et al* observations, the authors of RepeatModeler v2 also demonstrate the positive effect these implementations had on the identification of LTR elements. These improvements were expressed in the form of substantial increases to the total count of identified LTR elements, as well as a noticeable increase to the quality of sequences detected by these packages (*defined by the authors as >=95% sequence identity)* to reference TE libraries. The authors go on to note the extremely low false positive rate of RepeatModeler v2 when applied to simulated genomes lacking of TEs **[102]**.

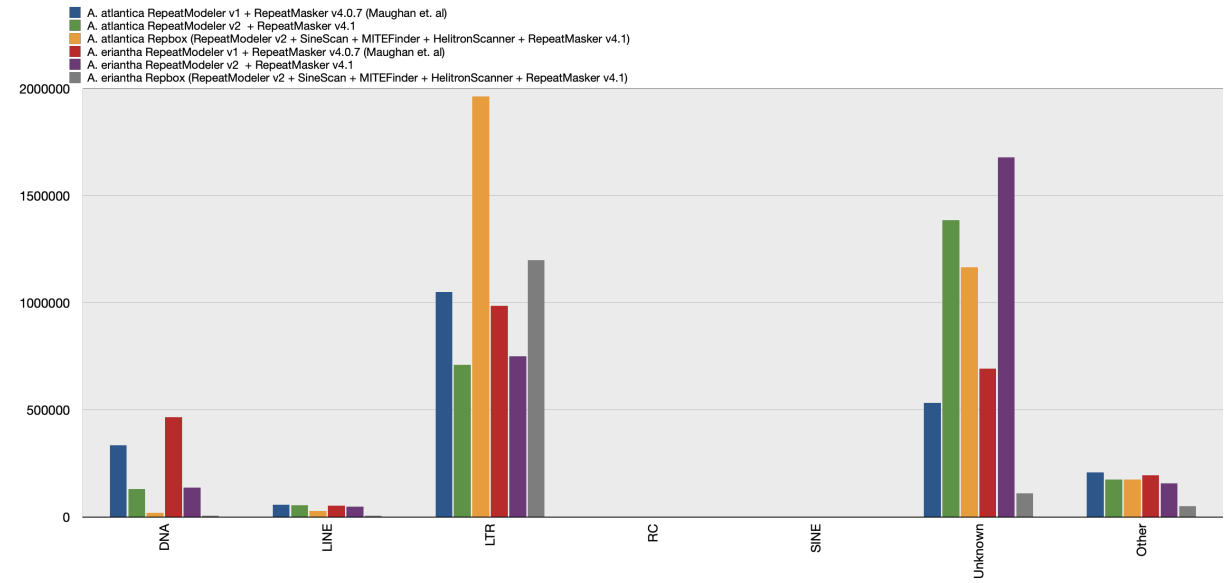
Both V2 and Repbox results, which use RepeatModeler v2 when compared to *Maughan et al*, produced strikingly different LTR identification results. The observed losses/reassignments of DNA and LTR elements in both *A. atlantica* and *A. eriantha* came as a surprise (*A. atlantica*: -340,018; *A. eriantha:* -235,306 elements) (*Figure 28*), as there was an expectation that the number of predicted LTRs would increase in V2 compare to V1 with the implementation of LTR

prediction methods in the new version of RepeatModeler. The decreases of these elements in both genomes are further confounded when comparing the later results of V2 and Repbox. In this analysis, the data demonstrates an additional drop in DNA and LINE elements; DNA: ~85.6% decrease in *A. atlantica* and ~95.6 % decrease in *A. eriantha*; LINE : ~49% decrease in *A. atlantica* and ~87% *A. eriantha*. Contrary to prior comparisons, LTR identification seems to reverse course, with Repbox producing a sizable increase to predicted LTR elements; ~176% increase in *A. atlantica* and ~59.9% in *A. eriantha*. In short, we observed a decrease in DNA and LTR elements between V1 and V2 analysis. We then observed a tremendous increase of LTR elements and continued to see a decrease of DNA elements between V2 and Repbox analyses. To investigate the major shifts in element classes between each annotation version (v1, v2 and Repbox), GFF output was analyzed to obtain positional data (i.e. start/end coordinates of annotations) and element classification derived from each analysis. This positional analysis was essentially comparing how TE family classification at each position in the genome differed between annotation sets. Analysis revealed that in comparisons between V1 and V2, *A. atlantica* had 130,426 elements reclassified into TE families that differed from its original designation in V1. Of that 130,426 elements, 19,894, or 15.25%, were reclassified into unknown (*Figure 29*). In *A. eriantha*, a similar pattern was observed, with 320,228 elements being reclassified into TE families that differed from original assignments in V1 analysis. Of the 320,228 elements that differed from V1 analysis of *A. eriantha*, unknown elements contributed 105,632 elements or 32.98%.

From V2 to Repbox, the counts of reclassified elements remains elevated, with 375,901 elements being reclassified in *A. atlantica* and 329,571 being reclassified in *A. eriantha*. Interestingly, most of these reclassifications are from Unknown to LTR/Gypsy, with 148,022 elements (~39.4%) in *A. atlantica*, and 264,641 elements (~80.3%) being reclassified in *A.*

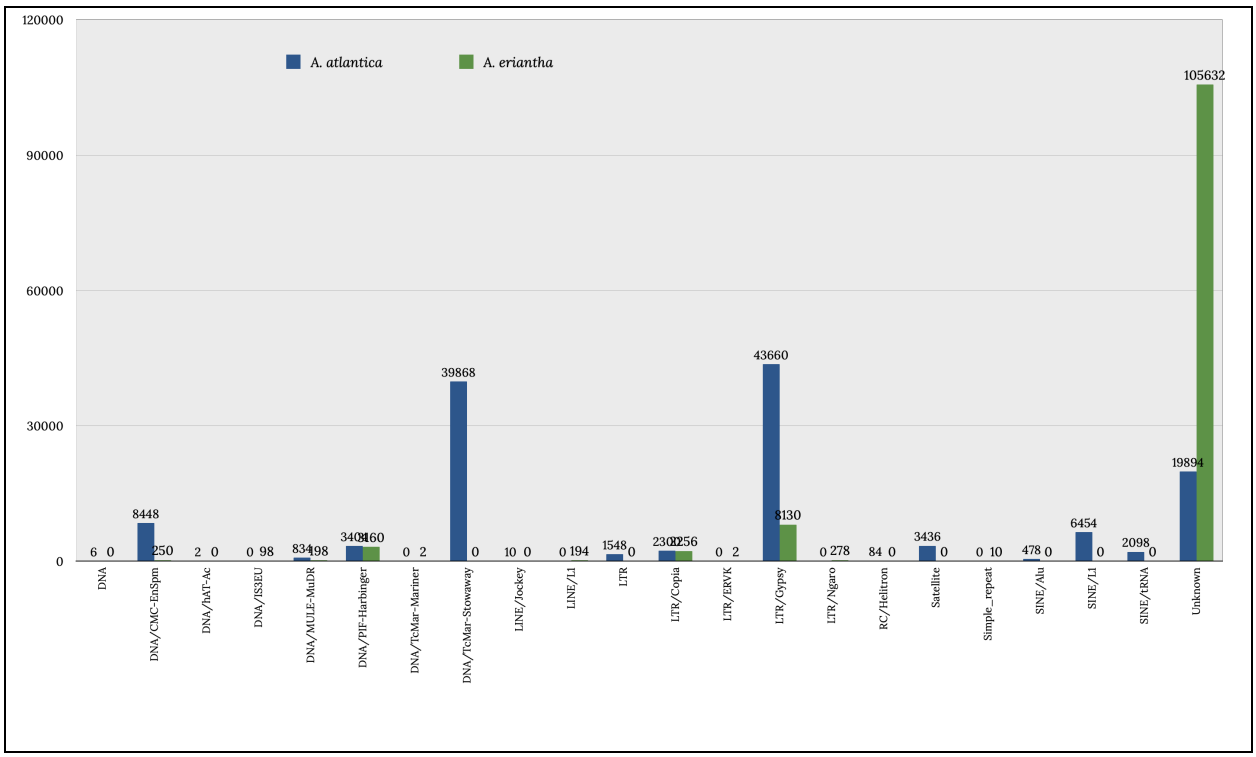
eriantha from Unknown to LTR/Gypsy. At least 130,426 elements and 320,228 elements are being reclassified into different elements between V1 and V2, and 375,901 and 329,571 elements between V2 and Repbox.

Figure 27, Elements Identified in A. atlantica and A. eriantha by Superfamily



	A. atlantica			*A. eriantha*		
	RepeatMasker v4.0 + RepeatModeler v1 (Maughan et al)	*RepeatModeler v2 + RepeatMasker v4.1*	*Repbox (RepeatModeler v2 + RepeatMasker v4.1 + SineScan + MITEFinder + HelitronScanner)*	*RepeatMasker v4.0 + RepeatModeler v1 (Maughan et al)*	*RepeatModeler v2 + RepeatMasker v4.1*	*Repbox (RepeatModeler v2 + RepeatMasker v4.1 + SineScan + MITEFinder + HelitronScanner)*
DNA	*335080*	*131030*	*18821*	*466469*	*137112*	*6038*
LINE	*55935*	*54099*	*27588*	*52843*	*48436*	*6288*
LTR	*1050693*	*710675*	*1963473*	*984983*	*749677*	*1198901*
RC	*0*	*0*	*615*	*1695*	*0*	*39*
SINE	*0*	*0*	*120*	*0*	*0*	*5*
Unknown	*533080*	*1386441*	*1165895*	*693112*	*1678345*	*110752*
Other	*207138*	*174397*	*173743*	*195248*	*157262*	*50542*

Figure 28, Counts of Superfamilies Reclassified to Unknown in A. atlantica and A. eriantha



Effect of Fragmentation on Annotations

Fragmentation, in the context of genome annotation is defined as the breaking up of repeat annotation features into smaller segments due to multiple insertion events in the same physical location. The result if fragmentation are smaller gene features that are abundant throughout the annotation that have the unfortunate effect of obscuring genomic features and characteristics **[94]**. Fragmentation is a likely candidate for some of the unexpected variances we observed in multiple annotations generated throughout the course of our analysis. We suspected fragmentation due to: (a) there was an increase in elements classified as unknown (b) we observed an increase in abundance of repetitive elements and (c) there was a significant decrease to the average length of elements annotated. From these observations, we concluded that fragmentation was likely occurring more frequently in RepeatModeler v2-derived annotations.

In both genome annotations of V2 and Repbox, the data reports a substantial shift in average feature length, however this is more pronounced in annotations of *A. atlantica*, with its features decreasing from 3533.9 bps to 1393.4 bps (60.57% decrease). *A.eriantha* reported opposing shifts in comparison to *A. atlantica*, displaying an increase average feature length from 1121.8 bps to 1577.7 bps. We feel this is attributed to the decrease in interspersed elements (~85.9 % to ~57.4%), and potentially indicates that fragmentation plays far less of a role in *A. eriantha*, highlighting that these two genomes, while both being from the Avena genus, possess different evolutionary trajectories with respect to repeat movement and accumulation.

Visual comparisons using IGV (version 2.8.2) *(Figures 30 and 31)*, as well as positional comparisons of feature files in R, reveals that structurally, less than 50% of elements identified between both RepeatModeler v2 and Repbox share loci positions (*Table 17*). Approximately 45.24% of *A. atlantica* and 10.55% of *A. eriantha* gene features overlap between the V2 and Repbox annotation pipelines. However, Repbox results show an increased number of repetitive elements identified overall *(Figures 30 and,31)*. As the annotation tools used in Repbox are more focused on smaller or unique repeat families (MITEs, etc), we hypothesize that feature entries are being segmented into several multiple features. For example, in Figures 29 and 30, instances of LTR-Gypsy elements annotated by RepeatModeler v2 + RepeatMasker v4.1, denoted by “V2” while elements of the same class detected in Repbox, denoted by “Repbox”, overlap but have clearly been fragmented in the Repbox annotation pipeline. This fragmentation on the genome scale, as noted in discussions above, can dramatically shift the repeat profile metrics, such as the average element length, resulting in the variances and unexpected results derived in our analysis above.

Figures 29, 30 Gene Feature File Comparison of A. atlantica and A. eriantha RepeatMasker v2 and Repbox Annotations


Avena_atlantica.fa
lcl_ScoFOjO_324_449
24_449:288,197,248-288,219,075
Go
21 kb
288,198 kb
288,200 kb
288,202 kb
288,204 kb
288,206 kb
288,208 kb
288,210 kb
288,212 kb
288,214 kb
288,216 kb
288,218 kb
V2)
Repbox)
Avena_eriantha.fa
lcl_SctCcfP_2331_4482
1_4482:358,525,830-358,589,653
Go
63 kb
358,530 kb
358,540 kb
358,550 kb
358,560 kb
358,570 kb
358,580 kb
V2)
Repbox)

Table 16, V2 (RepeatMasker v4.1 + RepeatModeler v2) & Repbox Element Comparisons

A. Atlantica

Analysis	*Element Count*	*Avg. Feature length*	*% of Genome*
V2 (RepeatMasker v4.1 + RepeatModeler v2)	*2,049,678*	*1393.42 bp*	*82.97%*
Repbox	*3,897,239*	*929.12 bp*	*85.74%*

A. Eriantha

Analysis	*Element Count*	*Avg. Feature length*	*% of Genome*
V2 (RepeatMasker v4.1 + RepeatModeler v2)	*3,037,939*	*1121.753 bp*	*83.67%*
Repbox	*1,375,505*	*1577.687 bp*	*57.40%*

Table 17, Overlapping Loci Between V2 (RepeatMasker v4.1 + RepeatModeler v2) & Repbox Annotations

Species	*A. atlantica*	*A. eriantha*
Count of overlapping loci	*960,005*	*425,200*
Count of BPs in overlapping loci	*1,661,681,792*	*398,556,665*
Total percentage of overlaps in BPs	*1,661,681,792/3,673,044,503 = 0.452399 * 100 = ~45.24%*	*398,556,665/3,776,743,233 = 0.1055292 * 100 = 10.55%*

The issue of fragmentation when performing *de novo* annotation or the use of multi-sourced gene features is not novel **[123]** as noted in *Salzberg et al*, and it represents one of the many challenges of characterizing genomes. When we consider how to prevent and/or correct some of these annotation issues, one approach is to implement context-awareness across all methods of detection. A defragmentation step following identification with de novo tools could improve the assigned boundaries of element candidates, as seen in the implementation of RECON **[87]** and RepeatScout **[82]**. In these software packages, consideration of context significantly improved identification of repetitive elements. In *Girgis et al* comparisons between RECON and RepeatScout revealed improvements to detection of TE sequences in *Homo sapiens* with RECON deriving a 55% sensitivity to TEs and RepeatScout deriving a 62% sensitivity to TE sequences**[86], [87], [102]**, based upon libraries of manually curated LTRs sequences as references. RepeatScout differs from RECON in its use of a fit-preferred alignment score that optimizes the boundaries of identified TEs, resulting in an improved consensus of TEs identified.

The current solution implemented in Repbox to address fragmentation is the removal of excessive sequences and protein-coding regions; however we believe Repbox could be further improved by adjusting its approach to mimic those as adopted by the authors RECON and RepeatScout.

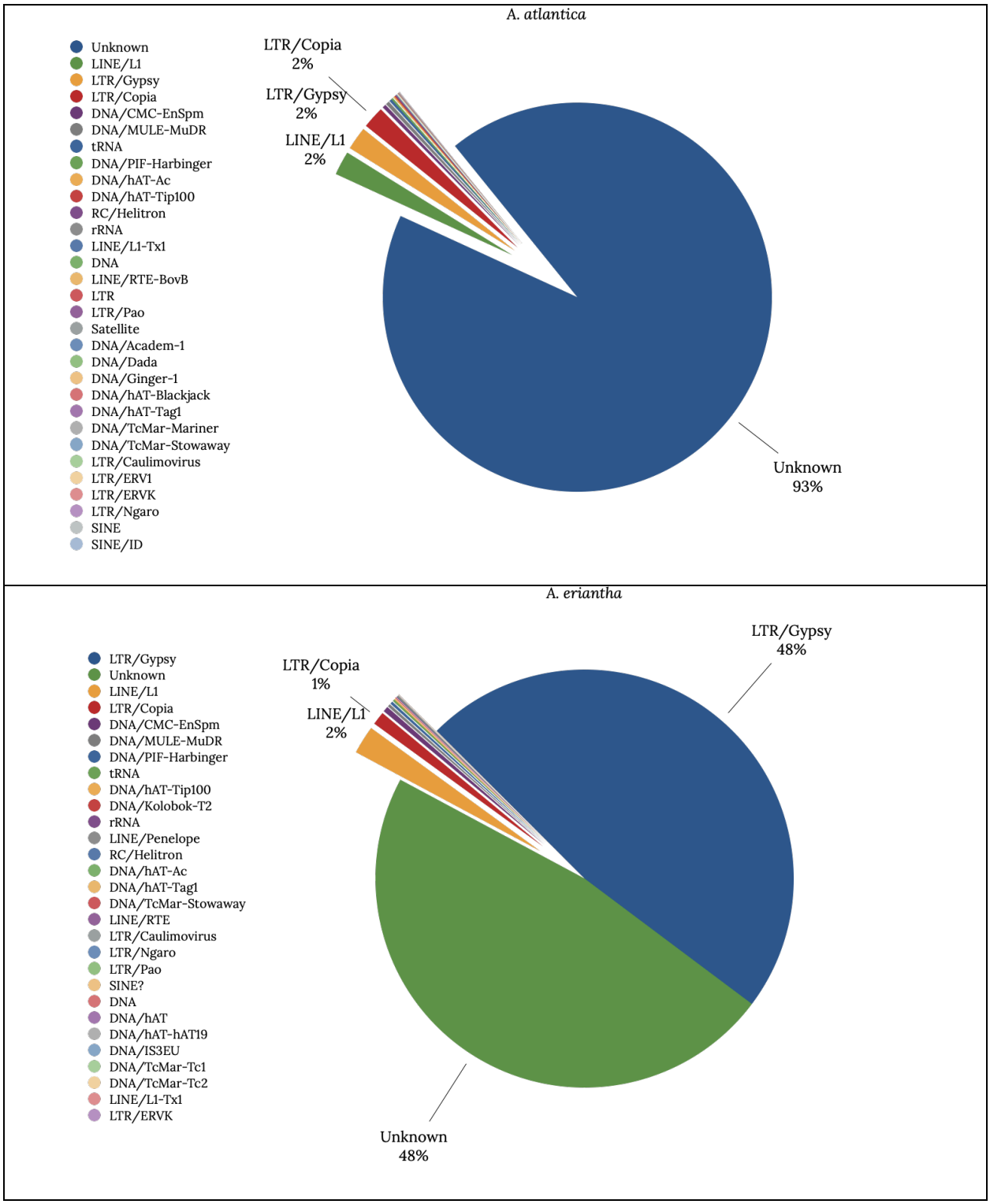
From our analysis, the modifications in RepeatModeler v2 have the potential to create unintended biases in the resulting repeat profiles and subsequent masking in RepeatMasker. We hypothesize the primary reasoning for these biases is fragmentation. In this context, fragmentation in TE annotation is the result of multiple annotation tools identifying candidate repeats, but differ in the feature coordinates that vary or are slightly different. As a result, these differences are carried along the pipeline, ending with unresolved elements in the formation of a consensus annotation that is fragmented **[123]**. Shorter sequences, such as those identified in the Repbox pipeline, are potentially the drivers of fragmentation, as many of these sequences are difficult to classify and are subsequently classified as unknown. These elements, though potentially fragments of actual TE families, due to their short sequence and fragmentation, are not merged into a complete repeat family in RepeatMasker, instead treated as a separate elements and unclassifiable. When there is an abundance of these sequences, the result is an annotation defined not only as fragmented, but one containing large counts of unknown repeats unable to be resolved or identified as was observed in RepeatModeler v2 annotations of *A. atlantica* and *A. eriantha*.

Effect of Clustering on Annotations

Another potential source of the shifts in repeat classifications observed in the Repbox pipeline from analysis using RepeatModeler V2, is the clustering required for removal of redundant sequences and creation of repeat families. As necessary as clustering analysis is for sequence redundancy, the process can further complicate fragmentation. Clustering can

complicate fragmentation, as clustering presents the potential for a loss of elements due to natural variances between families of sequences. This effect is illustrated in the implementation of VSEARCH within the Repbox pipeline. Sequences that possess lower sequence similarities are potentially lost in subsequent clustering steps of family creation, where families are generated using a 80% similarity to form a consensus that represents a TE family. Elements with a lower identity to the consensus sequence are rejected by VSEARCH due to its underlying heuristic centroid-based algorithm to derive clusters **[79]**. A large percentage of sequences rejected from clustering in VSEARCH subsequently create a singleton cluster, or a cluster consisting of a single sequence, with many of these singletons later classified at a lower rate. This is observed in data derived from *A. atlantica* and *A. eriantha* clustering analysis, with a total singleton count of 10,265 (13.5% of sequences and 71.3% of clusters) that resulted in 93% of singletons later classified as unknown in RepeatMasker (*Figure 30*). In *A. eriantha*, the effect of clustering is less noticeable, with a total singleton count of 14,980 (29.3% of sequences and 71.0% of clusters) and a significant proportion (48% of singletons) were subsequently classified as LTR elements. The classification of singleton clusters as unknown was still a large proportion of element assignment in *A. eriantha* only less pronounced in comparison to *A. atlantica*. Increasing the sequence identity required in VSEARCH will retain more variant sequences, however, could create additional singletons that are potentially classified as unknown. With this being the potential result of increasing identity, we choose to remain with an 80% sequence identity, as this is also a commonly utilized threshold by biologists as it confers a high degree of DNA sequence conservation.

Figure 31, Percentage of classifiable VSEARCH clusters in A. atlantica and A. eriantha


A. atlantica
Unknown
LINE/L1
LTR/Gypsy
LTR/Copia
DNA/CMC-EnSpm
DNA/MULE-MuDR
tRNA
DNA/PIF-Harbinger
DNA/hAT-Ac
DNA/hAT-Tip100
RC/Helitron
rRNA
LINE/L1-Tx1
DNA
LINE/RTE-BovB
LTR
LTR/Pao
Satellite
DNA/Academ-1
DNA/Dada
DNA/Ginger-1
DNA/hAT-Blackjack
DNA/hAT-Tag1
DNA/TcMar-Mariner
DNA/TcMar-Stowaway
LTR/Caulimovirus
LTR/ERV1
LTR/ERVK
LTR/Ngaro
SINE
SINE/ID
LTR/Copia 2%
LTR/Gypsy 2%
LINE/L1 2%
Unknown 93%
A. eriantha
LTR/Gypsy
Unknown
LINE/L1
LTR/Copia
DNA/CMC-EnSpm
DNA/MULE-MuDR
DNA/PIF-Harbinger
tRNA
DNA/hAT-Tip100
DNA/Kolobok-T2
rRNA
LINE/Penelope
RC/Helitron
DNA/hAT-Ac
DNA/hAT-Tag1
DNA/TcMar-Stowaway
LINE/RTE
LTR/Caulimovirus
LTR/Ngaro
LTR/Pao
SINE?
DNA
DNA/hAT
DNA/hAT-hAT19
DNA/IS3EU
DNA/TcMar-Tc1
DNA/TcMar-Tc2
LINE/L1-Tx1
LTR/ERVK
LTR/Gypsy 48%
LTR/Copia 1%
LINE/L1 2%
Unknown 48%

Phylogenetic Analysis of Gypsy Elements in *A. atlantica* and *A. eriantha*

LTR/Gypsy elements were among the most abundant TE family observed in both *A. atlantica* and *A. eriantha* across most analysis conducted using RepeatModeler (v1.0.11 & 2.0.1) and RepeatMasker (versions 4.07 & 4.1). LTR elements are generally one of the larger repeat classes identified in a given genome, but this is particularly consistent in higher order plants **[26], [29], [141], [142]**. In total, 296,460 elements were found to belong to LTR/Gypsy in Repbox annotations for *A. atlantica* and *A. eriantha*. Due to this abundance, we were motivated to investigate phylogenetic relationships between subsets of identified LTR/Gypsy elements in both *A. atlantica* and *A. eriantha* to understand how these elements proliferated in each genome. Due to the large number of elements identified in both species, a subset of sequences were analyzed between *A. atlantica and A. eriantha*. Sequences were determined by SW (Smith-Waterman) scores assigned by RepeatMasker that refers to the Smith-Waterman-Gotoh **[143]** alignment score (*default threshold = 250*) of a specific TE aligned to known elements referred to within RepBase and Dfam databases. The fifty top-scoring elements were extracted and aligned via multiple sequence alignment using MAFFT (7.453) **[138]**. From this alignment, phylogenetic trees were generated using FastTree (version 2.1.11) **[139]** to infer phylogenetic relationships. Parameters -nt -gtr, in reference to nucleotide and general time-reversibility assumptions were applied to tree generation, with final visualization utilizing ggtree (version 2.2.4) **[144]**. The derived tree is an unrooted tree (*Figure 33, Phylogenetic Analysis of Gypsy LTR elements*) constructed using a heuristic neighbor-joining method built into FastTree, incorporating tree length reduction and maximization of the tree's likelihood.

Table 18, Contig Representation of High-Scoring Sequences in A. Atlantica and A. Eriantha

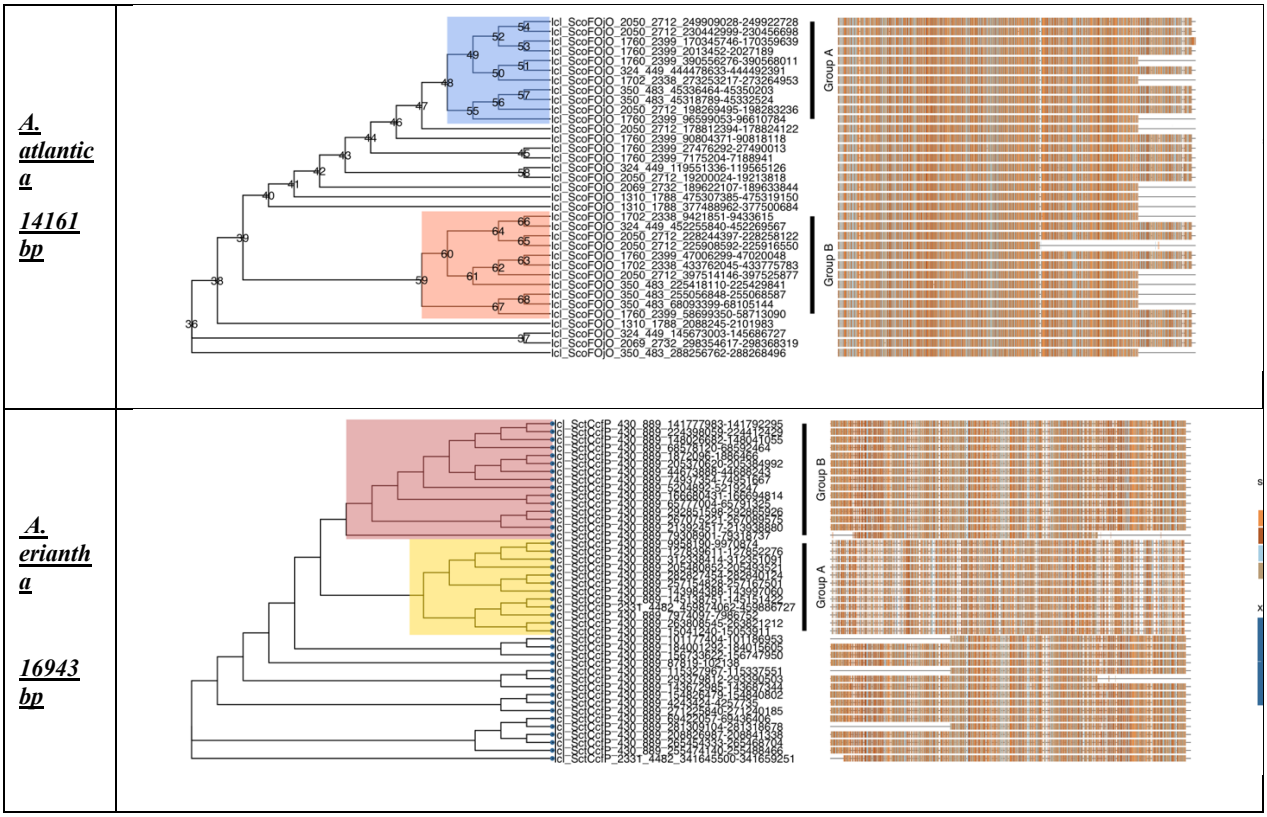
	Contig	Count
A. atlantica	*lcl_ScoFOjO_1310_1788*	*5*
	lcl_ScoFOjO_1702_2338	*5*
	lcl_ScoFOjO_1760_2399	*11*
	lcl_ScoFOjO_2050_2712	*12*
	lcl_ScoFOjO_2069_2732	*3*
	lcl_ScoFOjO_324_449	*4*
	lcl_ScoFOjO_350_483	*10*
A. eriantha	*lcl_SctCcfP_430_889*	*47*
	lcl_SctCcfP_2331_4482	*3*

Derived trees represent a subset of LTR/Gypsy elements possessing the highest SW (Smith-Waterman) alignment score assigned from alignments within RepeatMasker, and therefore represent a small representative perspective of elements analyzed in our analysis. High-scoring LTR/Gypsy elements were derived from a varied set of contigs in *A. atlantica*, (*Table 18*). The locations of LTR/Gypsy elements vary across the genome, potentially indicating a high rate of proliferation throughout the genome of *A. atlantica*, similar to prior work exploring proliferation patterns of LTR elements in plant species **[145], [146]**. Contrary to *A. atlantica*, *A. eriantha* produced less variance in the contig location of LTR/Gypsy elements, as only two contigs were represented in high-scoring elements filtered from our dataset. This potentially limits our phylogenetic analysis, as varied representation of elements across the genome can provide additional insight into polymorphisms and/or mutations in elements localized to specific contigs or chromosomes.

A. atlantica possessed clusters of sequences that are assumed to be evolutionarily close, however the unlocalized nature of these sequences indicates high dispersion throughout the genome, again potentially indicating an explosion of LTR proliferation. In comparison, *A. eriantha*'s tree topology was distinct, displaying single extending branch representing a group of

LTR/Gypsy elements (Group B/node 71), that is far removed from the majority of elements depicted in the derived tree. Elements contained within this far-removed cluster are almost exclusively located on chromosome 1 (contig lcl_SctCcfP_430_889). The distribution of LTR/Gypsy elements differs from patterns of dispersal observed in *A. atlantica* LTR, where elements originated from a greater diversity of contigs. The confinement of the majority of A. eriantha LTRs to the larger contigs potentially suggests biases as this is not representative of all contigs within our dataset. The observed divergence of this cluster of sequences potentially indicates distinct LTR/Gypsy elements present within the *A. eriantha* genome. MSA of coding regions revealed distinctive characteristics that were later used to group aligned sequences as in *A. atlantica*, where a CD-Search query predicted several domains (RNase_HI_like (cd09279); e-value of 2.58e-43. Visualization of this region of from the multiple sequence alignment is illustrated in *(Figure 33),* with a window 6200-6600 bps.

Figure 32, Phylogenetic Analysis of Gypsy LTR elements in A. atlantica and A. eriantha



Conclusions

Avena diploids *A. atlantica* and *A. eriantha* present a unique opportunity to gain further genetic insight into these recently assembled genomes, and with additional analysis by our novel pipeline Repbox, we gained an added understanding to the repetitive element composition of both *A. atlantica* and *A. eriantha* that extends beyond initial repeat characterization performed in *Maughan et al*. Our analysis revealed a relative shift in the repeat composition in these species in count and diversity, with both species presenting significant increases of LTR elements and gains in previously unobserved repeat families. These observed shifts in repeat family composition represent potentially uncharacterized repetitive structures and led insight to the overall relationships shared between members of the *Avena* genus as a whole. Improving the repeat annotation of a genome can potentially improve gene annotation as well. Our initial analysis into

the phylogenetic relationships of repeats found in these to distinct *Avena* diploid genomes show that proliferation of LTR elements is not necessarily consistent across the Avena species. This work underlies the importance of characterizing repetitive elements across *Avena* species, with the focus of future work revolving around investigations into the unexplored relationships repetitive elements across the all *Avena* diploid species.

Chapter 4: PHYLOGENETIC COMPARISON OF LTR RETROTRANSPOSONS IN *AVENA* DIPLOID GENOMES

Introduction

Class I transposable elements, also termed retrotransposons, are mobile segments of DNA characterized by their use of reverse transcription to proliferate and migrate throughout a given genome **[40]**. Categorically, retroelements are comprised of three primary element families, (1) Long terminal repeats (LTRs), (2) Long Interspersed Nuclear Elements (LINEs), and (3) Short Interspersed Nuclear Elements (SINEs). Structurally, LTRs are distinct from other retroelements in that they possess long terminal repeat regions that can extend upwards of several kilobases in length, making these elements some of the largest transposable elements ever observed **[37]**. Evolutionarily, these elements have been found to derive ancestry from retroviruses, with many genes located in the coding regions of these elements possessing high similarity to components observed within retroviral coding regions **[147]**. Scientifically, LTRs have long been of interest, as these sequences are commonly one of the largest classes observed in almost every genome **[27]**, with previous studies determining that these elements contribute to large proportions of many higher ploidy plants species, namely *O. sativa (17%)* **[148]**, *Z.mays (75%)* **[149]**, and *T. aestivum* (70%) **[141]**.

With advances in sequencing technologies, the opportunity to survey complex DNA sequences such as LTRs and other repetitive elements is more feasible allowing us to answer deeper biological questions surrounding these elements that were previously limited due to computational and biological limitations. Repetitive genome features can reveal relationships within previously uncharacterized organisms or the evolution of these mobile elements and the potential contributions they make to the overall genome evolution of a species. In this study, we

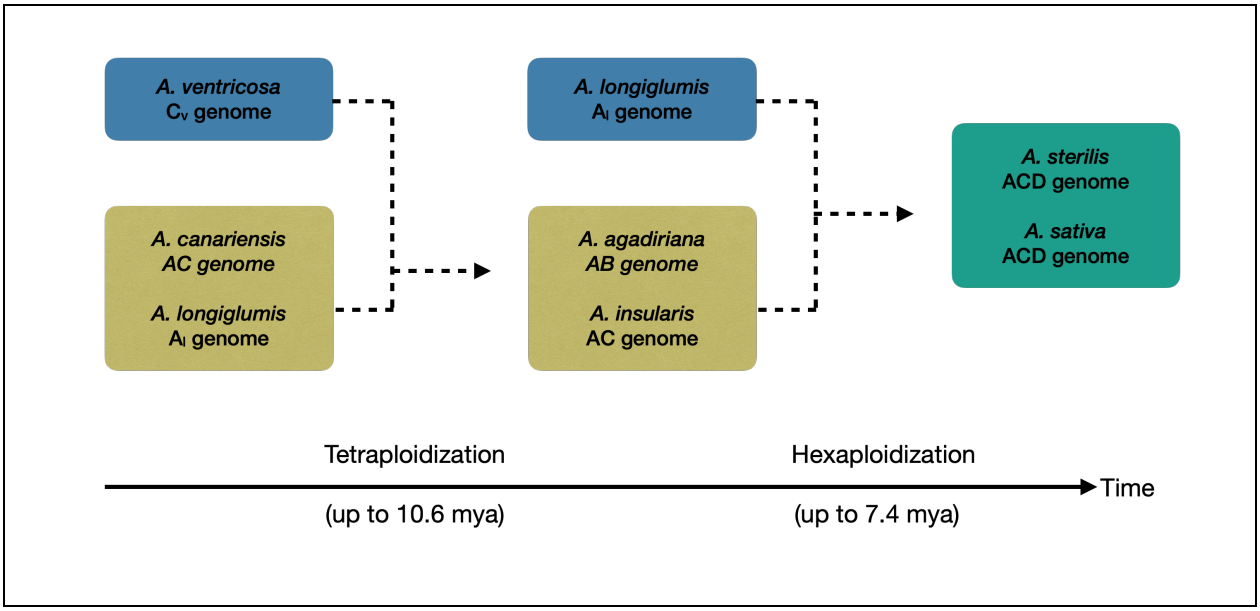
survey sequenced two *Avena* species, an A-genome diploid *A. longiglumis* and a C-genome diploid *A. ventricosa* These sequences were aligned to the diploid *A. atlantica* (A-genome) and *A. eriantha* (C-genome) reference genomes. These species allows us to explore family-specific repetitive elements within four *Avena* diploids, focusing on characterizing and investigating LTRs identified within these species.

Recent Phylogenetic Studies in Avena

The *Avena* genus is a diverse family comprised of diploid, tetraploid and hexaploid species of oat, featuring varieties that are both domesticated and undomesticated **[9]**. The *Avena* polyploidy complex consists of four distinct genome types: AA, BB, CC, and DD. The hexaploid, *A. sativa L.*, is composed of sub-genomes type A, C and D **[9]**. A progenitor DD diploid has not yet been identified, however current theories suggests that the D-genome may actually be a more A-like genome that diverged from the currently classified A-genomes and an extant ancestor of Avena; the genome sequencing efforts certainly support this theory **[9]**. Studies looking at the evolutionary relationships between *Avena* species postulate that the most likely timeline of divergence of species that led to subsequent polyploidization events occurred approximately 7.4Mya–10.6 Mya **[129]** *(Figure 33)*. In their analysis, *Fu et al* investigated lineages of twenty-five different species of *Avena* and created phylogenies based on SNP data from chloroplast, mitochondrial and hybrid chloroplast/mitochondrial DNA **[129]**. *Fu et al* also identified novel maternal pathways believed to represent a maternal source of the tetraploid progenitor from which *A. sativa L.* acquired a portion of its genome, as well as detailed divergence dating of several C-genome clades to approximately 8.5-9.5 and 19.9-21.1 Mya. Other phylogenetic studies include that of *Peng et al*, where relationships within *Avena* were investigated using protein-coding region data of a conserved gene, nuclear Pgk1 (nuclear plastid

3-phosphoglycerate kinase) **[150]**. This study revealed significant deletions in versions of this gene across species, and these deletions confirmed prior divergence estimates between the A and C subgenomes in Avena. In addition, *Peng et al* found that the C_p subgenomes and polyploid species are closer evolutionarily that previously determined

Figure 33, Proposed Scenario for the Maternal Origins of Hexaploid Oat



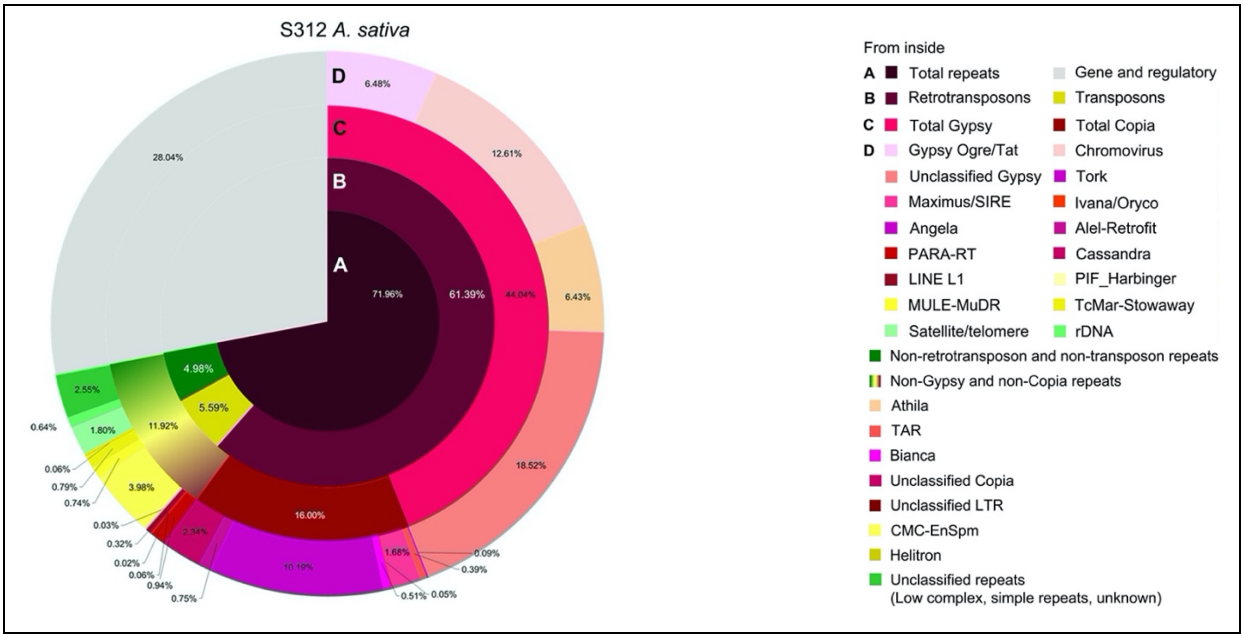
Fu et al (2018) [129]

Using repetitive sequence element information to study genome relationships between *Avena* species is limited due to a lack of genomic resources. However, more recent works such as a repeat analysis study by *Liu et al (2019)* sought to characterize the repeat landscape of several A_s *Avena* species (*A. brevis*, *A. strigosa*, and *A. hirtula)* to inform on the repeat composition of hexaploid oat *Avena sativa L.* (AACCDD) **[131]**. The repeat landscape of *A. sativa* was determined to be approximately 72% of the genome; of which 96% was determined to be related to known transposable elements and tandem repeat motifs (*Figure 34). K-mer* analysis revealed that repeat families specific to hexaploid *A. sativa* were unobserved in other diploid A-genomes (*A. brevis, A. strigosa,* and *A. hirtula*). A large portion of the *A. sativa* genome was determined to

contain retroelements (~61.4%), with LTR/Gypsy elements being the primary retroelement class observed *(Figure 34)*. This finding was consistent with studies of other polyploid plant systems with multiple subgenomes **[141]**.

Due to the abundance of repeats that span A and D chromosomes, *Liu et al* speculated that the diverged repeats of the D-genome in *A. sativa,* compared to those observed in A-genome diploids *A. strigosa*, *A. atlantica* and *A. brevis,* potentially represent a basis of evolutionary separation of A and D-genome progenitors. Given the increase in sequencing capabilities as well as the genome sequences of *A. atlantica* and *A. eriantha*, the phylogenetic relationships and repeat landscapes of family-specific repetitive elements, specifically those within the A-genome and C-genome *Avena* diploids, can be explored. These analyses provide additional insight into the evolutionary processes by which these elements diverged and became sub-genome-type specific.

Figure 34, Frequency of major repetitive DNA classes in Avena sativa



Liu et al [131]

Materials and Methods

A. atlantica (CN 7277) and *A. eriantha* (CN 19328) were sequenced using PacBio 122 RSII + Sequel and 54 SMART Sequel cells , generating a total of 31,544,396 and 28,257,346 PacBio reads and a coverage of approximately 84x and 71x coverage for *A. atlantica* and *A. eriantha* respectively **[130]**. Canu **[135]** was used in the assembly of *A. atlantica* and *A. eriantha*, with the resulting assemblies consisting of 3,914 and 8,067 contigs respectively, and an N50 of 5,544,947 and 1,385,002. Assemblies were further improved by Chicago + Dovetail Hi-C **[136]**, resulting in a scaffold N50 of 513.2Mb, and an L50 of 4, spanning a total sequence length of 3.685 Gb for *A. atlantica* and a scaffold N50 of 534.8 Mb, an L50 of 4, and spanned a total sequence length of 3.778 Gb for *A. eriantha* **[130]**. *A. longiglumis* (CN 58138),) and *A. ventricosa (BYU_143* sequencing data was derived from a prior diversity panel of Avena. Extracted DNA was sequenced at the Beijing Genomic Institute (BGI; Hong Kong, China) for 2 × 150 bp paired end (PE) sequencing from standard 500-bp insert libraries.

Generation of Reference-guided Assemblies

For analysis of *Avena* subgenomes, *A. atlantica* (CN 7277) and *A. eriantha* (CN 19328) were used as A-genome and C-genome references for generation of consensus sequences and identification of genome-specific repetitive elements of A-subgenome *A. longiglumis* (CN 58138), and C-subgenome *A. ventricosa (BYU_143)*. Quality control (QC) of raw reads from *A. longiglumis* and *A. ventricosa* were trimmed with adaptive-trimming tool Sickle (version 1.33), then aligned to *A. atlantica* and *A. eriantha* using Bowtie2 (version 2.4.1) **[151]** . Samtools (version 1.10) **[152]** and EMBOSS seqret (version 6.6.0) **[153]** were used to generate and convert final consensus sequences to fasta format. Final analysis of each genome consisted of RepeatModeler (version 2.0.1) **[102]** analysis for *de novo* identification of repeat families *(Code 15)*

Code 15, Commands for reference-guided annotation

```
INPUT1="LO_149_DSW61536-V_HL5TTCCXY_L1_1.fq"
INPUT2="LO_149_DSW61536-V_HL5TTCCXY _L1_2.fq"
BASE=$(basename $INPUT1 | cut -f 1 -d '.'| cut -f -4 -d '_')
sickle pe -f $INPUT1 -r $INPUT2 -t sanger -o $( basename $INPUT1 | cut -f 1 -d '.')_trimmed.fq
-p $( basename $INPUT2 | cut -f 1 -d '.')_trimmed.fq -s $BASE'_'trimmed.unpaired.fq -q 5
READ1=$( basename $INPUT1 | cut -f 1 -d '.')_trimmed.fq
READ2=$( basename $INPUT2 | cut -f 1 -d '.')_trimmed.fq
bowtie2-build $GENOME $INDEX
bowtie2 -p $THREAD -x $INDEX -1 $READ1 -2 $READ2 > $BASE.sam
samtools view -b -S $BASE.sam > $BASE.bam
samtools sort -m 1000000000 $BASE.bam -o $BASE.sorted.bam
samtools index $BASE.sorted.bam
##samtools bam2fq --no-aligned --force --strict -o readset_ref_unmapped#.fq
readset_ref_bwa.bam
samtools mpileup -E -uf $GENOME $BASE.sorted.bam | bcftools call -c | vcfutils.pl vcf2fq >
$BASE'_'cons.fq
seqret -osformat fasta $BASE'_'cons.fq -out2 $BASE'_'cons.fa
```

Extraction of family-specific sequences

In-depth repeat analysis was performed using RepeatModeler to derive predicted superfamilies of retrotransposons. Following formation of all representative repeat sequences, family-specific repetitive elements were compared based upon their genome-type (*A. eriantha*

compare to *A. ventricosa* and *A. atlantica,* compared to *A. longiglumis)*. The output of RepeatModeler, a classified fasta sequence, was separated into each TE family observed during annotation by a custom python script that uses BioPython (version 1.77) **[154]**. Analysis resulted in separated fasta sequences derived from the final annotation fasta that were grouped by family in a directory labeled 'families' (*Code 16, parsify.py - Utility script for parsing family-specific TEs*)

Code 16, parsify.py - Utility script for parsing family-specific TEs

```
# Parsify.py – a script for parsing Repbox annotation fasta into directories
from Bio.SeqIO.FastaIO import SimpleFastaParser
from Bio.SeqIO.QualityIO import FastqGeneralIterator
from Bio import SeqIO
import argparse
import numpy as np
import pandas as pd
import os
parser = argparse.ArgumentParser(description='parsify -i <file>.fasta')
parser.add_argument('mfilename', metavar='<filename>.fasta', type=str)
args = parser.parse_args()
sample = open(args.mfilename + '.table', 'w')
for record in SeqIO.parse(args.mfilename, "fasta"):
    listElement.append(record.id.split('#') [1])
families = (set(listElement))
for i in families:
    print(i + "," + str(listElement.count(i)), file=sample)
sample.close()
os.mkdir('families')
for i in families:
    filepath = os.path.join(os.getcwd() + '/families', str(args.mfilename) [0:16] + '_' +
str(i).replace("/","-") + '.fasta')
    for record in SeqIO.parse(args.mfilename, "fasta"):
        if record.id.split('#') [1] == str(i):
            print(">" + record.id, file=TE)
            print(record.seq, file=TE)
    TE.close()
```

Identification of Species-Specific LTRs

Families of LTRs investigated include *Ty3*-Gypsy and *Ty1*-Copia, as they were the most abundant retroelements identified in both diploid genomes (see Chapter 3). Genome-specific LTRs were identified via BLAST search against other *Avena* genomes using LTRs identified in each genome as query sequences. Code executing this analysis is outlined below *(Code 17,*

Identification of Genome-specific LTRs in A. atlantica, A. longiglumis, A eriantha and A. ventricosa).

Code 17, Identification of Genome-specific LTRs in A. atlantica, A. longiglumis, A eriantha and A. ventricosa.

```
# Generation of local BLAST Database for each genome
makeblastdb -in Avena_atlantica.fa -parse_seqids -title “Avena_atlantica” -dbtype nucl

# Query Identified Genome-specific LTRs to BLAST database
blastn -query ER.rename.fa -perc_identity 95 -db Avena_atlantica.fa -outfmt 6 >
ER_AT_results.out
```

Multiple Sequence Alignment & Nucleotide-based Phylogenies

Multiple sequence alignment of species-specific LTRs was performed using MAFFT (version 1.24) [138]. Parameters for MAFFT include setting –retree, the tree iteration parameter, to “2” for efficiency. Command line versions of both MAAFT and FastTree (version 2) [139] as executed using the following commands in *Code 18*. Comparisons were replicated and performed intraspecies, e.g. all LTR elements of all subfamilies within an organism, with interspecies comparisons consisting of counts, structural characteristics and potential homologous LTRs that present with high sequence similarity. Finally, all alignments were trimmed using trimAl (version 1.2rev59) [155] to remove overhang in the sequence alignments.

Code 18, Multiple Sequence Alignment of Sequences with MAFFT & Phylogenetic Analysis with FastTree

```
# Example commands used for LTR-Copia elements in Avena_eriantha
mafft --retree 2 Ae_Gypsy.fa > Ae_Gypsy_output_alignment
# Removal of invalid characters
grep -rl ":" Ae_Gypsy_output_alignment | xargs sed -i "" 's/:/-/g'

# Trimming of nucleotide MSA
trimAl -in Ae_Gypsy_output_alignment > Ae_Gypsy_output_alignment_trim

# Generation of Neighbor-joining Tree
FastTree -gtr -nt Ae_Gypsy_output_alignment_trim > Ae_tree
```

Prediction of Conserved Reverse Transcriptase Domain and Protein Phylogenies

LTR sequences derived from RepeatModeler v2 were translated with EMBOSS transeq (version 6.6.0.0) [153], followed by identification of conserved domain fragments using HMMER hmmalign (version 3.0) [156]. Hidden Markov Models specific to (RT) Reverse Transcriptase derived from GYDB (version 2.0) [157] were used to identify and isolate corresponding domains within extracted LTR sequences. Outputs from HMMER hmmalign were formatted as A2M, a derivative of fasta format *(Code 17, Commands for Isolation of Conserved Reverse Transcriptase)*. Sequences were trimmed using trimAl (version 1.2rev59) [155], sequences possessing more than a 50% fraction of gaps were removed from the final alignment, and resulting RT tree was generated using FastTree (version 2) [139].

Code 19, Commands for Isolation of Conserved Reverse Transcriptase

```
# Transeq translation of RepeatModeler-derived fasta files
Transeq LTR.fa > LTR.faa

# HMMAlign GYDB Gypsy HMM
hmmalign --trim --outformat A2M ~/GyDB_collection/profiles/RT_gypsy.hmm VE.faa > Gypsy_RT

# Trimming of Multiple Sequence Alignment, Removal of high-gap sequences
trimAl -in Gypsy_RT -gt 0.5> AT_output_alignment_trim

# Tree Generation
FastTree Gypsy_RT_trim_0.5.aln > Gypsy_trimmed_output_alignment_tree
```

Results and Discussion

Families of repeats identified by RepeatModeler v2 in *A. atlantica*, *A. eriantha*, *A. ventricosa*, and *A. longiglumis* were predominately classified as unknown, representing

approximately 75% of all elements identified in each genome *(Table 19)*. The high percentage of unknown elements was expected as our analysis only took into account *k*-mer counts and elements that were able to be classified by RepeatClassifier, the submodule within RepeatModeler responsible for classification of potential repeats. We choose this method as these *k*-mers represent the most abundant TEs encountered by RepeatModeler and allowed for rapid characterization of our genomes, as well as reducing the complexities in the data that would result from attempting to align all the shotgun resequencing data to reference genomes without repeat Masking. For consistency in interpreting the data, the *A. atlantica* and *A. eriantha* data were treated in the same manner. Excluding the unknown classification, LTR/Gypsy and LTR/Copia represented the largest superfamily across all species, with 107, 104, 158 and 53 distinct families of *Ty1*-Copia elements and 154, 174, 237, and 130 distinct families of *Ty3*-Gypsy elements (*Table 19*). Other repeat families reported include CMC-EnSpm, MULE-MuDr and L1 elements, with all three families contributing an average of 2.5%, 1.0% and 2.6% of repeats in each genome respectively. Total DNA elements, specifically CMC-EnSpm and MULE-MuDr (Mutator) were observed at a slightly lower percentage compared to previous works by *Liu et al* (2019) **[131]**, where DNA elements were observed to represent an average of ~4.5%, demonstrating a potential decrease in DNA elements predicted from our analysis. This may be an artifact of the methods used in each analysis. LINE elements contribute 2.58% on average to the repetitive space for all genomes *(Table 19*). Overall, the distribution of element classes is relatively consistent across each of the genomes.

Table 19, Count of Distinct Families Identified by RepeatModeler v2

	A-Genomes				***C- Genomes***				
TE Family	*AT*		*LO*		*ER*		*VE*		*Average*
	Count	*Percentage*	*Count*	*Percentage*	*Count*	*Percentage*	*Count*	*Percentage*	

DNA/CMC-EnSpm	*49*	*2.99%*	*56*	*2.63%*	*36*	*2.07%*	*25*	*2.59%*	*2.57%*
DNA/hAT-Ac	*3*	*0.18%*	*2*	*0.09%*	*0*	*0.00%*	*0*	*0.00%*	*0.07%*
DNA/hAT-Tip100	*8*	*0.49%*	*0*	*0.00%*	*5*	*0.29%*	*0*	*0.00%*	*0.19%*
DNA/IS3EU	*1*	*0.06%*	*1*	*0.05%*	*1*	*0.06%*	*0*	*0.00%*	*0.04%*
DNA/MULE-MuDR	*22*	*1.34%*	*14*	*0.66%*	*22*	*1.26%*	*8*	*0.83%*	*1.02%*
DNA/PIF-Harbinger	*13*	*0.79%*	*16*	*0.75%*	*19*	*1.09%*	*4*	*0.41%*	*0.76%*
DNA/TcMar-Mariner	*0*	*0.00%*	*0*	*0.00%*	*1*	*0.06%*	*0*	*0.00%*	*0.01%*
LINE/I	*1*	*0.06%*	*0*	*0.00%*	*0*	*0.00%*	*0*	*0.00%*	*0.02%*
LINE/L1	*60*	*3.67%*	*50*	*2.35%*	*50*	*2.87%*	*14*	*1.45%*	*2.58%*
LINE/L1-Tx1	*0*	*0.00%*	*0*	*0.00%*	*1*	*0.06%*	*0*	*0.00%*	*0.01%*
LINE/RTE-BovB	*1*	*0.06%*	*0*	*0.00%*	*1*	*0.06%*	*0*	*0.00%*	*0.03%*
Unclassified LTR	*4*	*0.24%*	*4*	*0.19%*	*0*	*0.00%*	*0*	*0.00%*	*0.11%*
LTR/Caulimovirus	*0*	*0.00%*	*0*	*0.00%*	*1*	*0.06%*	*0*	*0.00%*	*0.01%*
LTR/Copia	*107*	*6.54%*	*158*	*7.41%*	*104*	*5.97%*	*53*	*5.49%*	*6.35%*
LTR/ERV1	*0*	*0.00%*	*2*	*0.09%*	*0*	*0.00%*	*0*	*0.00%*	*0.02%*
LTR/ERVK	*0*	*0.00%*	*0*	*0.00%*	*1*	*0.06%*	*0*	*0.00%*	*0.01%*
LTR/Gypsy	*154*	*9.41%*	*237*	*11.12%*	*174*	*9.99%*	*130*	*13.47%*	*11.00%*
LTR/Ngaro	*1*	*0.06%*	*2*	*0.09%*	*3*	*0.17%*	*0*	*0.00%*	*0.08%*
RC/Helitron	*0*	*0.00%*	*3*	*0.14%*	*0*	*0.00%*	*1*	*0.10%*	*0.06%*
rRNA	*3*	*0.18%*	*5*	*0.23%*	*3*	*0.17%*	*0*	*0.00%*	*0.15%*
Satellite	*1*	*0.06%*	*0*	*0.00%*	*0*	*0.00%*	*0*	*0.00%*	*0.02%*
Simple repeat	*0*	*0.00%*	*1*	*0.05%*	*0*	*0.00%*	*0*	*0.00%*	*0.01%*
Unknown	*1209*	*73.85%*	*1580*	*74.14%*	*1319*	*75.76%*	*730*	*75.65%*	*74.85%*

Table 19. Count of Distinct Families identified by RepeatModeler v2*. AT (A. atlantica), LO (A. longiglumis), ER (A. eriantha), VE (A. ventricosa)*

Characterization Gypsy and Copia Retrotransposons

LTR elements in each of the four genomes are the largest and most abundant class of repeats observed. The number of sequences, average sequence length, minimum and maximum fasta entry length and total length of Gypsy and Copia-specific sequences were calculated by parsing RepeatModeler consensi.fa.classified fasta files and using seqkit (version 0.13.2) **[158]** prior to phylogenetic analysis (Table 20). Gypsy and Copia element length varied substantially, ranging from 72bp to 21,422bps. In comparison to the archetypal LTR structure, the upper range of

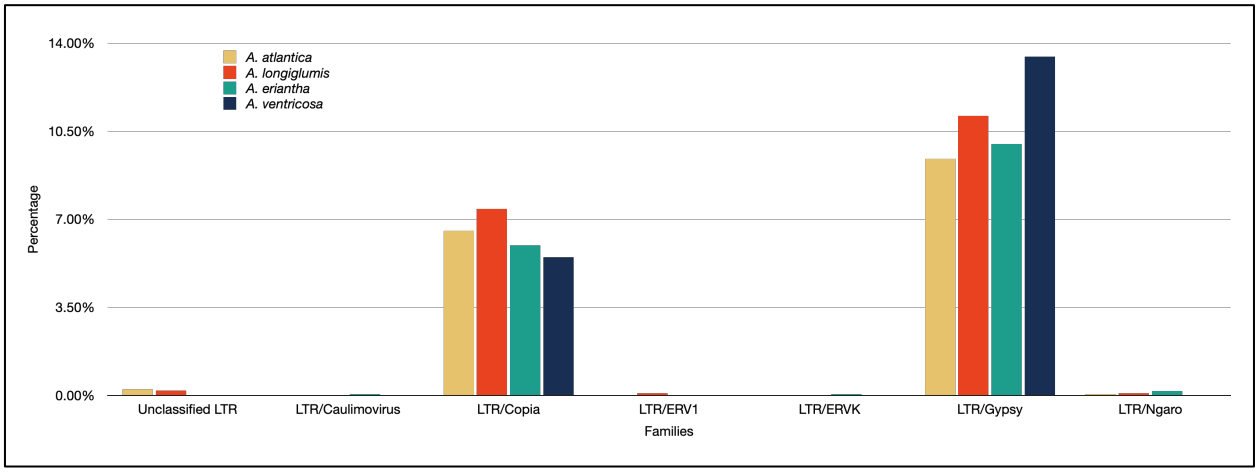
sequence length identified here is similar to some of the larger LTRs observed in other genomes, with the largest LTRs being reported to be upwards of 22kb in length **[37]**. The smaller elements however, we speculated represented partial alignments to LTRs identified by RepeatModeler. This was later confirmed in later nucleotide BLAST analysis seeking to identify genome-specific LTRs, where many of the shorter sequences were found to possess partial homology to larger protein-coding genes within LTRs (Genome specific-elements below). Multiple sequence alignment of LTRs were largely varied in terms of similarity, with the of alignments between sequences ranging widely consisting of regions that aligned well, and others that aligned poorly. element alignments *(multiple sequence alignments below)*. Percentage-wise LTR elements compose the largest proportion of identified repeat families, with Gypsy representing ~11% and Copia 6% on average across all genomes *(Figure 36)*.

Table 20, Gypsy/Copia Sequence Statistics

	file	*num_seqs*	*sum_len*	*min_len*	*avg_len*	*max_len*
A-Genome	*A. atlantica Gypsy*	*154*	*734,264*	*101*	*4,767.9*	*17,121*
	A. atlantica Copia	*107*	*380,745*	*77*	*3,558.4*	*21,422*
	A. longiglumis Gypsy	*237*	*259,861*	*46*	*1,096.5*	*6,131*
	A. longiglumis Copia	*158*	*161,522*	*60*	*1,022.3*	*5,843*
C- Genome	*A. eriantha Gypsy*	*174*	*773,281*	*87*	*4,444.1*	*15,212*
	A. eriantha Copia	*104*	*332,601*	*51*	*3,198.1*	*15,219*
	A. ventricosa Gypsy	*130*	*456,434*	*72*	*3,511*	*11,824*
	A. ventricosa Copia	*53*	*162,187*	*56*	*3,060.1*	*12,067*

Table 20, Gypsy/Copia Sequence Statistics. *Sequence statistics derived from A-genomes: A. atlantica, A. longiglumis; and C-genomes: A. eriantha and A. ventricosa. Headers are outputs from EMBOSS seqkit: num_seq = number of sequences in fasta file, sum_len = sum of all bps in file, min_len = shortest sequence length in fasta, max_len = longest sequence length in fasta file.*

Figure 35, Percentage of LTR Superfamily Elements in Diploid Avena



Phylogenetic Relationships Based on Nucleotide Data

Comparative analysis of A and C-genome *Avena* diploids was performed for a total of 422 Copia elements (A-genome: 265 , C-genome:157) and 695 Gypsy elements (A-genome: 391, C-genome: 304), with phylogenetic trees based on nucleotide and protein data. To observe potential relationships of LTR elements within species, several comparisons of elements derived from each organism was performed: (1) Multiple sequence alignment of LTR within individual species; (2) Analysis of paired A and C-genomes, i.e. *A. atlantica-A. longiglumis* and *A. eriantha-A.*

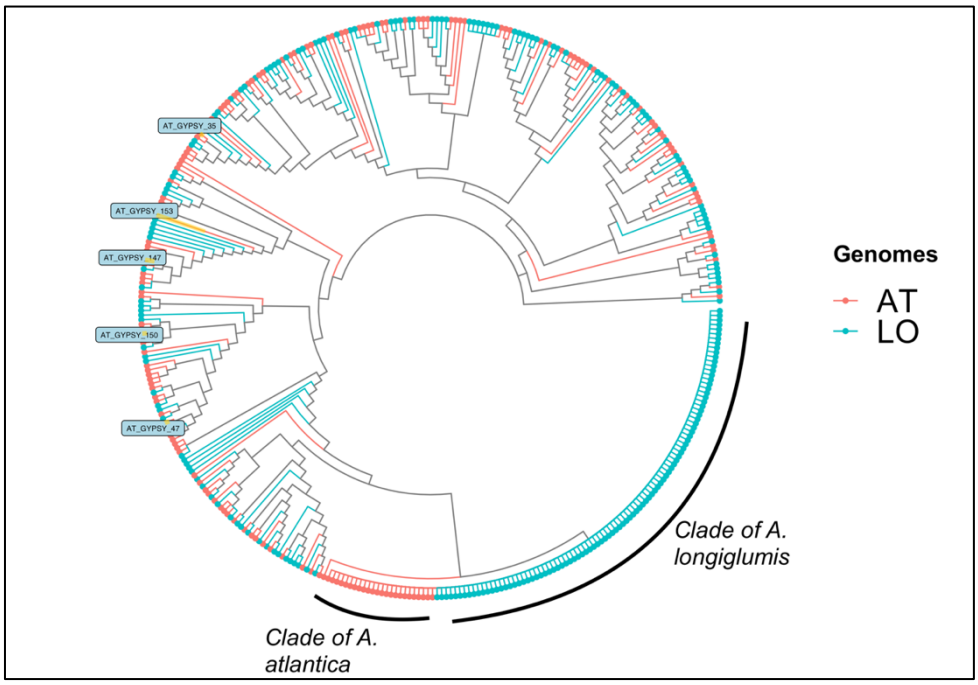
ventricosa with VSEARCH clustering. In both comparative analyses. Alignment was done with MAAFT, followed by trimming of alignments using trimAl, and trees generation using FastTree.

Results of A/C-genome specific comparisons, as illustrated in Figure 36, demonstrate LTR clades within each genome as well as LTRs determined to be genome-specific (labelled blue in Figure 36A, 36B, 36C, *36D*). Overall, genome-specific elements were not confined to clades or groups of elements, with many occurring in heterogenous clades that contained both genomes being compared (i.e. clades of A and C-genome LTRs). The best examples of this are observed in A-genome Copia and Gypsy (Sub-figures 36A and 36B), where genome specific LTRs such as AT_GYPSY_35, AT_GYPSY_150, and AT_GYPSY_47 are grouped among *A. longiglumis* elements, which is a surprising observation considering that elements that are genome-specific would be assumed to be less similar to elements observed in other species and grouped accordingly. Other general observations beyond grouping tendencies of genome-specific elements are the general sizes of clades. C-genome clades appear to be smaller on average than A-genomes (C-Genome; A-Genome:). This average is derived from the ratio of tips to clades that contained AT, ER, LO, or VE at a majority, 50% or more, and instances of this creates an average. A-genomes typically possessed ~17 elements (three large clades of *A. longiglumis across both A genomes in Gypsy and Copia*), while C-genomes typically contained ~3 elements (1 large clade, +20 clades of 3 or less across both C genomes in Gypsy and Copia). In addition, clades tended to be more heterogenous in C-genomes, and clades of elements consisting on only one genome, such as the large clade of *A. longiglumis* observed in Figure 36A , being less common. From this, we concluded that C-genome LTRs appear to be more variable in comparison to A-genomes LTRs, illustrated in Figures 36C,36D. The presence of clades within our phylogenetic tree prompted investigation into the degree of sequence similarity within LTR group and we chose to perform

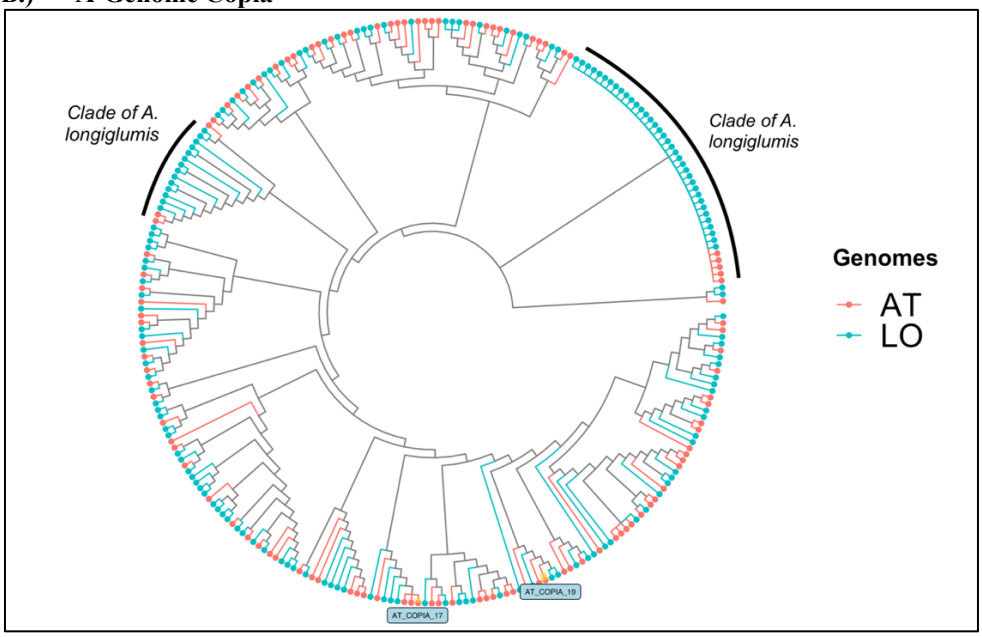
additional clustering of elements to evaluate these sequences. To investigate the extent of sequence similarity, VSEARCH was used to perform clustering. At a threshold of 97% sequence similarity, VSEARCH showed small distinct clusters of A and C-LTRs, with C-genomes *A. eriantha* and *A. ventricosa* possessing noticeably smaller sequence clusters, most existing as singletons (Copia: 109, 69.4% of seqs, 86.5% of clusters; Gypsy: 143, 47.0% of seqs, 73.7% of clusters). A-genomes *A. atlantica* and *A. longiglumis* were notably less variable (Copia: 133, 50.2% of seqs, 77.3% of clusters; Gypsy: 177, 45.3% of seqs, 72.5% of clusters). This was consistent with observation graphically within clades of Figure 36, where C-genome LTRs were variable. However, sequences within A-genomes were also variable, just not to the extent of C-genomes. Comparisons between are not meant to hypothesize ancestry, as these sequences lack support establishing any kind of phylogeny, however noting the presence of distinct groupings of sequences specific to each genome is worth mention, as knowledge of variability of LTRs, or any TE for that matter, can provide a bit more intuition into where we may observe novel/genome-specific sequences. This is discussed below, however, specific clusters from our analysis are highlighted in Figure 36.

Figure 36, Phylogenetic Trees of A/C-Genome Specific Elements in R

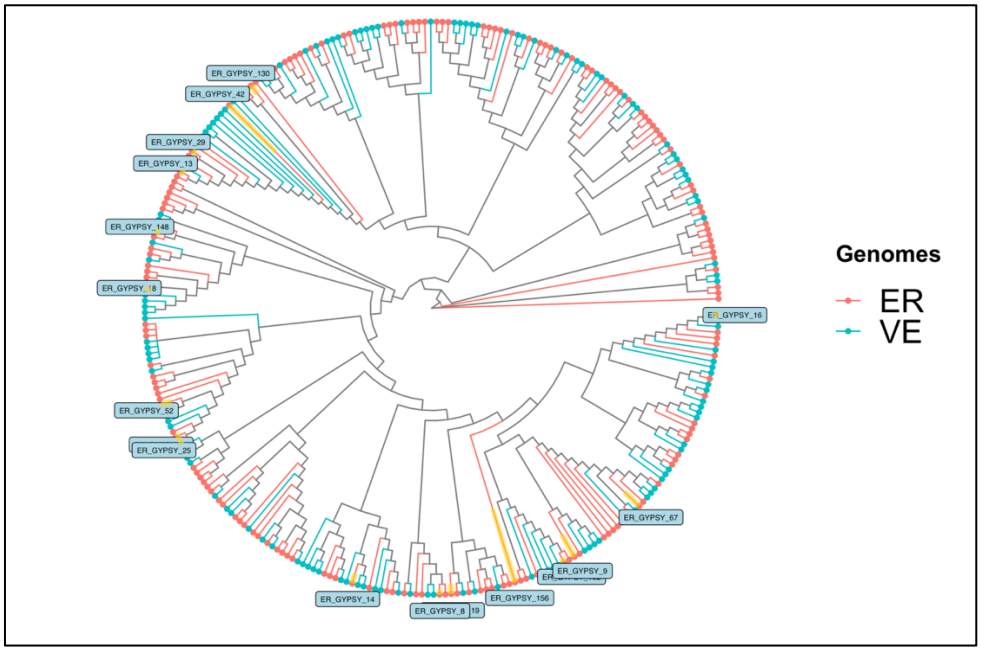
A.) A-Genome Gypsy



B.) A-Genome Copia



C.) C-Genome Gypsy



D.) C-Genome Copia

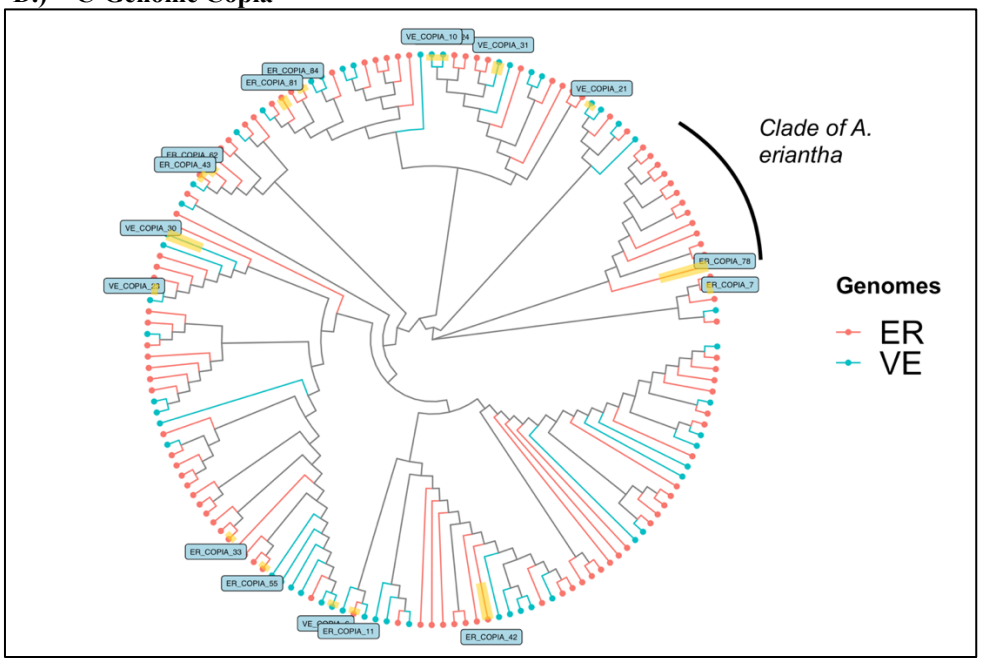
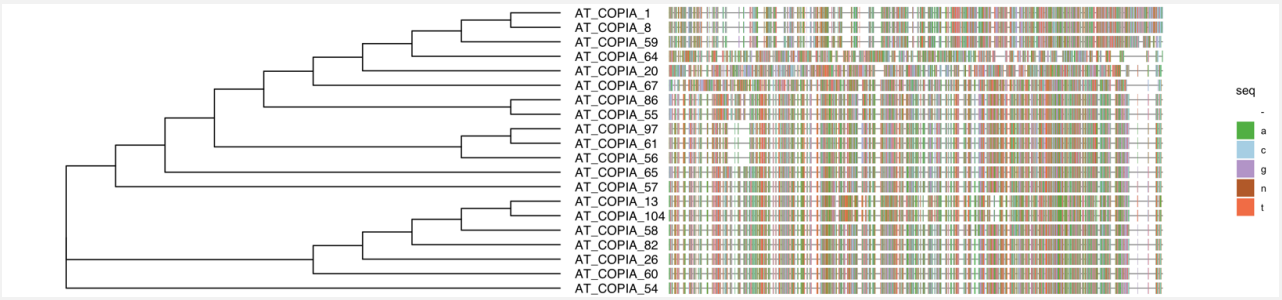


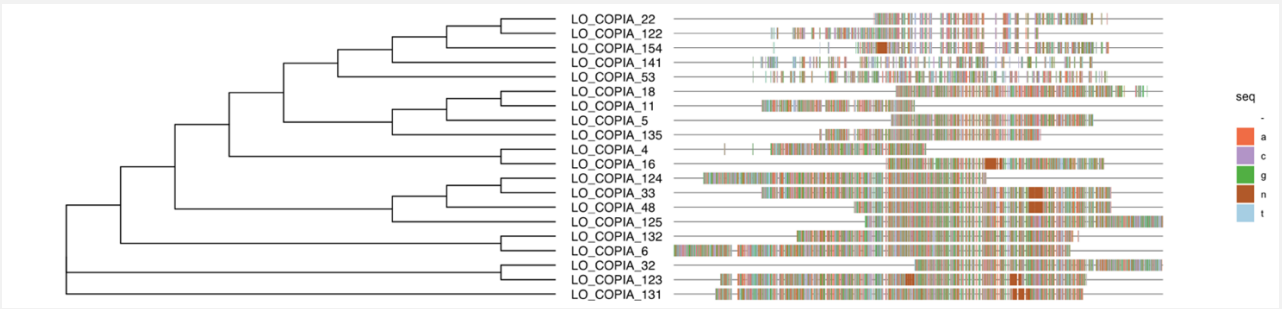
Figure 37, Ty1 Copia Retrotransposons. Phylogenetic Analysis of Nucleotide Sequences

A. atlantica



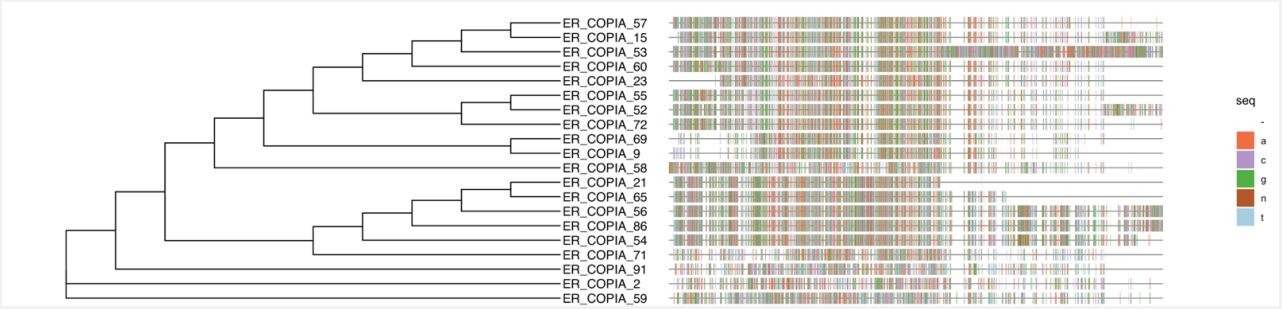
Original phylogenetic tree with 107 tips and 105 internal nodes. Pruned to 20 tips and 18 nodes; 33610 bp alignment, highest scoring hit for reverse transcriptase with RT_LTR (cd01647); e-value of 2.36e-85.

A. longiglumis



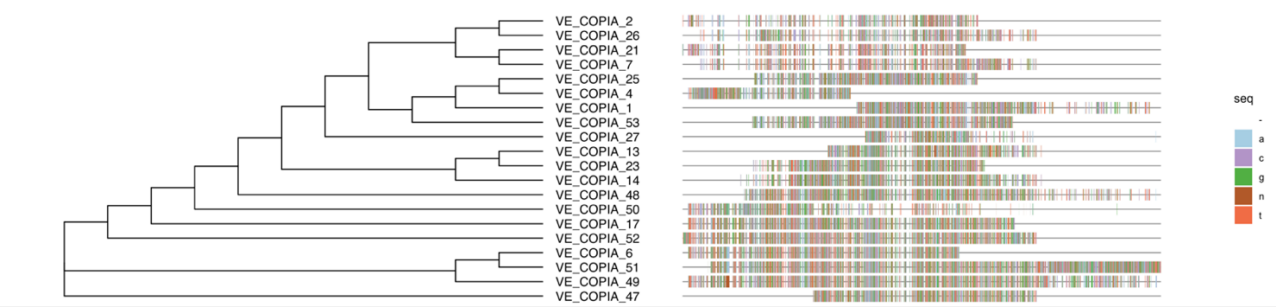
Original phylogenetic tree with 158 tips and 156 internal nodes. Pruned to 20 tips and 18 nodes.; 8208 bp alignment; Highest scoring hit for reverse transcriptase with RVT_2 (cl06662); e-value=1.03e-28.

A. eriantha



Original phylogenetic tree with 104 tips and 102 internal nodes. Pruned to 20 tips and 18 nodes.; 20089 bp alignment. Highest hit for reverse transcriptase with RVT_2 super family (cl06662); e-value= 5.23e-76.

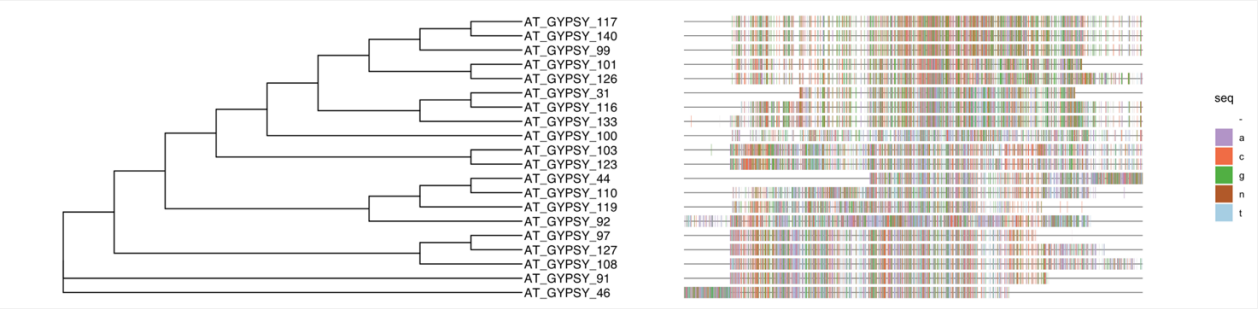
A. ventricosa



Original phylogenetic tree with 53 tips and 51 internal nodes. Pruned to 20 tips and 18 nodes; 16,564 bp alignment; Highest hit for reverse transcriptase with RVT_2 super family (cl06662); e-value= 4.10e-36.

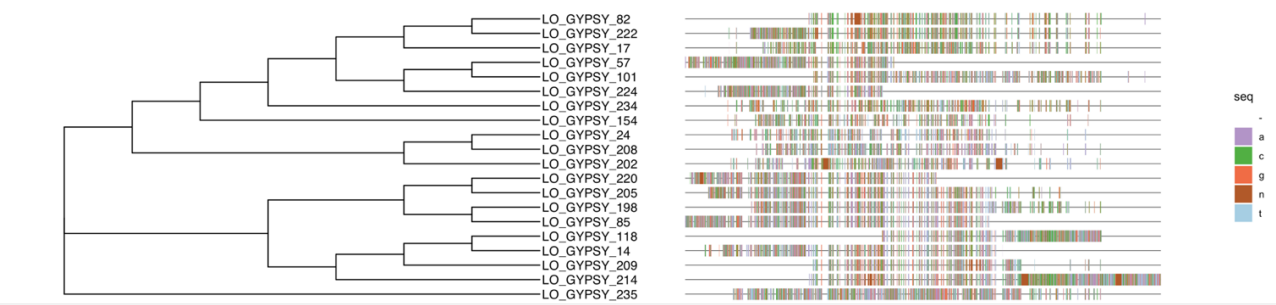
Figure 38, Ty3 Gypsy Retrotransposons. MSA of nucleotide sequences.

A. atlantica



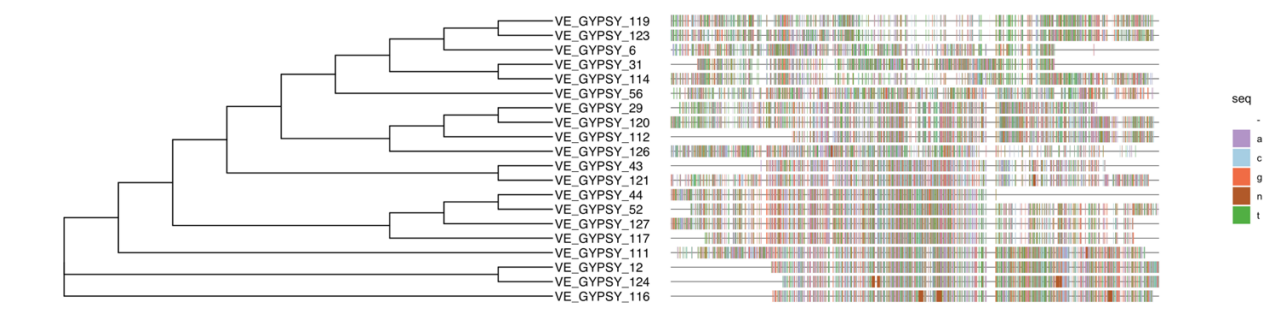
Original phylogenetic tree with 154 tips and 152 internal nodes. Pruned to 20 tips and 18 nodes.; 29844 bp alignment; Highest hit for reverse transcriptase with RT_LTR (cd01647); e-value= 1.10e-89.

A. longiglumis



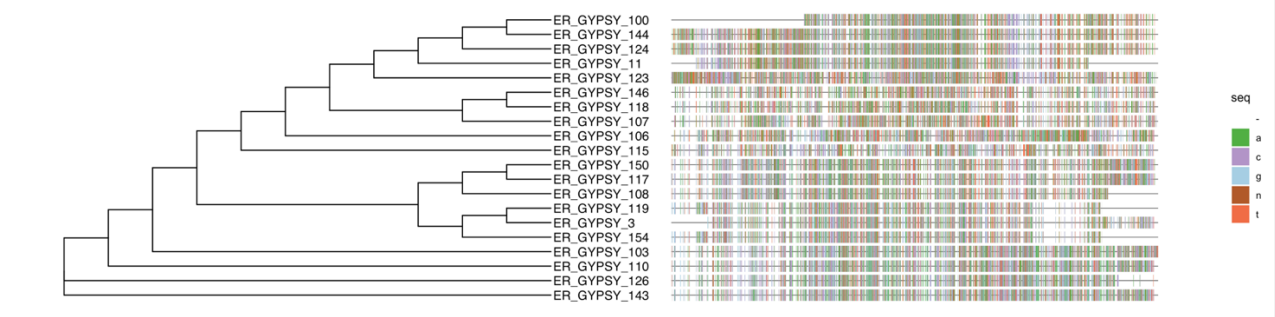
Original phylogenetic tree with 237 tips and 235 internal nodes. Pruned to 20 tips and 18 nodes.; 11971 bp alignment; Highest hit for reverse transcriptase with RT_LTR (cd01647); e-value= 6.91e-88.

A.eriantha



Original phylogenetic tree with 174 tips and 172 internal nodes. Pruned to 20 tips and 18 nodes.; 23162 bp alignment; Highest hit for reverse transcriptase with RT_LTR (cd01647); e-value= 2.44e-89.

A. ventricosa



Original phylogenetic tree with 130 tips and 128 internal nodes. Pruned to 20 tips and 18 nodes.; 18696 bp alignment; Highest hit for reverse transcriptase with RT_LTR (cd01647); e-value= 5.57e-71.

Phylogenetic Comparison of RT Domains

Reverse transcriptase (RT) domains are very well-characterized and highly conserved among LTR elements. As such, the RT domains were chosen to further characterize the identified elements in each genome **[40]**. Prior studies have identified primary LTR clades across various species, such as *Llorens et al* **[159]** and *Wicker and Keller* **[160]**. *Llorens et al* study included a total of 268 LTR sequences representing a diverse set of plants, animals and fungi, with sequences consisting of distinct Ty3-Gypsy and Retroviridae sequences derived from the Gypsy Database **[157]** and Ty1-Copia, Caulimoviridae, and Bel/Pao sequences being derived from the non-redundant NCBI database **[161]**. Five phylogenies inferring the evolution of Ty3-Gypsy, Ty1-Copia, Bel/Pao, Caulimoviridae and Retroviridae LTRs were produced, all based on conserved protein sequences for protease (PRT), reverse-transcriptase (RT), ribonuclease H (RH) and integrase (INT). Their work concludes with the identification of two Gypsy lineages, identified as Chromovirus and Tat/Athila, and five Ty3-Gypsy lineages identified as Oryco, Sire, Retrofit, Osser and Tork.

Wicker and Keller investigated Ty1-Copia elements identified in *A. thaliana* and *O. sativa*, with a total of 599 Copia elements and 68 distinct Ty1-Copia families being identified based on homologous families previously identified in *H. vulgare* (barley) and *T. aestivum* (wheat). Phylogenetic analysis revealed six lineages of Ty1-Copia that were not only highly conserved, as these phylogenies were built on prior knowledge of POL (PRT-RT-INT) protein domains, but these lineages also suggested the presence of these families of LTRs prior to the split of monocots and dicots **[162]**. Their study concludes with the identification of Ty1-Copia clades Maximus, Ivana, Ale, Angela, TAR and Bianca.

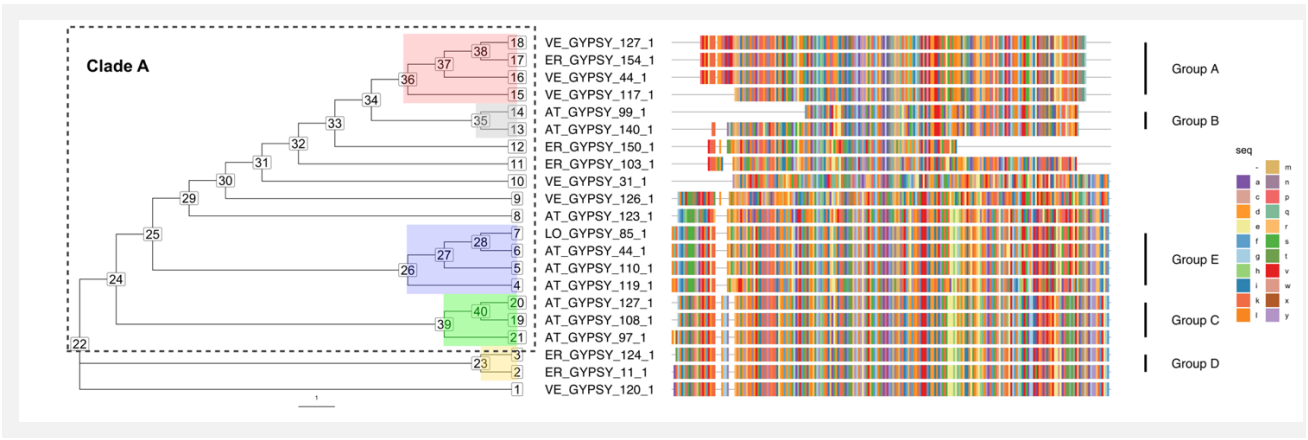
These studies demonstrate the large degree of diversity displayed within plant-specific LTRs. Our comparisons of RT domains displayed polymorphic in amino acid sequence between all species

for Copia and Gypsy, with frequent branch points displayed in Figure 39. To quantify observed polymorphisms, Chimera **[163]** was used to calculate average percent identities of sequences within the multiple sequence alignment; percent identity (Ty3-Gypsy: ~55% ; Ty1-Copia:~42%). Observation were mixed as some elements aligned well and some that did not, however this was somewhat expected as these sequences a derived from separate species and separated by millions of years. In that time an unknown degree of drift has likely occurred so we would expect to observe differences. The well aligning regions are also expected as A. eriantha and *A. atlantica* served as references for *A. ventricosa* and *A. longiglumis*, so these species forming clades of LTRs is logical. These points aside, we did observe patterns of genome-specific elements clustering into defined groups in when performing MSA analysis in (Figure 39).

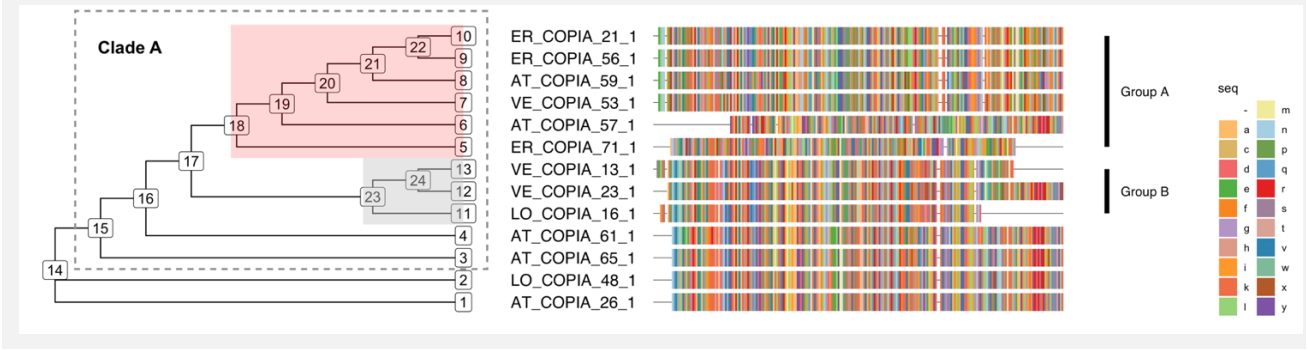
Ty3-Gypsy Reverse Transcriptase tree topology is relatively monophyletic, with relatively short branches (mean length ~0.634) and multiple nested clades. There appears to be a total of 5 groups: Group A, Group B, Group C, Group D and Group E, with one superclade containing 4 of 5 groups (denoted as Clade A). The most distant node appears to be Node 22, as Clade A, a small group at Node 23 and a singular sequence at Node 1 are descended from Node 22. Similar observations are made in regard to MSA of *Ty1*-Copia Reverse Transcriptase, as its derived tree also appears to be very monophyletic, and branches relatively are short (mean length ~0.533). There is one primary Clade (Clade A), consisting of two sister groups, Group A (red) and Group B (gray). Node 14 appears to be the most distant node, as Group A, Group B are derived from this node. There appears to be an occurrence of polytomy at Nodes 1 and 2, indicating a need for additional information to study these relationships further. Our comparisons of RT domains are general statements of relationships between observed elements and will not indicate ancestry, however it was interesting to observe the clustering tendencies between LTR elements, designated as groups in each figure.

Figure 39, Ty3-Gyspy & Ty1-Copia MSA

Ty3-Gypsy Reverse Transcriptase MSA



Ty1-Copia Reverse Transcriptase MSA

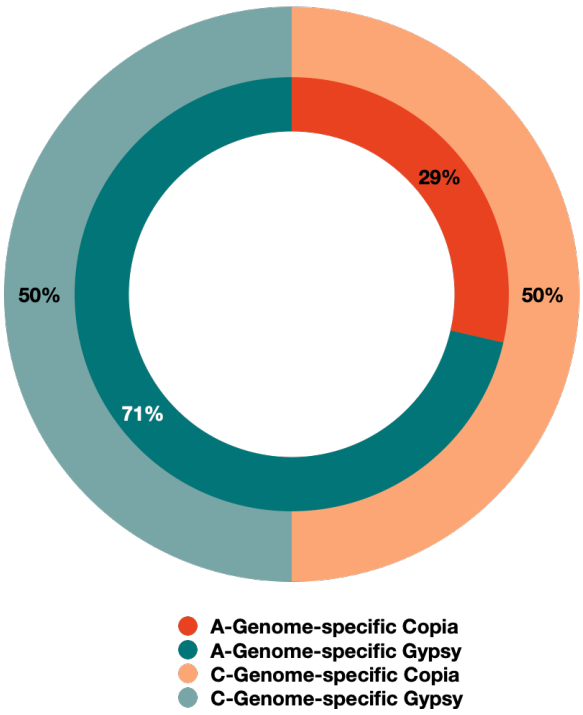


Identification & Characterization of A/C-Genome-Specific LTRs

Genome-specific LTRs were identified using BLAST, with analysis consisting of all *Avena*-specific LTR families from RepeatModeler v2 and individual databases derived from each *Avena* genome. Families of LTRs were queried against each genome; e.g. *A. atlantica* (A_s)-specific LTR families queried against *A. eriantha* genome, and queries were conducted using a 90% sequence identity to filter distant LTRs. In total, 4 separate BLAST analyses were performed, consisting of (1) *A. atlantica* genome against all non-Atlantica LTRs, (2) *A. longiglumis* genome against all non-longiglumis LTRs, (3) *A. eriantha* genome against all non-Eriantha LTRs, and (4) *A. ventricosa* genome against all non-Ventricosa LTRs. Further confirmation of potential genome-specific elements consisted of a confirmation BLAST analysis against all reference genomes, this time applying a 90% coverage cutoff for high-scoring sequences. These comparisons revealed several potential elements specific to each genome of *A. atlantica*, *A. eriantha*, and *A. ventricosa*; no genome-specific repeats were found for *A. longiglumis.*

At 90% sequence identity and a coverage requirement for high-scoring pairs, 7 sequences specific to *A.atlantica*, 27 sequences specific to *A. eriantha*, and 7 sequences specific to *A. ventricosa* were identified from BLAST analysis against each genome. In total, 41 sequences were identified as genome specific, with 19 elements classified as Ty1-Copia (46.34%) and 22 elements being classified as Ty3-Gypsy (53.66%) (*Figure 40*). Interestingly, of the 41 elements identified as genome-specific, ER_COPIA_55, ER_COPIA_78, and ER_COPIA_7 were identified within the larger clades of elements noted in previous phylogenetic analysis. This potentially indicates not only individual sequences that are distinct, but a clade of sequences that are specific to *A. eriantha*.

Figure 40, A-Genome-Specific & C-Genome-Specific LTRs



Genome Specific	*Ty1-Copia*	*Ty3-Gypsy*
A	*AT_COPIA_17*	*AT_GYPSY_147*
	AT_COPIA_19	*AT_GYPSY_150*
		AT_GYPSY_153
		AT_GYPSY_35
		AT_GYPSY_47
C	*ER_COPIA_11*	*ER_GYPSY_102*
	ER_COPIA_33	*ER_GYPSY_13*
	ER_COPIA_42	*ER_GYPSY_130*
	ER_COPIA_43	*ER_GYPSY_14*
	ER_COPIA_55	*ER_GYPSY_148*
	ER_COPIA_62	*ER_GYPSY_156*
	ER_COPIA_7	*ER_GYPSY_16*
	ER_COPIA_78	*ER_GYPSY_18*
	ER_COPIA_81	*ER_GYPSY_19*
	ER_COPIA_84	*ER_GYPSY_25*
	VE_COPIA_10	*ER_GYPSY_29*
	VE_COPIA_21	*ER_GYPSY_34*
	VE_COPIA_23	*ER_GYPSY_42*
	VE_COPIA_24	*ER_GYPSY_52*
	VE_COPIA_30	*ER_GYPSY_67*
	VE_COPIA_31	*ER_GYPSY_8*
	VE_COPIA_6	*ER_GYPSY_9*

CD-Search [164] queries were performed to identify which domains (GAG or POL) that each genome-specific sequence shared most homology (*Table 21*). Of the 41 genome-specific repeats, 6 were complete, defined as LTRs that possessed both GAG and POL ORFs, 15 sequences contained exclusively contained POL protein fragments, and 14 sequences contained GAG protein fragment exclusively. Complete LTRs were predominately of LTRs derived from *A. eriantha*. Multiple sequences alignments of elements containing to coding region corresponding

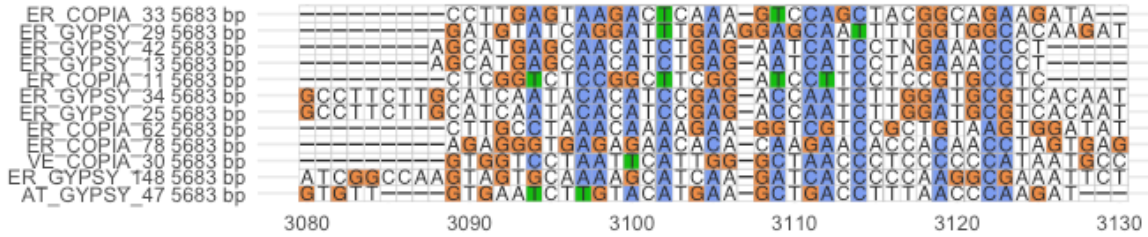
to GAG and POL and domains was performed to observe any potential similarities between genes. Many of these alignments were partial, containing regions that did not align well; Chimera was utilized to derive MSA percent Identities (GAG: ~37% ;and POL: ~40%). From this percentage implies distant relation between GAG and POL regions, however alignments overall possessed many polymorphisms between sequences. Notable regions of similarity in partial alignments of POL genes corresponding to are is displayed in Figure 41, A. with an alignment of ~5600bp in length. Alignment of GAG genes (~9000bp alignment) are displayed in Figure 42 B). We believe additional investigation is necessary to confirm the presence of these sequences in A or C-specific subgenomes, however these are potentially novel LTRs unique to specific *Avena* genome.

Table 21, CD-Search Homology of Genome-Specific elements to GAG and POL domains

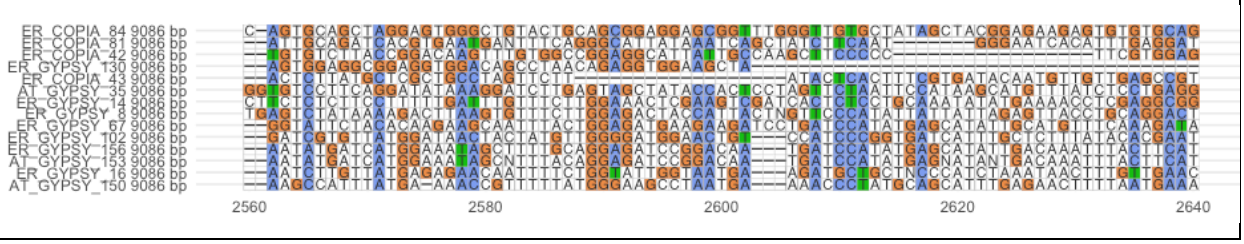
Protein-Domain	POL Domain	GAG Domain	Complete LTR Structure
Sequences	*AT_GYPSY_47*	*AT_GYPSY_150*	*ER_COPIA_7*
	ER_GYPSY_13	*AT_GYPSY_153*	*ER_COPIA_55*
	ER_GYPSY_25	*ER_GYPSY_14*	*ER_GYPSY_9*
	ER_GYPSY_34	*ER_GYPSY_16*	*ER_GYPSY_18*
	ER_GYPSY_42	*ER_GYPSY_67*	*ER_GYPSY_52*
	ER_GYPSY_29	*ER_GYPSY_102*	*VE_COPIA_21*
	ER_GYPSY_148	*AT_GYPSY_35*	*VE_COPIA_23*
	ER_COPIA_11	*ER_GYPSY_8*	*AT_COPIA_19*
	ER_COPIA_33	*ER_GYPSY_130*	
	ER_COPIA_62	*ER_GYPSY_156*	
	ER_COPIA_78	*ER_COPIA_42*	
	VE_COPIA_30	*ER_COPIA_43*	
		ER_COPIA_81	
		ER_COPIA_84	

Figure 41, Multiple Sequence Alignment of Genome-specific POL & GAG domains

A.) POL Alignment, Window interval = 3080,3130



B.) GAG Alignment; Window interval= 2560,2640



Conclusion

Repeat characterization of *Avena* species is an essential step in characterizing Avena genomes, as additional information of all elements within a given genome is required to gain greater understanding of *Avena* as a whole. With analysis of *A. atlantica*, *A. longiglumis, A. eriantha* and *A. ventricosa*, insight into to *Avena* has revealed potentially unique repetitive elements as well as an initial view of the repetitive element profile of each *Avena* species. Phylogenetic analysis of LTRs within A and C genomes reveal differing degrees of variability, that is, sequences that are distinct, within A and C, with C-genomes presenting as more variable and sequences that are less similar. This observation carried over to genome specific elements, where our study presents several potentially genome-specific elements identified to each subgenome. C-genome LTRs were represented at higher proportions that those of A-genome LTRs. Additional analysis into genic regions known to encode for GAG and POL genes was performed to assess the homology of these genes to known protein sequences corresponding to GAG and POL. As expected, genes were variable, with multiple sequences alignment of these genes revealing highly polymorphic regions in comparison to confirmed genes within BLAST. We believe this serves as a starting point for future studies and that this work provide insight into studies seeking to characterize or use repetitive elements as identifying markers for *Avena*.

Chapter 5: CONCLUSIONS

Our primary objective throughout this investigation sought to improve knowledge of the repeat landscape of oats, as characterization of these highly complex repetitive regions is essential if we are to continue growing our knowledge of *Avena*. In the process, we developed of a pipeline capable of offering enhanced resolution and detection of various repetitive elements ranging from Class I to Class II, but also gained insight into not only the repeat landscape but of provided an opportunity to explore in-depth inquiry into individual species of *Avena*. Additionally, development of our pipeline allowed for greater resolution of TE families and describes elements to an extent unobserved in prior studies of our benchmarking genomes, *A. thaliana* and *O. sativa*. With the advancement of repeat identification, our secondary objective was to capitalize on any novel information about the repeatome, or all repetitive elements in a given genome, and this allowed for further examination into transposable elements within two diploid *Avena* genomes, *A. atlantica* and *A. eriantha*. This analysis yielded not only more diverse characterizations of repeat elements within these species, but brought to light genomic difference highlighted by distinct differences observed in analysis of repetitive elements within these species. Using this analysis in conjunction with our previously developed pipeline, our final objective shifted towards beginning to understand the evolution of families of repetitive elements among *Avena* species, with the goal of gaining insight into the role transposable elements across *Avena* as a whole. With our analysis concluding with several potential genome-specific elements, our work serves as a starting point for future work striving to study repetitive elements to the extent we aspired to with this project, and we are optimistic that this work provide the potential of providing an additional resource to future evolutionary work within *Avena*.

REFERENCES

[1] A. Mushtaq, . Gul-Zaffar, A. D. Z., and H. Mehfuza, "A review on Oat (Avena sativa L.) as a dual-purpose crop," *Sci. Res. Essays*, vol. 9, no. 4, pp. 52–59, Feb. 2014, doi: 10.5897/SRE2014.5820.

[2] R. J. Moore-Colyer, "Oats and oat production in history and pre-history," in *The Oat Crop*, R. W. Welch, Ed. Dordrecht: Springer Netherlands, 1995, pp. 1–33.

[3] FAO, "Food and Agriculture Organization of the United Nations," *FAOSTAT*, 2017. http://www.fao.org/faostat/en/#data/QC (accessed Feb. 28, 2019).

[4] R. A. Othman, M. H. Moghadasian, and P. J. Jones, "Cholesterol-lowering effects of oat β-glucan," *Nutr. Rev.*, vol. 69, no. 6, pp. 299–309, Jun. 2011, doi: 10.1111/j.1753-4887.2011.00401.x.

[5] H. Boz, "Phenolic amides (avenanthramides) in oats – a review," *Czech J. Food Sci.*, vol. 33, no. No. 5, pp. 399–404, Jun. 2016, doi: 10.17221/696/2014-CJFS.

[6] R. Sur, A. Nigam, D. Grote, F. Liebel, and M. D. Southall, "Avenanthramides, polyphenols from oats, exhibit anti-inflammatory and anti-itch activity," *Arch. Dermatol. Res.*, vol. 300, no. 10, pp. 569–574, Nov. 2008, doi: 10.1007/s00403-008-0858-x.

[7] S. Bryngelsson, L. H. Dimberg, and A. Kamal-Eldin, "Effects of commercial processing on levels of antioxidants in oats (Avena sativa L.)," *J. Agric. Food Chem.*, vol. 50, no. 7, pp. 1890–1896, Mar. 2002, doi: 10.1021/jf011222z.

[8] M.-C. Carpentier *et al.*, "Retrotranspositional landscape of Asian rice revealed by 3000 genomes," *Nat. Commun.*, vol. 10, no. 1, p. 24, 03 2019, doi: 10.1038/s41467-018-07974-5.

[9] H. Yan *et al.*, "Genome size variation in the genus Avena," *Genome*, vol. 59, no. 3, pp. 209–220, Mar. 2016, doi: 10.1139/gen-2015-0132.

[10] B. McClintock, "Induction of Instability at Selected Loci in Maize," *Genetics*, vol. 38, no. 6, pp. 579–599, Nov. 1953.

[11] J. R. Cameron, E. Y. Loh, and R. W. Davis, "Evidence for transposition of dispersed repetitive DNA families in yeast," *Cell*, vol. 16, no. 4, pp. 739–751, Apr. 1979, doi: 10.1016/0092-8674(79)90090-4.

[12] T. Wicker *et al.*, "A unified classification system for eukaryotic transposable elements," *Nat. Rev. Genet.*, vol. 8, no. 12, pp. 973–982, Dec. 2007, doi: 10.1038/nrg2165.

[13] V. Walbot, "Saturation mutagenesis using maize transposons," *Curr. Opin. Plant Biol.*, vol. 3, no. 2, pp. 103–107, Apr. 2000.

[14] Y.-J. Kim, J. Lee, and K. Han, "Transposable Elements: No More 'Junk DNA,'" *Genomics Inform.*, vol. 10, no. 4, pp. 226–233, Dec. 2012, doi: 10.5808/GI.2012.10.4.226.

[15] S. Ohno, "So much 'junk' DNA in our genome," *Brookhaven Symp. Biol.*, vol. 23, pp. 366–370, 1972.

[16] D. Lisch, "Mutator and MULE Transposons," *Microbiol. Spectr.*, vol. 3, no. 2, pp. MDNA3-0032–2014, Apr. 2015, doi: 10.1128/microbiolspec.MDNA3-0032-2014.

[17] M. G. Kidwell and D. Lisch, "Transposable elements as sources of variation in animals and plants," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 94, no. 15, pp. 7704–7711, Jul. 1997, doi: 10.1073/pnas.94.15.7704.

[18] H. K. Dooner and O. E. Nelson, “Genetic control of UDPglucose:flavonol 3-O-glucosyltransferase in the endosperm of maize,” *Biochem. Genet.*, vol. 15, no. 5–6, pp. 509–519, Jun. 1977, doi: 10.1007/BF00520194.
[19] N. Jiang, Z. Bao, X. Zhang, S. R. Eddy, and S. R. Wessler, “Pack-MULE transposable elements mediate gene evolution in plants,” *Nature*, vol. 431, no. 7008, pp. 569–573, Sep. 2004, doi: 10.1038/nature02953.
[20] W. Wang *et al.*, “High Rate of Chimeric Gene Origination by Retroposition in Plant Genomes,” *Plant Cell*, vol. 18, no. 8, pp. 1791–1802, Aug. 2006, doi: 10.1105/tpc.106.041905.
[21] J. H. Werren, “Selfish genetic elements, genetic conflict, and evolutionary innovation,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108 Suppl 2, pp. 10863–10870, Jun. 2011, doi: 10.1073/pnas.1102343108.
[22] C. da Silva Linge *et al.*, “High-density multi-population consensus genetic linkage map for peach,” *PloS One*, vol. 13, no. 11, p. e0207724, 2018, doi: 10.1371/journal.pone.0207724.
[23] G. Ji *et al.*, “Construction of a high-density genetic map using specific-locus amplified fragments in sorghum,” *BMC Genomics*, vol. 18, no. 1, p. 51, 07 2017, doi: 10.1186/s12864-016-3430-7.
[24] A. Kumar and J. L. Bennetzen, “Plant Retrotransposons,” *Annu. Rev. Genet.*, vol. 33, no. 1, pp. 479–532, Dec. 1999, doi: 10.1146/annurev.genet.33.1.479.
[25] G. Loebenstein, “Local Lesions and Induced Resistance,” in *Advances in Virus Research*, vol. 75, Elsevier, 2009, pp. 73–117.
[26] J. Pellicer, O. Hidalgo, S. Dodsworth, and I. J. Leitch, “Genome Size Diversity and Its Impact on the Evolution of Land Plants,” *Genes*, vol. 9, no. 2, Feb. 2018, doi: 10.3390/genes9020088.
[27] S. Dodsworth, A. R. Leitch, and I. J. Leitch, “Genome size diversity in angiosperms and its influence on gene space,” *Curr. Opin. Genet. Dev.*, vol. 35, pp. 73–78, Dec. 2015, doi: 10.1016/j.gde.2015.10.006.
[28] K. M. Devos, J. K. M. Brown, and J. L. Bennetzen, “Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis,” *Genome Res.*, vol. 12, no. 7, pp. 1075–1079, Jul. 2002, doi: 10.1101/gr.132102.
[29] G. T. H. Vu *et al.*, “Comparative Genome Analysis Reveals Divergent Genome Size Evolution in a Carnivorous Plant Genus,” *Plant Genome*, vol. 8, no. 3, p. 0, 2015, doi: 10.3835/plantgenome2015.04.0021.
[30] L. Comai, “The advantages and disadvantages of being polyploid,” *Nat. Rev. Genet.*, vol. 6, no. 11, pp. 836–846, Nov. 2005, doi: 10.1038/nrg1711.
[31] M. C. Sattler, C. R. Carvalho, and W. R. Clarindo, “The polyploidy and its key role in plant breeding,” *Planta*, vol. 243, no. 2, pp. 281–296, Feb. 2016, doi: 10.1007/s00425-015-2450-x.
[32] B. E. Tropp, *Molecular biology 4e: genes to proteins internation edition.* Place of publication not identified: Jones & Bartlett Pubs.
[33] M. Muñoz-López and J. L. García-Pérez, “DNA transposons: nature and applications in genomics,” *Curr. Genomics*, vol. 11, no. 2, pp. 115–128, Apr. 2010, doi: 10.2174/138920210790886871.

[34] K. Hiom, M. Melek, and M. Gellert, “DNA transposition by the RAG1 and RAG2 proteins: a possible source of oncogenic translocations,” *Cell*, vol. 94, no. 4, pp. 463–470, Aug. 1998, doi: 10.1016/s0092-8674(00)81587-1.
[35] B. Alberts, Ed., *Molecular biology of the cell*, 4th ed. New York: Garland Science, 2002.
[36] S. Domingues, G. J. da Silva, and K. M. Nielsen, “Integrons: Vehicles and pathways for horizontal dissemination in bacteria,” *Mob. Genet. Elem.*, vol. 2, no. 5, pp. 211–223, Sep. 2012, doi: 10.4161/mge.22967.
[37] J. Macas and P. Neumann, “Ogre elements — A distinct group of plant Ty3/gypsy-like retrotransposons,” *Gene*, vol. 390, no. 1–2, pp. 108–116, Apr. 2007, doi: 10.1016/j.gene.2006.08.007.
[38] S. B. Sandmeyer and K. A. Clemens, “Function of a retrotransposon nucleocapsid protein,” *RNA Biol.*, vol. 7, no. 6, pp. 642–654, Nov. 2010, doi: 10.4161/rna.7.6.14117.
[39] A. Suoniemi, J. Tanskanen, and A. H. Schulman, “Gypsy-like retrotransposons are widespread in the plant kingdom,” *Plant J. Cell Mol. Biol.*, vol. 13, no. 5, pp. 699–705, Mar. 1998.
[40] Y. Xiong and T. H. Eickbush, “Origin and evolution of retroelements based upon their reverse transcriptase sequences.,” *EMBO J.*, vol. 9, no. 10, pp. 3353–3362, Oct. 1990, doi: 10.1002/j.1460-2075.1990.tb07536.x.
[41] V. F. Suguiyama, L. A. B. Vasconcelos, M. M. Rossi, C. Biondo, and N. de Setta, “The population genetic structure approach adds new insights into the evolution of plant LTR retrotransposon lineages,” *PLOS ONE*, vol. 14, no. 5, p. e0214542, May 2019, doi: 10.1371/journal.pone.0214542.
[42] A. H. Schulman, “Retrotransposon replication in plants,” *Curr. Opin. Virol.*, vol. 3, no. 6, pp. 604–614, Dec. 2013, doi: 10.1016/j.coviro.2013.08.009.
[43] T. H. Eickbush and V. K. Jamburuthugoda, “The diversity of retrotransposons and the properties of their reverse transcriptases,” *Virus Res.*, vol. 134, no. 1–2, pp. 221–234, Jun. 2008, doi: 10.1016/j.virusres.2007.12.010.
[44] Z. Schwarz-Sommer, L. Leclercq, E. Göbel, and H. Saedler, “Cin4, an insert altering the structure of the *A1* gene in *Zea mays* , exhibits properties of nonviral retrotransposons,” *EMBO J.*, vol. 6, no. 13, pp. 3873–3880, Dec. 1987, doi: 10.1002/j.1460-2075.1987.tb02727.x.
[45] D. A. Wright, N. Ke, J. Smalle, B. M. Hauge, H. M. Goodman, and D. F. Voytas, “Multiple non-LTR retrotransposons in the genome of Arabidopsis thaliana,” *Genetics*, vol. 142, no. 2, pp. 569–578, Feb. 1996.
[46] M. Komatsu, K. Shimamoto, and J. Kyozuka, “Two-step regulation and continuous retrotransposition of the rice LINE-type retrotransposon Karma,” *Plant Cell*, vol. 15, no. 8, pp. 1934–1944, Aug. 2003, doi: 10.1105/tpc.011809.
[47] V. V. Kapitonov and J. Jurka, “A novel class of SINE elements derived from 5S rRNA,” *Mol. Biol. Evol.*, vol. 20, no. 5, pp. 694–702, May 2003, doi: 10.1093/molbev/msg075.
[48] J.-M. Deragon and X. Zhang, “Short interspersed elements (SINEs) in plants: origin, classification, and use as phylogenetic markers,” *Syst. Biol.*, vol. 55, no. 6, pp. 949–956, Dec. 2006, doi: 10.1080/10635150601047843.

[49] R. T. M. Poulter and T. J. D. Goodwin, “DIRS-1 and the other tyrosine recombinase retrotransposons,” *Cytogenet. Genome Res.*, vol. 110, no. 1–4, pp. 575–588, 2005, doi: 10.1159/000084991.
[50] I. V. Nesmelova and P. B. Hackett, “DDE transposases: Structural similarity and diversity,” *Adv. Drug Deliv. Rev.*, vol. 62, no. 12, pp. 1187–1195, Sep. 2010, doi: 10.1016/j.addr.2010.06.006.
[51] M. B. Evgen’ev *et al.*, “Penelope, a new family of transposable elements and its possible role in hybrid dysgenesis in Drosophila virilis,” *Proc. Natl. Acad. Sci.*, vol. 94, no. 1, pp. 196–201, Jan. 1997, doi: 10.1073/pnas.94.1.196.
[52] I. R. Arkhipova, “Distribution and Phylogeny of Penelope-Like Elements in Eukaryotes,” *Syst. Biol.*, vol. 55, no. 6, pp. 875–885, Dec. 2006, doi: 10.1080/10635150601077683.
[53] V. V. Kapitonov and J. Jurka, “Molecular paleontology of transposable elements from Arabidopsis thaliana,” *Genetica*, vol. 107, no. 1–3, pp. 27–37, 1999.
[54] X. Lin, N. Faridi, and C. Casola, “An Ancient Transkingdom Horizontal Transfer of Penelope-Like Retroelements from Arthropods to Conifers,” *Genome Biol. Evol.*, vol. 8, no. 4, pp. 1252–1266, May 2016, doi: 10.1093/gbe/evw076.
[55] X. Zhang, C. Feschotte, Q. Zhang, N. Jiang, W. B. Eggleston, and S. R. Wessler, “P instability factor: an active maize transposon system associated with the amplification of Tourist-like MITEs and a new superfamily of transposases,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 98, no. 22, pp. 12572–12577, Oct. 2001, doi: 10.1073/pnas.211442198.
[56] R. H. Plasterk, Z. Izsvák, and Z. Ivics, “Resident aliens: the Tc1/mariner superfamily of transposable elements,” *Trends Genet. TIG*, vol. 15, no. 8, pp. 326–332, Aug. 1999.
[57] A. Palazzo *et al.*, “Transcriptionally promiscuous ‘blurry’ promoters in Tc1/mariner transposons allow transcription in distantly related genomes,” *Mob. DNA*, vol. 10, no. 1, p. 13, Dec. 2019, doi: 10.1186/s13100-019-0155-6.
[58] C. Ye, G. Ji, and C. Liang, “detectMITE: A novel approach to detect miniature inverted repeat transposable elements in genomes,” *Sci. Rep.*, vol. 6, p. 19688, Jan. 2016, doi: 10.1038/srep19688.
[59] F. Kempken and F. Windhofer, “The hAT family: a versatile transposon group common to plants, fungi, animals, and man,” *Chromosoma*, vol. 110, no. 1, pp. 1–9, Apr. 2001.
[60] E. Rubin, G. Lithwick, and A. A. Levy, “Structure and evolution of the hAT transposon superfamily,” *Genetics*, vol. 158, no. 3, pp. 949–957, Jul. 2001.
[61] M. N. Raizada, M.-I. Benito, and V. Walbot, “The MuDR transposon terminal inverted repeat contains a complex plant promoter directing distinct somatic and germinal programs: Transposon promoter expression pattern,” *Plant J.*, vol. 25, no. 1, pp. 79–91, Jul. 2008, doi: 10.1111/j.1365-313X.2001.00939.x.
[62] T. Wicker *et al.*, “DNA transposon activity is associated with increased mutation rates in genes of rice and other grasses,” *Nat. Commun.*, vol. 7, no. 1, p. 12790, Nov. 2016, doi: 10.1038/ncomms12790.
[63] B. Lewin, *Genes VI*. Oxford ; New York: Oxford University Press, 1997.
[64] L. Yang and J. L. Bennetzen, “Structure-based discovery and description of plant and animal Helitrons,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 31, pp. 12832–12837, Aug. 2009, doi: 10.1073/pnas.0905563106.

[65] J. Thomas, K. Vadnagara, and E. J. Pritham, “DINE-1, the highest copy number repeats in Drosophila melanogaster are non-autonomous endonuclease-encoding rolling-circle transposable elements (Helentrons),” *Mob. DNA*, vol. 5, no. 1, p. 18, 2014, doi: 10.1186/1759-8753-5-18.
[66] P. Heringer and G. Kuhn, “Exploring the Remote Ties between Helitron Transposases and Other Rolling-Circle Replication Proteins,” *Int. J. Mol. Sci.*, vol. 19, no. 10, p. 3079, Oct. 2018, doi: 10.3390/ijms19103079.
[67] W. Xiong, H. K. Dooner, and C. Du, “Rolling-circle amplification of centromeric *Helitrons* in plant genomes,” *Plant J.*, vol. 88, no. 6, pp. 1038–1045, Dec. 2016, doi: 10.1111/tpj.13314.
[68] I. Grabundzija, A. B. Hickman, and F. Dyda, “Helraiser intermediates provide insight into the mechanism of eukaryotic replicative transposition,” *Nat. Commun.*, vol. 9, no. 1, p. 1278, 29 2018, doi: 10.1038/s41467-018-03688-w.
[69] R. T. M. Poulter, T. J. D. Goodwin, and M. I. Butler, “Vertebrate helentrons and other novel Helitrons,” *Gene*, vol. 313, pp. 201–212, Aug. 2003, doi: 10.1016/S0378-1119(03)00679-6.
[70] S. Haapa-Paananen, N. Wahlberg, and H. Savilahti, “Phylogenetic analysis of Maverick/Polinton giant transposons across organisms,” *Mol. Phylogenet. Evol.*, vol. 78, pp. 271–274, Sep. 2014, doi: 10.1016/j.ympev.2014.05.024.
[71] S. Saha, S. Bridges, Z. V. Magbanua, and D. G. Peterson, “Computational Approaches and Tools Used in Identification of Dispersed Repetitive DNA Sequences,” *Trop. Plant Biol.*, vol. 1, no. 1, pp. 85–96, Mar. 2008, doi: 10.1007/s12042-007-9007-5.
[72] C. M. Bergman and H. Quesneville, “Discovering and detecting transposable elements in genome sequences,” *Brief. Bioinform.*, vol. 8, no. 6, pp. 382–392, May 2007, doi: 10.1093/bib/bbm048.
[73] R. Hubley *et al.*, “The Dfam database of repetitive DNA families,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D81–D89, Jan. 2016, doi: 10.1093/nar/gkv1272.
[74] T. Di Domenico *et al.*, “RepeatsDB: a database of tandem repeat protein structures,” *Nucleic Acids Res.*, vol. 42, no. D1, pp. D352–D357, Jan. 2014, doi: 10.1093/nar/gkt1175.
[75] W. Bao, K. K. Kojima, and O. Kohany, “Repbase Update, a database of repetitive elements in eukaryotic genomes,” *Mob. DNA*, vol. 6, p. 11, 2015, doi: 10.1186/s13100-015-0041-9.
[76] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, Oct. 1990, doi: 10.1016/S0022-2836(05)80360-2.
[77] C. A. Kerfeld and K. M. Scott, “Using BLAST to Teach ‘E-value-tionary’ Concepts,” *PLoS Biol.*, vol. 9, no. 2, p. e1001014, Feb. 2011, doi: 10.1371/journal.pbio.1001014.
[78] W. J. Kent, “BLAT--the BLAST-like alignment tool,” *Genome Res.*, vol. 12, no. 4, pp. 656–664, Apr. 2002, doi: 10.1101/gr.229202.
[79] T. Rognes, T. Flouri, B. Nichols, C. Quince, and F. Mahé, “VSEARCH: a versatile open source tool for metagenomics,” *PeerJ*, vol. 4, p. e2584, Oct. 2016, doi: 10.7717/peerj.2584.
[80] N. Volfovsky, B. J. Haas, and S. L. Salzberg, “A clustering method for repeat analysis in DNA sequences,” *Genome Biol.*, vol. 2, no. 8, p. research0027.1, 2001, doi: 10.1186/gb-2001-2-8-research0027.

[81] S. C. Manekar and S. R. Sathe, “A benchmark study of k-mer counting methods for high-throughput sequencing,” *GigaScience*, vol. 7, no. 12, 01 2018, doi: 10.1093/gigascience/giy125.
[82] A. L. Price, N. C. Jones, and P. A. Pevzner, “De novo identification of repeat families in large genomes,” *Bioinformatics*, vol. 21, no. Suppl 1, pp. i351–i358, Jun. 2005, doi: 10.1093/bioinformatics/bti1018.
[83] G. Benson, “Tandem repeats finder: a program to analyze DNA sequences,” *Nucleic Acids Res.*, vol. 27, no. 2, pp. 573–580, Jan. 1999, doi: 10.1093/nar/27.2.573.
[84] R. Li *et al.*, “ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun,” *PLoS Comput. Biol.*, vol. 1, no. 4, p. e43, Sep. 2005, doi: 10.1371/journal.pcbi.0010043.
[85] C. Chu, R. Nielsen, and Y. Wu, “REPdenovo: Inferring De Novo Repeat Motifs from Short Sequence Reads,” *PloS One*, vol. 11, no. 3, p. e0150719, 2016, doi: 10.1371/journal.pone.0150719.
[86] H. Z. Girgis, “Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale,” *BMC Bioinformatics*, vol. 16, p. 227, Jul. 2015, doi: 10.1186/s12859-015-0654-5.
[87] Z. Bao and S. R. Eddy, “Automated de novo identification of repeat sequence families in sequenced genomes,” *Genome Res.*, vol. 12, no. 8, pp. 1269–1276, Aug. 2002, doi: 10.1101/gr.88502.
[88] L. Malik, F. Almodaresi, and R. Patro, “Grouper: graph-based clustering and annotation for improved de novo transcriptome analysis,” *Bioinforma. Oxf. Engl.*, vol. 34, no. 19, pp. 3265–3272, 01 2018, doi: 10.1093/bioinformatics/bty378.
[89] P. Agarwal and D. J. States, “The Repeat Pattern Toolkit (RPT): analyzing the structure and evolution of the C. elegans genome,” *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, vol. 2, pp. 1–9, 1994.
[90] A. Caspi, “Identification of transposable elements using multiple alignments of related genomes,” *Genome Res.*, vol. 16, no. 2, pp. 260–270, Dec. 2005, doi: 10.1101/gr.4361206.
[91] E. Lyons, B. Pedersen, J. Kane, and M. Freeling, “The Value of Nonmodel Genomes and an Example Using SynMap Within CoGe to Dissect the Hexaploidy that Predates the Rosids,” *Trop. Plant Biol.*, vol. 1, no. 3–4, pp. 181–190, Dec. 2008, doi: 10.1007/s12042-008-9017-y.
[92] W. Makałowski, V. Gotea, A. Pande, and I. Makałowska, “Transposable Elements: Classification, Identification, and Their Use As a Tool For Comparative Genomics,” in *Evolutionary Genomics*, vol. 1910, M. Anisimova, Ed. New York, NY: Springer New York, 2019, pp. 177–207.
[93] D. Ellinghaus, S. Kurtz, and U. Willhoeft, “LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons,” *BMC Bioinformatics*, vol. 9, p. 18, Jan. 2008, doi: 10.1186/1471-2105-9-18.
[94] S. Ou and N. Jiang, “LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons,” *Plant Physiol.*, vol. 176, no. 2, pp. 1410–1422, Feb. 2018, doi: 10.1104/pp.17.01310.
[95] A. K. Konopka, “Sequence Complexity and Composition,” in *eLS*, John Wiley & Sons, Ltd, Ed. Chichester, UK: John Wiley & Sons, Ltd, 2005, p. a0005260.

[96] K. Hu *et al.*, “Helitron distribution in Brassicaceae and whole Genome Helitron density as a character for distinguishing plant species,” *BMC Bioinformatics*, vol. 20, no. 1, p. 354, Dec. 2019, doi: 10.1186/s12859-019-2945-8.
[97] W. Xiong, L. He, J. Lai, H. K. Dooner, and C. Du, “HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes,” *Proc. Natl. Acad. Sci.*, vol. 111, no. 28, pp. 10263–10268, Jul. 2014, doi: 10.1073/pnas.1410068111.
[98] R. Ragupathy, F. M. You, and S. Cloutier, “Arguments for standardizing transposable element annotation in plant genomes,” *Trends Plant Sci.*, vol. 18, no. 7, pp. 367–376, Jul. 2013, doi: 10.1016/j.tplants.2013.03.005.
[99] L. Stein, “Genome annotation: from sequence to biology,” *Nat. Rev. Genet.*, vol. 2, no. 7, pp. 493–503, Jul. 2001, doi: 10.1038/35080529.
[100] C. E. Cook, M. T. Bergman, R. D. Finn, G. Cochrane, E. Birney, and R. Apweiler, “The European Bioinformatics Institute in 2016: Data growth and integration,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D20–D26, Jan. 2016, doi: 10.1093/nar/gkv1352.
[101] “Repeat-Masker Open-4.0,” *RepeatMasker*. http://www.repeatmasker.org/.
[102] J. M. Flynn *et al.*, “RepeatModeler2 for automated genomic discovery of transposable element families,” *Proc. Natl. Acad. Sci.*, vol. 117, no. 17, pp. 9451–9457, Apr. 2020, doi: 10.1073/pnas.1921046117.
[103] S. Ou *et al.*, “Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline,” *Genome Biol.*, vol. 20, no. 1, p. 275, Dec. 2019, doi: 10.1186/s13059-019-1905-y.
[104] E. Lerat, “Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs,” *Heredity*, vol. 104, no. 6, pp. 520–533, Jun. 2010, doi: 10.1038/hdy.2009.165.
[105] E. Barghini *et al.*, “The peculiar landscape of repetitive sequences in the olive (Olea europaea L.) genome,” *Genome Biol. Evol.*, vol. 6, no. 4, pp. 776–791, Apr. 2014, doi: 10.1093/gbe/evu058.
[106] K. L. Howe *et al.*, “Ensembl Genomes 2020—enabling non-vertebrate genomic research,” *Nucleic Acids Res.*, vol. 48, no. D1, pp. D689–D695, Jan. 2020, doi: 10.1093/nar/gkz890.
[107] J. M. Crescente, D. Zavallo, M. Helguera, and L. S. Vanzetti, “MITE Tracker: an accurate approach to identify miniature inverted-repeat transposable elements in large genomes,” *BMC Bioinformatics*, vol. 19, no. 1, p. 348, Dec. 2018, doi: 10.1186/s12859-018-2376-y.
[108] J. Hu, Y. Zheng, and X. Shang, “MiteFinderII: a novel tool to identify miniature inverted-repeat transposable elements hidden in eukaryotic genomes,” *BMC Med. Genomics*, vol. 11, no. S5, p. 101, Nov. 2018, doi: 10.1186/s12920-018-0418-y.
[109] V. V. Kapitonov and J. Jurka, “Rolling-circle transposons in eukaryotes,” *Proc. Natl. Acad. Sci.*, vol. 98, no. 15, pp. 8714–8719, Jul. 2001, doi: 10.1073/pnas.151269298.
[110] J. Thomas and E. J. Pritham, “Helitrons, the Eukaryotic Rolling-circle Transposable Elements,” *Microbiol. Spectr.*, vol. 3, no. 4, Aug. 2015, doi: 10.1128/microbiolspec.MDNA3-0049-2014.
[111] H. Mao and H. Wang, “SINE_scan: an efficient tool to discover short interspersed nuclear elements (SINEs) in large-scale genomic datasets,” *Bioinformatics*, p. btw718, Jan. 2017, doi: 10.1093/bioinformatics/btw718.

[112] T. Wenke, T. Döbel, T. R. Sörensen, H. Junghans, B. Weisshaar, and T. Schmidt, "Targeted Identification of Short Interspersed Nuclear Element Families Shows Their Widespread Existence and Extreme Heterogeneity in Plant Genomes," *Plant Cell*, vol. 23, no. 9, pp. 3117–3128, Sep. 2011, doi: 10.1105/tpc.111.088682.
[113] N. S. Vassetzky and D. A. Kramerov, "SINEBase: a database and tool for SINE analysis," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D83–D89, Jan. 2013, doi: 10.1093/nar/gks1263.
[114] D. Swarbreck *et al.*, "The Arabidopsis Information Resource (TAIR): gene structure and function annotation," *Nucleic Acids Res.*, vol. 36, no. Database, pp. D1009–D1014, Dec. 2007, doi: 10.1093/nar/gkm965.
[115] A. R. Quinlan and I. M. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, no. 6, pp. 841–842, Mar. 2010, doi: 10.1093/bioinformatics/btq033.
[116] Python Software Foundation, "Python," *Welcome to Python*, Jun. 22, 2020. https://www.python.org/.
[117] Hadley Wickham, Romain François, Lionel Henry and Kirill Müller, *dplyr: A Grammar of Data Manipulation.* 2018.
[118] H. Wickham, stringr: Simple, Consistent Wrappers for Common String Operations. 2019.
[119] K. Ushey, reticulate: Interface to "Python." 2020.
[120] A. Coghlan, A. Coghlan, I. J. Tsai, and M. Berriman, "Creation of a comprehensive repeat library for a newly sequenced parasitic worm genome," *Protoc. Exch.*, May 2018, doi: 10.1038/protex.2018.054.
[121] J. T. Robinson *et al.*, "Integrative genomics viewer," *Nat. Biotechnol.*, vol. 29, no. 1, pp. 24–26, Jan. 2011, doi: 10.1038/nbt.1754.
[122] K. Y. Oróstica and R. A. Verdugo, "chromPlot: visualization of genomic data in chromosomal context," *Bioinformatics*, vol. 32, no. 15, pp. 2366–2368, Aug. 2016, doi: 10.1093/bioinformatics/btw137.
[123] S. L. Salzberg, "Next-generation genome annotation: we still struggle to get it right," *Genome Biol.*, vol. 20, no. 1, pp. 92, s13059-019-1715–2, Dec. 2019, doi: 10.1186/s13059-019-1715-2.
[124] K. M. Seibt, T. Wenke, K. Muders, B. Truberg, and T. Schmidt, "Short interspersed nuclear elements (SINEs) are abundant in Solanaceae and have a family-specific impact on gene structure and genome organization," *Plant J.*, vol. 86, no. 3, pp. 268–285, May 2016, doi: 10.1111/tpj.13170.
[125] Y. Han and S. R. Wessler, "MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences," *Nucleic Acids Res.*, vol. 38, no. 22, pp. e199–e199, Dec. 2010, doi: 10.1093/nar/gkq862.
[126] M. Boczkowska, M. Harasimiuk, and A. Onyśk, "Studies on genetic variation within old Polish cultivars of common oat," *Cereal Res. Commun.*, vol. 43, no. 1, pp. 12–21, Mar. 2015, doi: 10.1556/CRC.2014.0025.
[127] S. Leonova, T. Shelenga, M. Hamberg, A. V. Konarev, I. Loskutov, and A. S. Carlsson, "Analysis of Oil Composition in Cultivars and Wild Species of Oat (Avena sp.)," *J. Agric. Food Chem.*, vol. 56, no. 17, pp. 7983–7991, Sep. 2008, doi: 10.1021/jf800761c.

[128] H. Yan *et al.*, "High-density marker profiling confirms ancestral genomes of Avena species and identifies D-genome chromosomes of hexaploid oat," *TAG Theor. Appl. Genet. Theor. Angew. Genet.*, vol. 129, no. 11, pp. 2133–2149, Nov. 2016, doi: 10.1007/s00122-016-2762-7.
[129] Y.-B. Fu, "Oat evolution revealed in the maternal lineages of 25 Avena species," *Sci. Rep.*, vol. 8, no. 1, p. 4252, Mar. 2018, doi: 10.1038/s41598-018-22478-4.
[130] P. J. Maughan *et al.*, "Genomic insights from the first chromosome-scale assemblies of oat (Avena spp.) diploid species," *BMC Biol.*, vol. 17, no. 1, p. 92, Dec. 2019, doi: 10.1186/s12915-019-0712-y.
[131] Q. Liu *et al.*, "The repetitive DNA landscape in Avena (Poaceae): chromosome and genome evolution defined by major repeat classes in whole-genome sequence reads," *BMC Plant Biol.*, vol. 19, no. 1, p. 226, Dec. 2019, doi: 10.1186/s12870-019-1769-z.
[132] R. Solano, G. Hueros, A. Fominaya, and E. Ferrer, "Organization of repeated sequences in species of the genus Avena," *Theor. Appl. Genet.*, vol. 83, no. 5, pp. 602–607, Mar. 1992, doi: 10.1007/BF00226904.
[133] A. Katsiotis, "Repetitive DNA, Genome and Species Relationships in Avena and Arrhenatherum(Poaceae)," *Ann. Bot.*, vol. 86, no. 6, pp. 1135–1142, Dec. 2000, doi: 10.1006/anbo.2000.1284.
[134] C. Linares, E. Ferrer, and A. Fominaya, "Discrimination of the closely related A and D genomes of the hexaploid oat Avena sativa L.," *Proc. Natl. Acad. Sci.*, vol. 95, no. 21, pp. 12450–12455, Oct. 1998, doi: 10.1073/pnas.95.21.12450.
[135] S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and A. M. Phillippy, "Canu: scalable and accurate long-read assembly via adaptive *k* -mer weighting and repeat separation," *Genome Res.*, vol. 27, no. 5, pp. 722–736, May 2017, doi: 10.1101/gr.215087.116.
[136] N. H. Putnam *et al.*, "Chromosome-scale shotgun assembly using an in vitro method for long-range linkage," *Genome Res.*, vol. 26, no. 3, pp. 342–350, Mar. 2016, doi: 10.1101/gr.193474.115.
[137] M. Lawrence, rtracklayer: R interface to genome annotation files and the UCSC genome browser. .
[138] K. Katoh and D. M. Standley, "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability," *Mol. Biol. Evol.*, vol. 30, no. 4, pp. 772–780, Apr. 2013, doi: 10.1093/molbev/mst010.
[139] M. N. Price, P. S. Dehal, and A. P. Arkin, "FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments," *PLoS ONE*, vol. 5, no. 3, p. e9490, Mar. 2010, doi: 10.1371/journal.pone.0009490.
[140] M. Schreiber, N. Stein, and M. Mascher, "Genomic approaches for studying crop evolution," *Genome Biol.*, vol. 19, no. 1, p. 140, Dec. 2018, doi: 10.1186/s13059-018-1528-8.
[141] International Wheat Genome Sequencing Consortium *et al.*, "Impact of transposable elements on genome structure and evolution in bread wheat," *Genome Biol.*, vol. 19, no. 1, p. 103, Dec. 2018, doi: 10.1186/s13059-018-1479-0.

[142] S.-I. Lee and N.-S. Kim, “Transposable elements and genome size variations in plants,” *Genomics Inform.*, vol. 12, no. 3, pp. 87–97, Sep. 2014, doi: 10.5808/GI.2014.12.3.87.
[143] T. F. Smith and M. S. Waterman, “Identification of common molecular subsequences,” *J. Mol. Biol.*, vol. 147, no. 1, pp. 195–197, Mar. 1981, doi: 10.1016/0022-2836(81)90087-5.
[144] G. Yu, “Using ggtree to Visualize Data on Tree-Like Structures,” *Curr. Protoc. Bioinforma.*, vol. 69, no. 1, Mar. 2020, doi: 10.1002/cpbi.96.
[145] A. Zuccolo, J. S. S. Ammiraju, H. Kim, A. Sanyal, S. Jackson, and R. A. Wing, “Rapid and Differential Proliferation of the Ty3-Gypsy LTR Retrotransposon Atlantys in the Genus Oryza,” *Rice*, vol. 1, no. 1, pp. 85–99, Sep. 2008, doi: 10.1007/s12284-008-9002-y.
[146] A. Zuccolo *et al.*, “Transposable element distribution, abundance and role in genome size variation in the genus Oryza,” *BMC Evol. Biol.*, vol. 7, p. 152, Aug. 2007, doi: 10.1186/1471-2148-7-152.
[147] Y. Xiong and T. H. Eickbush, “Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns,” *Mol. Biol. Evol.*, vol. 5, no. 6, pp. 675–690, Nov. 1988, doi: 10.1093/oxfordjournals.molbev.a040521.
[148] L. Gao, E. M. McCarthy, E. W. Ganko, and J. F. McDonald, “Evolutionary history of Oryza sativa LTR retrotransposons: a preliminary survey of the rice genome sequences,” *BMC Genomics*, vol. 5, no. 1, p. 18, Dec. 2004, doi: 10.1186/1471-2164-5-18.
[149] R. S. Baucom *et al.*, “Exceptional Diversity, Non-Random Distribution, and Rapid Evolution of Retroelements in the B73 Maize Genome,” *PLoS Genet.*, vol. 5, no. 11, p. e1000732, Nov. 2009, doi: 10.1371/journal.pgen.1000732.
[150] Y. Peng *et al.*, “Phylogenetic relationships in the genus Avena based on the nuclear Pgk1 gene,” *PloS One*, vol. 13, no. 11, p. e0200047, 2018, doi: 10.1371/journal.pone.0200047.
[151] B. Langmead and S. L. Salzberg, “Fast gapped-read alignment with Bowtie 2,” *Nat. Methods*, vol. 9, no. 4, pp. 357–359, Apr. 2012, doi: 10.1038/nmeth.1923.
[152] H. Li *et al.*, “The Sequence Alignment/Map format and SAMtools,” *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009, doi: 10.1093/bioinformatics/btp352.
[153] P. Rice, I. Longden, and A. Bleasby, “EMBOSS: The European Molecular Biology Open Software Suite,” *Trends Genet.*, vol. 16, no. 6, pp. 276–277, Jun. 2000, doi: 10.1016/S0168-9525(00)02024-2.
[154] P. J. A. Cock *et al.*, “Biopython: freely available Python tools for computational molecular biology and bioinformatics,” *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, Jun. 2009, doi: 10.1093/bioinformatics/btp163.
[155] S. Capella-Gutiérrez, J. M. Silla-Martínez, and T. Gabaldón, “trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses,” *Bioinforma. Oxf. Engl.*, vol. 25, no. 15, pp. 1972–1973, Aug. 2009, doi: 10.1093/bioinformatics/btp348.
[156] R. D. Finn, J. Clements, and S. R. Eddy, “HMMER web server: interactive sequence similarity searching,” *Nucleic Acids Res.*, vol. 39, no. Web Server issue, pp. W29-37, Jul. 2011, doi: 10.1093/nar/gkr367.
[157] C. Llorens *et al.*, “The Gypsy Database (GyDB) of mobile genetic elements: release 2.0,” *Nucleic Acids Res.*, vol. 39, no. Database, pp. D70–D74, Jan. 2011, doi: 10.1093/nar/gkq1061.

[158] W. Shen, S. Le, Y. Li, and F. Hu, "SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation," *PLOS ONE*, vol. 11, no. 10, p. e0163962, Oct. 2016, doi: 10.1371/journal.pone.0163962.

[159] C. Llorens, A. Muñoz-Pomer, L. Bernad, H. Botella, and A. Moya, "Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees," *Biol. Direct*, vol. 4, p. 41, Nov. 2009, doi: 10.1186/1745-6150-4-41.

[160] P. Neumann, P. Novák, N. Hoštáková, and J. Macas, "Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification," *Mob. DNA*, vol. 10, no. 1, p. 1, Dec. 2019, doi: 10.1186/s13100-018-0144-1.

[161] NCBI Resource Coordinators *et al.*, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D8–D13, Jan. 2018, doi: 10.1093/nar/gkx1095.

[162] T. Wicker and B. Keller, "Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families," *Genome Res.*, vol. 17, no. 7, pp. 1072–1081, Jul. 2007, doi: 10.1101/gr.6214107.

[163] E. F. Pettersen *et al.*, "UCSF Chimera?A visualization system for exploratory research and analysis," *J. Comput. Chem.*, vol. 25, no. 13, pp. 1605–1612, Oct. 2004, doi: 10.1002/jcc.20084.

[164] S. Lu *et al.*, "CDD/SPARCLE: the conserved domain database in 2020," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D265–D268, 08 2020, doi: 10.1093/nar/gkz991.

Appendix

Appendix A, R Scripts for Processing Reference Annotations

Oryza sativa Processing Scripts

https://github.com/shelvasha/repbox/blob/master/test/Oryza.md.Rmd

Arabidopsis thaliana Processing Scripts

https://github.com/shelvasha/repbox/blob/master/test/Arabidopsis.md.Rmd

Appendix B, Overlap Regions Annotation

Oryza sativa – MITE overlap

https://github.com/shelvasha/repbox/blob/master/test/Oryza_MITEFinder-MITETracker_comparision.txt

Oryza sativa – Helitron overlap

https://github.com/shelvasha/repbox/blob/master/test/Oryza_EAHelitron-Helitronscanner_comparision.txt

Arabidopsis thaliana – MITE overlap

https://github.com/shelvasha/repbox/blob/master/test/Arabidopsos_MITEFinder-MITETracker_comparision.txt

Arabidopsis thaliana – Helitron overlap

https://github.com/shelvasha/repbox/blob/master/test/Arabidopsos_EAHelitron-Helitronscanner_comparision.txt

Appendix C, Excerpt of Positional Comparison of RepeatModeler v2 and Repbox in A. eriantha

Chrom A	*Type A*	*Start A*	*End A*	*Chrom B*	*Type B*	*Start B*	*End B*
lcl_SctCcfP_2331_4482	*Unknown*	*313702789*	*313703306*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313702365*	*313704190*
lcl_SctCcfP_2331_4482	*Unknown*	*313703291*	*313703886*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313702365*	*313704190*
lcl_SctCcfP_2331_4482	*LTR/Gypsy*	*313705774*	*313709269*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313706131*	*313715555*
lcl_SctCcfP_2331_4482	*LTR/Gypsy*	*313708940*	*313710644*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313706131*	*313715555*
lcl_SctCcfP_2331_4482	*Unknown*	*313710645*	*313711627*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313706131*	*313715555*
lcl_SctCcfP_2331_4482	*Unknown*	*313711588*	*313711676*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313706131*	*313715555*
lcl_SctCcfP_2331_4482	*Unknown*	*313711688*	*313711872*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313706131*	*313715555*
lcl_SctCcfP_2331_4482	*Unknown*	*313711872*	*313714061*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313706131*	*313715555*
lcl_SctCcfP_2331_4482	*Unknown*	*313714028*	*313714067*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313706131*	*313715555*
lcl_SctCcfP_2331_4482	*Unknown*	*313714781*	*313714919*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313706131*	*313715555*
lcl_SctCcfP_2331_4482	*Unknown*	*313714781*	*313714919*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313714597*	*313715568*
lcl_SctCcfP_2331_4482	*Unknown*	*313715122*	*313715280*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313706131*	*313715555*
lcl_SctCcfP_2331_4482	*Unknown*	*313715122*	*313715280*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313714597*	*313715568*
lcl_SctCcfP_2331_4482	*Unknown*	*313717093*	*313717695*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313716575*	*313717807*
lcl_SctCcfP_2331_4482	*Unknown*	*313717704*	*313720116*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313717857*	*313722677*
lcl_SctCcfP_2331_4482	*LTR/Gypsy*	*313718954*	*313721765*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313717857*	*313722677*
lcl_SctCcfP_2331_4482	*Unknown*	*313721722*	*313722316*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313721055*	*313723850*
lcl_SctCcfP_2331_4482	*Unknown*	*313721722*	*313722316*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313717857*	*313722677*
lcl_SctCcfP_2331_4482	*Unknown*	*313722313*	*313723199*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313721055*	*313723850*
lcl_SctCcfP_2331_4482	*Unknown*	*313723195*	*313723752*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313721055*	*313723850*
lcl_SctCcfP_2331_4482	*LTR/Gypsy*	*313738217*	*313738265*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313737674*	*313738949*
lcl_SctCcfP_2331_4482	*Unknown*	*313741998*	*313742127*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313741389*	*313744848*
lcl_SctCcfP_2331_4482	*LTR/Gypsy*	*313746814*	*313747205*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313745969*	*313749390*
lcl_SctCcfP_2331_4482	*Unknown*	*313747205*	*313748411*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313745969*	*313749390*
lcl_SctCcfP_2331_4482	*Unknown*	*313753096*	*313753160*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313753079*	*313753280*

lcl_SctCcfP_2331_4482	*Unknown*	*313755863*	*313755989*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313754728*	*313756678*
lcl_SctCcfP_2331_4482	*Unknown*	*313756749*	*313757136*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313756671*	*313757671*
lcl_SctCcfP_2331_4482	*Unknown*	*313758376*	*313758871*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313758219*	*313759254*
lcl_SctCcfP_2331_4482	*Unknown*	*313758376*	*313758871*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313758333*	*313761475*
lcl_SctCcfP_2331_4482	*Unknown*	*313759357*	*313759468*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313758333*	*313761475*
lcl_SctCcfP_2331_4482	*Unknown*	*313772530*	*313773047*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313770636*	*313781085*
lcl_SctCcfP_2331_4482	*Unknown*	*313773049*	*313773288*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313770636*	*313781085*
lcl_SctCcfP_2331_4482	*Unknown*	*313773267*	*313776051*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313770636*	*313781085*
lcl_SctCcfP_2331_4482	*Unknown*	*313775842*	*313776258*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313770636*	*313781085*
lcl_SctCcfP_2331_4482	*Unknown*	*313776213*	*313776468*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313770636*	*313781085*
lcl_SctCcfP_2331_4482	*Unknown*	*313776394*	*313778611*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313770636*	*313781085*
lcl_SctCcfP_2331_4482	*Unknown*	*313778612*	*313778772*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313770636*	*313781085*
lcl_SctCcfP_2331_4482	*Unknown*	*313778773*	*313779052*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313770636*	*313781085*
lcl_SctCcfP_2331_4482	*Unknown*	*313778939*	*313779575*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313770636*	*313781085*
lcl_SctCcfP_2331_4482	*LTR/Gypsy*	*313779555*	*313779986*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313770636*	*313781085*
lcl_SctCcfP_2331_4482	*Unknown*	*313781931*	*313782112*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313781934*	*313784066*
lcl_SctCcfP_2331_4482	*Unknown*	*313782042*	*313782187*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313781934*	*313784066*
lcl_SctCcfP_2331_4482	*Unknown*	*313782188*	*313782976*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313781934*	*313784066*
lcl_SctCcfP_2331_4482	*LTR/Gypsy*	*313782961*	*313784203*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313781934*	*313784066*
lcl_SctCcfP_2331_4482	*LTR/Gypsy*	*313792493*	*313796348*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313793259*	*313796700*
lcl_SctCcfP_2331_4482	*LTR/Gypsy*	*313796546*	*313801906*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313796701*	*313805093*
lcl_SctCcfP_2331_4482	*Unknown*	*313800954*	*313802039*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313796701*	*313805093*
lcl_SctCcfP_2331_4482	*Unknown*	*313802055*	*313803456*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313796701*	*313805093*
lcl_SctCcfP_2331_4482	*Unknown*	*313803653*	*313803980*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313796701*	*313805093*
lcl_SctCcfP_2331_4482	*LTR/Gypsy*	*313807419*	*313807885*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313804928*	*313807869*
lcl_SctCcfP_2331_4482	*LTR/Gypsy*	*313807884*	*313807970*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313807870*	*313808114*
lcl_SctCcfP_2331_4482	*Unknown*	*313807971*	*313807988*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313807870*	*313808114*

lcl_SctCcfP_2331_4482	*Unknown*	*313807971*	*313807988*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313807947*	*313808866*
lcl_SctCcfP_2331_4482	*LTR/Gypsy*	*313807989*	*313808023*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313807870*	*313808114*
lcl_SctCcfP_2331_4482	*LTR/Gypsy*	*313807989*	*313808023*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313807947*	*313808866*
lcl_SctCcfP_2331_4482	*LTR/Copia*	*313808024*	*313808605*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313807947*	*313808866*
lcl_SctCcfP_2331_4482	*Unknown*	*313812833*	*313814850*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313811959*	*313825369*
lcl_SctCcfP_2331_4482	*LTR/Copia*	*313814846*	*313823581*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313811959*	*313825369*
lcl_SctCcfP_2331_4482	*Unknown*	*313823582*	*313824551*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313811959*	*313825369*
lcl_SctCcfP_2331_4482	*Unknown*	*313824550*	*313825060*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313811959*	*313825369*
lcl_SctCcfP_2331_4482	*LTR/Gypsy*	*313825825*	*313825940*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313825370*	*313826743*
lcl_SctCcfP_2331_4482	*LTR/Gypsy*	*313826045*	*313826908*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313825370*	*313826743*
lcl_SctCcfP_2331_4482	*Unknown*	*313826909*	*313827036*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313826844*	*313829796*
lcl_SctCcfP_2331_4482	*Unknown*	*313827534*	*313828211*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313826844*	*313829796*
lcl_SctCcfP_2331_4482	*Unknown*	*313828199*	*313829781*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313826844*	*313829796*
lcl_SctCcfP_2331_4482	*Unknown*	*313828199*	*313829781*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313828148*	*313832051*
lcl_SctCcfP_2331_4482	*Unknown*	*313829782*	*313830381*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313828148*	*313832051*
lcl_SctCcfP_2331_4482	*Unknown*	*313830382*	*313830810*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313828148*	*313832051*
lcl_SctCcfP_2331_4482	*Unknown*	*313830811*	*313831685*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313828148*	*313832051*
lcl_SctCcfP_2331_4482	*Unknown*	*313834206*	*313835278*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313834161*	*313836304*
lcl_SctCcfP_2331_4482	*LTR/Copia*	*313835055*	*313835305*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313834161*	*313836304*
lcl_SctCcfP_2331_4482	*LTR/Gypsy*	*313840495*	*313841284*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313839807*	*313843604*
lcl_SctCcfP_2331_4482	*Unknown*	*313841284*	*313841951*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313839807*	*313843604*
lcl_SctCcfP_2331_4482	*Unknown*	*313841871*	*313842050*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313839807*	*313843604*
lcl_SctCcfP_2331_4482	*Unknown*	*313842051*	*313842180*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313839807*	*313843604*
lcl_SctCcfP_2331_4482	*Unknown*	*313842181*	*313842213*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313839807*	*313843604*
lcl_SctCcfP_2331_4482	*Unknown*	*313842193*	*313842471*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313839807*	*313843604*
lcl_SctCcfP_2331_4482	*Unknown*	*313842475*	*313843391*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313839807*	*313843604*
lcl_SctCcfP_2331_4482	*Unknown*	*313843392*	*313843428*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313839807*	*313843604*

lcl_SctCcfP_2331_4482	*Unknown*	*313844756*	*313844890*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313844749*	*313852651*
lcl_SctCcfP_2331_4482	*Unknown*	*313844900*	*313845008*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313844749*	*313852651*
lcl_SctCcfP_2331_4482	*Unknown*	*313845000*	*313845072*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313844749*	*313852651*
lcl_SctCcfP_2331_4482	*Simple_repeat*	*313845083*	*313845228*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313844749*	*313852651*
lcl_SctCcfP_2331_4482	*Unknown*	*313845229*	*313846438*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313844749*	*313852651*
lcl_SctCcfP_2331_4482	*Unknown*	*313846416*	*313846569*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313844749*	*313852651*
lcl_SctCcfP_2331_4482	*Unknown*	*313846448*	*313846592*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313844749*	*313852651*
lcl_SctCcfP_2331_4482	*Unknown*	*313846571*	*313846950*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313844749*	*313852651*
lcl_SctCcfP_2331_4482	*Unknown*	*313846960*	*313847076*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313844749*	*313852651*
lcl_SctCcfP_2331_4482	*Unknown*	*313847080*	*313847159*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313844749*	*313852651*
lcl_SctCcfP_2331_4482	*Unknown*	*313847161*	*313847310*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313844749*	*313852651*
lcl_SctCcfP_2331_4482	*Unknown*	*313847309*	*313848296*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313844749*	*313852651*
lcl_SctCcfP_2331_4482	*Unknown*	*313848296*	*313850478*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313844749*	*313852651*
lcl_SctCcfP_2331_4482	*Unknown*	*313850091*	*313850787*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313844749*	*313852651*
lcl_SctCcfP_2331_4482	*Unknown*	*313859573*	*313859820*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313858284*	*313860162*
lcl_SctCcfP_2331_4482	*LTR/Gypsy*	*313860654*	*313861284*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313860145*	*313863928*
lcl_SctCcfP_2331_4482	*LTR/Gypsy*	*313861291*	*313864023*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313860145*	*313863928*
lcl_SctCcfP_2331_4482	*LTR/Gypsy*	*313864024*	*313868061*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313864500*	*313872422*
lcl_SctCcfP_2331_4482	*Unknown*	*313868192*	*313868313*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313864500*	*313872422*
lcl_SctCcfP_2331_4482	*Unknown*	*313868334*	*313868946*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313864500*	*313872422*
lcl_SctCcfP_2331_4482	*Unknown*	*313868946*	*313869846*	*lcl_SctCcfP_2331_4482*	*LTR/Gypsy*	*313864500*	*313872422*

Appendix D, Excerpt of Positional Comparison of RepeatModeler v2 and Repbox in A. atlantica

Chrom A	*Type A*	*Start A*	*End A*	*Chrom B*	*Type B*	*Start B*	*End B*
lcl_ScoFOjO_1000_1280	*LTR/Copia*	*17*	*1299*	*lcl_ScoFOjO_1000_1280*	*LTR/Gypsy*	*16*	*1295*
lcl_ScoFOjO_1000_1280	*Unknown*	*2216*	*4538*	*lcl_ScoFOjO_1000_1280*	*LTR/Gypsy*	*2216*	*4386*
lcl_ScoFOjO_1000_1280	*Unknown*	*6115*	*47260*	*lcl_ScoFOjO_1000_1280*	*LTR/Gypsy*	*6115*	*45943*
lcl_ScoFOjO_1001_1282	*LTR/Copia*	*1*	*13624*	*lcl_ScoFOjO_1001_1282*	*LTR/Gypsy*	*1*	*12824*
lcl_ScoFOjO_1001_1282	*LTR/Copia*	*13625*	*13795*	*lcl_ScoFOjO_1001_1282*	*LTR/Gypsy*	*13625*	*13795*
lcl_ScoFOjO_1002_1284	*LTR/Gypsy*	*3791*	*7818*	*lcl_ScoFOjO_1002_1284*	*LTR/Gypsy*	*3807*	*7818*
lcl_ScoFOjO_1002_1284	*LTR/Gypsy*	*17422*	*18263*	*lcl_ScoFOjO_1002_1284*	*LTR/Gypsy*	*17415*	*18390*
lcl_ScoFOjO_1002_1284	*Unknown*	*18482*	*19033*	*lcl_ScoFOjO_1002_1284*	*LTR/Gypsy*	*18482*	*19033*
lcl_ScoFOjO_1002_1284	*LTR/Gypsy*	*20781*	*23378*	*lcl_ScoFOjO_1002_1284*	*LTR/Gypsy*	*20773*	*22878*
lcl_ScoFOjO_1002_1284	*LTR*	*26144*	*28688*	*lcl_ScoFOjO_1002_1284*	*LTR/Gypsy*	*26147*	*28684*
lcl_ScoFOjO_1003_1286	*DNA/PIF-Harbinger*	*1691*	*1812*	*lcl_ScoFOjO_1003_1286*	*Unknown*	*1694*	*1820*
lcl_ScoFOjO_1003_1286	*DNA/PIF-Harbinger*	*1691*	*1812*	*lcl_ScoFOjO_1003_1286*	*LTR/Gypsy*	*1699*	*1806*
lcl_ScoFOjO_1003_1286	*Unknown*	*27299*	*28481*	*lcl_ScoFOjO_1003_1286*	*LTR/Gypsy*	*27299*	*28717*
lcl_ScoFOjO_1003_1286	*LTR/Gypsy*	*29418*	*32789*	*lcl_ScoFOjO_1003_1286*	*LTR/Gypsy*	*28718*	*32799*
lcl_ScoFOjO_1003_1286	*LTR/Gypsy*	*32796*	*34198*	*lcl_ScoFOjO_1003_1286*	*LTR/Gypsy*	*32796*	*34198*
lcl_ScoFOjO_1003_1286	*Unknown*	*34604*	*35907*	*lcl_ScoFOjO_1003_1286*	*Unknown*	*34332*	*35694*
lcl_ScoFOjO_1003_1286	*LTR/Gypsy*	*37028*	*39572*	*lcl_ScoFOjO_1003_1286*	*LTR/Gypsy*	*37027*	*39572*
lcl_ScoFOjO_1004_1287	*LTR/Gypsy*	*1*	*427*	*lcl_ScoFOjO_1004_1287*	*LTR/Gypsy*	*1*	*430*
lcl_ScoFOjO_1004_1287	*LTR/Copia*	*2866*	*3016*	*lcl_ScoFOjO_1004_1287*	*LTR/Gypsy*	*2874*	*3022*
lcl_ScoFOjO_1004_1287	*LTR/Copia*	*3026*	*6179*	*lcl_ScoFOjO_1004_1287*	*LTR/Gypsy*	*3014*	*6184*

lcl_ScoFOjO_1004_1287	*LTR*	*6580*	*9782*	*lcl_ScoFOjO_1004_1287*	*Unknown*	*7214*	*9782*
lcl_ScoFOjO_1004_1287	*LTR/Copia*	*12519*	*14101*	*lcl_ScoFOjO_1004_1287*	*LTR/Gypsy*	*12525*	*14099*
lcl_ScoFOjO_1004_1287	*LTR/Gypsy*	*14102*	*18142*	*lcl_ScoFOjO_1004_1287*	*LTR/Gypsy*	*14095*	*18145*
lcl_ScoFOjO_1004_1287	*LTR/Copia*	*18143*	*21533*	*lcl_ScoFOjO_1004_1287*	*LTR/Gypsy*	*18143*	*21531*
lcl_ScoFOjO_1004_1287	*LTR/Copia*	*21589*	*29071*	*lcl_ScoFOjO_1004_1287*	*LTR/Gypsy*	*21644*	*29071*
lcl_ScoFOjO_1005_1288	*LTR/Copia*	*4*	*9052*	*lcl_ScoFOjO_1005_1288*	*LTR/Gypsy*	*5*	*9053*
lcl_ScoFOjO_1005_1288	*LTR/Gypsy*	*9047*	*10185*	*lcl_ScoFOjO_1005_1288*	*LTR/Gypsy*	*9053*	*10185*
lcl_ScoFOjO_1005_1288	*LTR/Gypsy*	*10185*	*12478*	*lcl_ScoFOjO_1005_1288*	*LTR/Gypsy*	*10185*	*12478*
lcl_ScoFOjO_1005_1288	*LTR/Copia*	*12476*	*14320*	*lcl_ScoFOjO_1005_1288*	*LTR/Gypsy*	*12476*	*14320*
lcl_ScoFOjO_1005_1288	*LTR/Gypsy*	*25624*	*27772*	*lcl_ScoFOjO_1005_1288*	*LTR/Gypsy*	*25624*	*27772*
lcl_ScoFOjO_1005_1288	*LTR/Gypsy*	*27772*	*30063*	*lcl_ScoFOjO_1005_1288*	*LTR/Gypsy*	*27772*	*30063*
lcl_ScoFOjO_1005_1288	*LTR/Copia*	*30061*	*31901*	*lcl_ScoFOjO_1005_1288*	*LTR/Gypsy*	*30061*	*31901*
lcl_ScoFOjO_1005_1288	*LTR/Copia*	*31900*	*33204*	*lcl_ScoFOjO_1005_1288*	*Unknown*	*31940*	*33205*
lcl_ScoFOjO_1005_1288	*LTR/Copia*	*33211*	*40867*	*lcl_ScoFOjO_1005_1288*	*LTR/Gypsy*	*33211*	*40867*
lcl_ScoFOjO_1006_1290	*LTR/Gypsy*	*3*	*2005*	*lcl_ScoFOjO_1006_1290*	*LTR/Gypsy*	*1*	*2027*
lcl_ScoFOjO_1006_1290	*LTR/Gypsy*	*2038*	*6324*	*lcl_ScoFOjO_1006_1290*	*LTR/Gypsy*	*2038*	*6322*
lcl_ScoFOjO_1006_1290	*LTR/Gypsy*	*6322*	*10176*	*lcl_ScoFOjO_1006_1290*	*LTR/Gypsy*	*6323*	*10176*
lcl_ScoFOjO_1006_1290	*LTR/Gypsy*	*10149*	*10436*	*lcl_ScoFOjO_1006_1290*	*LTR/Gypsy*	*10146*	*10436*
lcl_ScoFOjO_1006_1290	*LTR/Gypsy*	*10149*	*10436*	*lcl_ScoFOjO_1006_1290*	*LTR/Gypsy*	*10177*	*10438*
lcl_ScoFOjO_1006_1290	*LTR/Copia*	*10438*	*11650*	*lcl_ScoFOjO_1006_1290*	*LTR/Gypsy*	*10437*	*11650*
lcl_ScoFOjO_1006_1290	*LTR/Gypsy*	*11648*	*11954*	*lcl_ScoFOjO_1006_1290*	*LTR/Gypsy*	*11649*	*11957*
lcl_ScoFOjO_1006_1290	*LTR/Copia*	*11953*	*13655*	*lcl_ScoFOjO_1006_1290*	*LTR/Gypsy*	*11950*	*13661*

lcl_ScoFOjO_1006_1290	*LTR/Copia*	*13657*	*14677*	*lcl_ScoFOjO_1006_1290*	*LTR/Gypsy*	*13657*	*14677*
lcl_ScoFOjO_1006_1290	*Unknown*	*14700*	*16083*	*lcl_ScoFOjO_1006_1290*	*LTR/Gypsy*	*14677*	*16078*
lcl_ScoFOjO_1006_1290	*LTR/Copia*	*18066*	*20779*	*lcl_ScoFOjO_1006_1290*	*LTR/Gypsy*	*17418*	*20740*
lcl_ScoFOjO_1006_1290	*LTR/Copia*	*20772*	*21415*	*lcl_ScoFOjO_1006_1290*	*LTR/Gypsy*	*20772*	*21417*
lcl_ScoFOjO_1006_1290	*LTR/Gypsy*	*21416*	*22189*	*lcl_ScoFOjO_1006_1290*	*LTR/Gypsy*	*21416*	*22189*
lcl_ScoFOjO_1006_1290	*LTR/Copia*	*22180*	*29197*	*lcl_ScoFOjO_1006_1290*	*LTR/Gypsy*	*22259*	*29197*
lcl_ScoFOjO_1006_1290	*Unknown*	*29176*	*53848*	*lcl_ScoFOjO_1006_1290*	*Unknown*	*29197*	*53522*
lcl_ScoFOjO_1006_1290	*Simple_repeat*	*53900*	*53997*	*lcl_ScoFOjO_1006_1290*	*Simple_repeat*	*53900*	*53997*
lcl_ScoFOjO_1006_1290	*Unknown*	*76159*	*78529*	*lcl_ScoFOjO_1006_1290*	*LTR/Gypsy*	*76145*	*78527*
lcl_ScoFOjO_1006_1290	*Unknown*	*76159*	*78529*	*lcl_ScoFOjO_1006_1290*	*Unknown*	*76432*	*78529*
lcl_ScoFOjO_1006_1290	*Unknown*	*78516*	*80556*	*lcl_ScoFOjO_1006_1290*	*LTR/Gypsy*	*78502*	*80554*
lcl_ScoFOjO_1006_1290	*Unknown*	*78516*	*80556*	*lcl_ScoFOjO_1006_1290*	*Unknown*	*78516*	*80556*
lcl_ScoFOjO_1006_1290	*Unknown*	*80543*	*96887*	*lcl_ScoFOjO_1006_1290*	*LTR/Gypsy*	*80530*	*96758*
lcl_ScoFOjO_1006_1290	*Unknown*	*96920*	*97045*	*lcl_ScoFOjO_1006_1290*	*Unknown*	*96920*	*97045*
lcl_ScoFOjO_1006_1290	*Unknown*	*98391*	*98536*	*lcl_ScoFOjO_1006_1290*	*LTR/Gypsy*	*98391*	*98560*
lcl_ScoFOjO_1006_1290	*Simple_repeat*	*102662*	*102719*	*lcl_ScoFOjO_1006_1290*	*Simple_repeat*	*102662*	*102719*
lcl_ScoFOjO_1006_1290	*Simple_repeat*	*103231*	*103266*	*lcl_ScoFOjO_1006_1290*	*Simple_repeat*	*103231*	*103266*
lcl_ScoFOjO_1006_1290	*Simple_repeat*	*104874*	*104911*	*lcl_ScoFOjO_1006_1290*	*Simple_repeat*	*104874*	*104911*
lcl_ScoFOjO_1006_1290	*Simple_repeat*	*105227*	*105260*	*lcl_ScoFOjO_1006_1290*	*Simple_repeat*	*105227*	*105260*
lcl_ScoFOjO_1006_1290	*LTR/Copia*	*106615*	*115647*	*lcl_ScoFOjO_1006_1290*	*LTR/Gypsy*	*106665*	*115474*
lcl_ScoFOjO_1007_1292	*LTR/Copia*	*4907*	*13964*	*lcl_ScoFOjO_1007_1292*	*LTR/Gypsy*	*4877*	*13964*
lcl_ScoFOjO_1007_1292	*LTR/Copia*	*13965*	*14890*	*lcl_ScoFOjO_1007_1292*	*LTR/Gypsy*	*13965*	*14890*

lcl_ScoFOjO_1008_1294	*LTR/Gypsy*	*1*	*414*	*lcl_ScoFOjO_1008_1294*	*LTR/Gypsy*	*1*	*413*
lcl_ScoFOjO_1008_1294	*Unknown*	*1396*	*11455*	*lcl_ScoFOjO_1008_1294*	*LTR/Gypsy*	*1528*	*11454*
lcl_ScoFOjO_1009_1296	*LTR/Copia*	*1*	*8074*	*lcl_ScoFOjO_1009_1296*	*LTR/Gypsy*	*532*	*8022*
lcl_ScoFOjO_1009_1296	*Unknown*	*9655*	*10225*	*lcl_ScoFOjO_1009_1296*	*Unknown*	*9698*	*10224*
lcl_ScoFOjO_1009_1296	*Unknown*	*10227*	*10327*	*lcl_ScoFOjO_1009_1296*	*Unknown*	*10227*	*10351*
lcl_ScoFOjO_1009_1296	*Simple_repeat*	*13086*	*13166*	*lcl_ScoFOjO_1009_1296*	*Simple_repeat*	*13086*	*13166*
lcl_ScoFOjO_1009_1296	*Unknown*	*13999*	*14168*	*lcl_ScoFOjO_1009_1296*	*Unknown*	*13999*	*14168*
lcl_ScoFOjO_1009_1296	*LTR/Copia*	*14491*	*14593*	*lcl_ScoFOjO_1009_1296*	*LTR/Gypsy*	*14488*	*14593*
lcl_ScoFOjO_1009_1296	*LTR/Gypsy*	*14594*	*18212*	*lcl_ScoFOjO_1009_1296*	*LTR/Gypsy*	*14594*	*18212*
lcl_ScoFOjO_1009_1296	*LTR/Gypsy*	*19321*	*21750*	*lcl_ScoFOjO_1009_1296*	*LTR/Gypsy*	*19325*	*21750*
lcl_ScoFOjO_1009_1296	*Simple_repeat*	*23501*	*23535*	*lcl_ScoFOjO_1009_1296*	*Simple_repeat*	*23501*	*23535*
lcl_ScoFOjO_1009_1296	*Simple_repeat*	*23983*	*24036*	*lcl_ScoFOjO_1009_1296*	*Simple_repeat*	*23983*	*24036*
lcl_ScoFOjO_1009_1296	*Low_complexity*	*25263*	*25319*	*lcl_ScoFOjO_1009_1296*	*Low_complexity*	*25263*	*25319*
lcl_ScoFOjO_100_141	*LTR/Copia*	*2304*	*4195*	*lcl_ScoFOjO_100_141*	*LTR/Gypsy*	*2300*	*4195*
lcl_ScoFOjO_100_141	*Unknown*	*4195*	*14552*	*lcl_ScoFOjO_100_141*	*LTR/Gypsy*	*4203*	*14546*
lcl_ScoFOjO_100_141	*Unknown*	*15285*	*15683*	*lcl_ScoFOjO_100_141*	*Unknown*	*15285*	*15683*
lcl_ScoFOjO_100_141	*Unknown*	*16515*	*31519*	*lcl_ScoFOjO_100_141*	*LTR/Gypsy*	*16651*	*31517*
lcl_ScoFOjO_100_141	*Unknown*	*32351*	*32700*	*lcl_ScoFOjO_100_141*	*Unknown*	*32351*	*32700*
lcl_ScoFOjO_100_141	*Unknown*	*47018*	*52009*	*lcl_ScoFOjO_100_141*	*LTR/Gypsy*	*47150*	*51882*
lcl_ScoFOjO_100_141	*DNA/CMC-EnSpm*	*52219*	*52541*	*lcl_ScoFOjO_100_141*	*LTR/Gypsy*	*52154*	*52548*
lcl_ScoFOjO_100_141	*Simple_repeat*	*53015*	*53035*	*lcl_ScoFOjO_100_141*	*Simple_repeat*	*53015*	*53035*
lcl_ScoFOjO_100_141	*Low_complexity*	*53958*	*53996*	*lcl_ScoFOjO_100_141*	*Low_complexity*	*53958*	*53996*

lcl_ScoFOjO_100_141	*Simple_repeat*	*54276*	*54294*	*lcl_ScoFOjO_100_141*	*Simple_repeat*	*54276*	*54294*
lcl_ScoFOjO_100_141	*LTR/Copia*	*62945*	*64858*	*lcl_ScoFOjO_100_141*	*LTR/Gypsy*	*62941*	*64867*
lcl_ScoFOjO_100_141	*Unknown*	*64856*	*72748*	*lcl_ScoFOjO_100_141*	*LTR/Gypsy*	*64864*	*72746*
lcl_ScoFOjO_100_141	*Unknown*	*73585*	*73934*	*lcl_ScoFOjO_100_141*	*Unknown*	*73585*	*73934*
lcl_ScoFOjO_100_141	*Unknown*	*74825*	*123556*	*lcl_ScoFOjO_100_141*	*LTR/Gypsy*	*83035*	*123429*
lcl_ScoFOjO_1010_1297	*LTR/Gypsy*	*192*	*1272*	*lcl_ScoFOjO_1010_1297*	*LTR/Gypsy*	*192*	*1272*
lcl_ScoFOjO_1010_1297	*Unknown*	*1265*	*2185*	*lcl_ScoFOjO_1010_1297*	*LTR/Gypsy*	*1251*	*2181*
lcl_ScoFOjO_1010_1297	*Unknown*	*2364*	*4158*	*lcl_ScoFOjO_1010_1297*	*Unknown*	*2509*	*4158*
lcl_ScoFOjO_1010_1297	*Unknown*	*16306*	*18294*	*lcl_ScoFOjO_1010_1297*	*LTR/Gypsy*	*16306*	*18294*
lcl_ScoFOjO_1010_1297	*Simple_repeat*	*18308*	*18395*	*lcl_ScoFOjO_1010_1297*	*Simple_repeat*	*18308*	*18395*
lcl_ScoFOjO_1010_1297	*Unknown*	*19022*	*20383*	*lcl_ScoFOjO_1010_1297*	*Unknown*	*19007*	*20374*
lcl_ScoFOjO_1010_1297	*Unknown*	*19022*	*20383*	*lcl_ScoFOjO_1010_1297*	*LTR/Gypsy*	*19206*	*20383*
lcl_ScoFOjO_1010_1297	*LTR/Gypsy*	*22551*	*23614*	*lcl_ScoFOjO_1010_1297*	*LTR/Gypsy*	*22526*	*23614*