A Bioinformatics Pipeline to Study the Dynamics of Retrotransposons in Grass Evolution

by

Andrew E. Murray

Honors Thesis

Appalachian State University

Submitted to the Department of Biology in partial fulfillment of the requirements for the degree of

Bachelor of Arts in Biology

May 2017

Approved by:

_____

Matt C Estep, Ph.D., Thesis Director

_____

Brooke E Christian, Ph.D., Second Reader

_____

Lynn Siefferman, Ph.D., Departmental Honors Director

Abstract

Transposable elements (TEs) are repetitive sequences found within all eukaryotic genomes and possess the genetic mechanisms necessary to move locations within their host. The repetitive fraction of plant genomes has historically been viewed as “junk DNA”. A growing body of evidence suggests that repetitive sequences play a large role in the evolution of species and may influence diversification rates. Using bioinformatic approaches we mined nine million, 250 base pair illumina reads from four grass species closely related to maize (*Zea mays*) and have identified 121 Long Terminal Repeat-Retrotransposons (LTR-RTs). These include TE’s previously identified in maize (3) and novel LTR-RTs (118) using the 80:80:80 homology rules for identification. Using these identified TE’s we are able to calculate the abundance of each element within the sampled grass genomes and using a phylogenetic framework, map changes in copy number to examine the dynamics of LTR-RTs proliferation and extinction within the grass lineages. Using this bioinformatic approach we can begin to examine the complex relationship between grass diversification and the proliferation of the TEs contained within their genomes.

Introduction

Transposable elements (TEs) are mobile elements of DNA, which are the most common genetic component of many eukaryotic genomes (Feschotte and Pritham, 2007). Barbara McClintock originally described TEs in 1950 (McClintock, 1950). She was awarded the Nobel prize in 1983 for her discovery. The genomics era has shined a new light on the dynamics of TEs and their importance to eukaryotic genome evolution (Bennetzen 2014). TE proliferation can have dramatic effects on genome size, but also influences the regulation of gene expression and gene function, as well as creating new genes (Bennetzen and Wang, 2014). Gene duplication can be directly caused by retrotransposons through ectopic recombination or retrotransposition. The duplication of a gene can allow for neufunctionalization, wear duplicated genes can gain new function distinct from their ancestral gene through mutation without disrupting the original function (Conant and Wolfe, 2008). Promoters, enhancers, and silencers can be moved through retrotransposition, resulting in novel combinations or gene regulatory elements and genes (Sabot and Schulman, 2006). These changes can have massive effects on the overall function of an organism, and understanding the mechanisms that drive this change are key to understanding evolution as a whole.

TEs are extremely diverse; with thousands of different families described in plant genomes where they are best studied (Feschotte et al., 2002). TEs commonly constitute 80% or more of the total genomic DNA in plant genomes (Feschotte et al., 2002). They are usually less abundant in fungi and metazoans, comprising 3-20% of known fungal genomes (Daboussi and Capy, 2003) and anywhere from 3-45% of metazoan genomes (Hua-Van, A., et al., 2005).

The first step in any project looking at retrotransposon dynamics is the actual identification of the retrotransposons in a genome. In this work we attempt to create a pragmatic bioinformatics pipeline to identify novel TEs from raw sequencing data. We are attempting to study patterns of genome composition to identify evolutionarily significant trends. Creating a fast, user friendly, and accurate method of identifying the TEs is an essential first step.

Transposable Element Classification

TEs are classified into class I and class II transposable elements based on their replication mechanisms (Finnegan, 1989). Class II transposable elements, or DNA transposons, replicate via a “cut and paste” mechanism, where the element excises itself from the DNA at one location and reinserts itself at another location, usually nearby (Wicker et al., 2007, Sabot and Schulman, 2006). Due to the fact that these transposons cut and paste themselves, they are only able to increase their copy number if they transpose during S-phase of the cell cycle (Sabot and Schulman, 2006). DNA transposons are subdivided based on the number of DNA strands that are cut during transposition (1 or 2), and then by the variable length of the terminal inverted repeat (TIR) (Wicker et al., 2007).

Class I transposable elements, or retrotransposons, replicate via a “copy and paste” mechanism where the retrotransposon is first transcribed from a genomic copy, and then reverse-transcribed back into DNA by reverse transcriptase (Wicker et al., 2007). This mechanism allows retrotransposons to be the major contributor of repetitive content in large genomes (Kumar and Bennetzen, 1999). Retrotransposons can be further subdivided based on their structural features and reverse transcriptase

phylogeny (Wicker et al., 2007). The LTR-Retrotransposons (LTR-RT) possess long terminal repeats (LTRs), which can range from a few hundred base pairs to upwards of 5,000 base pairs, flanking internal structural sequences (Wicker et al., 2007). The internal sequences are composed of two polycistronic genes, GAG (which codes for proteins necessary for the virus like particle or VLP) and POL (which codes for proteins necessary for reverse transcription and integration) regions. The GAG domain only contains a single protein coding domain, while the POL codes for other proteins essential for retrotransposon function including: aspartic proteinase (AP), integrase (INT), reverse transcriptase (RT), RNase H (RH), and sometimes additional ORFs of known or unknown function (Wicker et al., 2007). The two main superfamilies of LTR-RTs are *Gypsy* (or TY1) and *Copia* (or TY3), which are distinguished by the order in which INT and RT appear within the LTR-RT (Wicker et al., 2007). LTR-RT and retroviruses are evolutionarily closely related, with retroviruses possessing an envelope protein (ENV) as well as some additional proteins and regulatory sequences (Wicker et al. 2007). An inactivated retrovirus can be changed into an LTR-RT through the inactivation of the ENV domain, which eliminates the ability of the virus to spread between cells. These elements can then propagate vertically through the germ line, and are labeled as endogenous retroviruses (ERVs) (Bannert, Norbert, and Kurth, 2006) and compose another LTR-RT superfamily (Wicker et al., 2007).

Retrotransposon Function

Retrotransposons within a genome can be classified as autonomous or non-autonomous. Non-autonomous elements are defined by the fact that they lack some or all of the protein coding domains that are required for transposition. However this does

not mean that they are unable to proliferate within a genome. They often share strong sequence conservation within the 5` UTR and terminal sequences with an autonomous element (due to their necessity in regards to functional transposition) and likely use the protein products transcribed by that autonomous element (Wicker et al., 2007). There are autonomous and non-autonomous partner TEs which likely function together and can possess sufficient sequence homology to be classified as subfamilies (Sabot and Schulman, 2006; Wicker et al., 2007). Regardless of their level of autonomy, all transposons require several proteins to successfully replicate within a genome.

One of the major protein products of the POL domain is Reverse Transcriptase (RT). RT leads to the production of complementary DNA (cDNA) from a single stranded RNA alongside the action of RNase H (Moelling and Broecker, 2015). RT is linked to RNase H by a linker domain, which is likely an inactive RNase H that was created via gene duplication (Malik and Eickbush, 2001). A trio of conserved amino acid residues, DDD in RT and DDE in Rnase H, coordinate divalent cations that are important for enzymatic activity and molecular specificity (Broecker et al., 2012). Rnase H is an endonuclease that is specific to DNA-RNA hybrids and degrades the RNA strand cDNA synthesis by RT (Coffin et al., 1997). The overall function of RT is initiated by the binding of a slightly modified tRNA to the primer binding site (PBS) on the RNA. Minus strand cDNA synthesis occurs until RT reaches the 5` end of the RNA, followed by Rnase H degrading the RNA complementary to the transcribed DNA sequence. This generates a negative sense DNA strand called minus-strand strong-stop DNA (-sssDNA) which is relatively short due to the binding location of the tRNA. The -sssDNA then binds to the 3` end of the RNA sequence, which is made possible due to the fact

that the LTR of the LTR-RT is identical at both the 5` and 3` end. Negative sense strand synthesis then proceeds alongside Rnase H degradation, leading to a complete negative sense DNA strand. The positive sense strand also contains a polypurine tract (PPT) which is resistant to degradation by Rnase H. The PPT is then used to initiate positive strand DNA synthesis, creating a fragment of positive sense DNA called plus-strand strong-stop DNA (+sssDNA). Rnase H then removes the tRNA primer, exposing the sequence on the negative sense DNA that is complementary to the +sssDNA near the 3` end. The +sssDNA then binds to the original PBS site and initiates positive sense DNA synthesis, resulting in two complementary DNA strands (fig. 1) (Coffin et al., 1997). RT does not require conservation of actual sequences, and only requires the presence of a PBS (Sabot and Schulman, 2006).

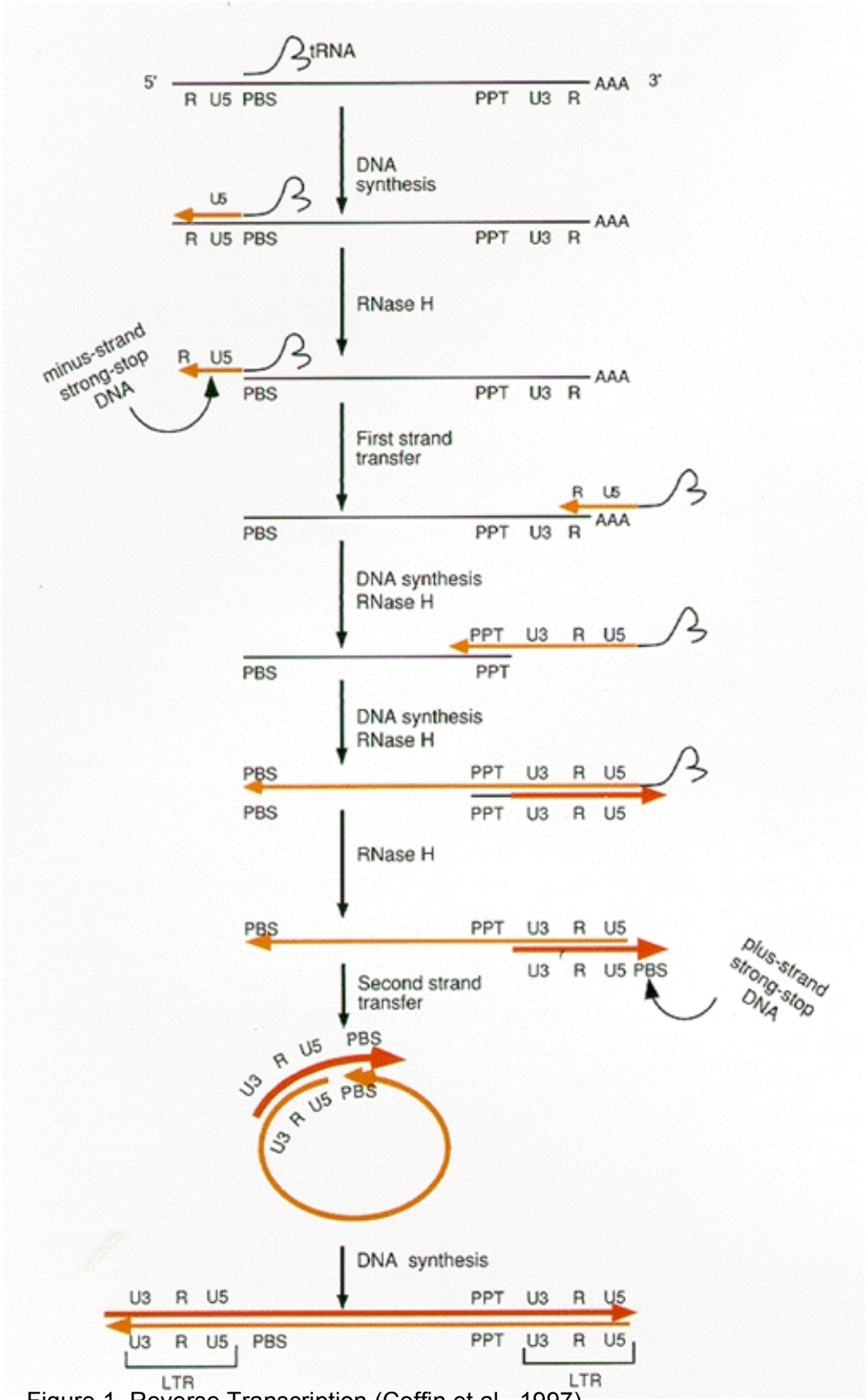

Figure 1. Reverse Transcription (Coffin et al., 1997)

RT occurs in all species, with an impressive 1021 different types identified in bacteria alone (Simon and Zimmerly, 2008). Many have some unknown functions and have often fused with other protein domains (Simon and Zimmerly, 2008). RT is of particular interest for studying early evolution and in the RNA world hypothesis is instrumental for explaining the transition for RNA to DNA (Moelling, 2013). RT shares structural similarity with parts of the spliceosome, and some have theorized that RNA splicing and introns are have been “invented” by mobile genetic elements harboring a RT (Pena et al., 2008; Nelson, Lehninger, Cox, 2008). RT also shares structural similarity and binding patterns with telomerase, which performs a similar RNA-DNA transcription process (Nelson, Lehninger, Cox, 2008).

The GAG domain contains three major functional domains: the Capsid, the Nucleocapsid and the Matrix domain (Sabot and Schulman, 2006). The Capsid domain is polymerized and forms a shell that is referred to as a virus like particle (VLP) in non-retroviral retrotransposons and is simple called a capsid on retroviruses. The Nucleocapsid interacts with the actual DNA and harbors zinc fingers and basic resides, that interact with the negatively charged phosphodiester backbone of DNA (Sabot and Schulman, 2006). Finally the Matrix domain interacts with the envelope protein, which when present and functional causes attachment of the virion or VLP to the cell membrane and subsequent budding (Sabot and Schulman, 2006; Adamson and Jones, 2004). The VLP is theoretically where reverse transcription is performed, and the nucleocapsid is thought to associate with the RNA that contains the sequence which originally coded for the GAG domain proteins used in VLP formation. While the mechanism for retroviral association with a Nucleocapsid is well understood via a PSI

(packaging signal sequence) the location of sequences that would associate with a Nucleocapsid in retrotransposons have not been verified in retrotransposons (Evans et al., 2004). However there does appear to be a high level of RNA structural conservation in areas near the PBS as well as family specific sequences that seem to strongly indicate a conserved mechanism (Sabot and Schulman, 2006). However it has been well established that non-autonomous sequences are able to effectively propagate (Kalendar et al., 2004). This likely indicates that they either contain a specific PSI that can interact with the products of an autonomous retrotransposon, a generalist PSI that can interact with many nucleocapsids, or the VLP is not as important as previously thought to retrotransposon activity (Sabot and Schulman, 2006).

The VLP is localized to the nucleus and the double stranded cDNA synthesized by RT is transferred into the nucleus, however the exact time point during RT action and cDNA formation when this occurs is unknown (Sabot and Schulman, 2006). Another major protein of the POL domain is Integrase (INT). This protein binds both regions of the LTR and facilitates insertion of the LTR-RT into the host genome. INT creates an asymmetric double-stranded break 2-16 BP long in the genomic DNA and inserts the double stranded cDNA. The double-stranded break is then repaired by DNA repair mechanisms, which leads to a target site duplication where overhanging BP are present from the asymmetric break. There does seem to be a strong bias for insertion of retrotransposons in non-genic DNA, and very few genetic mutations are associated with retrotransposon insertion, suggesting an epigenetic homing pattern (Schulman and Kalendar, 2005; Bennetzen, 2000). As with the GAG domain non-autonomous elements do seem to be able to proliferate using INT from other TE's. This likely means that they

either share sequence homology with autonomous elements allowing the use of their INT, or they possess generalist motifs that allow them to use INT from a wide variety of autonomous retrotransposons (Sabot and Schulman, 2006). INT is unique to the LTR-RT, though other elements do contain domains with similar function such as Tyrosine Recombinase in the DIRS superfamily (Sabot and Schulman, 2006; Wicker et al., 2007). In the last protein product from the POL domain is aspartic proteinase (AP), which is responsible for post translational processing of the POL protein product (Sabot and Schulman, 2006).

Genome Evolution

The most obvious way that retrotransposons contribute to genome evolution is through their size and replicative transposition. Roughly 40% of mammalian genomes are composed of retrotransposons, mainly LINEs (retrotransposons that lack an LTR) and SINEs (non-autonomous retrotransposons that lack an LTR and contain a Pol III promoter) (Finnegan, 2012; Wicker et al., 2007). In angiosperm (flowering plants) genomes TEs can contribute over 75% of the genome, and LTR-RT are often the most significant contributing factor (Schnable et al., 2009). Retrotransposons have been shown to possess the ability to rapidly accumulate over time, with evidence showing that LTR-RT have doubled the size of the Maize genome within the last 6 million years (SanMiguel et al., 1996; SanMiguel et al., 1998). The observed lack of correlation between organism complexity and genome size, termed the “C-value paradox”, can potentially be explained in the grass family with LTR-RT content (Feschotte, Jiang, & Wessler, 2002).

However despite the massive number of TEs in eukaryotic genomes, only a few seem to be active (Feschotte, 2002). Approximately 45% of the Human genomes is composed of L1 elements (LINE family), with around 500,000 identified copies. However only around 30-60 L1 elements appear to be active in human genomes today (Sassaman et al., 1997). Retrotransposons can be inactive during development, but become active due to biotic and abiotic stress later in life. This was originally demonstrated by the accidental activation of LTR-RT by exposing tobacco cells to a fungal extract in an attempt to degrade a cell wall to create a protoplast (Pouteau et al., 1994). Retrotransposons have also been shown to be activated by wounding, oxidative stress, pathogens and microbial stress (Chandler and Mahillon, 2002; Kapitonov and Jurka, 2003). While transcriptional regulation via chromatin remodeling and hypermethylation is common, the successful transcription of an LTR-RT does not necessarily guarantee insertion. Translation, reverse transcription, and integration processes can be inhibited by the host to reduce the rate of LTR-RT insertion (Feschotte, Jiang, & Wessler, 2002).

Retrotransposons impact eukaryotic genome evolution in ways other than just genome size. They can inactivate a gene by insertional mutagenesis, or separate a gene from its regulatory elements by inserting between them and the transcriptional start site. They can also bring a new gene regulatory product into the vicinity of another gene if it is contained within the sequence that is reverse transcribed, or even use the sequences contained within themselves to serve as enhancers or promoters (Slotkin and Martienssen, 2007). They can duplicate existing genes to generate new genes, by incorporating an adjacent sequence into a transposition intermediate followed by

insertion in a novel location within the genome. There is some evidence that retrotransposons can enter a genome horizontally (Finnegan, 2012). Evidence has come forward recently indicating that certain eukaryotic lineages have co-opted retrotransposons to perform highly specific regulatory functions (Bennetzen et al., 2014). They will likely prove to be powerful engines of genome evolution.

In order to study the effect of transposable elements we need to first identify them within a genome. Understanding the dynamics of retrotransposons in a genome requires us to identify the retrotransposon content across several different genomes. Because retrotransposons are error prone in replication and appear to exhibit little to no species level selection they evolve extremely rapidly, in particular in their LTR region (Sabot and Schulman, 2006). This requires us to study the genome dynamics of retrotransposons in closely related species in order to spot relationships.

Due to the fact that LTR-RTs appear so frequently in a genome we can find them with low coverage sequencing that would normally be required to continue constructing a genome. We have designed a bioinformatics pipeline that will rapidly analyze a large number of assembled raw sequence reads an allow us to analyze them for content. To that end we have also designed a bioinformatics process that will allow us to identify LTR sequences without direct sequence homology to an existing LTR or an LTR-RT. This will allow us to rapidly identify and assemble a large number of novel LTR-RT from closely related species so that we can examine the dynamics of the repetitive content within those genomes.

Methods

Database deconstruction and reconstruction

Long Terminal Repeat Retrotransposons (LTR-RT) were identified in the Maize transposable elements (TE) database (Baucom et al., 2009). Each LTR-RT was examined for protein coding domains and long terminal repeat (LTR) sequences using PFAM and NCBI (Finn et al., 2016, Johnson et al., 2008). The nucleotide positions of each protein-coding domain identified was recorded, and the sequence was copied to a unique database. A BLAST2Seq analysis was used to identify the LTR sequence, which was also copied to a unique database. A total of ten LTR-RT databases were created including: Known Retrotransposons (KN), Long Terminal Repeats (LTR), Capsid Proteins (GAG), Aspartic Proteinase (AP), Integrase (INT), Reverse Transcriptase (RT), RNase H (RH), and three Unique Protein Coding Domain Databases (U1-U3). Any protein, which did not fall into an existing category, was placed into a unique database.

Sequence Data

A total of 37,993,058 sequences from four grass species were analyzed for this study. The species include Carpetgrass (*Arthraxon prionodes*), *Chasmopodium caudatum*, Vetiver (*Chrysopogon zizanoides*), and Hippo grass (*Vossia cuspidate*), which were vouchered in previously published work (Table 1) (Estep et al., 2014). Briefly, CTAB extracted DNA was sequenced using an Illumina short sequence read by the lab PI during the summer of 2014 (Estep per. Comm.).

TE discovery

The raw data was analyzed using the contig software AAARF (DeBarry et al., 2008) using default parameters on a 2x Intel Xeon CPU E5-2697 v2 2.70GHz (16

hyperthreaded cores, 32 logical) with 256GB RAM, a 1.5TB RAID5 SSD and a 1.5TB RAID5 HDD (Biology High Performance Computer -BioHPC). The resulting pseudo-elements (contiguous sequences grouped together based on homology) were then analyzed using a custom python code and the annotation pipeline DAWGPAWS (Estill and Bennetzen, 2009). The pipeline runs a series of BLAST alignments for each pseudo-element constructed against the known LTR-RT created databases (described above), and outputs a .gff (general feature format) file summarizing the results. This file allowed the visulization of hundreds of BLAST results about each pseudo-element using simple text editing software. In order to reduce the search for novel LTR-RTs, pseudo-elements that were at least 4000 base pairs (BP) in length and contained hits to at least two protein-coding domains known to exist in LTR-RT, were selected for further analysis. Pseudo-elements that contained chloroplast sequences (also high copy) were removed from analysis using the NCBI blastn and the non-redundant nucleotide database (Johnson, Mark, et al., 2008).

The selected pseudo-elements were then aligned against the raw data and a custom R script was used to visualize a depth of coverage (DoC) graph (R Core Team, 2013). The script identified how many times any particular base pair was identified within the pseudo-element. The results were then graphed in excel to create the DoC graph. Each DoC graph was then hand analyzed to determine the probable location of the LTR sequence. Probable LTR regions identified met two qualifications; 1) at least double the background DoC and 2) region did not show homology to LTR-RT databases, except the LTR database. If LTR database homology was identified, the pseudo-element was annotated as an LTR-homologous LTR-RT.

To ensure that new LTR-RT identified through this process were unique, a nucleotide blast was performed using LTR sequences identified in other pseudo-elements. Duplicates were synonymized and a consensus sequence was chosen based on prevalence in the raw sequence reads.

Database update

Once pseudo-element annotation was completed each novel element was deconstructed (as above) and their sequences were placed in the appropriate LTR-RT database with the original Maize data. These databases are referred to as -v2.

Grass genome composition

The raw sequence data for each of the four grass species were then described with the LTR-RT-v2 databases using BLAST. Raw sequences were annotated (described/identified) using a competitive approach where the highest BLAST hit to the KN-v2 database was used for the sequence identity. The annotated raw sequences were than analyzed using a find command and counting the instances of a unique string for each individual LTR-RT in vim (a unix based text editor). Using this approach, LTR-RT sequences found within the sequence dataset could be counted. The percentages of each LTR-RT from the total reads were calculated and the total percentage of each grass genome was estimated. This percentage was the multiplied against the total genome size to estimate the total Mbp/1C of each LTR-RT in its respective genome.

Results

Database construction

A total of 616 transposons were extracted from the Maize database and deconstructed into ten databases (3049 entries). Nine of the databases (all but the KN database) were used for the actual annotation of pseudo-RT. These databases contained a total of 2433 sequences.

Table 1. Summary of Constructed Databases

Database Name	Total Sequences	Average Length (BP)
Known RT	616	5932
LTR	547	742
GAG	303	305
AP	101	322
RT	568	341
RH	3	366
INT	272	343
U1	392	401
U2	201	254
U3	46	225
Sequences used for annotation	2433	-
Total	3049	-

TE discovery

The annotation pipeline constructed here was able to identify and describe novel LTR-RT from within the ~9 million or so raw sequence reads (table 1). The AAARF algorithm constructed between 26,601 and 40,844 contigs for each grass genome. These were further reduced to between 1245 and 2820 pseudo-elements by alignment with at least one LTR-RT protein coding domain database. The remaining elements were manually reduced to between 41 and 67 pseudo-elements by size (> 4000bp) and homology (2 x database) criteria. The remaining pseudo-elements were then further reduced by identifying the LTR with an all vs. all blast comparison to between 24 and 36 pseudo-elements for each grass taxa.

Table 2. Annotation Data Flow

Species	Raw Reads	AAARF Pseudo-elements	Annotation Pipeline	Manual Review	All vs. All
Arthraxon prionodes	9,553,552	40,844	1,689	61	32
Chasmopodium caudatum	9,310,798	40,586	2,820	67	24
Chrysopogon zizanoides	9,037,374	36,751	1,245	41	30
Vossia cuspidata	10,091,334	26,601	1,714	57	36

Each pseudo-element identified as an LTR-RT was named based on LTR superfamily, species, and sequence homology. Any pseudo-element that shared sequence homology with the LTR database was named based on the highest BLAST hit. Identified pseudo-elements that did not have any sequence homology to the LTR

database were given a numerical placeholder name. Only one sequence met the 80-80-80 (minimum 80 base pairs with 80% sequence homology for at least 80% of the LTR or total sequence) criteria to be considered a novel element (Wicker et al., 2007). The only previously identified element was RLC_VoCu_n10, which contained 82.23% homology to the maize retrotransposon RLC_ibulaf_AC186801-1662 for 3720 base pairs, which is 80.21% of the length of the LTR-RT.

LTR Annotation Validation

Additionally, the depth of coverage graphs were validated with sequences that contained LTR sequence homology. Areas of LTR homology aligned with areas predicted to be an LTR by the depth of coverage analysis as shown in figure 1. For this particular sequence the LTR would be predicted to fall between base pairs 40 and 1726, with a second LTR sequence between base pairs 7027 and 8729. The LTR sequence homology occurs between base pairs 1328 and 1552, and between base pairs 8149 and 8406. Both of these ranges are contained within the predicted LTR region.

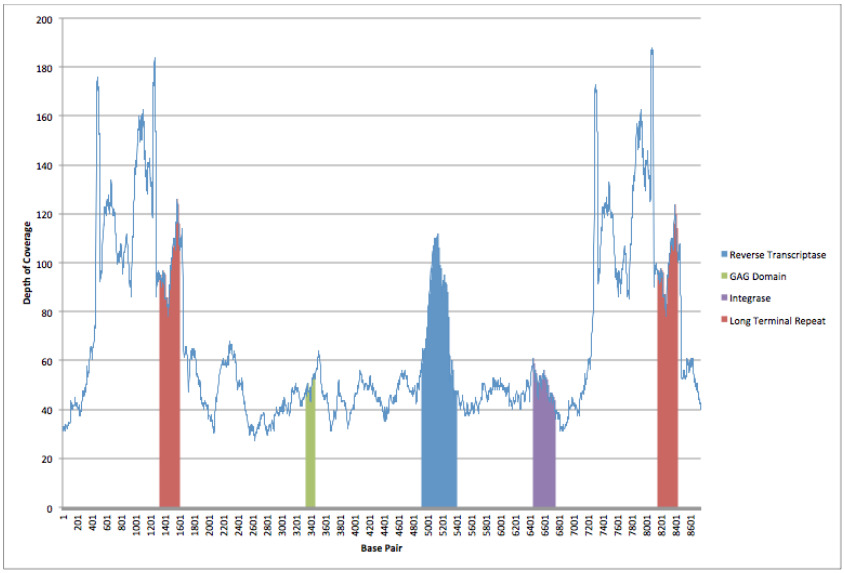

Figure 2. RLG_ArPr_guhis depth of coverage graph with sequence homology

LTR-RT Annotation

A total of 121 pseudo-RT were examined using a DoC graph to identify the LTR sequence and reconstruct the LTR-RT. Of the total LTR-RT analyzed, 55 (45%) were gypsy LTR-RTs, 64 (53%) were copia LTR-RTs and 2 were unknown superfamilies of LTR-RTs. Fifty seven of these pseudo-elements showed no sequence homology to any annotated LTR sequence, and were given a numerical name.

Table 3. Identified Retrotransposon Summary

Species	Identified Retrotransposons	Percent Genome	Genome Size (Mbp/1C)	Retrotransposon Size (Mbp/1C)
Arthraxon prionodes	32	9.76	2185	213.2
Chasmopodium caudatum	24	11.04	1675	184.9
Chrysopogon zizanoides	30	1.02	1058	10.75
Vossia cuspidata	36	8.49	Unknown	Unknown

Table 4. Identified Retrotransposon Final Count

Species	Average Length (BP)	Gypsy	Copia	Unknown	Novel LTR	Total
Arthraxon prionodes	8131	13	17	2	20	32
Chasmopodium caudatum	7608	12	12	0	10	24
Chrysopogon zizanoides	8019	13	17	0	9	30
Vossia cuspidata	7952	17	18	0	18	35
Total	7948	55	64	2	57	121

A total of 32 LTR-RT were identified in *Arthraxon prionodes* covering between 0.01% and 2.40% of the genome, and explaining between 0.04% and 24.61% of the total identified retrotransposons. The copy number of LTR-RT in this taxa clustered heavily around two LTR-RT, with RLG_ArPr_n1 was the most common identified retrotransposon in the genome followed by RLC_ArPr_dijap. RLC_ArPr_giepum_1 was also notable in that it explained 1.03% of the genome, which is significantly higher then any other LTR-RT present aside from the two previously mentioned.

Table 5. *Arthraxon prionodes* Novel Identified Retrotransposons, red coloring indicates higher percent genome

	Percent Annotated Retrotransposon	Percent Genome	Mbp/1C
RLC_ArPr_gudyeg	0.19%	0.02%	0.40
RLC_ArPr_giepum_1	10.56%	1.03%	22.51
RLC_ArPr_dijap	23.26%	2.27%	49.58
RLG_ArPr_guhis	0.20%	0.02%	0.42
RLG_ArPr_gymna	5.37%	0.52%	11.45
RLG_ArPr_n1	24.61%	2.40%	52.47
RLG_ArPr_n2	6.42%	0.63%	13.68
RLG_ArPr_n3	4.83%	0.47%	10.30
RLX_ArPr_milt	4.84%	0.47%	10.33
RLC_ArPr_wiwa	0.26%	0.03%	0.55
RLC_ArPr_machiavelli	0.14%	0.01%	0.29
RLC_ArPr_giepum_2	2.07%	0.20%	4.41
RLG_ArPr_CRM1	0.71%	0.07%	1.52
RLG_ArPr_xilon-diguus	1.50%	0.15%	3.20
RLX_ArPr_wihov	0.51%	0.05%	1.08
RLC_ArPr_n4	0.53%	0.05%	1.14
RLC_ArPr_n5	1.18%	0.11%	2.51
RLC_ArPr_n6	0.34%	0.03%	0.73
RLC_ArPr_n7	1.05%	0.10%	2.23
RLG_ArPr_n8	0.43%	0.04%	0.91
RLC_ArPr_n9	2.13%	0.21%	4.55
RLC_ArPr_n10	0.04%	0.00%	0.09
RLC_ArPr_n11	0.56%	0.05%	1.20
RLC_ArPr_n12	0.19%	0.02%	0.40

RLC_ArPr_n13	0.35%	0.03%	0.74
RLG_ArPr_n14	0.38%	0.04%	0.82
RLX_ArPr_n15	0.49%	0.05%	1.04
RLG_ArPr_n16	2.36%	0.23%	5.03
RLC_ArPr_n17	0.19%	0.02%	0.41
RLC_ArPr_n18	0.39%	0.04%	0.83
RLG_ArPr_n19	2.15%	0.21%	4.59
RLG_ArPr_n20	1.78%	0.17%	3.80

A total of 24 LTR-RT were identified in *Chasmopodium caudatum* covering between 0.01% and 2.73% of the genome, and explaining between 0.07% and 24.75% of the total identified retrotransposons. *Chasmopodium caudatum* had one element, RLG_ChCa_CRM4, which explained a very significant portion of the genome (2.7%), and was present in a much higher copy number then any other LTR-RT present. Many other LTR-RT were present that each individually explained roughly 1% of the genome, with the second most common LTR-RT being RLG_ChCa_n3.

Table 6. *Chasmopodium caudatum* Novel Identified Retrotransposons, red coloring indicates higher percent genome

	Percent Annotated Retrotransposon	Percent Genome	Mbp/1C
RLC_ChCa_giepum_1	10.30%	1.14%	19.04
RLC_ChCa_giepum_2	8.82%	0.97%	16.30
RLC_ChCa_fourf	0.66%	0.07%	1.21
RLG_ChCa_CRM4	24.75%	2.73%	45.75
RLG_ChCa_tekay_prem1	6.19%	0.68%	11.44
RLG_ChCa_tekay_ruda	5.88%	0.65%	10.86
RLG_ChCa_n1	4.72%	0.52%	8.73
RLG_ChCa_n2	7.99%	0.88%	14.77
RLG_ChCa_n3	11.49%	1.27%	21.24
RLG_ChCa_n4	5.02%	0.55%	9.28
RLC_ChCa_raider	0.27%	0.03%	0.49
RLC_ChCa_wiwa_1	3.19%	0.35%	5.90
RLC_ChCa_wiwa_2	0.37%	0.04%	0.69
RLC_ChCa_eninu	0.10%	0.01%	0.19
RLC_ChCa_n5	2.15%	0.24%	3.97

RLC_ChCa_n6	0.11%	0.01%	0.20
RLX_ChCa_wihov	0.50%	0.06%	0.92
RLX_ChCa_CRM2	0.07%	0.01%	0.14
RLG_ChCa_uwum_prem1	0.93%	0.10%	1.71
RLG_ChCa_n7	0.20%	0.02%	0.37
RLC_ChCa_giepum_3	5.02%	0.55%	9.28
RLG_ChCa_n8	0.39%	0.04%	0.73
RLG_ChCa_n9	0.80%	0.09%	1.47
RLG_ChCa_n10	0.11%	0.01%	0.20

A total of 30 LTR-RT were identified in *Chrysopogon zizanoides* covering between 0.01% and 0.13% of the genome, and explaining between 13.26% and 0.01% of the total identified retrotransposons. The annotated LTR-RT from *Chrysopogon zizanoides* explained a relatively small amount of the genome as compared to other species examined. A large number of LTR-RT explained roughly 0.12% of the genome, with RLC_ChZi_ji the most common identified retrotransposon in the genome followed by RLG_ChZi_n10.

Table 7. *Chrysopogon zizanoides* Novel Identified Retrotransposons, red coloring indicates higher percent genome

	Percent Annotated Retrotransposons	Percent Genome	Mbp/1C
RLG_ChZi_gymna	1.56%	0.02%	0.17
RLG_ChZi_pebi	0.09%	0.00%	0.01
RLC_ChZi_n1_cosmos	0.51%	0.01%	0.06
RLC_ChZi_dijap	4.65%	0.05%	0.50
RLG_ChZi_huck	2.35%	0.02%	0.25
RLC_ChZi_ji	13.26%	0.13%	1.43
RLG_ChZi_CRM4/3	3.27%	0.03%	0.35
RLC_ChZi_n2	0.20%	0.00%	0.02
RLC_ChZi_n3	1.17%	0.01%	0.13
RLC_ChZi_n4	0.33%	0.01%	0.04
RLC_ChZi_n5	1.66%	0.02%	0.18
RLC_ChZi_n6	1.15%	0.01%	0.12
RLC_ChZi_n7	0.27%	0.00%	0.03
RLC_ChZi_n8	0.87%	0.01%	0.09

RLC_ChZi_n9	0.17%	0.00%	0.02
RLG_ChZi_n10	12.02%	0.12%	1.29
RLG_ChZi_n11	0.46%	0.00%	0.05
RLG_ChZi_n12	0.60%	0.01%	0.06
RLG_ChZi_n13	1.30%	0.01%	0.14
RLC_ChZi_machiavelli	11.31%	0.11%	1.22
RLC_ChZi_wiwa_1	11.64%	0.12%	1.25
RLC_ChZi_wiwa_2	0.09%	0.00%	0.01
RLC_ChZi_giepum	10.15%	0.10%	1.09
RLC_ChZi_gudyeg	0.47%	0.01%	0.05
RLC_ChZi_wawo	0.01%	0.01%	0.00
RLX_ChZi_wihov	0.32%	0.00%	0.03
RLG_ChZi_CRM1	0.28%	0.00%	0.03
RLG_ChZi_CRM2	7.49%	0.08%	0.81
RLG_ChZi_guhis	1.59%	0.02%	0.17
RLG_ChZi_xilon-diguus	10.77%	0.11%	1.16

A total of 36 LTR-RT were identified in *Vossia cuspidata* covering between 1.85% and 0.01% of the genome, and explaining between 0.13% and 21.75% of the total identified retrotransposons. This taxa had one LTR-RT, RLG_VoCu_prem1, which explained a relatively large portion of the genome. RLG_VoCu_flip was the second most common LTR-RT seen.

Table 8. *Vossia cuspidata* Novel Identified Retrotransposons, red coloring indicates higher percent genome

	Percent Annotated Retrotransposon	Percent Genome
RLC_VoCu_giepum_1	6.75%	0.57%
RLC_VoCu_giepum_2	2.66%	0.23%
RLC_VoCu_dijap	7.85%	0.67%
RLG_VoCu_prem1	21.75%	1.85%
RLG_VoCu_flip	10.84%	0.92%
RLG_VoCu_guhis	0.21%	0.02%
RLC_VoCu_n1	5.10%	0.43%
RLC_VoCu_n2	0.45%	0.04%
RLG_VoCu_n3	0.89%	0.08%
RLG_VoCu_n4	8.12%	0.69%
RLC_VoCu_machiavelli_1	0.25%	0.02%

RLC_VoCu_machiavelli_2	1.58%	0.13%
RLX_VoCu_wiwa	2.30%	0.20%
RLC_VoCu_raider	0.09%	0.01%
RLC_VoCu_ji	1.04%	0.09%
RLG_VoCu_gymna_1	3.87%	0.33%
RLG_VoCu_gymna_2	5.53%	0.47%
RLG_VoCu_CRM4	0.74%	0.06%
RLG_VoCu_CRM2	1.56%	0.13%
RLX_VoCu_wihov	0.14%	0.01%
RLG_VoCu_huck	1.11%	0.09%
RLC_VoCu_n5	0.29%	0.02%
RLC_VoCu_n6	3.92%	0.33%
RLC_VoCu_n7	3.67%	0.31%
RLC_VoCu_n8	0.49%	0.04%
RLC_VoCu_n9	0.40%	0.03%
RLC_VoCu_n10	0.24%	0.02%
RLC_VoCu_n11	0.14%	0.01%
RLC_VoCu_n12	0.13%	0.01%
RLG_VoCu_n13	0.59%	0.05%
RLG_VoCu_n14	0.06%	0.00%
RLG_VoCu_n15	3.00%	0.25%
RLG_VoCu_n16	1.63%	0.14%
RLG_VoCu_n17	1.91%	0.16%
RLG_VoCu_n18	0.72%	0.06%

Discussion

Understanding the dynamics of LTR-RT within a genome is critical to understanding genome dynamics as a whole. Given their high copy number within a genome and their potential to amplify across a relatively short period of time, understanding their impact on genomes, both on an individual species level and across related species, will likely provide significant clues to some of evolutionary biology's most persistent questions.

The data presented validates the ability of our bioinformatics pipeline to identify novel LTR-RT using a previously validated method of using relatively low coverage of high throughput Illumina sequencing (Estep et al., 2013). The databases created from

the identified Maize LTR-RT were able to successfully identify protein-coding domains within contigs generated from raw data using the AAARF algorithm. Our DoC graphs were able to reliably predict the location of the LTR within a contig without any sequence homology to an existing LTR. Using these approaches we were able to successfully identify and annotate many novel LTR-RT within a plant genome. Lastly, we can predict the total amount of a genome occupied by any particular LTR-RT using the competitive blast and known genome size.

Better understanding of the dynamics of retrotransposons between separate species will help us understand how their movement correlates to evolution. This can help elucidate the role of retrotransposons dynamics of speciation, gene duplication, and genetic dysfunction and disorders. However, the first step in beginning to examine any of these phenomena is the identification of the retrotransposons in the genome, which this project achieves.

Future directions

In order to reduce the amount of hand annotation, we eliminated similar sequences using an all vs. all blast. This step may have eliminated too many sequences for final analysis, impacting the final results. While the step reduced the amount of manual annotation done by the researchers, initial examinations of this step indicate that it may have eliminated relevant sequences. In the future it might be wise to eliminate this reduction, and manually analyze a larger number of pseudo-elements.

The accuracy of this technique seems to be directly proportional to the amount of data included in the original databases, as well at the amount of data that is analyzed with those databases. The major bottlenecks in the annotation procedure and the

database generation procedure all revolve around manual action done by a researcher. To this end any steps that can be automated will greatly increase the accuracy of the final product simply due to the fact that they will be able to eliminate manual steps. For example, currently researchers manually generate the DoC graphs by inputting a series of variables into a custom python script. If this step could be incorporated into the pipeline the researchers would be able to more effectively analyze a larger volume of DoC graphs.

Accuracy of the product can also be improved by using more existing databases of annotated LTR-RT from well studied grass taxa. The more time spent by researchers identifying protein coding domains from a wider range of LTR-RT will also greatly benefit the final product. It may be possible to automate this step, which would allow researchers to more quickly build larger databases. These sequences will become more prevalent as more data is published over time.

References

Adamson, C. S., & Jones, I. M. (2004). The molecular basis of HIV capsid assembly—five years of progress. Reviews in medical virology, 14(2), 107-121.

Bannert, Norbert, and Reinhard Kurth. "The evolutionary dynamics of human endogenous retroviral families." Annu. Rev. Genomics Hum. Genet. 7 (2006): 149-173.

Baucom, R. S., Estill, J. C., Chaparro, C., Upshaw, N., Jogi, A., Deragon, J. M., ... & Bennetzen, J. L. (2009). Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome.PLoS Genet, 5(11), e1000732.

Bennetzen, J. L. (2000). Transposable element contributions to plant gene and genome evolution. Plant molecular biology, 42(1), 251-269.

Bennetzen, Jeffrey L., and Hao Wang. "The contributions of transposable elements to the structure, function, and evolution of plant genomes." Annual review of plant biology 65 (2014): 505-530.

Broecker, F., Andrae, K., & Moelling, K. (2012). Premature activation of the HIV RNase H drives the virus into suicide: a novel microbicide?. AIDS research and human retroviruses, 28(11), 1397-1403.

Chandler, M., & Mahillon, J. (2002). Insertion sequences revisited. In Mobile DNA II (pp. 305-366). American Society of Microbiology.

Coffin JM, Hughes SH, Varmus HE, editors. Retroviruses. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 1997. Overview of Reverse Transcription. Available from: https://www.ncbi.nlm.nih.gov/books/NBK19424/

Daboussi, Marie-Josee, and Pierre Capy. "Transposable elements in filamentous fungi." Annual Reviews in Microbiology 57.1 (2003): 275-299.

DeBarry, Jeremy D., Renyi Liu, and Jeffrey L. Bennetzen. "Discovery and assembly of repeat family pseudomolecules from sparse genomic sequence data using the Assisted Automated Assembler of Repeat Families (AAARF) algorithm." BMC bioinformatics 9.1 (2008): 235.

Estep, M. C., DeBarry, J. D., & Bennetzen, J. L. (2013). The dynamics of LTR retrotransposon accumulation across 25 million years of panicoid grass evolution. Heredity, 110(2), 194-204.

Estep, M. C., McKain, M. R., Diaz, D. V., Zhong, J., Hodge, J. G., Hodkinson, T. R., ... & Kellogg, E. A. (2014). Allopolyploidy, diversification, and the Miocene grassland expansion. Proceedings of the National Academy of Sciences, 111(42), 15149-15154.

Estill, James C., and Jeffrey L. Bennetzen. "The DAWGPAWS pipeline for the annotation of genes and transposable elements in plant genomes." Plant methods 5.1 (2009): 1.

Evans, M. J., Bacharach, E., & Goff, S. P. (2004). RNA sequences in the Moloney murine leukemia virus genome bound by the Gag precursor protein in the yeast three-hybrid system. Journal of virology, 78(14), 7677-7684.

Feschotte, C., Jiang, N., & Wessler, S. R. (2002). Plant transposable elements: where genetics meets genomics. Nature Reviews Genetics, 3(5), 329-341.

Feschotte, Cédric, and Ellen J. Pritham. "DNA transposons and the evolution of eukaryotic genomes." Annual review of genetics 41 (2007): 331.

Finn, Robert D., et al. "The Pfam protein families database: towards a more sustainable future." Nucleic acids research 44.D1 (2016): D279-D285.

Finnegan, D. J. (2012). Retrotransposons. Current Biology, 22(11), R432-R437.
Finnegan, David J. "Eukaryotic transposable elements and genome evolution." Trends in genetics 5 (1989): 103-107.

Grandbastien, M. A. (1998). Activation of plant retrotransposons under stress conditions. Trends in plant science, 3(5), 181-187.

Hua-Van, A., et al. "Abundance, distribution and dynamics of retrotransposable elements and transposons: similarities and differences."Cytogenetic and genome research 110.1-4 (2005): 426-440.

Johnson, Mark, et al. "NCBI BLAST: a better web interface." Nucleic acids research 36.suppl 2 (2008): W5-W9.

Kalendar, R., Vicient, C. M., Peleg, O., Anamthawat-Jonsson, K., Bolshoy, A., & Schulman, A. H. (2004). Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes.Genetics, 166(3), 1437-1450.

Kapitonov, V. V., & Jurka, J. (2003). Molecular paleontology of transposable elements in the Drosophila melanogaster genome. Proceedings of the National Academy of Sciences, 100(11), 6569-6574.

Kumar, Amar, and Jeffrey L. Bennetzen. "Plant retrotransposons." Annual review of genetics 33.1 (1999): 479-532.

Lwin, Aung Kyaw, et al. "Genomic skimming for identification of medium/highly abundant transposable elements in Arundo donax and Arundo plinii." Molecular Genetics and Genomics (2016): 1-15.

Malik, H. S., & Eickbush, T. H. (2001). Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. Genome Research, 11(7), 1187-1197.

McClintock, B. (1950). The origin and behavior of mutable loci in maize.Proceedings of the National Academy of Sciences, 36(6), 344-355.

Moelling, K. (2013). What contemporary viruses tell us about evolution: a personal view. Archives of virology, 158(9), 1833-1848.

Moelling, K., & Broecker, F. (2015). The reverse transcriptase–RNase H: from viruses to antiviral defense. Annals of the New York Academy of Sciences, 1341(1), 126-135.

Nelson, D. L., Lehninger, A. L., & Cox, M. M. (2008). Lehninger principles of biochemistry. Macmillan.

Pena, V., Rozov, A., Fabrizio, P., Lührmann, R., & Wahl, M. C. (2008). Structure and function of an RNase H domain at the heart of the spliceosome. The EMBO Journal, 27(21), 2929-2940.

Pouteau, S., Grandbastien, M. A., & Boccara, M. (1994). Microbial elicitors of plant defence responses activate transcription of a retrotransposon. The Plant Journal, 5(4), 535-542.

Sabot, F., & Schulman, A. H. (2006). Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. heredity, 97(6), 381-388.

SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y., & Bennetzen, J. L. (1998). The paleontology of intergene retrotransposons of maize. Nature genetics, 20(1), 43-45.

SanMiguel, P., Tikhonov, A., Jin, Y. K., & Motchoulskaia, N. (1996). Nested retrotransposons in the intergenic regions of the maize genome. Science,274(5288), 765.

Sassaman, D. M., Dombroski, B. A., Moran, J. V., Kimberland, M. L., Naas, T. P., DeBerardinis, R. J., ... & Kazazian, H. H. (1997). Many human L1 elements are capable of retrotransposition. Nature genetics, 16(1), 37-43.

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., ... & Minx, P. (2009). The B73 maize genome: complexity, diversity, and dynamics. science, 326(5956), 1112-1115.

Simon, D. M., & Zimmerly, S. (2008). A diversity of uncharacterized reverse transcriptases in bacteria. Nucleic acids research, 36(22), 7219-7229.

Slotkin, R. K., & Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. Nature Reviews Genetics, 8(4), 272-285.

Team, R. Core. "R: A language and environment for statistical computing." (2013): 409.
Wicker, Thomas, et al. "A unified classification system for eukaryotic transposable elements." Nature Reviews Genetics 8.12 (2007): 973-982.

Conant, G. C., & Wolfe, K. H. (2008). Turning a hobby into a job: how duplicated genes find new functions. Nature Reviews Genetics, 9(12), 938-950.