A TAXONOMY AND GENOME CARTOLOGY OF MAIZE LTR RETROTRANSPOSONS

by

JAMES CHRISTOPHER ESTILL

(Under the Direction of Jeffrey L. Bennetzen)

ABSTRACT

The genome cartology approach to genomics is a spatially explicit framework for the study of patterns in the distribution and abundance of sequence features within and among genomes. The ultimate goal of this approach is to identify the demographic and selective processes that have given rise to these extant patterns. However, tools for the accurate annotation and taxonomic assignment of sequence features must first be implemented before these ultimate goals can be realized. I have designed, implemented and assessed the accuracy of novel annotation and taxonomic classification software and applied these tools to a genome cartology of maize LTR retrotransposons (LRPs).

The DAWGPAWS pipeline facilitates combined evidence human curation of *ab initio* and similarity search based computational results. I verified the value of DAWGPAWS by using this pipeline to annotate genes and transposable elements in 220 BAC insertions from the hexaploid wheat genome. To illustrate that these techniques can scale to entire genomes, the pipeline was applied to the annotation of LRPs in the B73 maize genome and discovered over 31,000 intact elements.

The RepMiner suite of programs allows for the clustering of sequences into families based on networks of shared homology. I applied the RepMiner approach to the database of

intact maize LRPs annotated by DAWGPAWS. RepMiner further illuminated previously identified family relationships, indicated an unrecognized split in the *Huck* family, and recognized over 350 new families of LRPs. Affinity propagation based clustering of intact LRPs identified a subset of ~500 exemplar sequences that can serve as a representative database of all maize LRPs.

The exemplar database of maize LRPs was used to map the location of intact and fragmented LRPs in the assembled genome of maize. These LRPs comprised over 75% of the genome, and are nonrandomly distributed with a preferential accumulation in pericentromeric heterochromatin. Surprisingly, the regions of the genome with the highest accumulation of LRPs had the lowest diversity of LRP families. These results indicate that genome cartology will provide new insight into genome dynamics, and that continued development of this approach to genomics will further enlighten the study of genome evolution.

INDEX WORDS: Affinity Propagation, Bioinformatics, Biological Clustering, *Copia*, Genome Annotation, Genome Cartology, Genome Evolution, Graph Theory, *Gypsy*, LTR Retrotransposon, Maize, Poaceae, Transposon, Transposable Element, *Zea mays*

A TAXONOMY AND GENOME CARTOLOGY OF MAIZE LTR RETROTRANSPOSONS

by

JAMES CHRISTOPHER ESTILL

BS, Western Kentucky University, 1996

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2010

A TAXONOMY AND GENOME CARTOLOGY OF MAIZE LTR RETROTRANSPOSONS

by

JAMES CHRISTOPHER ESTILL

Major Professor: Jeffery L. Bennetzen

Committee: Kelly A. Dyer
Jim H. Leebens-Mack
Russell L. Malmberg
Xiaoyu Zhang

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2010

DEDICATION

To my family; Regina, Jack and Neko thank you. Regina has been both a patient spouse and an insightful scientific collaborator. Jack has provided numerous insights into life and science that are completely unequaled in their depth and candor. Neko motivates wakefulness in the morning, keeps me engaged in work throughout the day and has smiles for me through the night. My parents Frank and Libby Estill have always supported my schoolwork and science and for that I am very thankful.

ACKNOWLEDGEMENTS

I would first like to thank Jeff Bennetzen who served as my mentor and graduate committee chair. Jeff has been very supportive of my career and is one the most imaginative and knowledgeable scientist I have worked with. I would also like to thank Jim Leebens-Mack for fruitful collaboration over the past year. Each member of my committee has made valuable contributions to my dissertation and career development.

Hilmar Lapp and William Piel served as mentors for a phyloinformatics Perl coding project in the 2007 Google Summer of Code. They provided general programming examples and advice that were incorporated into the Perl command line and database interface projects. Multiple people also helped improve the software that are described in this dissertation by acting as early adopters of the programs as they were in development. I would like to thank Katrien Devos, Jim Leebens-Mack, Antonio Costa de Oliveira, Xiangyang Xu, Ansuya Jogi, Jennifer Hawkins and Hao Wang for their useful comments and bug reports that have been incorporated in the implementation of DAWGPAWS. Regina Baucom, Phillip SanMiguel, and Evan Staton provided helpful feedback in the early development of the RepMiner suite of programs. My dissertation research also benefited from stimulating conversations concerning the biology of transposable elements with fellow members of the Bennetzen lab group at UGA. This dissertation could not have been completed without the support of family and friends.

This work was supported by NSF grants DBI-0501814, DBI-0607123, and DBI-0821263.

This study was also supported in part by resources and technical expertise from the University of Georgia Research Computing Center, a partnership between the Office of the Vice President for Research and the Office of the Chief Information Officer.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Overview

This dissertation describes the design, implementation and assessment of genome annotation and sequence classification software and the application of these tools to a study of the genome cartology of maize long terminal repeat (LTR) retrotransposons. The eventual goals of this genome cartology are to describe the spatiotemporal patterns in the distribution, abundance and diversity of transposable elements (TEs) in the genomic landscape and to generate and test hypotheses concerning the demographic and selective processes that have contributed to these extant patterns. I am referring to this approach as genome cartology since it is, in essence, a study of spatial dependencies in high-resolution genome organization. This study differs from genome cartography, which has the objective of locating features in genome space and presenting these results as a visually informative map of the genome. I am using the term genome cartology to mean a spatially explicit study of genome evolution.

A suite of informatic tools needs to be developed before the genome cartology approach can be applied to any study of genome evolution, and consequently my dissertation research has informatic goals in addition to its biological objectives. The informatic goals and the software designed to meet these goals are 1) *DAWGPAWS* to annotate genes and transposable elements in plant genomes, 2) *RepMiner* to classify these sequence features into a useful taxonomy, and 3) *GenCart* to evaluate the spatiotemporal patterns of these features. The biological objectives of my dissertation are to use these informatic tools to 1) generate a family-level taxonomy for all of

the LTR retrotransposons in the B73 maize genome, and 2) study spatiotemporal patterns in the distribution and abundance of these families mapped in the B73 maize genome [1]. An overview of how the software tools generated for this dissertation will be used to meet these biological goals is presented in Figure 1.1.

It is expected that the insights gained from this research will shed new light on the biology of LTR retrotransposons as well as contribute to our understanding of maize genome evolution as influenced by these elements. Furthermore, the tools generated to make this approach possible can be widely applied to other classes of transposable elements as well as many other sequence features. An overview of the rationale behind this study, specifics of the implementation of the software packages, and an outline of this dissertation are provided below.

The Biology of Plant LTR Retrotransposons

Transposable elements (TEs) are mobile DNA sequences that are ubiquitous across the tree of life [2] and represent one of the most dominant features comprising eukaryotic genomes [3]. TEs can transpose either by directly transposing their DNA to a new location in the host genome, or by using reverse transcription of an RNA intermediate to make copies of the parent DNA in a new genomic location. Those elements that use an RNA intermediate are referred to as retrotransposons or class I elements, while those that directly move DNA copies within the genome are referred to as DNA transposons or class II elements [4].

Transposable elements are major contributors to eukaryotic genome evolution. At the whole genomic level, TE insertions have increased eukaryotic genome size [3,5,6] and are largely responsible for the 'C-value paradox' [7]. The replication of individual TEs also results in local structural rearrangements at the points of insertion or removal [8,9]. TEs contribute to

host genetic diversity by relocating gene fragments within a genome or by shuffling exons into entirely new genic arrangements [10,11,12,13]. TEs have also contributed directly to genic evolution through the neofunctionalization of TE genes into novel roles in the host genome [14] that can contributed directly to adaptive evolution [15]. TE insertions can also influence the nature and timing of expression of individual genes [16,17] and large bursts in TE activity can generate novel gene regulatory networks over short periods of time [18,19]. The ability of the host genome to control activity of TEs by epigenetic modification is the dominant factor influencing the distribution of heterochromatin and euchromatin within genomes [20] and small RNA-based silencing mechanisms have evolved as a genomic defense against TE proliferation [21,22]. The breakdown of epigenetic TE regulation under times of environmental stress [23] combined with the above mentioned contributions of TEs to genic evolution have led to speculation that TEs can be major players in genomic response to stress [24] and may have even contributed to the observation of punctuated equilibrium in the fossil record [25].

LTR (long terminal repeat) retrotransposons are an order of class I elements [26] that are one of the most abundant types of TEs in plant genomes [9,27]. Plant LTR retrotransposons are represented by the *Copia* [28] and *Gypsy* superfamilies [29] that are circumscribed based on shared sequence identity and the order of the reverse transcriptase protein with respect to the integrase domain in the *pol* ORF (Figure 1.2). LTR retrotransposons have also been placed in a viral taxonomic framework in which the *Copia* superfamily of TEs known as the Pseudoviridae family of viruses [30] and the *Gypsy* superfamily of TEs is placed in the Metaviridae [31]. Although this viral based taxonomy is informative for placing LTR retrotransposon in the context of *Gypsy* retroviral evolution, I will use the nomenclature of Wicker *et al.* [26] because

this system provides a unified classification system for naming all types of transposable elements.

Fully autonomous copies of plant LTR retrotransposons (Figure 1.2) contain *gag* and *pol* open reading frames (ORFs) that code for the proteins required for completion of the LTR retrotransposon life cycle in a host cell (Figure 1.3). These coding domains are flanked by the diagnostic LTRs for which this order is named. The 5' LTR contains the signals required for transcription of the element in the host genome (Figure 1.3 A), and LTR pairs are generated anew during the process of reverse transcription (Figure 1.3 C). Short 3-5 bp target site duplications (TSDs) flank the LTRs and are generated from the host genome at integration (Figure 1.3 D). A primer binding site (PBS) downstream of the 5' LTR initiates reverse transcription through priming by a host small RNA, usually a tRNA, and a similarly placed polypurine tract near the 3' LTR helps to initiate 2nd strand synthesis of DNA (Figure 1.3 C). The *gag* transcript generates the encapsulating particles (GAG) that comprise the nucleocapsid virus like particle (VLP) in the cytoplasm. The *pol* transcript is a polyprotein that is cleaved into the aspartic protease (PR), reverse transcriptase with RNaseH capacity (RVT) and integrase (INT) proteins that are required for reverse transcription and integration of new copies in the host genome. Gene products from the host are also required for the LTR retrotransposon's life cycle. The processes of LTR retrotransposon transcription and translation make use of the host machinery that is normally used for the production of host proteins.

In addition to the archetypal *Copia* and *Gypsy* autonomous models, LTR retrotransposons have structural variants with added coding capacity that can facilitate movement and targeted integration within the host genome (Figure 1.4). The *pol* ORF may contain a chromodomain in the C-terminal region that can target insertion to heterochromatin [32,33,34] and thus avoid

insertion into genes. LTR retrotransposons may also include an additional ORF that has been controversially argued to be similar to the *envelope* coding regions of retroviruses [35,36,37] that could potentially facilitate movement of LTR retrotransposons between cells [38,39,40]. This movement between cells could take place in an intermediate invertebrate host and facilitate horizontal transfer [36,40]. Both *Gypsy* and *Copia* superfamilies have representatives that included an *env*-like ORF, and LTR retrotransposons containing this ORF are widely distributed in plants with representatives in moncots, dicots and gymnosperms [41]. In maize, the *env*-like coding region was also found to be associated with an extended *gag* ORF with a domain that binds to the plant Light Chain 8 (LC8) protein family [39]. The LC8 family binds to cytoskeleton proteins and could thus facilitate the movement of virus like particles towards the nucleus within individual cells or could assist virus like particle formation [39]. LTR retrotransposons can also incorporate protein-coding regions from host genomes [42,43], probably through template switching between the retrotransposon RNA and a host mRNA during reverse transcription [44]. Replication of these elements within host genomes can results in TE lineages with stable incorporations of host genic sequences that can be laterally transmitted and present in multiple extant host species [43].

Like most classes of transposable elements, LTR retrotransposons may have nonautonomous deletion derivatives that are dependent on protein products from autonomous partners for their transposition [45] (Figure 1.4). Some of the nonautonomous derivatives of fully autonomous LTR retrotransposons have been structurally categorized [27] into LARDs (LArge Retrotransposon Derivates) [46] or TRIMs (Tandem Repeats In Miniature) [47]. LTR retrotransposons derivatives that can be described as LARDs and TRIMs do not share a monophyletic origin, but similar to MITEs these terms represent a practical way to structurally

describe a 'way of life' for groups of transposable elements [26]. TRIM LTRs are usually 100-250 bp long and surround a central region of about 300 bp, while LARD LTRs are around 4.5kb long and flank a central conserved domain of about 3.5kb. As an alternative to full dependence on gene products from an autonomous partner, deletion derivatives may also be mutually co-dependent. For example, the *Cinful* LTR retrotransposons in maize lack *gag* and *Zeon* LTR retrotransposons lack *pol*. However, members of both *Cinful* and *Zeon* have successfully transposed after the loss of these coding regions [48]. The members of these *Gypsy* elements are thus hypothesized to share gene products for successful transposition, and sequence similarity in termini of the LTRs indicates that the same integrase protein could be responsible for chaperoning their integration into the host genome [48].

Illegitimate recombinational mechanisms acting during the process of reverse transcription [49,50,51] or after insertion in the host genome [52] can generate structural genomic diversity within LTR retrotransposon families. Short repeats within an LTR retrotransposon can be a target site for illegitimate recombination [51,52] and result in the loss of DNA segments between the short repeats ranging in length from a few base pairs to nearly four kilobases [52,53]. Illegitimate recombination is a frequently observed recombinational outcome with multiple causative mechanisms and has been proposed to be the major mechanism of genome size reduction in plants [5,52] and animals [54,55].

Unequal homologous recombination also generates structural variants of LTR retrotransposon families. Unequal homologous recombination within an element can lead to the removal of the internal coding regions and one of the LTRs resulting in the formation of solo LTRs flanked by target site duplications (Figure 1.4). Unequal homologous recombination between two members of the same family will result in the removal of host DNA between the

recombined LTRs and will leave behind solo LTRs and full length elements that lack target site duplications [52]. Chimeric elements with three LTRs (Figure 1.4) may also result from unequal homologous recombination between elements [52,56] or through abnormal template switching during reverse transcription [50]. The formation of solo LTRs is a relatively common phenomenon in plant genomes while the three-LTR derivatives are rarely observed. Solo LTRs in rice have even been found to outnumber full-length elements by about 2 to 1 [57], while in Arabidopsis they are nearly even in number [52]. The ratio of observed solo LTRs to full-length elements have also been found to vary by more than an order of magnitude among families of LTR retrotransposons in maize [58].

In general, LTR retrotransposons preferentially accumulate in centromeric, intergenic and gene poor regions of grass genomes [6,59,60,61]. Since LTR retrotransposons and other transposable elements comprise these regions, transposition events often result in an LTR retrotransposon insertion into another element. Combined with DNA removal mechanisms, this can result in a variety of patterns of insertions such as solo LTRs nested within full-length elements, full-length elements nested within solo LTRs, or clusters of difficult-to-delineate truncated elements. Over evolutionary time this process has generated complex patterns of nested insertions (Figure 1.5) and has created a genome in maize that is a "sea" of nested LTR retrotransposons [59] interspersed with small "islands" of gene-containing segments [61]. The majority of LTR retrotransposons in many grass genomes are elements that are disrupted by inserted copies [50,58].

The insertion mechanism of LTR retrotransposons allows for a chronological reconstruction of their insertion history. The nested nature of many LTR retrotransposon insertions allows for the relative ordering of insertion events [59], and the long terminal repeats

of LTR retrotransposons also provide for a unique method to more precisely date the time since insertion [62]. The process of reverse transcription during LTR retrotransposon replication results in a pair of LTRs that usually have sequences that are identical at the time of insertion. If it is assumed that the accumulation of mutations between pairs of LTRs occurs in a clock-like manner, then the observed number of mutations can be combined with a mutation rate to generate an age for the insertion event [62]. Molecular dating of full-length LTR retrotransposon insertions near the *adh1*-F gene in the maize genome has indicated that the majority of these insertions occurred in the last three million years [62]. The authors achieved this estimate using a mutation rate of 6.5×10^{-9} substitutions per site per year that was derived from an the average substitution rate at synonymous sites of *adh*1 and *adh*2 loci in the grasses [63]. Later analysis of sequence variation across a large (>1MB) region among three rice genomes has indicated that nucleotide substitution in intergenic regions is approximately 2 times higher then genic regions, and that a more appropriate substitution rate to estimate for LTR retrotransposons insertion dates is 1.3×10^{-8} substitutions per site per year [64]. This refined mutation rate would place the majority of LTR retrotransposon insertions in the maize *adh*1-F region as occurring within the last two million years (Figure 1.5).

Intragenomic Distribution of Transposable Elements

The observed intragenomic patterns in the accumulation of transposable elements in extant genomes are largely the result of the processes of insertion and deletion generating structural variation that is sorted by population genetic processes [65]. Nonrandom distributions of TE accumulation will be observed in extant genomes if insertion and deletion occur in a spatially nonrandom pattern, or if the influence of population processes takes place in a spatially

nonrandom manner. In order to distinguish between the processes of insertion and removal for structuring patterns of intragenomic TE distribution, it is useful to study active elements that undergo transposition in experimental settings and then have their new location mapped onto a fully sequenced reference genome. These experiments have been undertaken in numerous cases as either the mapping of endogenous elements that are naturally active such as *Ac* and *Mu* elements in maize [66,67], or from the transfer of naturally active elements into novel genomes were they could transpose as a heterologous elements, such as the transfer of active maize elements into *Arabidopsis thaliana* [68,69]. Additional experimental systems have made use of the fact that TEs can be made active in tissue culture such as the *Tos17* family of *Copia* LTR retrotransposons in rice [70,71]. Endogenous elements can also be made active by mutations that interfere with the host epigenetic silencing of TE transposition such as the *decrease in DNA methylation 1* (*ddm1*) mutant of *Arabidopsis thaliana* [72]. In situations where actively transposing elements are not currently available for study, distinguishing between the processes of insertion and removal can be undertaken by studying differences in the distributions of insertion cohorts within extant genomes as has been done with the *Alu* elements in humans [73,74,75]. Such studies of extant elements are also used to infer the patterns of population genetic processes that have influenced the distribution of transposable elements [76,77]. Using a combination of the above approaches has revealed nonrandom spatial patterns in the distribution of transposable elements for at least some families of nearly all of the major categories of transposable elements that have been studied to date (Table 1.1).

A common pattern among many groups of DNA transposons is that transposition often occurs into a closely linked site resulting in a spatial clustering of elements. This has been observed in the active transposition of TE families within the *hAT* [68,78,79,80,81,82,83,84,85

], *Tc1-Mariner* [86,87,88], *P-element* [89], and *PiggyBac* [87] superfamilies of DNA transposons. The observed linkage in these superfamilies is possibly due the lack of a free extrachromosomal intermediate during transposition and the dependence on a physical linkage between the donor and recipient chromosomal locations at the time of transposition. Such a physical association is thought to explain the observed linkage of transposition events for *Ac* elements in maize [90] and *Tam3* elements in snapdragon [91].

There can be variation in the pattern of this linkage reported within individual families as has been shown with *PiggyBac* elements transposing as heterologous elements within *Drosophila melanogaster* and mouse cells lines. Insertion into genetically linked sites was reported in mouse cell lines [87] while this linkage was not observed in *Drosophila melanogaster* [92,93] or previous experiments in mouse [94]. These differences might be attributed to multiple rounds of transposition occurring within experimental systems before a mapping of the element insertion can occur [87]. Such an obfuscated spatial pattern of transpositions due to secondary transposition events was observed for the *Ac* element in tomato [95,96]. These differences in linkage can result in variable patterns of hotspots of insertion, and caution should be used when interpreting the lack of observed genetic linkage in an experimental system as a lack of spatial linkage in the underlying mechanism of transposition.

There can also be variation in the pattern of linkage among donor sites within the same genome as was observed in the spatial distribution of transposition events of heterologous *Ac* elements in tobacco [81]. Some of the transposed *Ac* elements showed tight clustering around the donor locus, while others show a more dispersed pattern. While these patterns could be the result of more than one transposition event occurring before mapping of the insertion, they could also be due to differences in chromatin organization among loci. Dooner *et al.* [81] suggest that the

observed differences in spatial patterns of insertion among *Ac* elements were due to the transposition machinery requiring donor and recipient sites that undergo replication at the same time in S phase. They did recognize, however, that these differences could also be due to differences between physical distance and genetic distance in their study system.

Hotspots of insertion are also associated with specific DNA regions such as the large subunit ribosomal RNA in arthropods. These rRNA genes exist in tandemly repeated units, with genomes carrying more copies than required for RNA synthesis [97]. This overabundance of loci makes these rRNA genes potential sites where insertions of transposable elements might not cause deleterious fitness consequences to the host and thus serve as an ideal niche for TE persistence [97]. An accumulation of *R1* and *R2* elements in ribosomal RNA is a general feature found in arthropod genomes [98,99]. The *R1* and *R2* elements mapped in *Drosophila* insert into specific sequences that are 74 base pairs apart in 28s ribosomal RNA genes [100,101,102]. This region is also a hotspot of insertion for the *Pokey PiggyBac* element in *Daphnia* [103]. This element inserts into the sequence TTAA in multiple genomic locations [104,105,106], but has an enrichment for a specific TTAA sequence in the large subunit ribosomal rRNA gene a few bases away from *R1* and *R2* insertions [103]

Similar to the *Pokey* element, other DNA transposons show a targeting for a specific consensus sequence, indicating a direct interaction between the DNA and transposase [107]. The *Tc1* and *Tc3* families in *C. elegans* insert specifically into the sequence TA [108], the *PiggyBac* superfamily inserts exclusively into the sequence TTAA [103,104,105,106,109,110], and the *PIF* family in maize inserts into the TTA of the more generalized consensus CWCTTAGWG [111,112,113]. In other cases, the insertion preference is not into a specific consensus sequence, but into more generalized patterns in DNA structure such as the preference of *P-elements* in

Drosophila to insert into a 14 bp palindromic hydrogen bonding pattern in the DNA major grove [114]. Even in cases where specific sequences are the target of insertion, not all possible target sites are used, indicating additional factors (*e.g.*, other DNA structural features or chromatin availability) that may influence the spatial distribution of insertion sites.

DNA transposon insertions in maize have revealed both global and local patterns in insertion preference. Recent high resolution analysis of over 2000 *Dissociation* (*Ds*) insertion events in maize has revealed that, although no target site consensus sequence exists for *Ds* insertions in maize [115,116], specific DNA structural variation surrounding the observed target site duplication was associated with *Ds* insertion sites [67]. Over 1600 of these insertions could be mapped to specific maize BACs, and the global distribution of these maize BACs in the maize genome [1] allowed for high resolution mapping of these insertions. These mapping data revealed that, in addition to the previously observed preference for insertion near the donor site [78,79], *Ds* elements had a global insertion preference for telomeric regions [67]. This distribution can be interpreted to be consistent with a preferred insertion of *Ds* into hypomethylated [117] and gene rich [118] regions of the genome [67].

Similar to *Ds* elements[67], *Mutator* (*Mu*) DNA transposon insertions are non-randomly distributed in the maize genome, and preferentially insert into genes [119,120,121]. However, while *Ds* elements exhibit a preference for insertion into the introns and exons of genes [67], the *Mu* elements exhibit a preference for promoters and 5' UTRs [66]. *Mu* insertion sites in maize are also positively associated with hypomethylated regions and histone modifications that are signatures of open chromatin [66]. Additionally, *Mu* insertions occur in regions of high rates of recombination even when gene density is taken into account [66]. The authors speculate that

similar processes direct both the insertion preference of *Mu* elements and locations of recombination.

Miniature inverted repeat transposable elements (MITEs) are a structurally defined group of deletion derivates of full length DNA transposons that were first discovered in association with genes in plants [122,123,124]. Genomic sequence analysis has shown that MITEs accumulate in genic regions, with a preference for the 5' and 3' region and where matrix attachment regions are present [9,125,126]. The recently active maize MITE *Heartbreaker* (*Hbr*) was found to preferentially insert into genic regions of maize [127]. The MITE family *miniature Ping* (*mPing*) is a currently active element in rice [128,129,130] that has preferentially inserted into genic regions in rice [128,131] and Arabidopsis [132].

The *Helitron* rolling circle DNA transposons have been show to accumulate in gene-poor regions of *C. elegans*, rice, and Arabidopsis [133], but accumulate in gene-rich regions in maize [134]. This observation is perhaps due to gene-rich regions of the large maize genome having chromatin structural features that are similar to the gene-poor regions of the more compact genomes [134]. *Helitrons* in maize were also shown to accumulate in regions of the genome that were near the same family or subfamily, indicating that family level insertional specificities may exist [134]. In general *Helitrons* insert in A/T rich regions of the genome [134] and preferentially accumulate near other *Helitrons* or other DNA transposons [134].

Similar to DNA transposons, retrotransposons show specific spatial patterns in insertion that result in hotspots of accumulation. A trio of *Drosophila melanogaster* LINE elements almost exclusively transpose to the termini of chromosomes [135,136] with no copies present in the sequenced gene-rich euchromatic region of the genome [136,137]. The *TART*, *HeT-A* and *TAHRE* families belong to the *Jockey* superfamily of LINEs and insert at broken chromosome

ends taking on telomeric maintenance functions [138,139,140]. Gag proteins target these families to the telomeres in an interdependent fashion. The *TART* and *TAHRE* Gag proteins require the presence of *HeT-A* Gag proteins for this targeting [136,141,142] while *HeT-A* is presumably dependent on *TART* for production of its reverse transcriptase [136]. Variants of *HeT-A* were also identified in the centromere of *Drosophila* Y chromosomes, indicating the potential for a telomeric origin of this centromere as well as a potential role of this element in centromeric function [143,144].

Centromeric enrichment has also been well documented in plant LTR retrotransposons. Centromeric retrotransposons (CRs) are core components of the centromeres in the Poaceae [145,146,147,148,149,150,151]. These include multiple families of centromeric retrotransposons in rice (CRRs) [152] and centromeric retrotransposons in maize (CRMs) [60,153]. Interestingly, differences in centromeric enrichment exist between autonomous and nonautonomous members of the same family [152]. The enrichment of these elements in the centromeres suggests a targeting of insertion for centromeric heterochromatin [152]. The C-terminal region of CR integrase contains a conserved chromodomain-like region referred to as the CR motif that potentially directs integration of LTR retrotransposons to heterochromatin [33,34]. Furthermore, the CRM elements have been shown to interact specifically with centromeric histone H3 [150], providing a potential mechanism for directed integration to centromeric histones.

The targeting of TE integration by the recognition of epigenetic features is a potential mechanism of targeted insertion in other chromodomain-carrying *Gypsy* LTR retrotransposons [34,154]. The *MAGGY* elements in the rice blast fungus genome are spatially clustered with other transposable elements in the host genome [155,156,157] and are located in relatively gene-poor regions [33]. A chromodomain in the C terminal end of integrase targets insertion to histone

H3 methylated at lysine-9 (H3 K9), an epigenetic marker of heterochromatin [33,158]. The *Tf1* family in fission yeast preferentially inserts into a narrow window 100-400 bp upstream of open reading frames which corresponds to the promoter region of polymerase II transcribed genes [159,160,161,162,163]. This insertion preference involves the interaction of a chromodomain in the integrase and chromatin features specific to pol II promoters [164,165]. These results illustrate that related families of LTR retrotransposons have the ability to target different epigenetic features using variations of the same underlying mechanism.

The *Ty* elements in yeast provide one of the best studied systems for the mechanisms of directed LTR retrotransposon integration. The *Ty3 Gypsy* element of *Saccharomyces cerevisiae* has an insertion preference into a narrow window one to two base pairs upstream of RNA polymerase III transcription initiation [166,167,168]. This element thus targets upstream of tRNA genes and has a periodicity of insertion of 80bp. Targeting to this region is potentially due to a direct recognition of the polymerase transcription complex or to chromatin states associated with this pol III transcription [168,169,170]. Similarly, the *Ty1 Copia* element in *S. cerevisiae* has an insertion preference that is upstream of RNA polymerase III transcribed genes [167,171,172] including tRNA genes, snRNA genes and 5s rRNA [171,173,174]. This targeted insertion is also probably due to the targeted integration at chromatin features associated with RNA polymerase III transcription [170,173,175].

The *Ty5 Copia* family in *S. cerevisiae* has an insertion preference into telomeres, and silent mating loci with chromatin states that are similar to heterochromatin in eukaryotes [167,176,177,178]. This insertion is chaperoned by an interaction between the targeting domain of the *Ty5* integrase and the heterochromatin protein encoded by *silent information regulator* (*Sir4*) which tethers the integration complex to the target site [179,180,181,182], with the *Ty5*

integrase mimicking a host protein associated with the inner nuclear membrane [183]. The targeting domain of *Ty5* requires phosphorylation by the host this for this interaction with the Sir4 protein to occur [184]. The transposition rate decreases and targeting of insertion becomes more spatially random when this phosphorylation is absent in the targeting domain [184]. Since targeting domain phosphorylation is reduced under physiological stress conditions, this loss of target specificity will increase the probability of insertion into genic regions in conditions where genic changes could be adaptive.

In general, it could be maladaptive for elements to target integration to genes. The targeting of *Ty* elements to regions upstream of pol III transcribed genes or targeting to heterochromatic DNA has been interpreted as a mechanism to target integration to benign regions of the genome that avoid disruption of genes [107,185,186]. These *Ty* elements also provide an interesting example in which the intragenomic distribution of elements does not have similar spatial patterns due simply to relatedness of the elements. While the *Ty1* family is more closely related to *Ty5,* the *Ty1* distribution is more similar in pattern to the *Ty3* family [167]. The *Ty* elements of *S. cerevisiae* thus also illustrate that phylogenetically divergent families can target spatially related regions using similar targeting mechanisms. This raises the possibility that widely divergent families can evolve similar mechanisms for increasing their fitness in host genomes.

In addition to the role of targeted integration to structuring the spatial distribution of TEs, spatial differences in removal and retention of TEs will also influence the intragenomic distribution of TEs. The major mechanisms for removing TEs from plant genomes are unequal intrastrand homologous recombination (UHR) and illegitimate recombination (IR) [5,52,187]. In

rice, these processes are estimated to have removed more than 190MB of DNA derived from LTR retrotransposons [64].

UHR can remove one of two copies of tandemly arranged sequence segments as well as the DNA between the tandemly arranged sequences. For example, intrastrand UHR between the two LTRs of a single LTR retrotransposons will lead to the removal of one of the LTRs as well as the region between the two LTRs. This results in a solo LTR flanked by the original TSDs UHR between two copies of an LTR retrotransposons can also result in removal of three of the LTRs, as well as the coding regions of both copies, resulting in a solo LTRs without matching TSDs [52,188].

Since UHR intrinsically requires recombination, one would expect the process of UHR would be most effective in regions of the genome that have higher rates of recombination. This would be observed as an increase in solo LTR formation as the rate of recombination increases across a genome. The processes of DNA removal via UHR does appear to be more frequent in gene rich regions of the genome [57,64,189,190]. Ma and Bennetzen [189] observed a rate of solo LTR formation that was approximately twice as fast in euchromatic regions of rice than it was in the CEN8 pericentromeric region. However, the authors also observed a high rate of solo-LTR formation in a small region within the CENH3 binding domain which is more difficult to explain. Additional studies in rice have also shown that LTR retrotransposons and fragments are shorter in regions with higher recombination [190], and similar results have also been shown for human endogenous retroviruses [191,192]. Thus, at least for elements with long terminal repeats, it has been well documented that spatial differences in recombination and gene density will influence the persistence of inserted elements in genomes. These spatial differences in recombination may also influence the efficacy of selection acting on element insertions [65,193]

and could in part influence the accumulation of TEs in low recombination regions of the genome [193].

Rationale for a Genome Cartology of Maize LTR Retrotransposons

LTR retrotransposons comprise over 75% of the maize genome, and are thus the major contributors to DNA content in maize [1,194]. All datable LTR retrotransposons in maize have inserted within the last four million years and the majority of the LTR retrotransposon copies have inserted in replicative bursts over the last two million years [194]. These saltatory replication events have been taxonomically non-random with the five top families estimated to contribute between 60% and 80% of the LTR retrotransposon content of the maize genome [58,194]. This large proportion of TEs in the genomic DNA of maize has allowed for analyses of the spatial distribution of the high-copy-number and medium-copy-number families of LTR retrotransposons in the maize genome using wet lab methods and analysis of partial genomic sequence datasets.

The accumulation of some of the high-copy-number families is spatially nonrandom within the maize genome [118,195,196]. A low resolution study of physically mapped BAC ends indicated that the *Cinful*, *Zeon* and *Prem-1* families are preferentially accumulated near the centromere, while the *Ji*, *Huck* and *Opie* families are preferentially depleted near the centromere [118]. The same study indicated that the *Grande* family was uniformly distributed across the genome. These findings are generally consistent with previous results derived from fluorescent *in situ* hybridization studies [146,197] and the bioinformatic annotation of transposable elements in randomly selected genomic contigs [61,198]. These later efforts have suggested that LTR retrotransposons are more effectively removed in gene-rich regions of the genome [53,61].

A group of medium-copy-number LTR retrotransposons are also found to be broadly conserved in the grasses, with a general trend of centromeric specificity [60,147,148]. The centromeric retrotransposons in maize (CRMs) are *Gypsy* elements related to a larger group of retrotransposons with chromodomains [199,200]. The C-terminal region of the CRM integrase contains a conserved chromodomain-like region referred to as the CR motif that potentially directs integration of LTR retrotransposons to heterochromatin [33,34]. The CRM elements have been shown to interact specifically with centromeric histone H3 [150] providing a potential mechanism for directed integration. There are four named clades of CRM elements within the maize genome [60]. Three of the CRM clades are associated with active centromeres in maize while a fourth is comprised of inactive elements that are not associated with currently active centromeres [60].

The extant patterns of LTR retrotransposon accumulation in genomes result from the interplay of directed insertion and differential removal that generate genetic diversity that is sorted by demographic and selective processes acting at the population genetic level. Elucidating how these processes have contributed to the observed extant diversity of LTR retrotransposons in maize will require a high resolution spatial analysis of the full suite of LTR retrotransposon diversity that goes beyond the description of high- and medium-copy number elements that have been studied to date. The recent completion of the draft genome sequence assembly of the maize B73 inbred line [1] allows for the first comprehensive study of the full ensemble of LTR retrotransposon diversity in maize. Furthermore, the accessioned golden path assembly [201] produced for maize will allow the location of these sequences to be mapped at previously unavailable spatial resolutions. These emerging genomic data will thus allow for the first comprehensive genome cartology of LTR retrotransposons in maize.

At approximately 2,300 MB in length, the B73 maize inbred line has the largest plant genome sequenced to date [1]. Since over 75% of this genome is comprised of LTR retrotransposon insertions, this assembly contains an unprecedented volume of LTR retrotransposon data. The fact that these TE sequences exist in a complex sea of nested insertions further complicates their description. The comprehensive analysis of the genome cartology of maize TEs in the assembled maize genome will thus require the development of new bioinformatics tools for high-throughput annotation and taxonomic assignment of TE sequences.

The Annotation of Maize LTR Retrotransposon with DAWGPAWS

LTR retrotransposons can be computationally located in genomic DNA sequences using *ab initio* techniques, similarity based searches, or by mapping the positions of mathematically defined repeats. Several *ab initio* annotation programs have been published that search genomes for the presence of the diagnostic structural features of individual LTR retrotransposon insertions [202,203,204,205,206]. These features include the long terminal repeats flanked by target site duplications, and the protein coding regions found in full-length autonomous copies. Programs have also been developed that use similarity searches against a database of previously identified TEs [58,207,208,209]. These programs use local alignment search tools such as BLAST [210] to find regions of conservation, and then join these fragments into contiguous TE models. Because databases of TE sequences are often large and contain redundant copies of the same family, comprehensive databases are often reduced to a smaller set of representative sequences [211,212], a consensus sequence for each family [213], or a set of computationally generated chimeric sequences that capture the structural diversity of a family [214]. Finally, in addition to full-length similarity-based annotation using these databases, it is also possible to visualize the

frequency of individual oligomer 'words' in the genome to facilitate annotation of repeats by defining regions of mathematically defined overrepresentation [215,216]. These various methods can be combined together through manual curation or computational integration to generate a robust annotation of LTR retrotransposons.

The DAWGPAWS program for plant genome annotation [217] supports multiple similarity-search based programs and *ab initio* methods, as well as the use of mathematically defined repeats. DAWGPAWS will be used here for LTR retrotransposon annotation in maize. The specific annotation process that will be used for this dissertation is outlined in Figure 1.1. I will first make use of structure-based annotation of assembled bacterial artificial chromosomes (BACs) to find the full-length copies of LTR retrotransposons in the genome. An all-by-all BLAST search of the 5'-LTR (Figure 1.2) of all full length LTR retrotransposons derived from this search will be used to generate a matrix of similarity. This similarity matrix will be used to cluster individual LTR retrotransposon copies into families using the graph theory based methods of connected component clustering and Markov clustering [218]. Representative exemplar sequences from these families will be selected using affinity propagation-based clustering [219]. These exemplar models will then be manually curated and used for similarity-based annotation of the assembled B73 maize genome with the RepeatMasker program [209]. These results will be computationally post-processed in a custom-designed MySQL database to remove overlapping ends resulting from RepeatMasker annotation. The resulting database will serve as the platform to query the location of LTR retrotransposons in the maize genome.

Network Methods for Analysis of Maize LTR Retrotransposons in RepMiner

The taxonomic assignment of maize LTR retrotransposons into families will make use of graph theory based methods implemented in the RepMiner package [220]. One useful component of this package is the ability to visualize the networks of similarity in large datasets of LTR retrotransposons. The use of physically constructed network models for modeling biological relationships were abandoned early in the development of numerical taxonomy due to the difficulty in constructing and communicating these models [221,222]. However, modern software tools have revitalized the network approach to visualizing data relationships [223]. Network methods allow for dense representations not possible in dendrogram views. For instance, modern computational methods can allow over 1 million OTUs (operational taxonomic units) to be represented in a single visualization [224].

In computational approaches to modeling networks, individual OTUs are treated as virtual nodes that can be connected by edges that represent the similarity among the data. Algorithms that combine mathematical models from topology, graph theory and geometry can be applied to this network to reposition the nodes into meaningful representations in two or three-dimensional space [225,226]. In addition to positioning nodes into meaningful clusters of similarity, features relating to the nodes or edges can be mapped onto these networks. Node attributes such as taxonomic membership or other biologically relevant features may be mapped onto these networks by modifying node shape or color (Figure 1.6-B). Edge attributes such as a measure of similarity among nodes may also be represented as color or line thickness to provide useful visualizations of the individual relationships among pairs of nodes.

The visualization approach used in the RepMiner package takes matrices of similarities among nucleotide sequence strings as input, and visualizes these as networks using the

Cytoscape program [227]. These similarity matrices are generated from all-by-all BLAST searches of the sequences that are being clustered into taxonomic groups. For LTR retrotransposons, the 5' LTR was selected as the diagnostic sequence to represent the rest of the full-length LTR retrotransposon. Using the 5' LTR captures the region that is present in nearly all structural variants of LTR retrotransposons (Figure 1.4), and allows for the assignment of solo LTRs to a family using the same criteria as full-length elements [52]. RepMiner can also process biological feature data such as protein domains or known family membership and convert these data into a format that can be visualized as node color or shape on these networks (Figure 1.6-B).

In additional to network visualization, the RepMiner suite also facilitates using graph theory-based clustering methods for grouping these networks into groups representing putative families. The currently implemented methods include connected component clustering (CC), Markov clustering (MCL) [218,228], and affinity propagation (AP) [219]. The connected component method simply joins all connected nodes into a single family. One drawback of this method is that any chimeric LTR retrotransposon model that contains parts of two separate families will falsely join two separate groups into a single family. This is a particular problem for LTR retrotransposons where computational annotation results will contain many chimeric annotations. This can be alleviated to some degree be selecting only the 5' LTR for clustering, but chimeric LTR annotations will still contribute to the problem of false joins.

An alternative to CC clustering is the use of data driven unsupervised clustering software such as the MCL program that takes a graph-theory based approach to partitioning a dataset into "natural" classes. The MCL approach clusters data by using a matrix implementation of flow simulation that includes an expansion step where flow among nodes over edges is made more homogenous, and an inflation step that models a contraction of this flow and results in regions of

high and low currents (http://www.micans.org/mcl/#questions). This process is repeated iteratively until natural clusters are determined where the flow is spread out within natural clusters and evaporated between unconnected clusters. MCL results are robust to false edges that result from chimeric annotations [229,230], and thus represent a powerful supplement to connected component based results. MCL-based clustering has been used previously for clustering of protein families [228], but this study represents the first time the MCL algorithm has been used to generate families of transposable elements.

Once all of the full-length copies of LTR retrotransposons are clustered into families, it is useful to define representative sequences from within these families to generate a nonredundant database for similarity-based annotation. This can be done by aligning all members of a family to generate a single consensus sequence, or by selecting a subset of the extant representatives from each family to serve as exemplar sequences. The RepMiner program facilitates the use of AP clustering [219] to find exemplar sequences. AP implements a search algorithm similar to MCL clustering, but with the goal of finding specific sequences that can represent each cluster. These exemplar sequences are distributed throughout the sequence similarity network, and the affinity propagation mode of distilling TE databases generates representative sequence sets that are more representative than alternative methods. These exemplar LTR retrotransposon sequences can undergo additional manual curation for quality by removing short repetitive sequences or insertions of other elements into the target family. This will result in an exemplar database suitable for similarity-based annotation that will give higher quality results than using the comprehensive full-length TE database, and the concise database will be computationally faster to search against.

Genome Cartology of Maize LTR Retrotransposons with GenCart

The primary annotation results for LTR retrotransposons in the maize genome will represent one of the largest datasets of sequence features for any eukaryotic genome sequenced to date. The results of the RepeatMasker-based annotation will generate many overlaps in the resulting output that must be trimmed before an analysis of these data can be undertaken. These overlapping annotations result from situations where the RepeatMasker program did not assign a particular sequence segment to any single family, but assigned the fragment with equal probability to two adjacent families. These overlaps are particularly prevalent in the maize genome where thousands of nested insertions makes the similarity-based annotation results difficult to deconvolute. Additionally, the RepeatMasker results will be analyzed in combination with over 30,000 full-length LTR retrotransposons annotated in the genome. These data management issues have required that the genome cartology of maize LTR retrotransposons incorporate a relational database management system (RDBMS) to facilitate the organization and querying of this massive data set. For my dissertation, this RDBMS will be implemented in the MySQL database program using a custom database format.

The MySQL database of annotated LTR retrotransposons in maize will serve as the main tool to query the location of LTR retrotransposons in the fully assembled maize genome. This database is stored in a custom GenCart database schema that has been optimized for spatially related queries. Multiple PERL programs interface with this database to import data, remove overlaps in sequence features, and provide summarizing queries of the distribution of sequence features in the genome. These query results are then imported into the R statistical computing framework [231] for visualization and analysis of the spatial distribution of sequence features using custom tools.

Overview of the Dissertation Chapters

A major component of this dissertation is a description of the bioinformatic tools described briefly above, as well as the use of these tools for analysis of LTR retrotransposons in maize. Chapter 2 describes in detail the DAWGPAWS suite of programs for plant genome annotation. Chapter 3 describes the RepMiner network visualization program and its use for LTR retrotransposon analysis in maize, while chapter 4 provides an overview of the use of an affinity propagation algorithm for the generation of a database of representative exemplars sequences of transposable elements suitable for similarity search-based genome annotation. Chapter 5 describes the use of this suite of tools in the annotation of retrotransposons in maize, and the analysis of the spatial patterns in the distribution of LTR retrotransposons. Chapter 6 concludes the dissertation by summarizing these results, and indicating how these informatic tools can be applied to future studies of transposable element diversity and distribution.

Table 1.1. Intragenomic spatial patterns in eukaryotic transposable element distribution. Specific examples of the spatial genomic patterns observed in accumulation, insertion, removal and natural selection on transposable elements that have shaped extant distribution patterns in eukaryotes are listed below. The classification of elements follows Wicker *et al* [26]. The local spatial pattern indicates the genomic nature of the individual points of insertion while the global spatial pattern refers to the general trends observed for the genome as a whole. The three-letter code following the host genome name indicates the experimental system used. The first letter indicates the element source and whether it is endogenous (E) or heterologous (H). The second letter indicates the activity of the element family as currently transposing (C) or active in the past (P). The third letter indicates the cells in which transposition took place as being *in vivo* (V) or grown in tissue culture (T).

Classification			Host Genome	Local Spatial Pattern	Global Spatial Pattern
Order	**Superfamily**	**Family**			
Retrotransposons					
LTR	*Copia*	*Ty1*	Yeast (ECV)	Insertion does not target a specific consensus sequence [175], rather insertion preference is upstream of RNA polymerase III-transcribed genes [167,171,172] including tRNA genes, snRNA genes and 5s rRNA [171,173,174], probably due to the targeted integration at chromatin features associated with RNA polymerase III transcription [170,173,175].	Global spatial patterns are at least partially due to local targeting of RNA polymerase III transcribed genes [167].
LTR	*Copia*	*Ty5*	Yeast (ECV)	Insertion does not target a specific consensus sequence but targets epigenetic signals [167,176].	Insertion preference into telomeres, and silent mating loci with chromatin states that are similar to heterochromatin in eukaryotes [167,176,177,178]. Insertion is chaperoned by an interaction between the targeting domain of the *Ty5* integrase and the heterochromatin protein *silent information regulator* (*Sir4*) [179,180,181,182], with the *Ty5* integrase mimicking a host protein associated with the inner nuclear membrane [183]. Transposition rate decreases and targeting of insertion becomes more spatially random under stress conditions [184].
LTR	*Copia*	*Tos17*	Rice (ECT)	Insertion preference into genes [70,71]. Insertion preference into protein kinase and disease resistance genes, this preference due in part to GC pattern specificity of the insertion site [70,71].	Insertion preference into distal regions of chromosomes avoiding retrotransposon-rich pericentromeric regions [70,71].

Classification Order	Superfamily	Family	Host Genome	Local Spatial Pattern	Global Spatial Pattern
LTR	*Gypsy*	*Ty3*	Yeast (ECV)	Insertion preference into a narrow window one to two base pairs upstream of RNA polymerase III transcription initiation [166,167,168]. Targeting to this region is potentially due to a direct recognition of the polymerase transcription complex [168,169,170].	Global patterns are at least partially due to the local targeting of RNA polymerase III-transcribed genes [167].
LTR	*Gypsy*	*CRs* Multiple famlies	Maize (EPV) Rice (EPV)	Centromeric retrotransposons (CRs) accumulate preferentially in centromeric DNA, suggesting specific targeting of centromeric heterochromatin [152].	CRs are core components of the centromeres in the Poaceae [145,146,147,148,149,150,151]. These include multiple families of centromeric retrotransposons in rice (CRRs) [152] and centromeric retrotransposons in maize (CRMs) [60,153]. Interestingly, differences in centromeric enrichment exist between autonomous and nonautonomous members of the same family [152]. The enrichment of these elements in the centromeres suggests a targeting of insertion for centromeric heterochromatin [152]. The C-terminal region of CR integrase contains a conserved chromodomain-like region referred to as the CR motif that potentially directs integration of LTR retrotransposons to heterochromatin [33,34], and the CRM elements have been shown to interact specifically with centromeric histone H3 [150] providing a potential mechanism for directed integration.
LTR	*Gypsy*	*MAGGY*	Rice Blast Fungus (EPV, ECV)	Chromodomains in the C terminal end of integrase target insertion at histone H3 methylated at lysine-9 (H3 K9), an epigenetic marker of heterochromatin [33].	Spatially clustered with other transposable elements in the host genome [155,156,157] and are located in relatively gene poor regions [33].

Classification Order	Superfamily	Family	Host Genome	Local Spatial Pattern	Global Spatial Pattern
LTR	*Gypsy*	*Tf1*	Fission Yeast (EPV, ECV)	Preferentially inserts into a narrow window 100-400 bp upstream of open reading frames, which corresponds to the promoter region of polymerase II-transcribed genes [159,160,161,162,163]. This insertion preference involves the interaction of a chromodomain in the integrase and chromatin features specific to pol II promoters [164,165].	Insertion preference at intergenic regions since the global distribution is due to local targeting of polymerase II promoters [159,163]. There is no correlation between the level of transcription and the rate of insertion, and genes that are transcribed under stress conditions are preferentially targeted [163].
LINE	*R2*	*R2*	Arthropods (EPV) Drosophila (EPV, ECV)	*R1* and *R2* insert into specific sequences that are 74 base pairs apart in 28S ribosomal RNA genes [100,101,102].	Accumulation in the 28 S RNA genes of arthropods [98,99].
LINE	*Jockey*	*Jockey*	Drosophila (EPV)	No sequence motif patterns or bias in physical properties of the DNA at points of insertion [232].	Accumulation preference in euchromatic regions [233,234].
LINE	*Jockey*	*TART* *HeT-A* *TAHRE*	Drosophila (ECV, EPV)	Insertion at broken chromosome ends [138,139,140].	Exclusively transpose to the termini of chromosomes [135,136] with no copies present in the sequenced gene-rich euchromatic region of the genome [136,137] although these regions of the telomeres are arguably euchromatic in nature [235]. Gag proteins target these families to the telomere; *TART* and *TAHRE* Gag proteins require the presence of *HeT-A* Gag proteins for this targeting [136,141,142] while *HeT-A* is presumably dependent on *TART* for production of its reverse transcriptase [136].

Classification Order	Superfamily	Family	Host Genome	Local Spatial Pattern	Global Spatial Pattern
LINE	*L1*	*L1 (LINE1)*	Human (ECT, ECV, EPV) Mouse (EPV)	Endonuclease directs insertion into poly AT sequences [236,237] that meet specific requirements of minor grove width. In general, the first L1 endonuclease nicking targets the TTAAAA consensus sequence [238] while the second nicking site has the general consensus of ANTNTNAA [239] Accumulation in introns of genes [240]. Insertion preference for antisense orientation with respect to mRNA transcription within the introns of the genes they insert into [240].	Chromatization generally represses nicking by L1 endonuclease although some sequences showed enhanced nicking when packaged in histones [241].
SINE	*7SL*	*Alu*	Human (EPV)	Due to a dependence on L1 endonuclease for insertion [242], the first nick for insertion is usually at the TTAAAA (T_mA_n) consensus sequence. [243,244] Accumulation preference in the introns of genes [240].	Enriched accumulation in gene rich, high GC content regions of the genome in general, but younger *Alu* elements have a more uniform distribution [73,74,75] with very young elements showing an enrichment in regions of low GC content [245]. Very young *Alu* elements are more enriched on the Y chromosome than the X chromosome or autosomes [246,247,248]. *Alu* deletion rate is fastest on the Y chromosome, intermediate on autosomes, and slowest on the X chromosome [247,248].
SINE	*tRNA*	*ID1 (Identifier)*	Rat (EPV)	No documented pattern.	Both younger and older *ID* elements are enriched in gene rich, high GC content regions [245,249,250]. Accumulation of *ID* SINES in rat have the same genomic distributions as accumulation of the independently inserted and unrelated *B1* SINEs in mouse [250].

Classification Order	Superfamily	Family	Host Genome	Local Spatial Pattern	Global Spatial Pattern
DNA Transposons					
TIR	*Tc1-Mariner*	*Tc1*	Nematode (EPV)	Insertion into the sequence TA [108]. Selection against TE insertions within coding regions [77].	Insertion into regions of high recombination [77]. Selection against TE insertions does not increase with recombination rate [77].
TIR	*Tc1-Mariner*	*Sleeping Beauty* reconstructed element	Human (HCT) Mouse (HCT)	Insertion into the central TA of the sequence ATATATAT, a bendable string of AT repeats with a symmetrical pattern in hydrogen bonding [251,252,253].	Insertion near donor site (mouse) [86,87,88]. No bias for insertion near genes [88].
TIR	*hAT*	*Ac/Ds*	Maize (ECV) Arabidopsis (HCV) Tobacco (HCV) Tomato (HCV)	No consensus sequence preference for target site integration [254,255] although specific structural DNA properties are preferred [255]. Insertion into both the introns and exons of genes in rice and maize [67,256] with an observed insertion preference near the ATG start of translation in Arabidopsis [68].	Insertion preference near donor site [68,78,79,81,82]. Insertion preference into low copy DNA and coding regions [256,257,258]. Insertion preference into areas near chromosome ends in maize [255].
TIR	*hAT*	*Tol2*	Zebrafish (HCV)	Insertion preference in AT-rich regions with weak palindromic core consensus [84]. Insertion preference into transcription units [84]. Insertion into DNA transposons but not retrotransposons [84].	Insertion preference near donor site [83,84]. Generally no trend for insertion into specific regions of the chromosome [259].
TIR	*hAT*	*hobo*	Drosophila (EPV, ECV)	Consensus sequence of NTNNNNAC at the point of integration [260,261] with an integration hotspot at the sequence ATCCTCAC [261,262]. Although the sequences present in the 31 bp flanking the integration hot spot influence the availability of the sequence for integration, there is no common sequence motif for this pattern suggesting a role for structural characteristics of the DNA or chromatin state [261].	Insertion preference near donor site [85,263]. Accumulation distribution positively correlated with recombination rate [264]. Older elements have accumulated in pericentromeric regions but recent insertions are more evenly distributed [85].

Classification Order	Superfamily	Family	Host Genome	Local Spatial Pattern	Global Spatial Pattern
TIR	*hAT*	*Tam3*	Snapdragon (EPV, ECV)	Insertion preference for noncoding regions of genes such as promoters, 5' UTRs and introns [265].	Insertion preference near the donor site [80] probably due to physical association between target and donor sites at time of transposition [91]. Insertion preference into low copy DNA [266].
TIR	*Mutator*	*Mu1, MuDR (and others)*	Maize (ECV)	Insertion preference into promoters and 5' UTRs of genes [66]. Insertion preference into specific DNA structural features but not a specific DNA sequence [267].	Insertion not linked to donor site [268]. Insertion preference into low copy DNA and genic regions [119,258,267]. Insertion into regions of high recombination [66].
TIR	*Mutator*	*AtMu1*	Arabidopsis (ECV*) **ddm1 mutant*	No bias for insertion into repeats, promoters or retrotransposons [72].	Insertion not linked to donor site [72].
TIR	*P*	*P*	Drosophila (ECV)	Although insertion preferences into GC-rich sequences similar to GGCCAGAC have been reported [269,270], more comprehensive analyses have revealed no clear consensus sequence in the 8bp target site duplications [114]. Rather, insertion preference is into specific DNA structural features with a 14 bp palindromic hydrogen-bonding pattern in the major groove [114] and an insertion preference at the 5' end of genes near the transcription start site [114,271].	Insertion preference near the donor site in some cases [89,272] and an insertion preference into euchromatin over heterochromatin [114,273].
TIR	*PiggyBac*	*PiggyBac*	Mouse (HCT) Drosophila (HCV)	Insertion exclusively into TTAA [104,105,106]. Insertion preference into genes or near genes [87,88,92,94]. Insertion preference near 5' end of gene [88,92]	Insertion preference near donor site in some cases [87] but often observed to be spatially random with respect to donor site [88,92].
TIR	*PiggyBac*	*Pokey*	Daphnia (EPV)	Inserts exclusively into TTAA [103,109,110] with a high number of insertions into the large subunit ribosomal rRNA gene a few bases away from *R1* and *R2* insertions [103].	Although this family inserts into multiple locations in the *Daphnia* genome, it occurs in a high frequency in large subunit rRNA genes in multiple *Daphnia* species [103].
TIR	*PIF-Harbinger*	*PIF*	Maize (ECV)	Preference for insertion at 9bp sequence CWCTTAGWG [111,112,113].	Insertion preference for genic regions [113].

Classification Order	Superfamily	Family	Host Genome	Local Spatial Pattern	Global Spatial Pattern
TIR	*PIF-Harbinger*	*mPing MITE*	Rice (EPV, ECV) Arabidopsis (HCV) Yeast (HCV)	Inserts at TAA/TTA with the 9bp target consensus sequence YTMTWAKAR in yeast and rice, and YTMTWAKAD in Arabidopsis [131,132,274]. Insertion preference in or near genes [131] but generally avoids insertion into exons [19].	Insertion preference into low copy DNA and genic regions [19,128,131,132]. Insertion preference for euchromatic regions [19].
TIR	*CACTA*	*En/Spm*	Maize (ECV) Arabidopsis (HCV) Rice (HCV) Tobacco (HCV)	Insertion preference into genes, but no bias for insertion into specific region of genes (rice) [275]. No sequence specificity of insertion reported.	Insertion preference into low copy DNA (maize) [258] Insertion preference near donor site (maize,tobacco) [276,277]. Insertion preference away from centromere (Arabidopsis) [69].
TIR	*CACTA*	*CACTA1*	Arabidopsis (ECV*,EPV) **ddm1 mutant*	No sequence specificity of insertion reported.	Accumulation in pericentromeric DNA and transposon-rich heterochromatic regions [278,279]. Integration was found to not target heterochromatin in natural populations [280], implicating the role of spatial differences in selection or removal that are responsible for current accumulation patterns.
TIR	*CACTA*	*Rim2/Hipa*	Rice (EPV)	Accumulation preference in genes including introns, exons and UTRs [281]. Accumulation preference in AT-rich regions [281].	Accumulation in centromeric regions of chromosomes [281,282].
Helitron	*Helitron*	Multiple families	Arabidopsis (EPV) Nematode (EPV) Rice (EPV) Sorghum (EPV) Maize (EPV)	Insertion preference in AT rich DNA with insertion orientation bias (Arabidopsis, nematode, maize, rice) [133,134]. Accumulation near other *Helitrons* and DNA transposons [134].	Accumulation in gene-poor regions (Arabidopsis, Nematode, Rice)[133]. Accumulation in gene-rich regions (Maize) [134]. Insertion preference into regions of the genome containing the same family or subfamily of *Helitrons* [134].
Unclassified MITE	Unknown	Multiple families	Maize (EPV) Sorghum (EPV)		MITEs colocalized with matrix attachment regions [283]

Classification			Host Genome	Local Spatial Pattern	Global Spatial Pattern
Order	**Superfamily**	**Family**			
Unclassified MITE	Unknown	*Heartbreaker* MITE	Maize (EPV)	Accumulation preference in or near genes [127]	Insertion preference into low copy DNA and genic regions [127] Recently active copies are evenly distributed across maize chromosomes [284]
Unclassified MITE	Unknown	*Emigrant* MITE	Arabidopsis (EPV)	TA target site preference [285,286]	Insertions are far from genes [285,286]. Young elements insert far from ORFs but elements linked to genes are more frequently maintained [287]

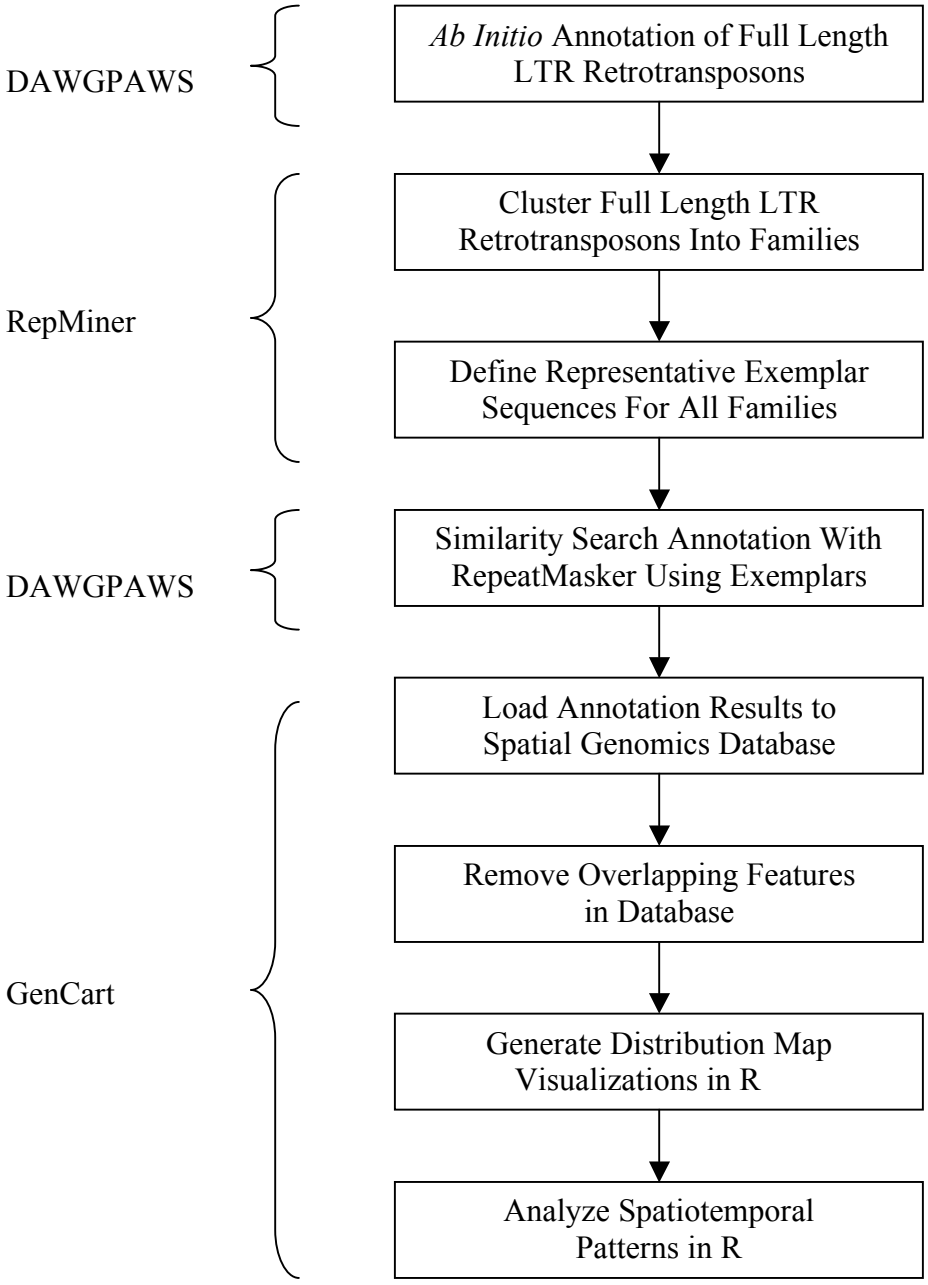


Figure 1.1. Overview of the informatic pipeline used to study the distribution and abundance of LTR retrotransposons in the B73 maize genome. The boxes on the right describe the basic process, and the specific package used to implement the process is indicated on the left.

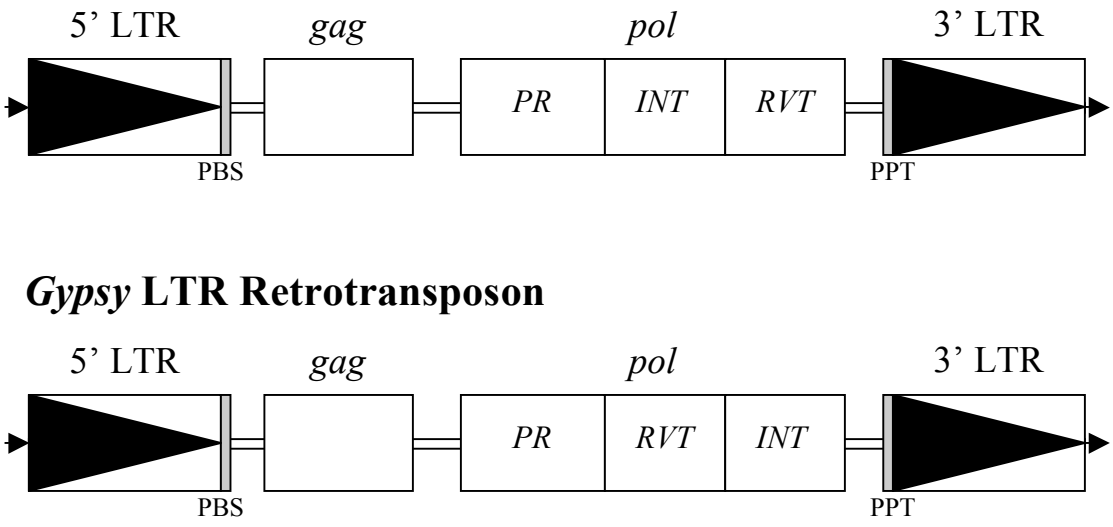


Figure 1.2. The structural features of the typical full length Long Terminal Repeat (LTR) retrotransposon for the *Gypsy* and *Copia* superfamilies. The LTRs are labeled above and the target site duplications are represented as arrows flanking the LTRs. The regions indicated as *gag* and *pol* are open reading frames that code for the two transcripts derived from an autonomous copy of an LTR retrotransposon. The *gag* transcript produces the GAG protein that contributes to the virus like particle in the host cytoplasm. The *pol* transcript is a polyprotein that is processed by an aspartic protease into an aspartic protease (PR), integrase (INT) and a reverse transcriptase (RVT) that also includes a RNaseH domain. The diagnostic structural difference between the *Gypsy* and *Copia* superfamily is the order of RVT and INT in the *pol* transcript. The primer binding site (PBS) and the polypurine tract (PPT) are important binding domains involved in reverse transcription and are indicated as gray boxes.

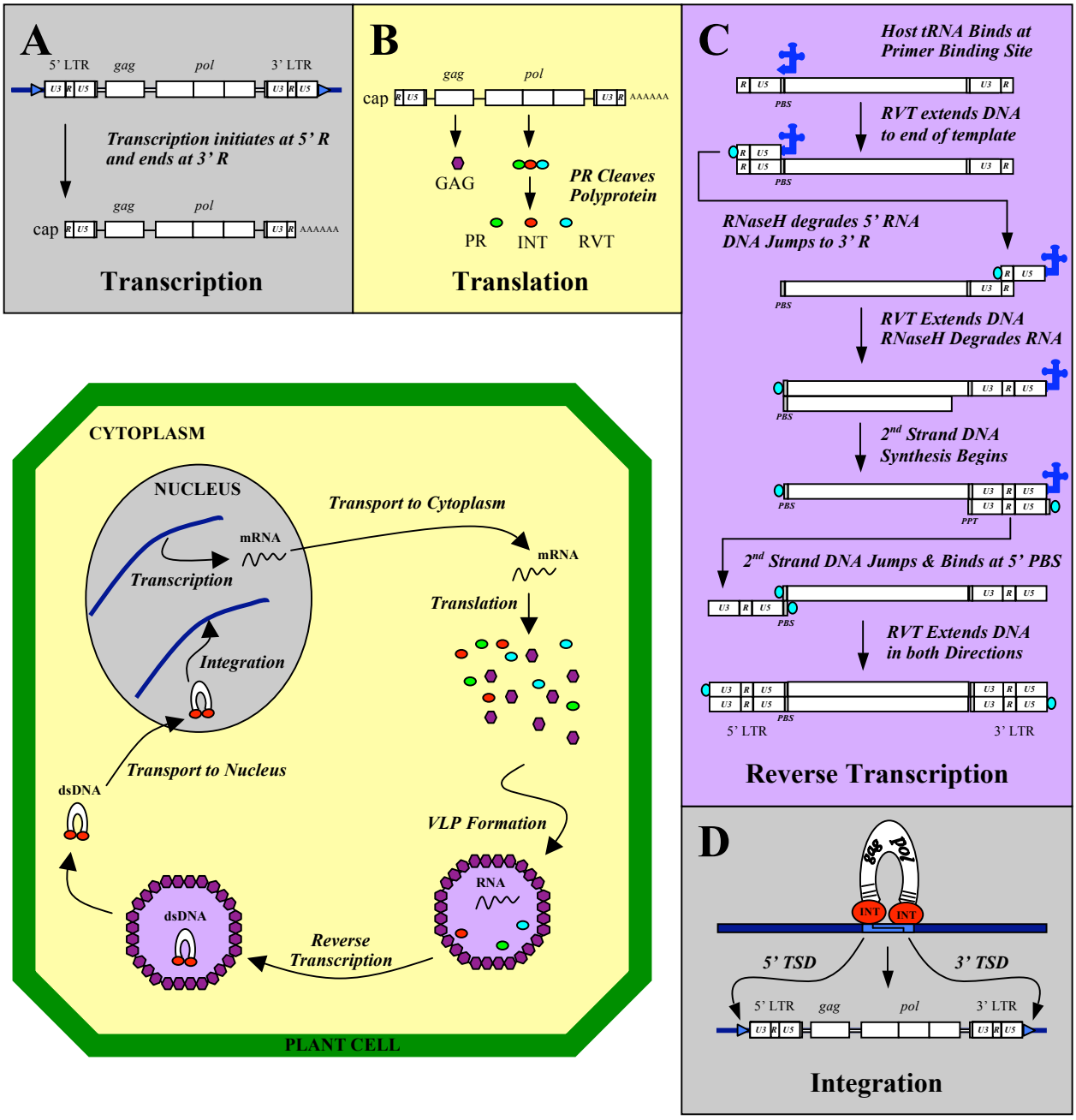


Figure 1.3. The life cycle of LTR retrotransposons. This cycle includes processes that take place in the host nucleus indicated with a gray background, the host cytoplasm indicated by the yellow background, and the virus like particle (VLP) indicated by the purple background. A) Transcription of mRNA from genomic DNA makes use of host transcription machinery and results in a capped RNA with a polyadenlyated tail. B) Protein production makes use of host translation machinery, resulting in a GAG protein and a polyprotein that is cleaved by PR to produce PR-Aspartic Protease, INT – Integrase, and RVT-Reverse Transcriptase. C) Reverse transcription of single stranded RNA to double stranded DNA is primed by a host tRNA with reverse transcriptase serving as the polymerase. D) Integration of double-stranded DNA into a new location in the host genome is mediated by integrase and results in a target site duplication (TSD) generated from DNA repair of the cut site in the host genome.

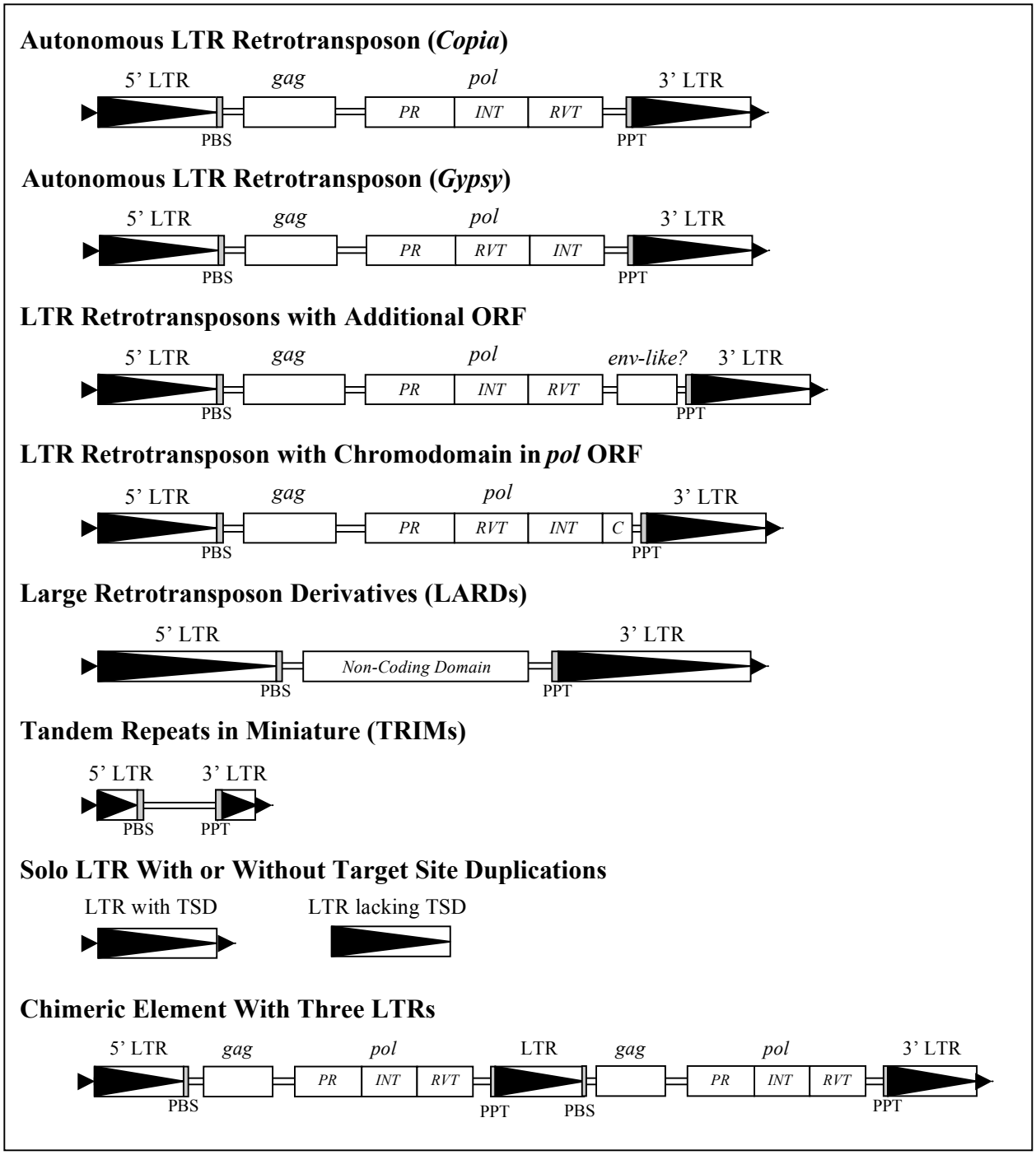


Figure 1.4. Structural diversity of LTR retrotransposon in plant genomes.

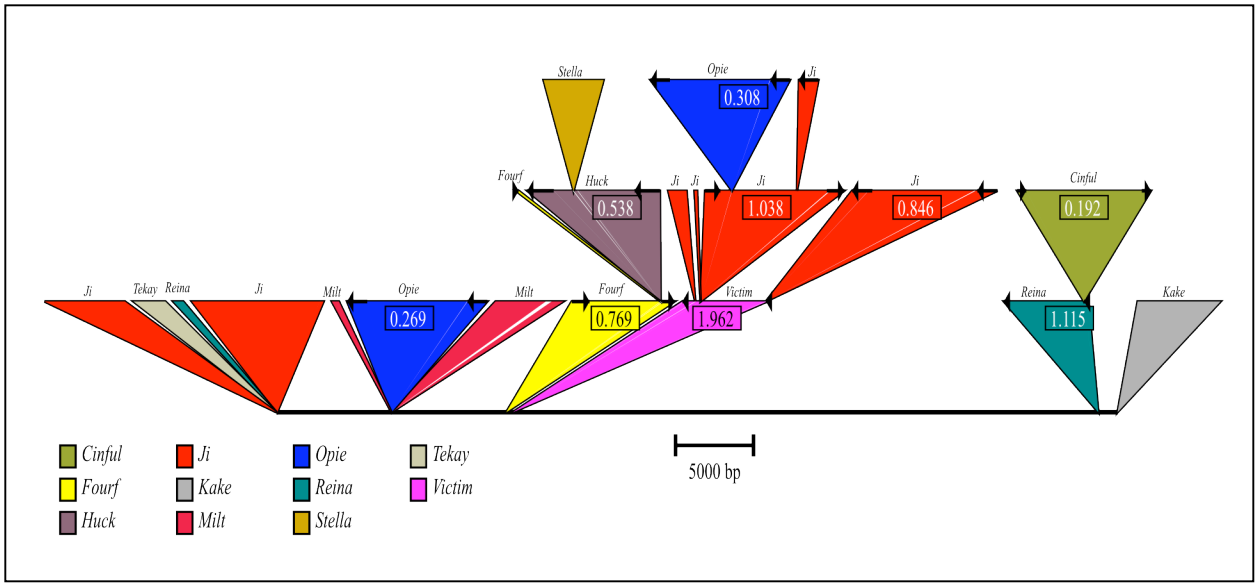


Figure 1.5 The nested LTR retrotransposons discovered a 160 kb region flanking the *adh1*-F gene in the maize genome. LTR location and direction are indicated as arrows, and the estimated date of insertion (million years ago) is indicated numerically on insertions with paired LTRs. This image was generated with the TEnest program [58] using published sequence data [288](Genbank accession AF123535.1) from a subregion of the *adh1*-F yeast artificial chromosome clone that was originally described by SanMiguel *et al.* [59].

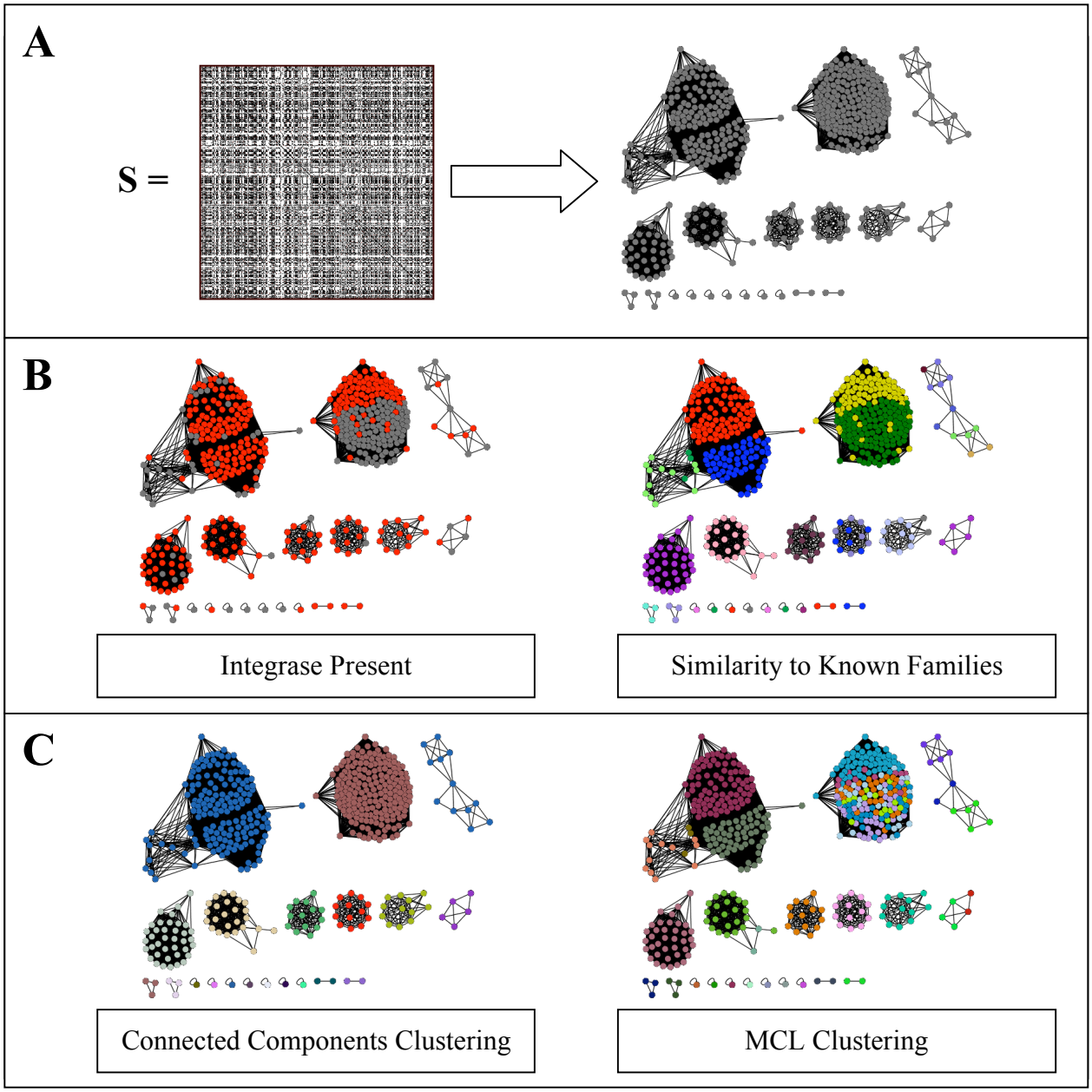


Figure 1.6. Network based methods of visualizing and clustering sequences. A) An all-by-all BLAST search of 500 LTR sequences generates a matrix of similarity (S). This matrix can be visualized as a network where nodes represent individual sequences, and the edges connecting the nodes are local alignments that exceed the significance threshold set by the user in BLAST (*i.e.,* $e<1x10^{-10}$). B) Biological sequence features mapped onto the network. The presence of a recognizable integrase domain is mapped as red nodes, and the similarity to known families is represented as different node colors. C) Clustering results mapped as node color onto the network. Results are shown for connected components clustering and MCL clustering of the similarity matrix.

References

1. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, et al. (2009) The B73 maize genome: complexity, diversity and dynamics Science 326: 1112-1115.
2. Feschotte C, Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet 41: 331-368.
3. Kidwell MG (2002) Transposable elements and the evolution of genome size in eukaryotes. Genetica 115: 49-63.
4. Finnegan DJ (1989) Eukaryotic transposable elements and genome evolution. Trends in Genetics 5: 103-107.
5. Bennetzen JL, Ma JX, Devos K (2005) Mechanisms of recent genome size variation in flowering plants. Annals of Botany 95: 127-132.
6. Sanmiguel P, Bennetzen JL (1998) Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Annals of Botany 82: 37-44.
7. Thomas CA, Jr. (1971) The genetic organization of chromosomes. Annu Rev Genet 5: 237-256.
8. McClintock B (1947) Cytogenetic studies of maize and *Neurospora*. Carnegie Inst Washington Year Book 46: 146-152.
9. Bennetzen JL (2000) Transposable element contributions to plant gene and genome evolution. Plant Mol Biol 42: 251-269.
10. Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR (2004) *Pack-MULE* transposable elements mediate gene evolution in plants. Nature 431: 569-573.
11. Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, et al. (2005) Gene duplication and exon shuffling by *helitron*-like transposons generate intraspecies diversity in maize. Nature Genetics 37: 997-1002.
12. Bennetzen JL (2005) Transposable elements, gene creation and genome rearrangement in flowering plants. Curr Opin Genet Dev 15: 621-627.
13. Wang W, Zheng H, Fan C, Li J, Shi J, et al. (2006) High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell 18: 1791-1802.
14. Sekita Y, Wagatsuma H, Nakamura K, Ono R, Kagami M, et al. (2008) Role of retrotransposon-derived imprinted gene, *Rtl1*, in the feto-maternal interface of mouse placenta. Nature Genetics 40: 243-248.
15. Gonzalez J, Petrov DA (2009) The adaptive role of transposable elements in the *Drosophila* genome. Gene 448: 124-133.
16. Mcclintock B (1951) Chromosome organization and genic expression. Cold Spring Harb Symp Quant Biol 16: 13-47.
17. Gogvadze E, Buzdin A (2009) Retroelements and their impact on genome evolution and functioning. Cell Mol Life Sci.
18. Feschotte C (2008) Transposable elements and the evolution of regulatory networks. Nat Rev Genet 9: 397-405.
19. Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, et al. (2009) Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature 461: 1130-1134.
20. Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, et al. (2004) Role of transposable elements in heterochromatin and epigenetic control. Nature 430: 471-476.

21. Malone CD, Hannon GJ (2009) Small RNAs as guardians of the genome. Cell 136: 656-668.
22. Obbard DJ, Gordon KH, Buck AH, Jiggins FM (2009) The evolution of RNAi as a defence against viruses and transposable elements. Philos Trans R Soc Lond B Biol Sci 364: 99-115.
23. Wessler SR (1996) Turned on by stress. Plant retrotransposons. Curr Biol 6: 959-961.
24. McClintock B (1984) The significance of responses of the genome to challenge. Science 226: 792-801.
25. Zeh DW, Zeh JA, Ishida Y (2009) Transposable elements and an epigenetic basis for punctuated equilibria. Bioessays 31: 715-726.
26. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, et al. (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8: 973-982.
27. Sabot F, Schulman AH (2006) Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. Heredity 97: 381-388.
28. Flavell AJ, Pearce SR, Heslop-Harrison P, Kumar A (1997) The evolution of *Ty1-copia* group retrotransposons in eukaryote genomes. Genetica 100: 185-195.
29. Suoniemi A, Tanskanen J, Schulman AH (1998) *Gypsy*-like retrotransposons are widespread in the plant kingdom. Plant J 13: 699-705.
30. Boeke J, Eickbush T, Sandmeyer SB, Voytas DF (2005) Pseudoviridae. In: Fauquet CM, Mayo MA, Maniloff J, Desselberger U, Ball LA, editors. Virus Taxonomy: Classification and Nomenclature of Viruses: Eighth Report of the International Committee on the Taxonomy of Viruses. San Diego, CA: Elsevier. pp. 397-407.
31. Eickbush T, Boeke JD, Sandmeyer SB, Voytas DF (2005) Metaviridae. In: Fauquet CM, Mayo MA, Maniloff J, Desselberger U, Ball LA, editors. Virus Taxonomy: Classification and Nomenclature of Viruses: Eighth Report of the International Committee on the Taxonomy of Viruses. San Diego, CA: Elsevier. pp. 409-420.
32. Malik HS, Eickbush TH (1999) Modular evolution of the integrase domain in the *Ty3*/*Gypsy* class of LTR retrotransposons. J Virol 73: 5186-5190.
33. Gao X, Hou Y, Ebina H, Levin HL, Voytas DF (2008) Chromodomains direct integration of retrotransposons to heterochromatin. Genome Res 18: 359-369.
34. Novikova O (2009) Chromodomains and LTR retrotransposons in plants. Commun Integr Biol 2: 158-162.
35. Wright DA, Voytas DF (1998) Potential retroviruses in plants: *Tat1* is related to a group of *Arabidopsis thaliana Ty3/gypsy* retrotransposons that encode envelope-like proteins. Genetics 149: 703-715.
36. Peterson-Burch BD, Wright DA, Laten HM, Voytas DF (2000) Retroviruses in plants? Trends in Genetics 16: 151-152.
37. Vicient CM, Kalendar R, Schulman AH (2001) Envelope-class retrovirus-like elements are widespread, transcribed and spliced, and insertionally polymorphic in plants. Genome Res 11: 2041-2049.
38. Lerat E, Capy P (1999) Retrotransposons and retroviruses: analysis of the envelope gene. Mol Biol Evol 16: 1198-1207.
39. Havecker ER, Gao X, Voytas DF (2005) The Sireviruses, a plant-specific lineage of the *Ty1/copia* retrotransposons, interact with a family of proteins related to dynein light chain 8. Plant Physiol 139: 857-868.

40. Bousios A, Darzentas N, Tsaftaris A, Pearce SR (2010) Highly conserved motifs in non-coding regions of Sirevirus retrotransposons: the key for their pattern of distribution within and across plants? BMC Genomics 11: 89.
41. Marco A, MarÌn I (2005) Retrovirus-like elements in plants. Recent Res Devel Plant Sci 3: 15-24.
42. Bureau TE, White SE, Wessler SR (1994) Transduction of a cellular gene by a plant retroelement. Cell 77: 479-480.
43. Jin YK, Bennetzen JL (1994) Integration and nonrandom mutation of a plasma membrane proton ATPase gene fragment within the *Bs1* retroelement of maize. Plant Cell 6: 1177-1186.
44. Swain A, Coffin JM (1992) Mechanism of transduction by retroviruses. Science 255: 841-845.
45. Jin YK, Bennetzen JL (1989) Structure and coding properties of *Bs1*, a maize retrovirus-like transposon. Proc Natl Acad Sci U S A 86: 6235-6239.
46. Kalendar R, Vicient CM, Peleg O, Anamthawat-Jonsson K, Bolshoy A, et al. (2004) Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. Genetics 166: 1437-1450.
47. Witte CP, Le QH, Bureau T, Kumar A (2001) Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. Proc Natl Acad Sci U S A 98: 13778-13783.
48. Sanz-Alferez S, SanMiguel P, Jin YK, Springer PS, Bennetzen JL (2003) Structure and evolution of the *Cinful* retrotransposon family of maize. Genome 46: 745-752.
49. Marillonnet S, Wessler SR (1998) Extreme structural heterogeneity among the members of a maize retrotransposon family. Genetics 150: 1245-1256.
50. Sabot F, Schulman AH (2007) Template switching can create complex LTR retrotransposon insertions in Triticeae genomes. BMC Genomics 8: 247.
51. Moisy C, Blanc S, Merdinoglu D, Pelsy F (2008) Structural variability of *Tvv1* grapevine retrotransposons can be caused by illegitimate recombination. Theoretical and Applied Genetics 116: 671-682.
52. Devos KM, Brown JK, Bennetzen JL (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome Res 12: 1075-1079.
53. Vitte C, Bennetzen JL (2006) Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proceedings of the National Academy of Sciences of the United States of America 103: 17638-17643.
54. Petrov DA, Lozovskaya ER, Hartl DL (1996) High intrinsic rate of DNA loss in *Drosophila*. Nature 384: 346-349.
55. Petrov DA, Hartl DL (1997) Trash DNA is what gets thrown away: high rate of DNA loss in *Drosophila*. Gene 205: 279-289.
56. Vicient CM, Kalendar R, Schulman AH (2005) Variability, recombination, and mosaic evolution of the barley *BARE-1* retrotransposon. J Mol Evol 61: 275-291.
57. Ma J, Devos KM, Bennetzen JL (2004) Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res 14: 860-869.
58. Kronmiller BA, Wise RP (2008) TEnest: Automated chronological annotation and visualization of nested plant transposable elements. Plant Physiol 146: 45-59.

59. SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, et al. (1996) Nested retrotransposons in the intergenic regions of the maize genome. Science 274: 765-768.
60. Sharma A, Presting GG (2008) Centromeric retrotransposon lineages predate the maize/rice divergence and differ in abundance and activity. Molecular Genetics and Genomics 279: 133-147.
61. Liu R, Vitte C, Ma J, Mahama AA, Dhliwayo T, et al. (2007) A GeneTrek analysis of the maize genome. Proc Natl Acad Sci U S A 104: 11844-11849.
62. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. Nature Genetics 20: 43-45.
63. Gaut BS, Morton BR, McCaig BC, Clegg MT (1996) Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. Proc Natl Acad Sci U S A 93: 10274-10279.
64. Ma J, Bennetzen JL (2004) Rapid recent growth and divergence of rice nuclear genomes. Proc Natl Acad Sci U S A 101: 12404-12410.
65. Tenaillon MI, Hollister JD, Gaut BS (2010) A triptych of the evolution of plant transposable elements. Trends in Plant Science.
66. Liu S, Yeh CT, Ji T, Ying K, Wu H, et al. (2009) *Mu* transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. PLoS Genet 5: e1000733.
67. Vollbrecht E, Duvick J, Schares JP, Ahern KR, Deewatthanawong P, et al. (2010) Genome-wide distribution of transposed *Dissociation* elements in maize. Plant Cell.
68. Raina S, Mahalingam R, Chen F, Fedoroff N (2002) A collection of sequenced and mapped *Ds* transposon insertion sites in *Arabidopsis thaliana*. Plant Molecular Biology 50: 93-110.
69. Schneider A, Kirch T, Gigolashvili T, Mock HP, Sonnewald U, et al. (2005) A transposon-based activation-tagging population in *Arabidopsis thaliana* (TAMARA) and its application in the identification of dominant developmental and metabolic mutations. FEBS Lett 579: 4622-4628.
70. Miyao A, Tanaka K, Murata K, Sawaki H, Takeda S, et al. (2003) Target site specificity of the *Tos17* retrotransposon shows a preference for insertion within genes and against insertion in retrotransposon-rich regions of the genome. Plant Cell 15: 1771-1780.
71. Piffanelli P, Droc G, Mieulet D, Lanau N, Bes M, et al. (2007) Large-scale characterization of *Tos17* insertion sites in a rice T-DNA mutant library. Plant Molecular Biology 65: 587-601.
72. Singer T, Yordan C, Martienssen RA (2001) Robertson's *Mutator* transposons in *A. thaliana* are regulated by the chromatin-remodeling gene *Decrease in DNA Methylation* (*DDM1*). Genes Dev 15: 591-602.
73. Arcot SS, Shaikh TH, Kim J, Bennett L, Alegria-Hartman M, et al. (1995) Sequence diversity and chromosomal distribution of "young" *Alu* repeats. Gene 163: 273-278.
74. Arcot SS, Adamson AW, Lamerdin JE, Kanagy B, Deininger PL, et al. (1996) *Alu* fossil relics--distribution and insertion polymorphism. Genome Res 6: 1084-1092.
75. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.
76. Rizzon C, Marais G, Gouy M, Biemont C (2002) Recombination rate and the distribution of transposable elements in the *Drosophila melanogaster* genome. Genome Res 12: 400-407.

77. Rizzon C, Martin E, Marais G, Duret L, Segalat L, et al. (2003) Patterns of selection against transposons inferred from the distribution of *Tc1*, *Tc3* and *Tc5* insertions in the *mut-7* line of the nematode *Caenorhabditis elegans*. Genetics 165: 1127-1135.
78. Greenblatt IM (1984) A chromosome replication pattern deduced from pericarp phenotypes resulting from movements of the transposable element, *modulator*, in maize. Genetics 108: 471-485.
79. Dooner HK, Belachew A (1989) Transposition Pattern of the Maize Element *Ac* from the *Bz-M2*(*ac*) Allele. Genetics 122: 447-457.
80. Hudson AD, Carpenter R, Coen ES (1990) Phenotypic effects of short-range and aberrant transposition in *Antirrhinum majus*. Plant Molecular Biology 14: 835-844.
81. Dooner HK, Keller J, Harper E, Ralston E (1991) Variable patterns of transposition of the maize element *Activator* in tobacco. Plant Cell 3: 473-482.
82. Kuromori T, Hirayama T, Kiyosue Y, Takabe H, Mizukado S, et al. (2004) A collection of 11 800 single-copy *Ds* transposon insertion lines in *Arabidopsis*. Plant J 37: 897-905.
83. Urasaki A, Asakawa K, Kawakami K (2008) Efficient transposition of the *Tol2* transposable element from a single-copy donor in zebrafish. Proc Natl Acad Sci U S A 105: 19827-19832.
84. Kondrychyn I, Garcia-Lecea M, Emelyanov A, Parinov S, Korzh V (2009) Genome-wide analysis of *Tol2* transposon reintegration in zebrafish. BMC Genomics 10: 418.
85. Zakharenko L, Perepelkina M, Afonnikov D (2009) The importance of transpositions and tecombination to genome instability according *hobo*-element distribution pattern in completely sequenced genome of *Drosophila melanogaster*. Evolutionary Biology: 127-138.
86. Carlson CM, Dupuy AJ, Fritz S, Roberg-Perez KJ, Fletcher CF, et al. (2003) Transposon mutagenesis of the mouse germline. Genetics 165: 243-256.
87. Wang W, Lin C, Lu D, Ning Z, Cox T, et al. (2008) Chromosomal transposition of *PiggyBac* in mouse embryonic stem cells. Proc Natl Acad Sci U S A 105: 9290-9295.
88. Liang Q, Kong J, Stalker J, Bradley A (2009) Chromosomal mobilization and reintegration of *Sleeping Beauty* and *PiggyBac* transposons. Genesis 47: 404-408.
89. Tower J, Karpen GH, Craig N, Spradling AC (1993) Preferential transposition of *Drosophila P* elements to nearby chromosomal sites. Genetics 133: 347-359.
90. Brink RA, Williams E (1973) Mutable *R-navajo* alleles of cyclic origin in maize. Genetics 73: 273-296.
91. Robbins TP, Carpenter R, Coen ES (1989) A chromosome rearrangement suggests that donor and recipient sites are associated during *Tam3* transposition in *Antirrhinum majus*. EMBO J 8: 5-13.
92. Hacker U, Nystedt S, Barmchi MP, Horn C, Wimmer EA (2003) *piggyBac*-based insertional mutagenesis in the presence of stably integrated *P* elements in *Drosophila*. Proc Natl Acad Sci U S A 100: 7720-7725.
93. Thibault ST, Singer MA, Miyazaki WY, Milash B, Dompe NA, et al. (2004) A complementary transposon tool kit for *Drosophila melanogaster* using *P* and *piggyBac*. Nature Genetics 36: 283-287.
94. Ding S, Wu X, Li G, Han M, Zhuang Y, et al. (2005) Efficient transposition of the *piggyBac* (*PB*) transposon in mammalian cells and mice. Cell 122: 473-483.

95. Osborne BI, Corr CA, Prince JP, Hehl R, Tanksley SD, et al. (1991) *Ac* transposition from a T-DNA can generate linked and unlinked clusters of insertions in the tomato genome. Genetics 129: 833-844.
96. Belzile F, Yoder JI (1992) Pattern of somatic transposition in a high copy *Ac* tomato line. Plant J 2: 173-179.
97. Zhang X, Eickbush MT, Eickbush TH (2008) Role of recombination in the long-term retention of transposable elements in rRNA gene loci. Genetics 180: 1617-1626.
98. Jakubczak JL, Burke WD, Eickbush TH (1991) Retrotransposable elements *R1* and *R2* interrupt the rRNA genes of most insects. Proc Natl Acad Sci U S A 88: 3295-3299.
99. Burke WD, Malik HS, Lathe WC, 3rd, Eickbush TH (1998) Are retrotransposons long-term hitchhikers? Nature 392: 141-142.
100. Roiha H, Miller JR, Woods LC, Glover DM (1981) Arrangements and rearrangements of sequences flanking the two types of rDNA insertion in *D. melanogaster*. Nature 290: 749-753.
101. Jakubczak JL, Xiong Y, Eickbush TH (1990) Type I (*R1*) and type II (*R2*) ribosomal DNA insertions of *Drosophila melanogaster* are retrotransposable elements closely related to those of *Bombyx mori*. Journal of Molecular Biology 212: 37-52.
102. Perez-Gonzalez CE, Eickbush TH (2002) Rates of *R1* and *R2* retrotransposition and elimination from the rDNA locus of *Drosophila melanogaster*. Genetics 162: 799-811.
103. Penton EH, Crease TJ (2004) Evolution of the transposable element *Pokey* in the ribosomal DNA of species in the subgenus *Daphnia* (Crustacea: Cladocera). Mol Biol Evol 21: 1727-1739.
104. Cary LC, Goebel M, Corsaro BG, Wang HG, Rosen E, et al. (1989) Transposon mutagenesis of baculoviruses: analysis of *Trichoplusia ni* transposon *IFP2* insertions within the *FP*-locus of nuclear polyhedrosis viruses. Virology 172: 156-169.
105. Wang HH, and M. J. Fraser. (1993) TTAA serves as the target site for *TFP3* lepidopteran insertions in both nuclear polyhedrosis virus and *Trichoplusia ni* genomes. Insect Molecular Biology 1: 109-116.
106. Sarkar A, Sim C, Hong YS, Hogan JR, Fraser MJ, et al. (2003) Molecular evolutionary analysis of the widespread *piggyBac* transposon family and related "domesticated" sequences. Molecular Genetics and Genomics 270: 173-180.
107. Craig NL (1997) Target site selection in transposition. Annu Rev Biochem 66: 437-474.
108. Ketting RF, Fischer SE, Plasterk RH (1997) Target choice determinants of the *Tc1* transposon of *Caenorhabditis elegans*. Nucleic Acids Res 25: 4041-4047.
109. Sullender BW, Crease TJ (2001) The behavior of a *Daphnia pulex* transposable element in cyclically and obligately parthenogenetic populations. J Mol Evol 53: 63-69.
110. Penton EH, Sullender BW, Crease TJ (2002) *Pokey*, a new DNA transposon in *Daphnia* (cladocera: crustacea). J Mol Evol 55: 664-673.
111. Walker EL, Eggleston WB, Demopulos D, Kermicle J, Dellaporta SL (1997) Insertions of a novel class of transposable elements with a strong target site preference at the *r* locus of maize. Genetics 146: 681-693.
112. Zhang X, Feschotte C, Zhang Q, Jiang N, Eggleston WB, et al. (2001) *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 98: 12572-12577.

113. Zhang X, Jiang N, Feschotte C, Wessler SR (2004) *PIF*- and *Pong*-like transposable elements: distribution, evolution and relationship with *Tourist*-like miniature inverted-repeat transposable elements. Genetics 166: 971-986.
114. Liao GC, Rehm EJ, Rubin GM (2000) Insertion site preferences of the *P* transposable element in *Drosophila melanogaster*. Proc Natl Acad Sci U S A 97: 3347-3351.
115. Grotewold E, Athma P, Peterson T (1991) A possible hot spot for *Ac* insertion in the maize *P* gene. Mol Gen Genet 230: 329-331.
116. Dellaporta S, Moreno M (1994) Gene tagging with *Ac*/*Ds* elements in maize. In: Freeling M, Walbot V, editors. The Maize Handbook. Berlin: Springer-Verlag. pp. 219-233.
117. Rabinowicz PD, Schutz K, Dedhia N, Yordan C, Parnell LD, et al. (1999) Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. Nature Genetics 23: 305-308.
118. Fengler K, Allen SM, Li BL, Rafalski A (2007) Distribution of genes, recombination, and repetitive elements in the maize genome. Crop Science 47: S83-S95.
119. Cresse AD, Hulbert SH, Brown WE, Lucas JR, Bennetzen JL (1995) *Mu1*-related transposable elements of maize preferentially insert into low copy number DNA. Genetics 140: 315-324.
120. Settles AM, Latshaw S, McCarty DR (2004) Molecular analysis of high-copy insertion sites in maize. Nucleic Acids Res 32: e54.
121. Fernandes J, Dong Q, Schneider B, Morrow DJ, Nan GL, et al. (2004) Genome-wide mutagenesis of *Zea mays* L. using *RescueMu* transposons. Genome Biol 5: R82.
122. Bureau TE, Wessler SR (1992) *Tourist*: a large family of small inverted repeat elements frequently associated with maize genes. Plant Cell 4: 1283-1294.
123. Bureau TE, Wessler SR (1994) *Stowaway*: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. Plant Cell 6: 907-916.
124. Bureau TE, Wessler SR (1994) Mobile inverted-repeat elements of the *Tourist* family are associated with the genes of many cereal grasses. Proc Natl Acad Sci U S A 91: 1411-1415.
125. Avramova Z, Tikhonov A, Chen M, Bennetzen JL (1998) Matrix attachment regions and structural colinearity in the genomes of two grass species. Nucleic Acids Res 26: 761-767.
126. Bennetzen JL (2000) Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. Plant Cell 12: 1021-1029.
127. Zhang Q, Arbuckle J, Wessler SR (2000) Recent, extensive, and preferential insertion of members of the miniature inverted-repeat transposable element family *Heartbreaker* into genic regions of maize. Proc Natl Acad Sci U S A 97: 1160-1165.
128. Jiang N, Bao Z, Zhang X, Hirochika H, Eddy SR, et al. (2003) An active DNA transposon family in rice. Nature 421: 163-167.
129. Kikuchi K, Terauchi K, Wada M, Hirano HY (2003) The plant MITE *mPing* is mobilized in anther culture. Nature 421: 167-170.
130. Nakazaki T, Okumoto Y, Horibata A, Yamahira S, Teraishi M, et al. (2003) Mobilization of a transposon in the rice genome. Nature 421: 170-172.
131. Naito K, Cho E, Yang G, Campbell MA, Yano K, et al. (2006) Dramatic amplification of a rice transposable element during recent domestication. Proc Natl Acad Sci U S A 103: 17620-17625.

132. Yang G, Zhang F, Hancock CN, Wessler SR (2007) Transposition of the rice miniature inverted repeat transposable element *mPing* in *Arabidopsis thaliana*. Proc Natl Acad Sci U S A 104: 10962-10967.
133. Yang L, Bennetzen JL (2009) Structure-based discovery and description of plant and animal *Helitrons*. Proc Natl Acad Sci U S A 106: 12832-12837.
134. Yang L, Bennetzen JL (2009) Distribution, diversity, evolution, and survival of *Helitrons* in the maize genome. Proc Natl Acad Sci U S A 106: 19922-19927.
135. Sheen FM, Levis RW (1994) Transposition of the LINE-like retrotransposon *TART* to *Drosophila* chromosome termini. Proc Natl Acad Sci U S A 91: 12510-12514.
136. Fuller AM, Cook EG, Kelley KJ, Pardue ML (2010) *Gag* proteins of *Drosophila* telomeric retrotransposons: collaborative targeting to chromosome ends. Genetics 184: 629-636.
137. Celniker SE, Wheeler DA, Kronmiller B, Carlson JW, Halpern A, et al. (2002) Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. Genome Biol 3: RESEARCH0079.
138. Biessmann H, Valgeirsdottir K, Lofsky A, Chin C, Ginther B, et al. (1992) *HeT-A*, a transposable element specifically involved in "healing" broken chromosome ends in *Drosophila melanogaster*. Mol Cell Biol 12: 3910-3918.
139. Melnikova L, Georgiev P (2005) *Drosophila* telomeres: the non-telomerase alternative. Chromosome Res 13: 431-441.
140. Savitsky M, Kwon D, Georgiev P, Kalmykova A, Gvozdev V (2006) Telomere elongation is under the control of the RNAi-based mechanism in the *Drosophila* germline. Genes Dev 20: 345-354.
141. Rashkova S, Athanasiadis A, Pardue ML (2003) Intracellular targeting of *Gag* proteins of the *Drosophila* telomeric retrotransposons. J Virol 77: 6376-6384.
142. Rashkova S, Karam SE, Kellum R, Pardue ML (2002) *Gag* proteins of the two *Drosophila* telomeric retrotransposons are targeted to chromosome ends. J Cell Biol 159: 397-402.
143. Danilevskaya O, Lofsky A, Kurenova EV, Pardue ML (1993) The Y chromosome of *Drosophila melanogaster* contains a distinctive subclass of *Het-A*-related repeats. Genetics 134: 531-543.
144. Agudo M, Losada A, Abad JP, Pimpinelli S, Ripoll P, et al. (1999) Centromeres from telomeres? The centromeric region of the Y chromosome of *Drosophila melanogaster* contains a tandem array of telomeric *HeT-A*- and *TART*-related sequences. Nucleic Acids Res 27: 3318-3324.
145. Jiang J, Nasuda S, Dong F, Scherrer CW, Woo SS, et al. (1996) A conserved repetitive DNA element located in the centromeres of cereal chromosomes. Proc Natl Acad Sci U S A 93: 14210-14213.
146. Ananiev EV, Phillips RL, Rines HW (1998) Chromosome-specific molecular organization of maize (*Zea mays* L.) centromeric regions. Proc Natl Acad Sci U S A 95: 13073-13078.
147. Miller JT, Dong F, Jackson SA, Song J, Jiang J (1998) Retrotransposon-related DNA sequences in the centromeres of grass chromosomes. Genetics 150: 1615-1623.
148. Presting GG, Malysheva L, Fuchs J, Schubert I (1998) A *Ty3*/*gypsy* retrotransposon-like sequence localizes to the centromeric regions of cereal chromosomes. Plant J 16: 721-728.
149. Cheng Z, Dong F, Langdon T, Ouyang S, Buell CR, et al. (2002) Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. Plant Cell 14: 1691-1704.

150. Zhong CX, Marshall JB, Topp C, Mroczek R, Kato A, et al. (2002) Centromeric retroelements and satellites interact with maize kinetochore protein CENH3. Plant Cell 14: 2825-2836.
151. Nagaki K, Song J, Stupar RM, Parokonny AS, Yuan Q, et al. (2003) Molecular and cytological analyses of large tracks of centromeric DNA reveal the structure and evolutionary dynamics of maize centromeres. Genetics 163: 759-770.
152. Nagaki K, Neumann P, Zhang D, Ouyang S, Buell CR, et al. (2005) Structure, divergence, and distribution of the *CRR* centromeric retrotransposon family in rice. Mol Biol Evol 22: 845-855.
153. Wolfgruber TK, Sharma A, Schneider KL, Albert PS, Koo DH, et al. (2009) Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic loci shaped primarily by retrotransposons. PLoS Genet 5: e1000743.
154. Kordis D (2005) A genomic perspective on the chromodomain-containing retrotransposons: Chromoviruses. Gene 347: 161-173.
155. Nitta N, Farman M, Leong S (1997) Genome organization of *Magnaporthe grisea*: integration of genetic maps, clustering of transposable elements and identification of genome duplications and rearrangements. Theoretical and Applied Genetics 95: 20-32.
156. Thon MR, Martin SL, Goff S, Wing RA, Dean RA (2004) BAC end sequences and a physical map reveal transposable element content and clustering patterns in the genome of *Magnaporthe grisea*. Fungal Genet Biol 41: 657-666.
157. Dean RA, Talbot NJ, Ebbole DJ, Farman ML, Mitchell TK, et al. (2005) The genome sequence of the rice blast fungus *Magnaporthe grisea*. Nature 434: 980-986.
158. Nakayashiki H, Awa T, Tosa Y, Mayama S (2005) The C-terminal chromodomain-like module in the integrase domain is crucial for high transposition efficiency of the retrotransposon *MAGGY*. FEBS Lett 579: 488-492.
159. Behrens R, Hayles J, Nurse P (2000) Fission yeast retrotransposon *Tf1* integration is targeted to 5' ends of open reading frames. Nucleic Acids Res 28: 4709-4716.
160. Singleton TL, Levin HL (2002) A long terminal repeat retrotransposon of fission yeast has strong preferences for specific sites of insertion. Eukaryot Cell 1: 44-55.
161. Bowen NJ, Jordan IK, Epstein JA, Wood V, Levin HL (2003) Retrotransposons and their recognition of pol II promoters: a comprehensive survey of the transposable elements from the complete genome sequence of *Schizosaccharomyces pombe*. Genome Res 13: 1984-1997.
162. Leem YE, Ripmaster TL, Kelly FD, Ebina H, Heincelman ME, et al. (2008) Retrotransposon *Tf1* is targeted to Pol II promoters by transcription activators. Mol Cell 30: 98-107.
163. Guo Y, Levin HL (2010) High-throughput sequencing of retrotransposon integration provides a saturated profile of target activity in *Schizosaccharomyces pombe*. Genome Res 20: 239-248.
164. Chatterjee AG, Leem YE, Kelly FD, Levin HL (2009) The chromodomain of *Tf1* integrase promotes binding to cDNA and mediates target site selection. J Virol 83: 2675-2685.
165. Hizi A, Levin HL (2005) The integrase of the long terminal repeat-retrotransposon *Tf1* has a chromodomain that modulates integrase activities. J Biol Chem 280: 39086-39094.
166. Chalker DL, Sandmeyer SB (1992) *Ty3* integrates within the region of RNA polymerase III transcription initiation. Genes Dev 6: 117-128.

167. Kim JM, Vanguri S, Boeke JD, Gabriel A, Voytas DF (1998) Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. Genome Res 8: 464-478.
168. Yieh L, Kassavetis G, Geiduschek EP, Sandmeyer SB (2000) The *Brf* and TATA-binding protein subunits of the RNA polymerase III transcription factor IIIB mediate position-specific integration of the *gypsy*-like element, *Ty3*. J Biol Chem 275: 29800-29807.
169. Kirchner J, Connolly CM, Sandmeyer SB (1995) Requirement of RNA polymerase III transcription factors for *in vitro* position-specific integration of a retroviruslike element. Science 267: 1488-1491.
170. Bachman N, Gelbart ME, Tsukiyama T, Boeke JD (2005) TFIIIB subunit Bdp1p is required for periodic integration of the *Ty1* retrotransposon and targeting of Isw2p to *S. cerevisiae* tDNAs. Genes Dev 19: 955-964.
171. Devine SE, Boeke JD (1996) Integration of the yeast retrotransposon Ty1 is targeted to regions upstream of genes transcribed by RNA polymerase III. Genes Dev 10: 620-633.
172. Hani J, Feldmann H (1998) tRNA genes and retroelements in the yeast genome. Nucleic Acids Res 26: 689-696.
173. Ji H, Moore DP, Blomberg MA, Braiterman LT, Voytas DF, et al. (1993) Hotspots for unselected Ty1 transposition events on yeast chromosome III are near tRNA genes and LTR sequences. Cell 73: 1007-1018.
174. Bolton E, Boeke J (2003) Transcriptional interactions between yeast tRNA genes, flanking genes and Ty elements: a genomic point of view. Genome Research 13: 254.
175. Mou Z, Kenny AE, Curcio MJ (2006) *Hos2* and *Set3* promote integration of Ty1 retrotransposons at tRNA genes in *Saccharomyces cerevisiae.* Genetics 172: 2157-2167.
176. Zou S, Ke N, Kim JM, Voytas DF (1996) The *Saccharomyces* retrotransposon *Ty5* integrates preferentially into regions of silent chromatin at the telomeres and mating loci. Genes Dev 10: 634-645.
177. Zou S, Voytas DF (1997) Silent chromatin determines target preference of the *Saccharomyces* retrotransposon *Ty5*. Proc Natl Acad Sci U S A 94: 7412-7416.
178. Zhu Y, Zou S, Wright DA, Voytas DF (1999) Tagging chromatin with retrotransposons: target specificity of the *Saccharomyces Ty5* retrotransposon changes with the chromosomal localization of *Sir3p* and *Sir4p*. Genes Dev 13: 2738-2749.
179. Gai X, Voytas DF (1998) A single amino acid change in the yeast retrotransposon *Ty5* abolishes targeting to silent chromatin. Mol Cell 1: 1051-1055.
180. Xie W, Gai X, Zhu Y, Zappulla DC, Sternglanz R, et al. (2001) Targeting of the yeast *Ty5* retrotransposon to silent chromatin is mediated by interactions between integrase and Sir4p. Mol Cell Biol 21: 6606-6614.
181. Zhu Y, Dai J, Fuerst PG, Voytas DF (2003) Controlling integration specificity of a yeast retrotransposon. Proc Natl Acad Sci U S A 100: 5891-5895.
182. Brady TL, Schmidt CL, Voytas DF (2008) Targeting integration of the *Saccharomyces Ty5* retrotransposon. Methods Mol Biol 435: 153-163.
183. Brady TL, Fuerst PG, Dick RA, Schmidt C, Voytas DF (2008) Retrotransposon target site selection by imitation of a cellular protein. Mol Cell Biol 28: 1230-1239.
184. Dai J, Xie W, Brady TL, Gao J, Voytas DF (2007) Phosphorylation regulates integration of the yeast *Ty5* retrotransposon into heterochromatin. Mol Cell 27: 289-299.
185. Boeke JD, Devine SE (1998) Yeast retrotransposons: finding a nice quiet neighborhood. Cell 93: 1087-1089.

186. Bushman FD (2003) Targeting survival: integration site selection by retroviruses and LTR-retrotransposons. Cell 115: 135-138.
187. Shirasu K, Schulman AH, Lahaye T, Schulze-Lefert P (2000) A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. Genome Res 10: 908-915.
188. Roeder GS, Fink GR (1980) DNA rearrangements associated with a transposable element in yeast. Cell 21: 239-249.
189. Ma J, Bennetzen JL (2006) Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. Proc Natl Acad Sci U S A 103: 383-388.
190. Tian Z, Rizzon C, Du J, Zhu L, Bennetzen JL, et al. (2009) Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? Genome Res 19: 2221-2230.
191. Belshaw R, Watson J, Katzourakis A, Howe A, Woolven-Allen J, et al. (2007) Rate of recombinational deletion among human endogenous retroviruses. J Virol 81: 9437-9442.
192. Katzourakis A, Pereira V, Tristem M (2007) Effects of recombination rate on human endogenous retrovirus fixation and persistence. J Virol 81: 10712-10717.
193. Gaut BS, Wright SI, Rizzon C, Dvorak J, Anderson LK (2007) Recombination: an underappreciated factor in the evolution of plant genomes. Nat Rev Genet 8: 77-84.
194. SanMiguel P, Vitte C (2008) The LTR-retrotransposons of maize. In: Bennetzen JL, Hake SC, editors. Handbook of Maize: Genetics and Genomics Springer.
195. Meyers BC, Tingey SV, Morgante M (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res 11: 1660-1676.
196. Messing J, Bharti AK, Karlowski WM, Gundlach H, Kim HR, et al. (2004) Sequence composition and genome organization of maize. Proc Natl Acad Sci U S A 101: 14349-14354.
197. Mroczek RJ, Dawe RK (2003) Distribution of retroelements in centromeres and neocentromeres of maize. Genetics 165: 809-819.
198. Haberer G, Young S, Bharti AK, Gundlach H, Raymond C, et al. (2005) Structure and architecture of the maize genome. Plant Physiol 139: 1612-1624.
199. Gorinsek B, Gubensek F, Kordis D (2004) Evolutionary genomics of chromoviruses in eukaryotes. Mol Biol Evol 21: 781-798.
200. Gorinsek B, Gubensek F, Kordis D (2005) Phylogenomic analysis of chromoviruses. Cytogenet Genome Res 110: 543-552.
201. Institute AG (2009).
202. McCarthy EM, McDonald JF (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics 19: 362-367.
203. Kalyanaraman A, Aluru S (2006) Efficient algorithms and software for detection of full-length LTR retrotransposons. J Bioinform Comput Biol 4: 197-216.
204. Rho M, Choi JH, Kim S, Lynch M, Tang H (2007) *De novo* identification of LTR retrotransposons in eukaryotic genomes. BMC Genomics 8: 90.
205. Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35: W265-268.
206. Ellinghaus D, Kurtz S, Willhoeft U (2008) LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. BMC Bioinformatics 9: 18.
207. Jurka J, Klonowski P, Dagman V, Pelton P (1996) CENSOR--a program for identification and elimination of repetitive elements from DNA sequences. Comput Chem 20: 119-121.

208. Kohany O, Gentles AJ, Hankus L, Jurka J (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 7: 474.
209. Smit A, Hubley R, Green P (1996-2004) RepeatMasker Open-3.0.
210. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.
211. Gundlach H (2009) MIPS Repeat Element Database (mips-REdat) and Catalog (mips-REcat).
212. Wicker T (2009) TREP, the Triticeae Repeat Sequence Database.
213. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110: 462-467.
214. Buisine N, Quesneville H, Colot V (2008) Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. Genomics 91: 467-475.
215. Wicker T, Narechania A, Sabot F, Stein J, Vu GT, et al. (2008) Low-pass shotgun sequencing of the barley genome facilitates rapid identification of genes, conserved non-coding sequences and novel repeats. BMC Genomics 9: 518.
216. Kurtz S, Narechania A, Stein JC, Ware D (2008) A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics 9: 517.
217. Estill JC, Bennetzen JL (2009) The DAWGPAWS pipeline for the annotation of genes and transposable elements in plant genomes. Plant Methods 5: 8.
218. Dongen SMv (2000) Graph Clustering by Flow Simulation. Amsterdam: University of Utrecht. 169 p.
219. Frey BJ, Dueck D (2007) Clustering by passing messages between data points. Science 315: 972-976.
220. Estill JC, Baucom RS, Bennetzen JL (2009) RepMiner : http://repminer.sourceforge.net/.
221. Lysenko O, Sneath PH (1959) The use of models in bacterial classification. J Gen Microbiol 20: 284-290.
222. Sokal RR, Sneath PHA (1963) Principles of numerical taxonomy. San Francisco,: W. H. Freeman. xvi, 359 p. p.
223. Jünger M, Mutzel P (2004) Graph drawing software. Berlin: Springer. xii, 378 p. p.
224. Auber D (2004) Tulip - A Huge Graph Visualization Framework. In: Jünger M, Mutzel P, editors. Graph Drawing Software. Berlin: Springer. pp. 105-126.
225. Herman I, Melançon G, Marshall. MS (2000) Graph visualization and navigation in information visualisation: a survey. IEEE Transactions on Visualization and Computer Graphics 6: 24-43,.
226. Jünger M, Mutzel P (2004) Technical Foundations. In: Jünger M, Mutzel P, editors. Graph Drawing Software. Berlin: Springer. pp. 9-53.
227. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498-2504.
228. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30: 1575-1584.
229. Brohee S, van Helden J (2006) Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics 7: 488.

230. Vlasblom J, Wodak SJ (2009) Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. BMC Bioinformatics 10: 99.
231. Team RDC (2005) R: A language and environment for statistical computing. 1.8.1 ed. Vienna, Austria: R Foundation For Statistical Computing.
232. Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, et al. (2002) The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. Genome Biol 3: RESEARCH0084.
233. Terrinoni A, Franco CD, Dimitri P, Junakovic N (1997) Intragenomic distribution and stability of transposable elements in euchromatin and heterochromatin of *Drosophila melanogaster*: non-LTR retrotransposon. J Mol Evol 45: 145-153.
234. Berezikov E, Bucheton A, Busseau I (2000) A search for reverse transcriptase-coding sequences reveals new non-LTR retrotransposons in the genome of *Drosophila melanogaster*. Genome Biol 1: RESEARCH0012.
235. Biessmann H, Prasad S, Walter MF, Mason JM (2005) Euchromatic and heterochromatic domains at *Drosophila* telomeres. Biochem Cell Biol 83: 477-485.
236. Cost G, Boeke J (1998) Targeting of human retrotransposon integration Is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. Biochemistry 37: 18081-18093.
237. Takahashi H, Fujiwara H (2002) Transplantation of target site specificity by swapping the endonuclease domains of two LINEs. EMBO J 21: 408-417.
238. Jurka J (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc Natl Acad Sci U S A 94: 1872-1877.
239. Gentles AJ, Kohany O, Jurka J (2005) Evolutionary diversity and potential recombinogenic role of integration targets of non-LTR retrotransposons. Mol Biol Evol 22: 1983-1991.
240. Sela N, Mersch B, Gal-Mark N, Lev-Maor G, Hotz-Wagenblatt A, et al. (2007) Comparative analysis of transposed element insertion within human and mouse genomes reveals *Alu*'s unique role in shaping the human transcriptome. Genome Biol 8: R127.
241. Cost GJ, Golding A, Schlissel MS, Boeke JD (2001) Target DNA chromatinization modulates nicking by L1 endonuclease. Nucleic Acids Res 29: 573-577.
242. Dewannieux M, Esnault C, Heidmann T (2003) LINE-mediated retrotransposition of marked *Alu* sequences. Nature Genetics 35: 41-48.
243. Toda Y, Saito R, Tomita M (2000) Characteristic sequence pattern in the 5- to 20-bp upstream region of primate *Alu* elements. J Mol Evol 50: 232-237.
244. Batzer MA, Deininger PL (2002) *Alu* repeats and human genomic diversity. Nat Rev Genet 3: 370-379.
245. Yang S, Smit AF, Schwartz S, Chiaromonte F, Roskin KM, et al. (2004) Patterns of insertions and their covariation with substitutions in the rat, mouse, and human genomes. Genome Res 14: 517-527.
246. Jurka J, Krnjajic M, Kapitonov VV, Stenger JE, Kokhanyy O (2002) Active *Alu* elements are passed primarily through paternal germlines. Theor Popul Biol 61: 519-530.
247. Jurka J, Kohany O, Pavlicek A, Kapitonov VV, Jurka MV (2004) Duplication, coclustering, and selection of human *Alu* retrotransposons. Proc Natl Acad Sci U S A 101: 1268-1272.
248. Jurka J (2004) Evolutionary impact of human *Alu* repetitive elements. Curr Opin Genet Dev 14: 603-608.
249. Ono T, Kondoh Y, Kagiyama N, Sonta S, Yoshida MC (2001) Genomic organization and chromosomal distribution of rat *ID* elements. Genes Genet Syst 76: 213-220.

250. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, et al. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428: 493-521.
251. Vigdal TJ, Kaufman CD, Izsvak Z, Voytas DF, Ivics Z (2002) Common physical properties of DNA affecting target site selection of sleeping beauty and other *Tc1*/*mariner* transposable elements. Journal of Molecular Biology 323: 441-452.
252. Ivics Z, Kaufman CD, Zayed H, Miskey C, Walisko O, et al. (2004) The *Sleeping Beauty* transposable element: evolution, regulation and genetic applications. Curr Issues Mol Biol 6: 43-55.
253. Yant SR, Wu X, Huang Y, Garrison B, Burgess SM, et al. (2005) High-resolution genome-wide mapping of transposon integration in mammals. Mol Cell Biol 25: 2085-2094.
254. Ito T, Motohashi R, Kuromori T, Noutoshi Y, Seki M, et al. (2005) A resource of 5,814 *dissociation* transposon-tagged and sequence-indexed lines of *Arabidopsis* transposed from start loci on chromosome 5. Plant Cell Physiol 46: 1149-1153.
255. Vollbrecht E, Duvick J, Schares JP, Ahern KR, Deewatthanawong P, et al. (2010) Genome-wide distribution of transposed *Dissociation* elements in maize. Plant Cell 22: 1667-1685.
256. Kolesnik T, Szeverenyi I, Bachmann D, Kumar CS, Jiang S, et al. (2004) Establishing an efficient *Ac/Ds* tagging system in rice: large-scale analysis of *Ds* flanking sequences. Plant J 37: 301-314.
257. Enoki H, Izawa T, Kawahara M, Komatsu M, Koh S, et al. (1999) *Ac* as a tool for the functional genomics of rice. The Plant Journal 19: 605-613.
258. Kunze R, Weil C (2002) The *hAT* and *CACTA* superfamilies of plant transposons. Mobile DNA II 2: 565-610.
259. Kawakami K (2007) *Tol2*: a versatile gene transfer vector in vertebrates. Genome Biol 8 Suppl 1: S7.
260. Streck RD, Macgaffey JE, Beckendorf SK (1986) The structure of *hobo* transposable elements and their insertion sites. EMBO J 5: 3615-3623.
261. Saville KJ, Warren WD, Atkinson PW, O'Brochta DA (1999) Integration specificity of the *hobo* element of *Drosophila melanogaster* is dependent on sequences flanking the integration site. Genetica 105: 133-147.
262. O'Brochta DA, Warren WD, Saville KJ, Atkinson PW (1994) Interplasmid transposition of *Drosophila hobo* elements in non-drosophilid insects. Mol Gen Genet 244: 9-14.
263. Newfeld SJ, Takaesu NT (1999) Local transposition of a *hobo* element within the *decapentaplegic* locus of *Drosophila*. Genetics 151: 177-187.
264. Hoogland C, Biemont C (1996) Chromosomal distribution of transposable elements in *Drosophila melanogaster*: test of the ectopic recombination model for maintenance of insertion site number. Genetics 144: 197-204.
265. Uchiyama T, Fujino K, Ogawa T, Wakatsuki A, Kishima Y, et al. (2009) Stable transcription activities dependent on an orientation of *Tam3* transposon insertions into *Antirrhinum* and yeast promoters occur only within chromatin. Plant Physiol 151: 1557-1569.
266. Lister C, Martin C (1989) Molecular analysis of a transposon-induced deletion of the *nivea* locus in *Antirrhinum majus*. Genetics 123: 417-425.
267. Lisch D, Jiang N (2009) *Mutator* and *MULE* transposons. Handbook of Maize: 277-306.

268. Lisch D, Chomet P, Freeling M (1995) Genetic characterization of the *Mutator* system in maize: behavior and regulation of *Mu* transposons in a minimal line. Genetics 139: 1777-1796.
269. O'Hare K, Rubin GM (1983) Structures of *P* transposable elements and their sites of insertion and excision in the *Drosophila melanogaster* genome. Cell 34: 25-35.
270. O'Hare K, Driver A, McGrath S, Johnson-Schiltz DM (1992) Distribution and structure of cloned *P* elements from the *Drosophila melanogaster* P strain pi 2. Genet Res 60: 33-41.
271. Spradling AC, Stern DM, Kiss I, Roote J, Laverty T, et al. (1995) Gene disruptions using *P* transposable elements: an integral component of the *Drosophila* genome project. Proc Natl Acad Sci U S A 92: 10824-10830.
272. Zhang P, Spradling AC (1993) Efficient and dispersed local *P* element transposition from *Drosophila* females. Genetics 133: 361-373.
273. Berg CA, Spradling AC (1991) Studies on the rate and site-specificity of *P* element transposition. Genetics 127: 515-524.
274. Hancock CN, Zhang F, Wessler SR (2010) Transposition of the *Tourist*-MITE *mPing* in yeast: an assay that retains key features of catalysis by the class 2 *PIF*/*Harbinger* superfamily. Mobile DNA 1.
275. Kumar CS, Wing RA, Sundaresan V (2005) Efficient insertional mutagenesis in rice using the maize *En*/*Spm* elements. Plant J 44: 879-892.
276. Gierl A, Saedler H (1989) The *En*/*Spm* transposable element of *Zea mays*. Plant Molecular Biology 13: 261-266.
277. Cardon GH, Frey M, Saedler H, Gierl A (1993) Definition and characterization of an artificial En/Spm-based transposon tagging system in transgenic tobacco. Plant Molecular Biology 23: 157-178.
278. Miura A, Yonebayashi S, Watanabe K, Toyama T, Shimada H, et al. (2001) Mobilization of transposons by a mutation abolishing full DNA methylation in *Arabidopsis*. Nature 411: 212-214.
279. Miura A, Kato M, Watanabe K, Kawabe A, Kotani H, et al. (2004) Genomic localization of endogenous mobile *CACTA* family transposons in natural variants of *Arabidopsis thaliana*. Molecular Genetics and Genomics 270: 524-532.
280. Kato M, Takashima K, Kakutani T (2004) Epigenetic control of *CACTA* transposon mobility in *Arabidopsis thaliana*. Genetics 168: 961-969.
281. Wang GD, Tian PF, Cheng ZK, Wu G, Jiang JM, et al. (2003) Genomic characterization of *Rim2*/*Hipa* elements reveals a *CACTA*-like transposon superfamily with unique features in the rice genome. Molecular Genetics and Genomics 270: 234-242.
282. Tian PF (2006) Progress in plant *CACTA* elements. Yi Chuan Xue Bao 33: 765-774.
283. Tikhonov AP, Bennetzen JL, Avramova ZV (2000) Structural domains and matrix attachment regions along colinear chromosomal segments of maize and sorghum. Plant Cell 12: 249-264.
284. Casa AM, Brouwer C, Nagel A, Wang L, Zhang Q, et al. (2000) Inaugural article: the MITE family heartbreaker (*Hbr*): molecular markers in maize. Proc Natl Acad Sci U S A 97: 10083-10089.
285. Casacuberta E, Casacuberta JM, Puigdomenech P, Monfort A (1998) Presence of miniature inverted-repeat transposable elements (MITEs) in the genome of *Arabidopsis thaliana*: characterisation of the *Emigrant* family of elements. Plant J 16: 79-85.

286. Le QH, Wright S, Yu Z, Bureau T (2000) Transposon diversity in *Arabidopsis thaliana*. Proc Natl Acad Sci U S A 97: 7376-7381.
287. Santiago N, Herraiz C, Goni JR, Messeguer X, Casacuberta JM (2002) Genome-wide analysis of the *Emigrant* family of MITEs of *Arabidopsis thaliana*. Mol Biol Evol 19: 2285-2293.
288. Tikhonov AP, SanMiguel PJ, Nakajima Y, Gorenstein NM, Bennetzen JL, et al. (1999) Colinearity and its exceptions in orthologous *adh* regions of maize and sorghum. Proc Natl Acad Sci U S A 96: 7409-7414.

CHAPTER 2

THE DAWGPAWS PIPELINE FOR THE ANNOTATION OF GENES AND TRANSPOSABLE ELEMENTS IN PLANT GENOMES[1]

[1] Estill, J.C. and J.L. Bennetzen. 2009. *Plant Methods*. 5:8.
Reprinted here with permission of publisher.

Abstract

Background

High quality annotation of the genes and transposable elements in complex genomes requires a human-curated integration of multiple sources of computational evidence. These evidences include results from a diversity of *ab initio* prediction programs as well as homology-based searches. Most of these programs operate on a single contiguous sequence at a time, and the results are generated in a diverse array of readable formats that must be translated to a standardized file format. These translated results must then be concatenated into a single source, and then presented in an integrated form for human curation.

Results

We have designed, implemented, and assessed a Perl-based workflow named DAWGPAWS for the generation of computational results for human curation of the genes and transposable elements in plant genomes. The use of DAWGPAWS was found to accelerate annotation of 80-200 kb wheat DNA inserts in bacterial artificial chromosome (BAC) vectors by approximately twenty-fold and to also significantly improve the quality of the annotation in terms of completeness and accuracy.

Conclusion

The DAWGPAWS genome annotation pipeline fills an important need in the annotation of plant genomes by generating computational evidences in a high throughput manner, translating these results to a common file format, and facilitating the human curation of these computational results. We have verified the value of DAWGPAWS by using this pipeline to annotate the genes and transposable elements in 220 BAC insertions from the hexaploid wheat genome (*Triticum aestivum* L.). DAWGPAWS can be applied to annotation efforts in other plant

genomes with minor modifications of program-specific configuration files, and the modular design of the workflow facilitates integration into existing pipelines.

Background

Genomic sequence assemblies are rapidly being published for a great number of species [1,2]. The sequence data used to produce genome assemblies are being generated at ever-increasing rates for reduced costs [3], indicating that the genomes of many more plant species will be *de novo* sequenced in coming years. The relative value of these sequencing efforts is a direct function of the accuracy of the annotation of the resultant sequence assemblies. Genome annotation seeks to delineate the sequence features that occur on the genome, thereby permitting definition of the biological processes responsible for these features [4]. In plants, the sequence characteristics that are most critical to our interpretation of gene function and genome evolution include both genes and transposable elements (TEs) [5,6].

Identification of the genes that have been uncovered in assembled genome sequence data can utilize evidence from both *ab initio* gene annotation programs as well as sequence similarity searches against databases of previously identified proteins and expressed RNA [4,7,8]. The *ab initio* gene finding programs derive full gene models from DNA sequence data based solely on knowledge of the sequence features associated with protein coding domains. Sequence alignments can refine the exon-intron boundaries of these models and provide evidence that computationally predicted genes are actually transcribed *in vivo*. Existing software can automatically synthesize these data to derive combined evidence gene models [9,10].

While this combination of *ab initio* and homology-based approaches have been used to accurately annotate genes in a number of eukaryotic genomes, plant genome annotation efforts

cannot focus solely on the annotation of genes due to the risk of conflating genes with transposable elements [11]. Many TEs contain open reading frames (ORFs) that generate the proteins required for TE transposition. The *ab initio* gene annotation programs will often annotate these TE ORFs as genes. Since most TE genes are expressed and represented in cDNA libraries, homology-based searches will indicate that these ORFs are transcribed and they thus may be considered legitimate gene predictions. Simply removing the high-copy-number candidate genes does not alleviate this problem because some true gene families are highly abundant while not all transposable elements are highly repetitive [12]. These erroneous gene annotations are especially problematic in plant genomes where transposable elements make up the majority of sequenced genome space. Since these false positive gene predictions cannot be mitigated by gene prediction methods alone, plant genome annotation must directly annotate TEs in order to remove them from the gene candidate list.

Similar to the prediction of genes, accurate identification of the TEs in genomic sequence data combines homology-based searches and *ab initio* results [13,14,15]. Tools for *ab initio* transposable element discovery can exploit the fact that many families of TEs occur in high copy number within a host genome [16,17,18], or they can utilize diagnostic structural features such as tandem inverted repeats (TIRs) or long terminal repeats (LTRs) that delineate an individual TE insertion [19,20,21]. Homology-based searches of transposable elements are facilitated by specialized tools [22,23,24,25] that make use of databases of previously identified TEs [26,27,28,29] or leverage repetitive data from the sequenced genome [30,31,32].

The gold standard of genome annotation is the integration and curation of multiple computational results by a knowledgeable biologist [11]. This approach has been advocated for the structural annotation of genes [4,11], as well as transposable elements [33]. A limitation of

the manually-curated multiple-evidences approach is that the process requires the combination of computational results from a disparate set of independent annotation programs. The output of this software has been designed to maximize readability by humans and not to facilitate integration of results across programs. Furthermore, these tools are often designed to work on a single contiguous sequence (contig) at a time, while many annotation efforts require the generation of computational results for thousands of assembled contigs. Computational workflow suites that seek to aid in plant genome annotation must therefore overcome these limitations while facilitating the human interpretation of the computational results contributing to a biological annotation.

Here, we introduce an annotation suite that allows for computational evidences to be generated in an automated fashion, integrates the results from multiple programs and facilitates the human curation of these computational results. This suite was designed to assist a Distributed Annotation Working Group (DAWG) approach for a Pipeline to Annotate Wheat Sequences (PAWS), and we hereafter refer to this effort as DAWGPAWS.

Implementation

The DAWGPAWS workflow (Figure 2.1) is distributed as a suite of individual command line interface (CLI) programs written in the Perl programming language. Generally, each program is tailored for an individual step in the annotation process, and it can be used independently of all other programs in the package. This allows users to design an individualized annotation pipeline by selecting those computational components that are most appropriate to their annotation efforts. This modular design also facilitates using DAWGPAWS in a high throughput cluster-computing framework. Large-scale annotation jobs can be split across

compute nodes by contigs being annotated as well as by the computational process used to generate computational results.

A common thread to each component of the DAWGPAWS package is that computational evidences are translated from the native annotation program output into the standard general feature format (GFF) [34]. The GFF file format facilitates integration of multiple computational results. This format can be directly curated by any biologist using standard sequence curation and visualization tools such as Apollo [35], Artemis [36], GBrowse [37], the UCSC genome browser [38] or the Ensembl Genome Browser [39]. The GFF files also provide a standard format for loading annotation results to relational database schemas such as BioSQL [40] or CHADO [41].

One of the main sets of scripts in the DAWGPAWS package is the batch run program set (Table 2.1). All of these scripts are designed to run individual annotation programs in a high throughput batch mode. They take as their input a directory of sequence files that are to be annotated and a configuration file describing the sets of parameters to use for each sequence file. The output of these batch scripts includes the original output from the annotation program as well as this output translated to the GFF format. The resulting files are stored in a predefined directory structure that allows users to quickly locate the original annotation results as well as the GFF copy. These batch programs exist for both gene and TE annotation results. The *ab initio* gene annotation programs supported by these scripts include EuGène [9], GeneID [42], GeneMark.hmm [43], and Genscan [44]. The *ab initio* TE annotation programs that can be run in batch mode are Find_LTR [45], LTR_STRUC [20], LTR_FINDER [21], LTR_seq [46], FINDMITE [19], and Tandem Repeats Finder [47]. Batch mode scripts also support TE annotation using HMMER [48], NCBI-BLAST [49], RepeatMasker [22], and TEnest [24]. The

full set of gene and TE annotation programs that can be run in batch mode are summarized in Table 2.1.

In addition to the batch run programs, scripts that convert an individual annotation program output to GFF are also available (Table 2.2). These programs allow an existing annotation result to be specified, or they can take advantage of UNIX standard streams. If an input file is not specified, the conversion scripts will expect input from the standard input stream. Likewise, if the output path is not specified, these programs will write the output to a standard output stream. Accepting standard input and output streams facilitates using these programs as supplements to an existing workflow. For example, data can be piped directly from the output stream of an annotation program to a DAWGPAWS converter, and then piped on to a parser that loads the GFF formatted result to a database. These conversion programs provide the ability to support conversion of output from programs such as FGENESH [50] [51] and RepSeek [52] that are not supported by batch scripts in DAWGPAWS.

The DAWGPAWS suite also includes specialized tools for TE annotation. For identification of the highly repetitive regions of a contig, the seq_oligocount.pl program can count the occurrence of oligomers in the query sequence against an index of random shotgun sequences. This program generates all oligomers of length k from the query sequence, and uses the vmatch program [53] to determine the number of these k-mers that occur in a random shotgun sequence data set generated by mkvtree [53]. The output of this program is a GFF file indicating the count of these k-mers in the shotgun sequence dataset. These results may be used to identify the mathematically defined repeats in the query sequence, as well as provides a means to visualize low-copy-number runs in the query sequence [54].

In addition to the gene and TE annotation-specific scripts included in the DAWGPAWS package, helper applications are also included (Table 2.3). These CLI programs fulfill needs that occur when generating annotation results. They allow for file conversion such as the conversion of GFF to game.xml format or the conversion of a lowercase masked sequence file to a hard masked sequence file. They also prepare the sequence files for annotation by shortening FASTA headers as required by some programs, or by splitting a single FASTA file containing multiple records into multiple FASTA files containing single record files. The ability to generate Euler Diagrams is also supported via the vennseq.pl conversion script that formats GFF file data for input into the VennMaster program [55].

A CLI interface was selected for DAWGPAWS to facilitate the use of our applications in a cluster-computing environment, and to provide stability in program interface across multiple operating systems. While command line interface programs may be daunting to some users, every effort has been made to simplify their use. All of the CLI programs included in the DAWGPAWS suite follow consistent protocols for command line options (Table 2.4). Help files or full program manuals are available from the command line within all programs by invoking the --help or --man options. These application manuals are also available in HTML form on the DAWGPAWS website along with a general program manual describing the installation and use of a local implementation of the DAWGPAWS package [56]. This documentation is also included in the downloadable release of DAWGPAWS.

Results and Discussion

The computational annotation results generated by DAWGPAWS can be directly imported into any genome annotation program that supports GFF. We have used the Apollo

program [35] to visualize and curate our results for genes and transposable elements in the wheat genome (Figure 2.2). Since the game xml file format is the most stable way to store annotation results in Apollo, it is generally useful to first convert GFF files to the game xml format before beginning curation of computational results. The visual display of computational results in Apollo is modified by a tiers configuration file. This file controls how and where individual computational and annotation results are drawn on the annotation pane. The tiers file used in these annotation efforts is included in the DAWGPAWS download package, and it can serve as a starting point for generating individualized tier files for other plant annotation efforts. As an alternative to Apollo, it is also possible to curate computational results using the Artemis sequence visualization program [36].

The GBrowse package [37] can also visualize GFF formatted annotations, and has proven to be a useful method for visualizing TE results. GBrowse makes use of core images called glyphs that are used to draw sequence features along a genome. The available glyphs in GBrowse can be supplemented by writing additional Perl modules, and we have generated TE glyphs that allow visualization of the biologically relevant features of TEs. GBrowse also has the capability to draw histograms along the sequence contigs. GBrowse can thus combine TE glyphs and histograms to provide an informative visualization of the distribution of mathematically defined repeats and the structural features of TEs (Figure 2.3). The current drawback to visualizations in GBrowse is that the program is intended to serve as a static visualization tool, and does not provide the means for the curation and combination of computational results. It would therefore be helpful if the current curation programs for gene annotation, such as Apollo or Artemis, directly addressed the needs of TE annotation curation and developed glyphs for the major classes of TEs.

In addition to visualization and curation of the annotated DNA, it is also possible to transfer the DAWGPAWS results into existing database schema. For example, the CHADO database [41] can make use of the gmod_bulk_load_gff3.pl program [57] that can load GFF3 format files into a CHADO database. In the DAWGPAWS package, the GFF3 format files from curated results can be generated with the cnv_game2gff3.pl program. These curated results could then be stored in a local implementation of the CHADO database. The BioSQL database schema [40] also includes a bp_load_gff.pl script that can load GFF results into the database schema.

The DAWGPAWS annotation framework has a number of features that make it a good choice to facilitate the workflow in plant genome annotation. The use of configuration files makes it fairly easy to modify the annotation workflow for the species of interest. The configuration files also makes it quite easy to generate results with multiple parameter sets for an individual program. Using multiple parameter sets will be especially useful when working with a genome that has not been annotated before, and for which appropriate annotation parameters have not been identified. Also, while previous annotation pipelines have focused on gene annotation, the DAWGPAWS suite maximizes the quality of TE annotation results. Most plants contain genomes with sizes >5000 Mb [58], and are therefore expected to contain more than 80% TEs [59], so efficiently dealing with this large number and diverse set of mobile DNAs is necessary for effective genome annotation.

The current focus of DAWGPAWS in our laboratory is the structural annotation of the genes and TEs in a genome using methods and applications tuned to the Triticeae. In annotation of 220 BACs from hexaploid bread wheat, we found that the DAWGPAWS pipeline increased the rate of individual BAC annotations by twenty-fold. Due to the time required to manually generate annotation results, this previous annotation effort was limited to using the FGENESH

annotation program combined with a BLAST search of predicted models against known transposable elements and protein databases [60]. Using this method, annotators could annotate a single BAC in one to two days. The implementation of the DAWGPAWS pipeline increased the speed of annotation to ten-fifteen BACs per person per day. Furthermore, the quality of both TE and gene prediction were also seen to improve with the use of DAWGPAWS. This was due, at least in part, to the larger number of complementary programs for TE and gene discovery that could be conveniently employed in each BAC annotation. Specifically, the inclusion of *ab initio* TE prediction programs allowed for the identification of new families of LTR retrotransposons that would have been missed in our previous annotation efforts. Predicted gene models that span these newly discovered families would not have been identified as TEs in the exclusively homology-based searches that were previously used.

Future development of DAWGPAWS will incorporate tools for the functional annotation of the predicted genes. Currently, functional annotation can be done within the Apollo program by manually selecting individual gene models and BLASTing these results against appropriate databases. A batch run support for additional local alignment search tools will also be added. The use of NCBI-BLAST is sufficient for most comparisons of sequence contigs against reference databases, but programs such as BLAT [61] or sim4 [62] are designed specifically to align ESTs and flcDNAs against assembled genomes. While output from these local alignment tools can be converted to GFF using the existing cnv_blast2gff.pl program in DAWGPAWS, it would be useful to use these packages in a batch run framework similar to the batch_blast.pl program.

Support for additional *ab initio* gene annotation programs will also be added to future releases of DAWGPAWS. Augustus [63] is an *ab initio* annotation program that will be useful for gene annotation that seeks to identify all transcripts derived from a single locus. Support for

GENEZILLA [64] and GlimmerHMM [64] gene annotation packages will also be added to future releases of DAWGPAWS. The SNAP program [65] will be added to support the annotation of genomes that have been sequenced *de novo* and lack species-specific HMM model parameterizations. The addition of the PASA [66] program would assist in the annotation of genomes that have large transcript databases that can assist genome annotation. As additional fully-sequenced genomes are added to the plant genomics literature, we can make use of syntenic comparisons and multiple alignments to aid in gene annotation [67] as well as TE annotation [68]. Future development of DAWGPAWS will incorporate syntenic alignment and prediction programs such as SGP2 [69], SLAM [70], and TWINSCAN [71] as they become increasingly relevant to plant genome annotation.

Conclusions

The DAWGPAWS annotation workflow provides a suite of command line interface programs that can generate computational evidences for human curation in a high-throughput fashion. We have used the DAWGPAWS pipeline to annotate 220 randomly selected BACs with wheat DNA inserts for both gene and TE content. Our curation efforts on the DAWGPAWS output are implemented in the Apollo program. The tiers file used for visualization of this curation are available as part of the DAWGPAWS package.

DAWGPAWS represents an efficient tool for genome annotation in the Triticeae, and can be used in its current form to generate gene and TE computational results for other grass genomes. Minor modifications to the configuration files used by DAWGPAWS can make this program suitable to the generation of computational annotation results for any plant genome. The TE annotation capabilities of DAWGPAWS exceeds any other current genome annotation suite,

and makes this package particularly valuable for the great majority of plant genomes, such as wheat or maize, that contain a diverse arrays of TEs that comprise the majority of the nuclear genome.

The DAWGPAWS program has been specifically designed to facilitate use of individual component scripts outside of the entire package. Each script can function independently of all other applications in the package, and programs make use of standard input and standard output streams when possible to facilitate integration into existing pipelines. Since this package is being released under the open source GPL (version 3), the suite and its individual components can be used and modified under the terms of the GPL. Template batch run and conversion scripts are provided in a boilerplate format to facilitate extending DAWGPAWS to additional annotation tools. Furthermore, since we have selected the Perl language for the implementation of our package, the addition of new annotation tools can leverage existing modules in the BioPerl toolkit [72]. These modules include parsers for computational tools useful for predicting alternative splicing [61,62] as well as interfaces for transfer RNA prediction [73]. We also formally invite collaboration in the development of additional DAWGPAWS applications under the auspices of the GNU GPL, as facilitated by the SourceForge subversion repository of the DAWGPAWS source code. Interested collaborators may contact the authors or become member developers of the DAWGPAWS SourceForge project [74].

Availability and Requirements

Project Name: DAWGPAWS Plant Genome Annotation Pipeline

Project Home Page: http://dawgpaws.sourceforge.net/

Operating System: Platform Independent

Programming Language: Perl

Other Requirements: BioPerl 1.4, as well as the annotation programs that scripts are dependent upon.

License: GNU General Public License 3

Any restrictions to use by non-academics: No restrictions

List of Abbreviations

BAC: Bacterial Artificial Chromosome; cDNA: complementary DNA; CLI: Command Line Interface; EST: Expressed Sequence Tag; flcDNA: full-length complementary DNA; GFF: General Feature Format; GPL: General Public License; HMM: Hidden Markov Model; LTR: Long Terminal Repeat; ORF: Open Reading Frame; pHMM: Profile Hidden Markov Model; TE: Transposable Element

Table 2.1. DAWGPAWS annotation scripts for generating computational annotation results in batch mode. These scripts operate on a directory of FASTA files, and generate the native results of the annotation program as well as the GFF file format. The exception is the batch_ltrstruc.vbs visual basic script that must be used in conjunction with cnv_ltrstruc2gff.pl to generate results in GFF.

Annotation Program	Result Type	DAWGPAWS Script
EuGène [9]	Gene *ab initio* and automated combined evidence	batch_eugene.pl
GeneID [42]	Gene *ab initio*	batch_geneid.pl
GeneMark.hmm [43]	Gene *ab initio*	batch_genemark.pl
Genscan [44]	Gene *ab initio*	batch_genescan.pl
Find_LTR [45]	TE *ab initio*	batch_findltr.pl*
LTR_STRUC [20]	TE *ab initio*	batch_ltrstruc.vbs
LTR_FINDER [21]	TE *ab initio*	batch_ltrfinder.pl*
LTR_seq [46]	TE *ab initio*	batch_ltrseq.pl*
FINDMITE [19]	TE *ab initio*	batch_findmite.pl*
Tandem Repeats Finder [47]	Repeat *ab initio*	batch_trf.pl
HMMER [48]	TE homology	batch_hmmer.pl*
NCBI-BLAST [49]	TE and gene homology	batch_blast.pl*
RepeatMasker [22]	TE homology	batch_repmask.pl*
TEnest [24]	TE homology	batch_tenest.pl

* Indicates programs that make use of a configuration file. The nature and format of the configuration file for these programs is described in the individual help file for those programs.

Table 2.2. DAWGPAWS scripts for conversion of annotation results from native program output to GFF.

Annotation Program	Result Type	DAWGPAWS Script
FGENESH [50] [51]	Gene *ab initio*	cnv_fgenesh2gff.pl
GeneMark.hmm [43]	Gene *ab initio*	cnv_genemark2gff.pl
Find_LTR [45]	TE *ab initio*	cnv_findltr2gff.pl
LTR_FINDER [21]	TE *ab initio*	cnv_ltrfinder2gff.pl
LTR_seq [46]	TE *ab initio*	cnv_ltrseq2gff.pl
LTR_STRUC [20]	TE *ab initio*	cnv_ltrstruc2gff.pl
RepSeek [52]	TE *ab initio*	cnv_repseek2gff.pl
NCBI-BLAST [49]	TE and gene homology	cnv_blast2gff.pl
RepeatMasker [22]	TE homology	cnv_repmask2gff.pl
TEnest [24]	TE homology	cnv_tenest2gff.pl

Table 2.3. Additional helper scripts included in the DAWGPAWS package.

DAWGPAWS Script	Purpose
cnv_gff2game.pl	Converts GFF files to the game.xml format.
cnv_game2gff3.pl	Converts game.xml files to the GFF3 format.
batch_hardmask.pl	Given a directory of lowercase masked sequence files, this will replace lowercase residues with an N or X to indicate masking.
dir_merge.pl	Given annotation results scattered across multiple directories, this program can merge the results into subdirectories in a single parent directory.
vennseq.pl	Given GFF annotation results from multiple methods, this program generates a Euler Diagram of these features using the VennMaster program [55]
batch_findgaps.pl	This program will annotate gaps in the query sequences in the input directory.
clust_write_shell.pl	This program writes shell scripts to run DAWGPAWS in a cluster environment running the Platform LSF queuing system.
cnv_seq2dir.pl	Given a FASTA file with multiple sequence files, this program generates a separate FASTA file for each sequence record. The sequence files produced are named using the sequence ID in the FASTA header in the input file.
fasta_merge.pl	This program merges all FASTA files in a directory into a single FASTA file.
fasta_shorten.pl	This program shortens the FASTA header by limiting the header length, or splitting the header by a delimiting character. Some annotation programs are limited by the length of the FASTA header that is accepted, and this programs allows input files to meet this limitation.
fetch_tenest.pl	Fetches multiple results from the Plant GDB TEnest server and converts the results to GFF.
gff_seg.pl	Given a GFF file that contains point or segment data, this will extract segments with score values that exceed a threshold value.
ltrstruc_prep.pl	Because the LTR_STRUC program only runs under the windows environment, this program converts FASTA sequences in UNIX to DOS line endings and generates the files name and flist file required for LTR_STRUC.
seq_oligiocount.pl	This program allows for the generation of a GFF file that counts the number of times an oligomer in the genomic contig occurs in a reference shotgun sequence database.

Table 2.4. Common command line options used throughout the DAWGPAWS suite of programs.

Option	Description
--indir *or* --infile	For batch scripts, this indicates the input directory containing the FASTA files to annotate. For conversion scripts, this indicates the input file to convert from the native format to the GFF format.
--outdir *or* --outfile	For batch scripts, this indicates the output directory containing the annotation results for the program and the GFF results. For conversion scripts, this indicates the path to the GFF output file.
--config	For programs that make use of a configuration file, this indicates the path to the configuration file to use.
--seqname	For conversion scripts, this indicates the sequence id to use in the GFF output file.
--param	For conversion scripts, this indicates the name of that parameter set used with the annotation program. This option allows the user to distinguish among multiple parameter sets for the same annotation program, and this parameter name is appended to the source column of the GFF output file.
--program	For conversion scripts, this indicates the name of the program used to generate the annotation result.
--version	Print the current version of the script.
--usage	Print a short program usage message.
--help	Print a short help message including the common usage and all program options available at the command line.
--man	Print the full program manual.
--verbose	This will run the program with maximum verbosity. This option will generate status updates while the program is running, and will maximize the error reporting functions of the script. All verbose statements are written to the standard error output stream.

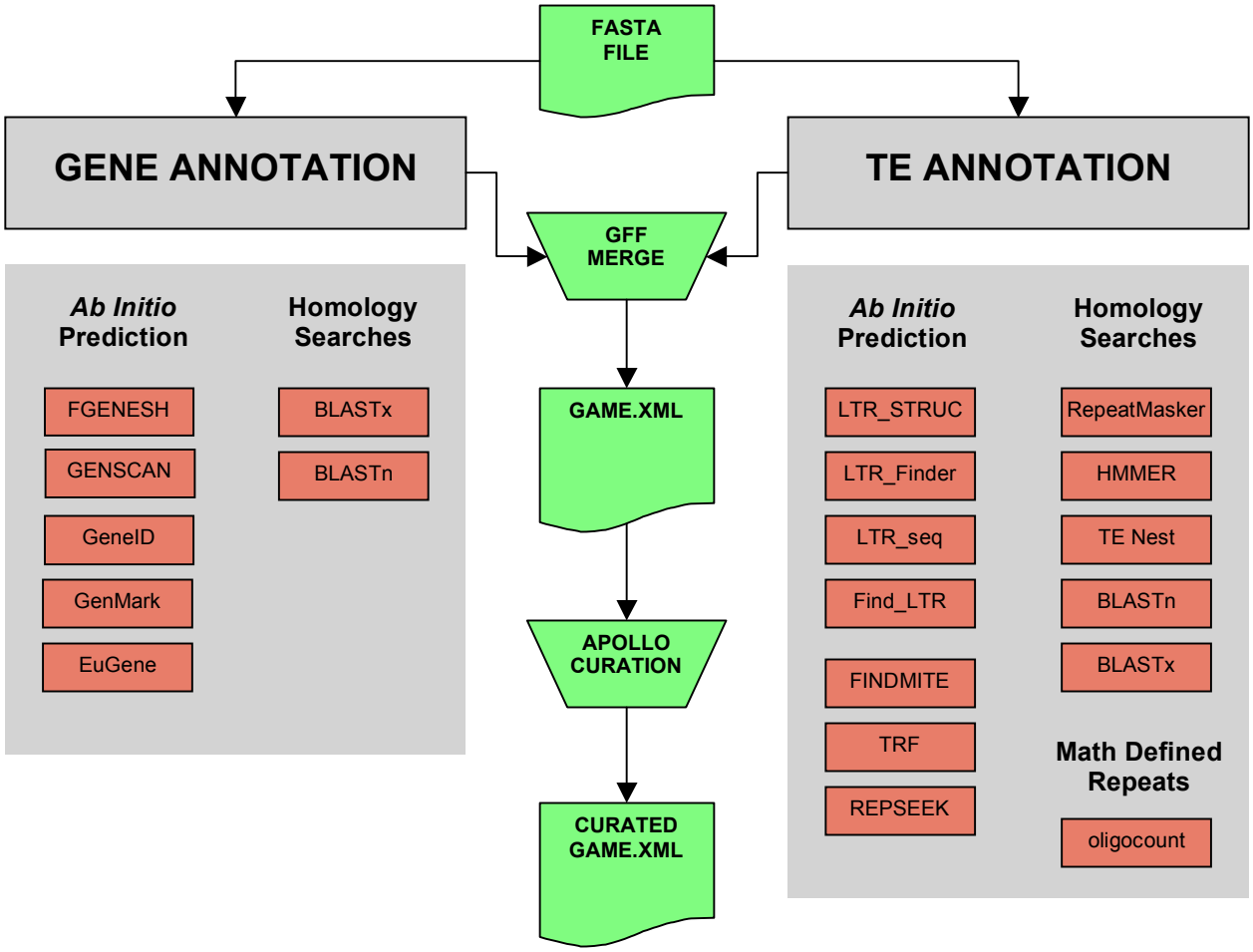


Figure 2.1. An overview of the workflow supported by the current version of the DAWGPAWS suite of programs.

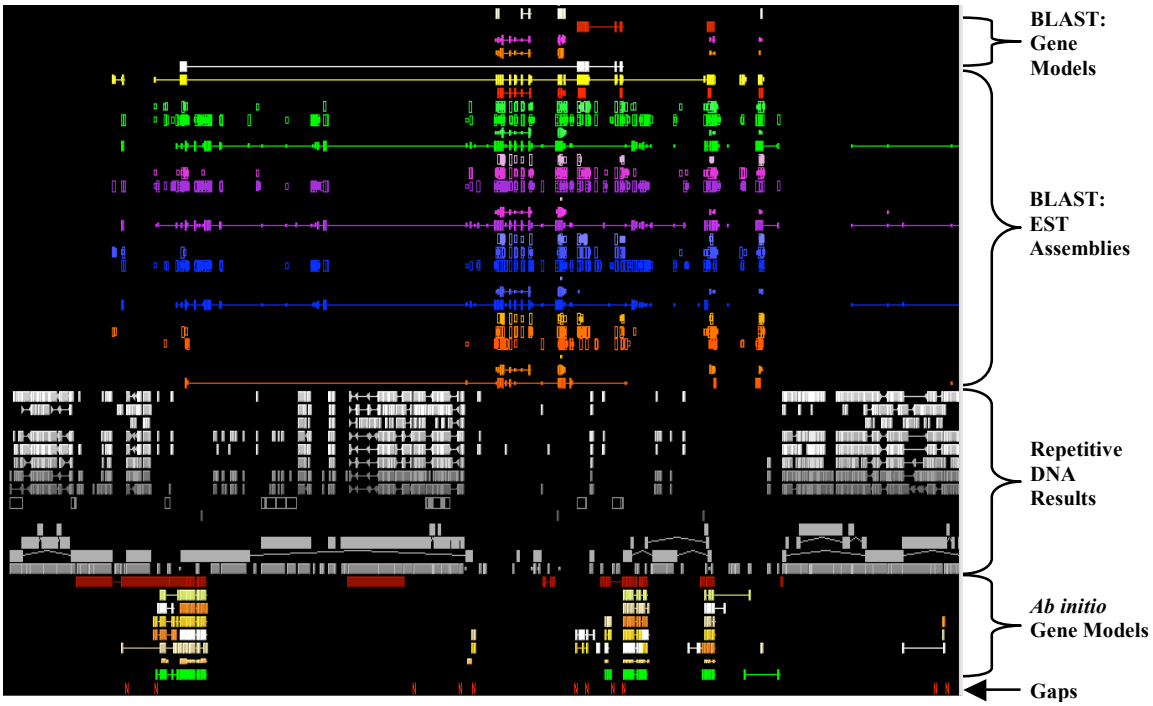


Figure 2.2. Screen capture image of gene and TE annotation results visualized in the Apollo genome annotation program. This example shown is for a wheat BAC that has been annotated and curated with the assistance of DAWGPAWS.

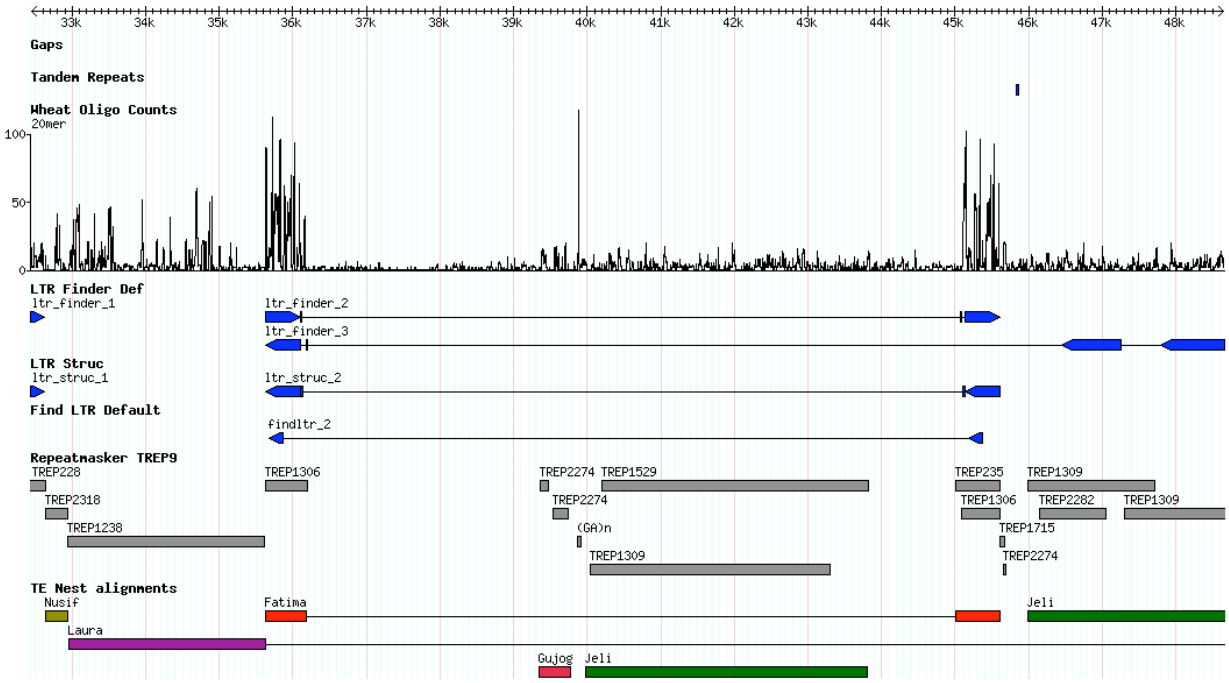


Figure 2.3. Screen capture image of the TE annotation results and oligomer counts visualized in the GBrowse genome annotation visualization program. The example shown is for a 15 kb segment of a BAC with a wheat DNA insert.

References

1. Liolios K, Mavromatis K, Tavernarakis N, Kyrpides NC (2008) The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 36: D475-479.
2. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, et al. (2008) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 36: D13-21.
3. Pop M, Salzberg SL (2008) Bioinformatics challenges of new sequencing technology. Trends Genet 24: 142-149.
4. Stein L (2001) Genome annotation: from sequence to biology. Nat Rev Genet 2: 493-503.
5. Bennetzen JL (2000) Transposable element contributions to plant gene and genome evolution. Plant Mol Biol 42: 251-269.
6. Bennetzen JL (2005) Transposable elements, gene creation and genome rearrangement in flowering plants. Curr Opin Genet Dev 15: 621-627.
7. Wang Z, Chen Y, Li Y (2004) A brief review of computational gene prediction methods. Genomics Proteomics Bioinformatics 2: 216-221.
8. Do JH, Choi DK (2006) Computational approaches to gene prediction. J Microbiol 44: 137-144.
9. Schiex T, Moisan A, Rouzé. P (2001) EuGene: An Eucaryotic Gene Finder that combines several sources of evidence. . In: Gascuel O, Sagot M-F, editors. Computational Biology. pp. 111-125.
10. Allen JE, Salzberg SL (2005) JIGSAW: integration of multiple sources of evidence for gene prediction. Bioinformatics 21: 3596-3603.
11. Bennetzen JL, Coleman C, Liu R, Ma J, Ramakrishna W (2004) Consistent over-estimation of gene number in complex plant genomes. Curr Opin Plant Biol 7: 732-736.
12. Sanmiguel P, Bennetzen JL (1998) Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Annals of Botany 82: 37-44.
13. Bergman CM, Quesneville H (2007) Discovering and detecting transposable elements in genome sequences. Brief Bioinform 8: 382-392.
14. Feschotte C, Pritham EJ (2007) Computational analysis and paleogenomics of interspersed repeats in eukaryotes. In: Stojanovic N, editor. Computational Genomics: Current Methods: Taylor and Francis. pp. 31-54.
15. Saha S, Bridges S, Magbanua ZV, Peterson DG (2008) Computational Approaches and Tools Used in Identification of Dispersed Repetitive DNA Sequences Tropical Plant Biology 1: 85-96.
16. Bao Z, Eddy SR (2002) Automated *de novo* identification of repeat sequence families in sequenced genomes. Genome Res 12: 1269-1276.
17. Edgar RC, Myers EW (2005) PILER: identification and classification of genomic repeats. Bioinformatics 21 Suppl 1: i152-158.
18. Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. Bioinformatics 21 Suppl 1: i351-358.
19. Tu Z (2001) Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, Anopheles gambiae. Proc Natl Acad Sci U S A 98: 1699-1704.
20. McCarthy EM, McDonald JF (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics 19: 362-367.

21. Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35: W265-268.
22. Smit A, Hubley R, Green P (1996-2004) RepeatMasker Open-3.0.
23. Kohany O, Gentles AJ, Hankus L, Jurka J (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 7: 474.
24. Kronmiller BA, Wise RP (2008) TEnest: Automated chronological annotation and visualization of nested plant transposable elements. Plant Physiol 146: 45-59.
25. Pereira V (2008) Automated paleontology of repetitive DNA with REANNOTATE. BMC Genomics 9: 614.
26. Wicker T, Matthews DE, Keller B (2002) TREP: a database for Triticeae repetitive elements. Trends in Plant Science 7: 561-562.
27. Ouyang S, Buell CR (2004) The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. Nucleic Acids Res 32: D360-363.
28. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110: 462-467.
29. Spannagl M, Noubibou O, Haase D, Yang L, Gundlach H, et al. (2007) MIPSPlantsDB--plant database resource for integrative and comparative plant genome research. Nucleic Acids Res 35: D834-840.
30. Li R, Ye J, Li S, Wang J, Han Y, et al. (2005) ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. PLoS Comput Biol 1: e43.
31. DeBarry JD, Liu R, Bennetzen JL (2008) 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: 235.
32. Kurtz S, Narechania A, Stein JC, Ware D (2008) A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics 9: 517.
33. Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, et al. (2005) Combined evidence annotation of transposable elements in genome sequences. PLoS Comput Biol 1: 166-175.
34. Durbin R, Haussler D, Stein L, Lewis S, Krogh A (2000) GFF Format Specifications. http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml.
35. Lewis SE, Searle SM, Harris N, Gibson M, Lyer V, et al. (2002) Apollo: a sequence annotation editor. Genome Biol 3: RESEARCH0082.
36. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, et al. (2000) Artemis: sequence visualization and annotation. Bioinformatics 16: 944-945.
37. Donlin MJ (2007) Using the Generic Genome Browser (GBrowse). Curr Protoc Bioinformatics Chapter 9: Unit 9 9.
38. Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, et al. (2009) The UCSC Genome Browser Database: update 2009. Nucleic Acids Res 37: D755-761.
39. Stalker J, Gibbins B, Meidl P, Smith J, Spooner W, et al. (2004) The Ensembl Web site: mechanics of a genome browser. Genome Res 14: 951-955.
40. Lapp H (2008) BioSQL. http://www.biosql.org.
41. Zhou P, Emmert D, Zhang P (2006) Using Chado to store genome annotation data. Curr Protoc Bioinformatics Chapter 9: Unit 9 6.

42. Parra G, Blanco E, Guigo R (2000) GeneID in Drosophila. Genome Res 10: 511-515.
43. Lukashin AV, Borodovsky M (1998) GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res 26: 1107-1115.
44. Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268: 78-94.
45. Rho M, Choi JH, Kim S, Lynch M, Tang H (2007) *De novo* identification of LTR retrotransposons in eukaryotic genomes. BMC Genomics 8: 90.
46. Kalyanaraman A, Aluru S (2006) Efficient algorithms and software for detection of full-length LTR retrotransposons. J Bioinform Comput Biol 4: 197-216.
47. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27: 573-580.
48. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14: 755-763.
49. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.
50. Solovyev VV, Salamov AA, Lawrence CB (1995) Identification of human gene structure using linear discriminant functions and dynamic programming. Proc Int Conf Intell Syst Mol Biol 3: 367-375.
51. Salamov AA, Solovyev VV (2000) Ab initio gene finding in Drosophila genomic DNA. Genome Res 10: 516-522.
52. Achaz G, Boyer F, Rocha EP, Viari A, Coissac E (2007) Repseek, a tool to retrieve approximate repeats from large DNA sequences. Bioinformatics 23: 119-121.
53. Kurtz S (2004) vmatch. http://www.vmatch.de
54. Wicker T, Narechania A, Sabot F, Stein J, Vu GT, et al. (2008) Low-pass shotgun sequencing of the barley genome facilitates rapid identification of genes, conserved non-coding sequences and novel repeats. BMC Genomics 9: 518.
55. Kestler HA, Muller A, Gress TM, Buchholz M (2005) Generalized Venn diagrams: a new method of visualizing complex genetic set relations. Bioinformatics 21: 1592-1595.
56. Estill JC DAWGPAWS User Manual. http://dawgpaws.sourceforge.net/man.html.
57. Osborne B How to Load GFF Into Chado. http://gmod.org/wiki/Load_GFF_Into_Chado.
58. Zonneveld BJM, Leitch IJ, Bennett MD (2005) First nuclear DNA amounts in more than 300 angiosperms. Annals of Botany 96: 229-244.
59. Flavell RB, Bennett MD, Smith JB, Smith DB (1974) Genome Size and Proportion of Repeated Nucleotide-Sequence DNA in Plants. Biochemical Genetics 12: 257-269.
60. Devos KM, deOliveira AC, X Xu, Estill JC, Estep M, et al. Structure and organization of the wheat genome – the number of genes in the hexaploid wheat genome. In: Lynne RAREELPLMM, editor; 2008; Brisbane. Sydney University Press.
61. Kent WJ (2002) BLAT--the BLAST-like alignment tool. Genome Res 12: 656-664.
62. Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res 8: 967-974.
63. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, et al. (2006) AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res 34: W435-439.
64. Majoros WH, Pertea M, Salzberg SL (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics 20: 2878-2879.
65. Korf I (2004) Gene finding in novel genomes. BMC Bioinformatics 5: 59.

66. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Jr., et al. (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res 31: 5654-5666.
67. Dubchak I (2007) Comparative analysis and visualization of genomic sequences using VISTA browser and associated computational tools. Methods Mol Biol 395: 3-16.
68. Caspi A, Pachter L (2006) Identification of transposable elements using multiple alignments of related genomes. Genome Res 16: 260-270.
69. Parra G, Agarwal P, Abril JF, Wiehe T, Fickett JW, et al. (2003) Comparative gene prediction in human and mouse. Genome Res 13: 108-117.
70. Alexandersson M, Cawley S, Pachter L (2003) SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model. Genome Res 13: 496-502.
71. Korf I, Flicek P, Duan D, Brent MR (2001) Integrating genomic homology into gene structure prediction. Bioinformatics 17 Suppl 1: S140-148.
72. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, et al. (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome Res 12: 1611-1618.
73. Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25: 955-964.
74. Estill JC DAWGPAWS SourceForge Project Page. http://sourceforge.net/projects/dawgpaws.

CHAPTER 3

REPMINER AND ITS USE FOR LTR RETROTRANSPOSON ANALYSIS IN MAIZE[1]

[1] Estill, J.C., R.S. Baucom and J.L. Bennetzen. To be submitted to *Genome Research.*

Abstract

Discerning the relationships among transposable elements (TEs) and circumscribing family identities is a fundamental step of TE annotation that provides the foundation for the study of TE demography and evolution. We have developed a graph theory-based strategy for the taxonomic assignment of DNA sequence features. This data mining approach allows for the clustering of TEs into families based on networks of shared homology. A key component of this approach is network visualization. Biologically relevant sequence features can be mapped onto networks, and cluster-based names can be compared to existing canonical databases. Subsets of sequences can also be selected for further phylogenetic analyses to test for monophyly of named groups. We have applied the RepMiner tools to an analysis of maize long terminal repeat (LTR) retrotransposons to illustrate the utility of this program. RepMiner succinctly illustrates and further illuminates previously identified relationships among the *Ji*/*Opie* and *Cinful*/*Zeon* families. In addition, RepMiner identifies numerous inconsistencies in classification and/or biology, including a previously unrecognized split in the *Huck* family indicating that *Huck* should be treated as two separate families. RepMiner can be applied to all classes of TEs or other repeats in all species and will be particularly useful to characterize TEs in species that do not have existing repeat databases.

Introduction

Transposable elements (TEs) are ubiquitous and widespread among eukaryotic genomes and have had profound impacts on genome evolution and genome architecture [1]. These mobile DNA sequences can undergo rapid episodes of proliferation that have resulted in plant genomes like maize that are composed of ~85% transposable element DNA [2]. TE insertions lead to

structural variation among genomes within species [3,4] that can accumulate as significant population variation [5,6]. TEs insertions may also directly influence host gene evolution [7,8], and can drive genome rearrangement [9]. One step in understanding how TEs impact genomes is a reasonable taxonomy for assigning TE families. Such taxonomies can organize assessment of general taxonomic diversity of TEs and its relationship to genome evolution.

The observation that the DNA sequences comprising the repetitive fraction of the genome can be classified into families were first made with early reassociation kinetics studies of entire genomes [10]. The DNA sequencing of entire genomes allows us to explore this TE diversity at the highest resolution possible, and allows us to fully describe family relationships. Much like the study of other levels of biological organization, the study of TE biology will require a consistent system of nomenclature. Unfortunately, the systems of nomenclature that have been developed to sort TEs into families have not been consistent among classes of TEs or even among different host genomes for the same class. These inconsistencies are a barrier to the basic study of the distributions, interactions, activities and abundances of TEs. Furthermore, most systems of nomenclature and databases of TEs are focused on the high copy "repetitive" members, and often ignore the contribution of the low copy elements that often comprise a majority of the TE diversity. The study of genome evolution thus needs a system of TE nomenclature that can be applied at the whole genome scale, that can be used consistently across TE classes, and that can be applied consistently across host genomes.

Current approaches to the taxonomic assignment of sequences features in genomes are generally geared toward the assignment of protein-coding regions of the genome into protein families or subdomains, so many published classification systems for protein families exist [11,12,13,14,15,16,17,18,19,20,21,22]. Family assignment of proteins using these query

databases can be undertaken by searching an unknown protein against one of these canonical databases. Protein coding domains have also been assigned to families using unsupervised classification programs such as MCL [23,24] that generate *de novo* classifications within a given dataset using an "all-by-all" similarity search of all the sequences comprising the database.

Some of the earliest sequence-based studies of the repeat content of genomes, based on reassociation kinetics, proposed that the repetitive fraction of genomes are organized into families that are derived from saltatory replication events [10]. The sequence of and assembly of entire genomes have allowed us to increase the resolution and specificity of the analysis of repeats within genomes, and have allowed us to identify a great diversity of transposable elements in eukaryotes. Various attempts have been made to classify these repeats and TEs into hierarchical classification systems [25], and previous attempts have been made to group individual classes of TEs into families. Most of these early attempts have focused on a classification of the high-copy-number repeats, with the goal of generating consensus sequences. For example, the Repeat pattern toolkit [26] used single linkage clustering and minimum spanning trees to generate family level taxonomies of repeats in the *C. elegans* genome. The RECON program [27] is another repeat-based system that uses single linkage clustering of local pairwise alignments to generate families of TEs. The RepeatGluer program [28] also attempts *de novo* classification of repeat families using mosaics of sub-repeats and relies on single linkage clustering for its taxonomic assignment. More recent advances in repeat classification have attempted to expand graph theory-based approaches to the assignment of TE families. These include the use of the "BAG" algorithm [29,30] that makes use of the graph properties of biconnected components and articulation points to find family clusters within connected

components, and the MUST system [31] for MITE identification that makes use of MCL to cluster putative MITEs into groups for analysis.

A weakness of the current approaches to family assignment of TEs is that most existing technologies are focused on assignment of high-copy-number repeats and cannot be readily extended to low-copy-number elements. Because the majority of TE families that occur in genomes are at low to medium copy number, a comprehensive assignment of TE diversity must include the ability to assign low- and medium-copy-number TEs to families. Such a comprehensive approach to TE family assignment will facilitate analyses regarding the distribution, inter-relationships and diversity of TEs in the full nuclear genome. Furthermore, it would be difficult to understand why high-copy-number families are at a high copy number without simultaneously considering the processes that have kept most families at low to moderate numbers in host genomes. An additional problem of current systems is that they are generally limited to single-species studies, and are not generally applicable across classes of TEs. Current systems of nomenclature thus are a barrier to studies that cross host genomes, and generally do not allow diversity to be compared across classes of transposable elements within single host genomes.

We propose a graph theory-based data mining approach to family assignment of TEs. A graph theory-based approach to TE family assignment has a number of strengths. First, this approach can leverage existing tools for the visualization of networks that have been designed for the display of protein interaction networks or other domains of biological knowledge [32,33]. Second, this approach can also make use of existing graph theory-based clustering methods that have been developed for protein family classification [23,24] or for the general algorithms designed for the identification of representative exemplars from larger collections of data

[34,35]. Finally network approaches have been shown to be useful models for the study of any biological system that shows power law-like behavior in scale-free networks [36]. The distribution of family size in TEs certainly appears to fit this general model of power law-like behavior with a few members being in very high copy number within genomes, while the majority of the families exist in low copy number [1].

The goals of this chapter are to describe the implementation of a graph theory-based data mining approach called RepMiner, and to apply RepMiner to the exploration of the relationships of the TEs in a genome. This approach will be illustrated using a dataset of maize LTR retrotransposons, and will highlight a previously undescribed split in the *Huck* family that indicates that there should be two separate families of *Huck* recognized in the maize genome. The tools generated to make this approach possible are being distributed as open source code, and should be applicable to many other types of sequence data that are in need of classification.

Materials and Methods

Implementation of the RepMiner Package

The RepMiner package was written in the Perl programming language and makes use of existing bioinformatics modules available from BioPerl [37]. The user interface for building network visualizations is through a series of command line programs that allow the user to design and implement a workflow suitable to their individual needs for the clustering and visualization of similarity networks (Figure 3.1). The individual programs accept standard input and standard output streams to facilitate building these workflows. The command line options for the individual programs conform to GNU coding standards for command line software [38], and all programs include options for obtaining command line help information or the full user

manual. Documentation on how to use individual programs is also being made available from the RepMiner SourceForge web site (http://sourceforge.net/projects/repminer/), along with examples of how to use the program for analyses of transposable element classifications.

RepMiner currently supports generation of similarity matrices using the NCBI-BLAST and FASTA algorithms. Clustering these matrices can make use of the connected components algorithm of single linkage clustering, MCL clustering [23,24], as well as affinity propagation clustering [34]. RepMiner facilitates using these clustering results for taxonomic assignment by mapping cluster groupings onto network visualizations as node attributes visualized as color or node shape. The score of similarity among individual nodes or other biological attributes of relatedness can also be visualized as edge attributes, such as edge color or edge width. These networks can be decorated with node attributes generated from multiple sources, including nearest-neighbor BLAST analysis to compare newly discovered elements to known element families, as well as biological annotation of sequence components such as the presence or absence of particular genes in the network. This is particularly useful for the identification of autonomous and nonautonomous partners within genomes. These sequence attributes can be visualized as node color or node shape in the network visualization. The network files produced from the RepMiner analyses are designed for visualization in the Cytoscape biological network visualization program [32].

Because RepMiner has been used primarily for annotation of LTR retrotransposon, the RepMiner suite also includes high throughput structural annotation and visualization tools for LTR retrotransposon. RepMiner is designed to interface with the DAWGPAWS suite of plant genome annotation programs [39], and could be used in conjunction with these programs for curation and taxonomic assignment of other classes of repeats or other sequence feature data.

Annotation of Maize LTR Retrotransposons

To avoid misclassifications of families due to any errors in genome assembly that could create false chimeric elements, only annotations derived from high quality assembled regions of the maize genome were used for the analysis of taxonomic membership. The annotation of maize LTR retrotransposons took place as detailed in a previous manuscript [40] and only a general overview is provided here. The program LTR_Struc [41] was used to identify LTR retrotransposons in the ~15,000 maize BACs that were sequenced in depth and masked for the version 1 assembly of the B73 maize genome [2]. The sequences derived from the LTR_Struc program were mapped onto their locations in these BACs and the coding domains within the LTR retrotransposon models were annotated using the *cnv_ltrstruc2ann.pl* program of the DAWGPAWS package [39]. These annotation results and coding domain sequences were stored in a Microsoft Access database for further analysis.

Clustering of Maize LTR Retrotransposons

The similarity matrix used for the following clustering experiments was generated by a BLAST search of all 5'LTRs against all 5' LTRs. An e-value threshold of $e<1x10^{-10}$ was used, and all significant BLAST hits were returned. The BLAST bitscore of the tiled high-scoring segment pairs (HSPs) was determined by the RepMiner program *cnv_blast2sim.pl* program and was used as the similarity metric for clustering. This method sums the bitscores across all of the locally aligned segments between two sequences to produce a single similarity value.

The connected components cluster of the BLAST-based similarity matrix used the *jabablast.pl* program in RepMiner to generate connected component clusters. This program uses simple single linkage clustering as determined by the Graph CPAN PERL module

(http://search.cpan.org/~jhi/Graph-0.94/). The MCL clusters were generated using version 1.007 of the MCL program [23] with default inflation parameters.

Estimation of Insertion Dates

The estimation of insertion dates of LTR retrotransposons follows now-standard protocols first described in maize [42], in which the time since insertion is deduced by the accumulation of mutations that have occurred between the 5' and 3' long terminal repeats since insertion. Perl programs have been included in the RepMiner package to automate this process for high-throughput analysis of all annotated LTR retrotransposons within genomes. These programs align LTRs using ClustalW [43], and estimate divergence between the two LTRs with the PAML baseml module [44]. The substitution rate of 1.3 $X10^{-8}$ per site per year was then used to estimate time since insertion [45]. These results were uploaded to the MS Access database of all maize LTR retrotransposons for further analysis.

Additional Clustering Analysis of* Huck*-Like Sequences

Since the *Huck*-like sequences clustered into two separate sets (Figure 3.5), the original description of the *Huck* LTR retrotransposon from the *Adh1-F* region of maize [46] was used to assign the *Huck* name to the cluster with the highest similarity to the original description of the family. The DNA sequence of the original *Huck* was downloaded from GenBank (AF123535:50734..63094) and included in an All-By-All BLAST analysis of the 5' LTRs of all the *Huck*-like sequences using BLASTn with an e-value threshold of $e < 1 \times 10^{-10}$ and the top 250 hits reported. These results were converted to files suitable for visualization in Cytoscape [32] using the RepMiner tools workflow (Figure 3.1). This permitted a visual identification of the more abundant and most recently active cluster as containing the original *Huck* element (Figure 3.5).

Phylogenetic Analysis of Huck*-Like Sequences*

The reverse transcriptase-encoding regions of maize LTR retrotransposons were identified by BLAST-based alignment between a database of reverse transcriptase sequences downloaded from the PFAM database [21]. A multiple alignment of the DNA sequences for the reverse transcriptase of the *Huck*-like family were performed in Muscle [47] using default settings. Maximum likelihood phylogenies were generated in the MPI version of RaxML [48] using the general time reversible (GTR) model of nucleotide substitution with *Gamma* distributed rate heterogeneity. The best tree of 1000 runs is reported here (Figure 3.5) as drawn by version 1.3.1 of the FigTree program (http://tree.bio.ed.ac.uk/software/figtree/).

Mapping Huck *and* Puck *to the Maize Genome Assembly*

The identification of the BAC contigs containing the predicted full-length LTR retrotransposons from the LTR_Struc annotations [41] were incorporated into the database of maize LTR retrotransposons using the RepMiner program *ltrann2db.pl*. The BAC contigs corresponding to the each set of *Huck* and *Puck* MCL clusters were identified in the database, and these BACs were mapped onto the maize genome assembly using the BAC mapping tool at maizesequence.org (http://www.maizesequence.org).

Identification of the Huck *Repeat Unit*

The *Huck* repeat unit was identified by a combination of dot plot visualization using the program dottup (http://emboss.bioinformatics.nl/cgi-bin/emboss/dottup) and the program vmatch (http://www.vmatch.de/) to identify putative repeated sequences within the LTRs of representative *Huck* sequences. The frame of the repeated unit was verified using the program Muscle [47] to align 5' LTRs of representative *Huck* sequences. These analyses detected the *Huck* repeat unit (HuRU) as the following 57 bp sequence:

CCCGACCCCAGGGCTCGGACTCGGGCTAAGACCCGGAAGACGGCGAACTCCGCTCCG.

The HuRU sequence string was searched against the full set of *Huck* sequences using exact string search in the vmatch program (http://www.vmatch.de/). This search identified HuRU in approximately 60% of all *Huck* sequences, while HuRU was not detectable in any of the *Puck* sequences analyzed. Those *Huck* sequences that did contain the HuRU sequences correspond to the younger of the two MCL clusters identified in the MCL clustering analysis of maize LTRs.

Results

The *RepMiner* package has been implemented in the Perl programming language and is being made available as a set of command line programs. These programs will work on multiple operating systems using flat text files to pass information among programs, and can alternatively make use of a MySQL database to hold annotation and classification information (Figure 3.1). RepMiner is particularly useful for visualizing the best matches to known elements mapped onto the derived networks of shared similarity among structurally annotated TEs (Figure 3.2). Family identification of elements can also make use of connected components approaches in conjunction with MCL results to assign newly discovered clusters to family sets. In some cases, these MCL-based clusterings will even identify putative insertion cohorts within families (Figure 3.3). For LTR retrotransposons, it is possible to visualize these insertion dates across the entire network of elements using tools provided in the RepMiner package (Figure 3.4).

For LTR retrotransposons, using just the 5'LTR for all-by-all BLAST analysis does provide useful taxonomies that are generally consistent with phylogeny-based family assessment (Figure 3.5). For visualization purposes, it is useful to just include the top 250 best hits for connected component graph visualizations of these results. Additional edges made visualization computationally infeasible for large datasets, and additional edges made it more difficult to

visualize structured relationships for the moderate to small datasets that could be visualized in this manner. However, it is relevant to use the full similarity matrix for computationally based clustering analysis of similarity networks such as MCL clustering.

The specific network visualization of the maize LTR retrotransposons used in this analysis showed general support for previous classifications of *Ji* and *Opie* [49] as well as previously unknown relationships between *Giepum*, *Ruda* and the *Ji*/*Opie* group. The large connected network cluster for all maize LTR retrotransposons (Figure 3.2) also nicely illustrates problems with using only connected components clustering as a tool for family identification in elements, because the entire *Ji*/*Opie* group and most high-copy-number families would have been grouped as a single family using a single linkage approach. However, even visualizing the network in a spring-embedded framework, as shown in Figure 3.2, easily shows that these families are clustering together within the larger network. And these visual clusters are supported by MCL cluster assignment of these data. Some of the edges in this larger network are due to real homology among LTRs included in the network visualization, while others are due simply to chimeric annotations in the dataset.

Another interesting result from this visualization is the observation that the previously named *Huck* group should potentially be treated as two separate families. These can be seen as the two separate clusters for all families in Figure 3.3, and these results from the LTR-based clustering are generally supported by a phylogenetic analysis (Figure 3.5) of the reverse transcriptase regions for these LTR retrotransposons. A MCL clustering of the maize genome groups these elements into 5 clusters. These elements generally show the same distribution across the maize chromosomes (Figure 3.6), and the MCL clusters within the two connected component clusters of the *Huck*-like correspond to trends in LTR length and age between the two

groups (Figure 3.7). Adding the original description of the *Huck* element in maize [46] to this network indicated that the *Huck* name is appropriately assigned to the more recently active of the two groups, and the second group is here named *Puck* to acknowledge it as a separate family that has a historical relationship with the type *Huck* group (Figure 3.7). A structurally diagnostic feature that distinguishes between at least some members of the *Huck* and *Puck* families is the existence of a 57 bp repeated unit refereed to here as the *Huck* repeat unit (HuRU). This repeat unit is associated with the youngest two MCL clusters of the *Huck* group, and expansion of this repeated unit is largely responsible for the observed expansion in the *Huck* LTR length (Figure 3.8).

Discussion

The application of the RepMiner approach, even to well-characterized genomes, can provide new insights into the relationship among families of TEs. This is due in part to the fact that many current methods (*i.e.,* RECON) for TE identification and family assignment have a tendency to focus on the high-copy-number fraction of genomes at the expense of ignoring the low-copy-number families of TEs. The RepMiner methods allow these low copy members to be included in an overall taxonomy of elements when these TEs can be structurally identified. The ability to include evidence from clustering programs such as MCL also allows RepMiner to avoid the problem of falsely joining together families due to chimeric annotations that can arise from nested insertions, as observed with strictly connected component clustering of genomes [29]. The ability to map known families onto the networks generated by RepMiner also emphasizes novel insights highlighted in the context of previously known elements.

RepMiner is perhaps even more useful when used in conjunction with structure-based TE annotation tools for newly sequenced genomes for which there are not yet known TEs. For example, RepMiner can be used to quickly cluster and visualize the relationship among elements discovered by structure-based annotation methods such as LTR_Struc [41] or LTR_Finder [50]. These elements can be compared to known transposable elements from canonical TE databases since RepMiner includes the means to use existing databases as a means for classifying unknown TEs with respect to already established taxonomies. Comparison to known families in closely related species can highlight novel elements in the genome being studied, and it is even possible to generate entirely new family relationships when working with novel genomes. RepMiner can even be used to visualize clustering of sample sequence data from novel genomes, thus serving as a potential adjunct to repeat assembly tools such as AAARF [51].

The RepMiner code used to generate the figures and analyses described in this manuscript are available from the RepMiner SourceForge website (http://sourceforge.net/projects/repminer/). An online user manual is also being developed that provides a walkthrough of using RepMiner for annotation, clustering, exemplar discovery and classification. Although the storage of RepMiner results in relational databases is currently supported, a database back end is not currently required for use of this release of RepMiner. Rather, the package focuses on command line scripts that can make use of flat text files that can be passed between individual components of the program (Figure 3.1). This database-independent format facilitates the use of the program in distributed computing environments, and minimizes the dependence on external software. Future development of the RepMiner program will implement a tighter integrated database design to allow for scalability to larger datasets that can include multiple genomes in a single analysis. This development will also seek to include an

online graphical user interface to the RepMiner database and tools to greatly reduce the learning curve for applying the RepMiner approach.

In our application of RepMiner to classification of LTR retrotransposons in the maize genome, it is clear that RepMiner can further illuminate existing insights into the relationships among transposable elements. For example, while the relationship of a shared integrase protein between the *Cinful* and *Zeon* families has been previously proposed [52], RepMiner nicely visualized how the LTRs of these families show clear networks of similarity (Figure 3.2). Furthermore, the ability of RepMiner to cluster sequences within families, using tools such as MCL, allows for the identification of putative insertion cohorts within individual families (Figure 3.3). While it has long been suggested that repeat elements in genomes proliferate in bursts of proliferation [10], RepMiner provides the capability to identify the cohorts of those individual bursts and follow their evolution across the genome. Studies of the differences in distribution between older and more recent bursts of transpositions could provide additional insight into the interplay of insertion and deletion in shaping the current arrangement of transposable elements in the maize genome. Such studies of insertion cohorts have provided unique insight into the evolution of *Alu* elements in the human genome [53,54,55] and have shown that spatial differences in selection and removal mechanisms are often important factors in structuring current patterns of distribution.

The application of RepMiner to maize LTR retrotransposon analysis also illustrates the strength of a visually based multiple evidences approach to family assignment. While connected component clustering would have artificially included unrelated members in the same family of LTR retrotransposons [29], MCL based clustering is able to help deconvolute these instances. MCL clustering has already been shown to be robust to false edges in protein interaction graphs

[56,57], but this is the first application of this tool for deconvoluting chimeric annotations in transposable elements. These chimeric annotations could exist due to misassembles in the underlying sequence data, or could result for biological phenomena such as the propensity of some classes of transposable elements to insert into TE-rich regions of the genome.

RepMiner can also distinguish among families that were previously lumped together into a single family, as the *Huck* example illustrates. When canonical databases are used to assign newly discovered TEs to an existing nomenclature, simply taking the nearest neighbor BLAST match from the database, and assigning the newly discovered TE to that family can lead to false lumping of TEs into groups. Such nearest-neighbor based assignment will assign members of a previously undetected clade to a previously described sister clade. This appears to have been the case with the *Huck* family. As new sequence data became available for the maize genome, the family named *Puck* in this RepMiner analysis was simply assigned to its nearest neighbor BLAST hit. This assignment was then added to the canonical database, further propagating this improper assignment of family identity. As a number of online transposable element databases serve existing data without extensive curation of taxonomies scraped from other sources, this misidentification has been propagated to multiple online databases of transposable elements. The LTRs of the *Huck* and *Puck* groups are however more divergent from one another than the widely recognized *Ji, Opie* and *Giepum* families and it seems quite prudent to split *Huck* and *Puck* into different families.

The previous lumping together of the *Huck* and *Puck* families may also have missed the opportunity to discover interesting aspects of the biology of these families. The *Huck* and *Puck* families do share similar general genomic niches, as can be seen from the distribution of these families in the maize genome (Figure 3.6), but higher resolution information on these insertions

is required to specify the precise genomic context of the insertions. While the *Huck* family appears to has been recently active in the maize genome, the *Puck* family appears to have ceased being mobile more than one million years ago. A compelling interpretation of these data is that *Huck* and *Puck* represent sister clades of LTR retrotransposons, with *Puck* no longer being an active family in the maize genome, possibly as a result of competitive exclusion by *Huck*. Alternatively, something in the structure of current *Puck* members may have made them unable to encode or respond to trans-acting transposition factors that have continued to mobile *Huck* elements in recent times. The 57 bp *Huck* repeat unit could have played a role in the proliferation of *Huck* over *Puck*. The acquisition of this repeated unit in *Huck* is generally concurrent with the apparent loss of *Puck* activity in the maize genome (Figure 3.7), and this repeated unit is directly responsible for the observed increase in *Huck* LTR length (Figure 3.8). The possibility also exists that the longer *Huck* LTRs appear younger than they actually are if the longer LTRs have allow for increased gene conversion between LTRs. Gene conversion between LTRs has been shown to reduce the estimated time since insertion for primates and rodents, and LTR length could influence the probability of this gene conversion [58]. Regardless of the biological interpretation of these data, it is clear that a RepMiner approach to data visualization and interpretation greatly facilitated discovery and exploration of these trends and that the biology of this group warrants further study across additional maize genomes.

Future reexamination of TEs in sequenced genomes with RepMiner will allow us to describe the spectrum of diversity in transposable elements that can be identified by structure-based criteria. Even carefully curated database such as exist for rice TEs [59] have not adequately or comprehensively assigned family membership to low copy elements. The graph theory-based methods of clustering implemented in RepMiner will allow an exploration of the

biodiversity in these genomes that takes into account rare TEs as well as high-copy-number transposable elements. Also, since RepMiner is not intrinsically a repeat-based family assignment program, transposable elements from genomes that are only partially sequenced can be included in universal analysis of LTR retrotransposon diversity. The production of reduced exemplar databases using RepMiner tools will also facilitate these intergenomic analyses by reducing the large databases of transposable elements that can occur within genomes to a more manageable number of TEs that can be compared across genomes. The graph-based visualizations produced by RepMiner for these reduced intergenomic comparisons will further illustrate relationships among multiple genomes, and provide representative sequences for comprehensive phylogenies of plant LTR retrotransposons.

Acknowledgements

This development of the RepMiner tool benefited from helpful comments from Phillip SanMiguel and Evan Staton. This work was supported by NSF grants DBI-0501814, DBI-0607123, and DBI-0821263. This study was also supported in part by resources and technical expertise from the University of Georgia Research Computing Center, a partnership between the Office of the Vice President for Research and the Office of the Chief Information Officer.

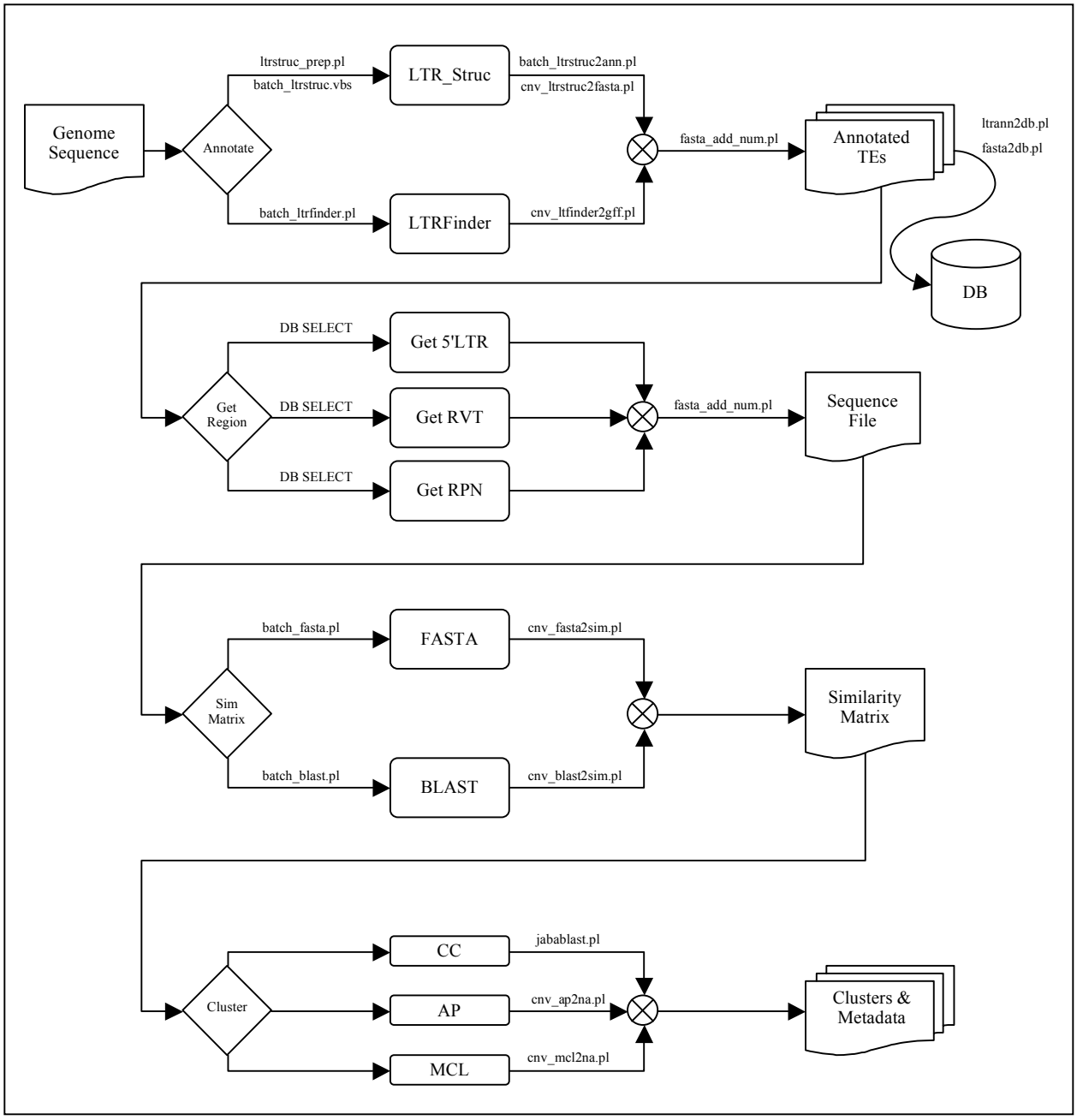


Figure 3.1. Workflow for LTR retrotransposon annotation and clustering using the RepMiner suite of programs. A diagram of possible workflows using the RepMiner suite and the specific programs used in each step is shown above. This suite of programs takes an unannoated genome sequence file as its intput, and uses this to generate a database of annotated LTR retrotransposons. The database is used to select regions of the LTR retrotransposon to use for clustering and produces a FASTA format sequence file. This sequence files is used to generate a similarity matrix using either FASTA or NCBI-BLAST. The similarity files can also be clustered using connected components clustering (CC), MCL clustering or affinity propagation-based clustering (AP). The files of clusters and metadata that are produced by this process can be visualized in the cytoscape network visualization program.

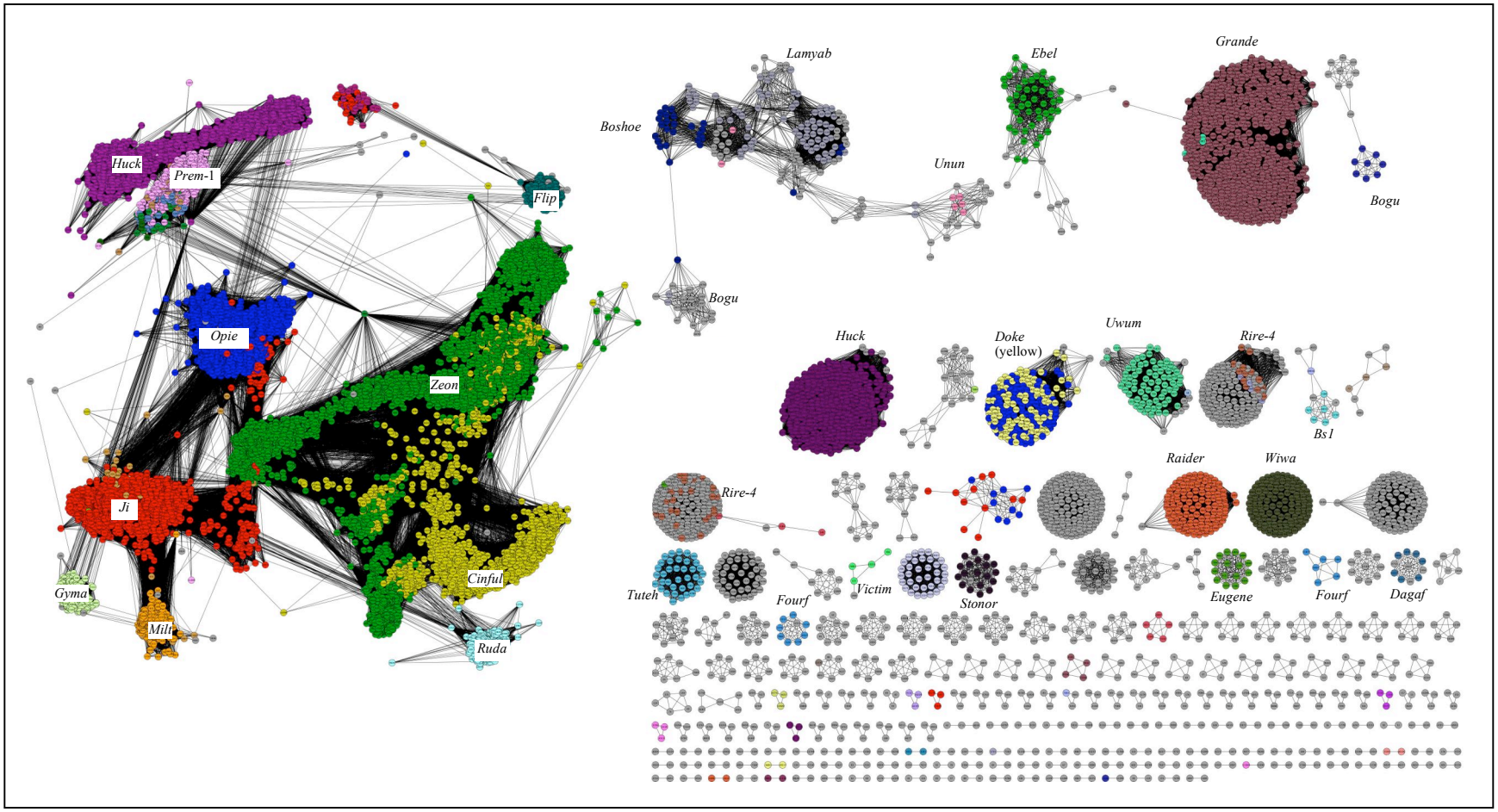


Figure 3.2. Previously named families mapped onto the network of full-length maize LTR retrotransposons. Nodes that are gray indicate previously unknown families. Relationships among the LTRs of *Cinful* and *Zeon* are easily visualized, and occasional chimeric LTRs can be seen to artificially connect some families. This result also illustrates that some families such as the previously assigned *Huck* group should be considered as more than one family.

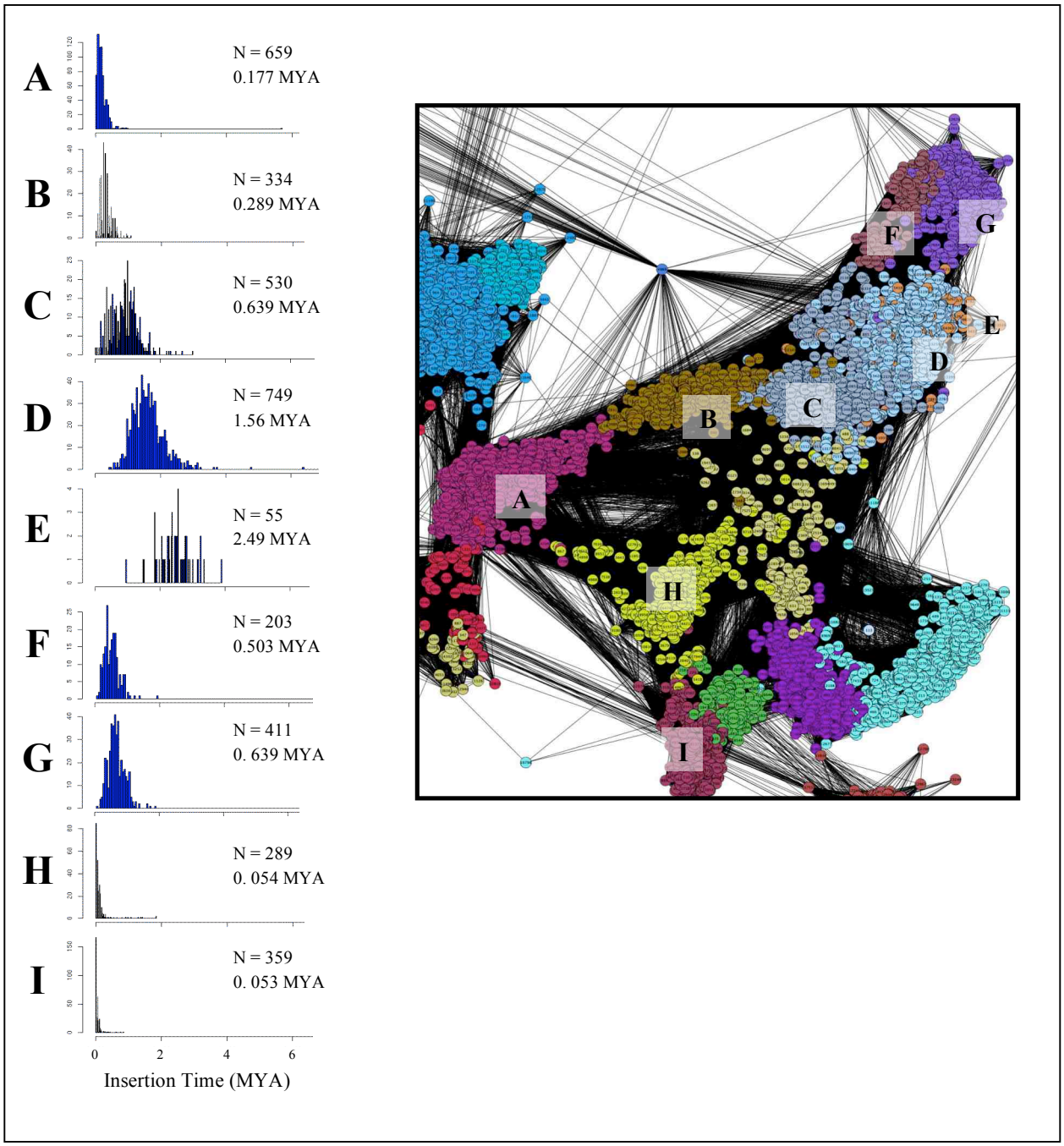


Figure 3.3. The insertion time distribution for MCL-identified clusters in the *Zeon* group of maize LTR retrotransposons. Colors on the network on the right indicate the MCL cluster that the node is assigned to, and the letter indicates the label of the MCL cluster. A histogram for the age of insertion was generated for the members of each cluster, and is shown on the left. The histograms include the number of LTR retrotransposons represented (N) and the median time of insertion is indicated in million years ago (MYA).

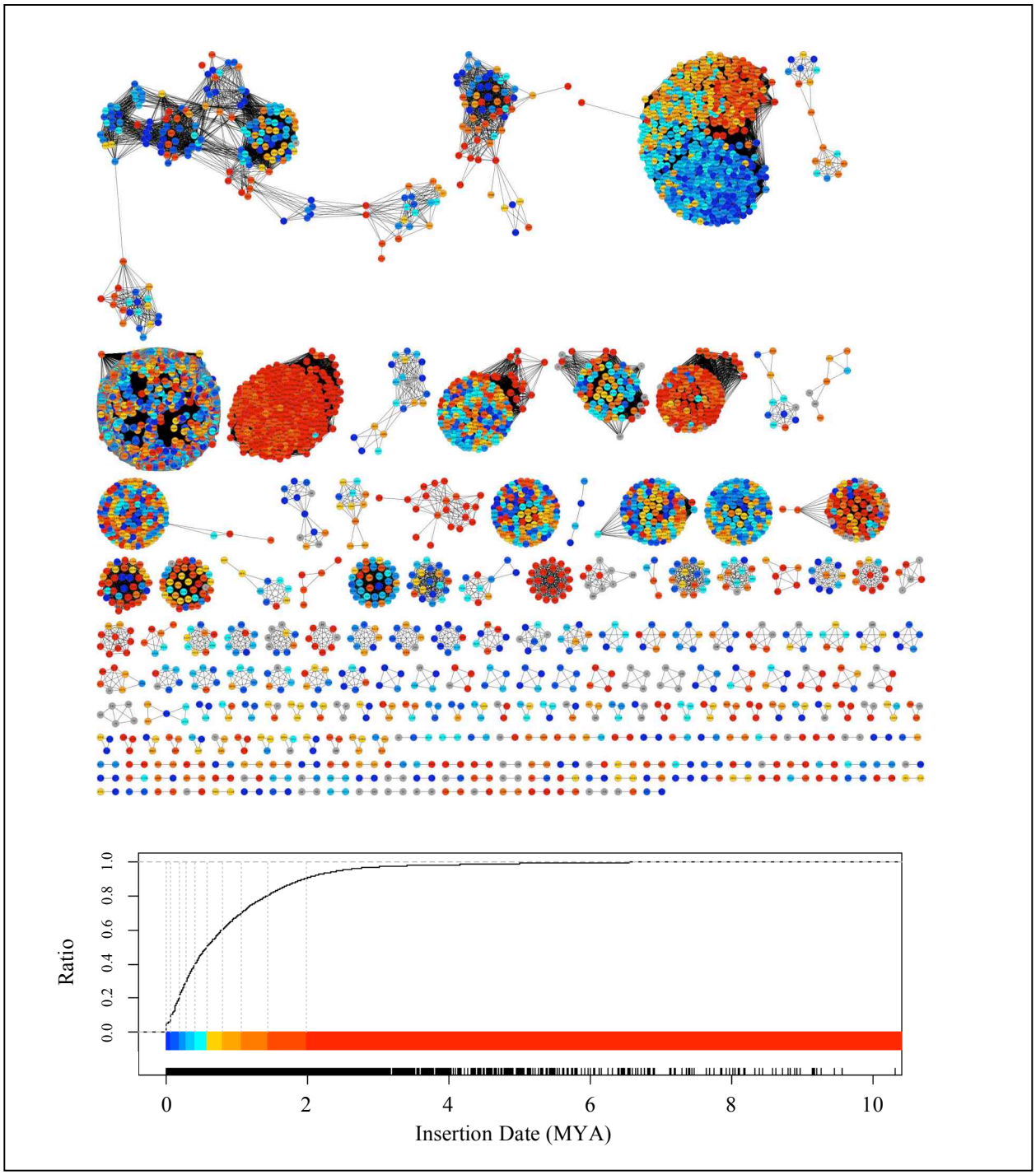


Figure 3.4. Mapping of the time since insertion onto the entire network of full-length maize LTR retrotransposons. The network above shows the mapping of insertion date as a color heatmap onto the network of full-length maize LTR retrotransposons. The graph below is the empirical cumulative distribution of all of the estimated insertion dates, with the color assigned to the dates indicated in along the bottom of the graph. Colors are assigned by sorting the distribution of insertion dates into 10 equal sized quantiles, with each quantile containing the same number of LTR retrotransposons.

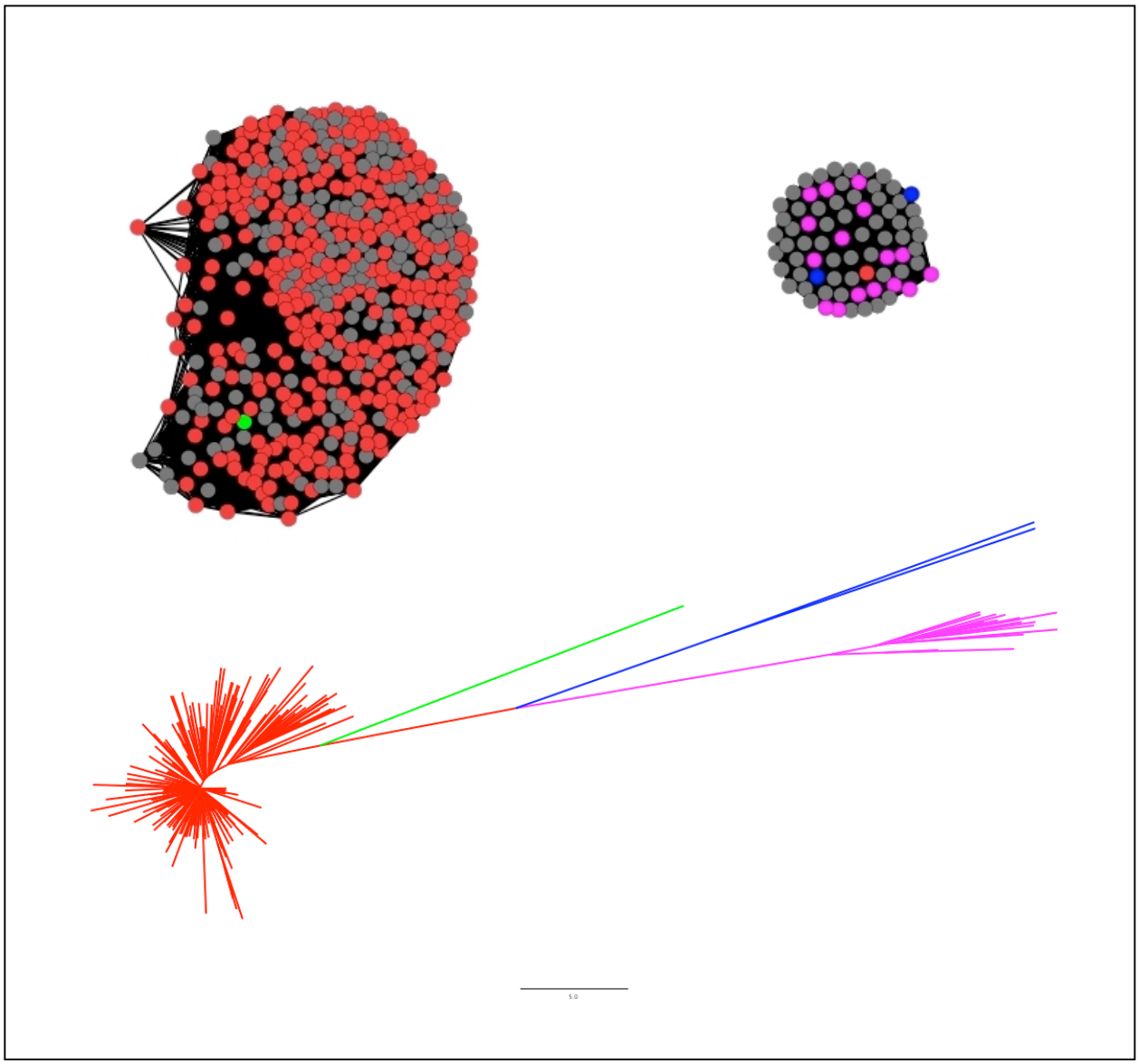


Figure 3.5 Phylogeny of reverse transcriptase of representative *Huck*-like sequences in the maize genome. The RepMiner-derived clustering of the 5' LTRs of the *Huck*-like sequences are shown in the graph clusters above. The LTR retrotransposons containing a reverse transcriptase sequence that could be aligned are indicated as colored nodes on this network. These same colors are used to indicate branches in the unrooted phylogeny of these reverse transcriptase loci.

Figure 3.6. Genomic distribution of *Huck* and *Puck* elements on the 10 chromosomes of the maize genome. The location of full length insertions of the three MCL clusters within the *Huck* family are shown (A,B,C) as are the full length insertions of the two MCL clusters within the *Puck* family (D). Chromosomes are drawn as ideograms with gray banding representing staining patterns, and the red triangles represent individual insertions on the 10 maize chromosomes.

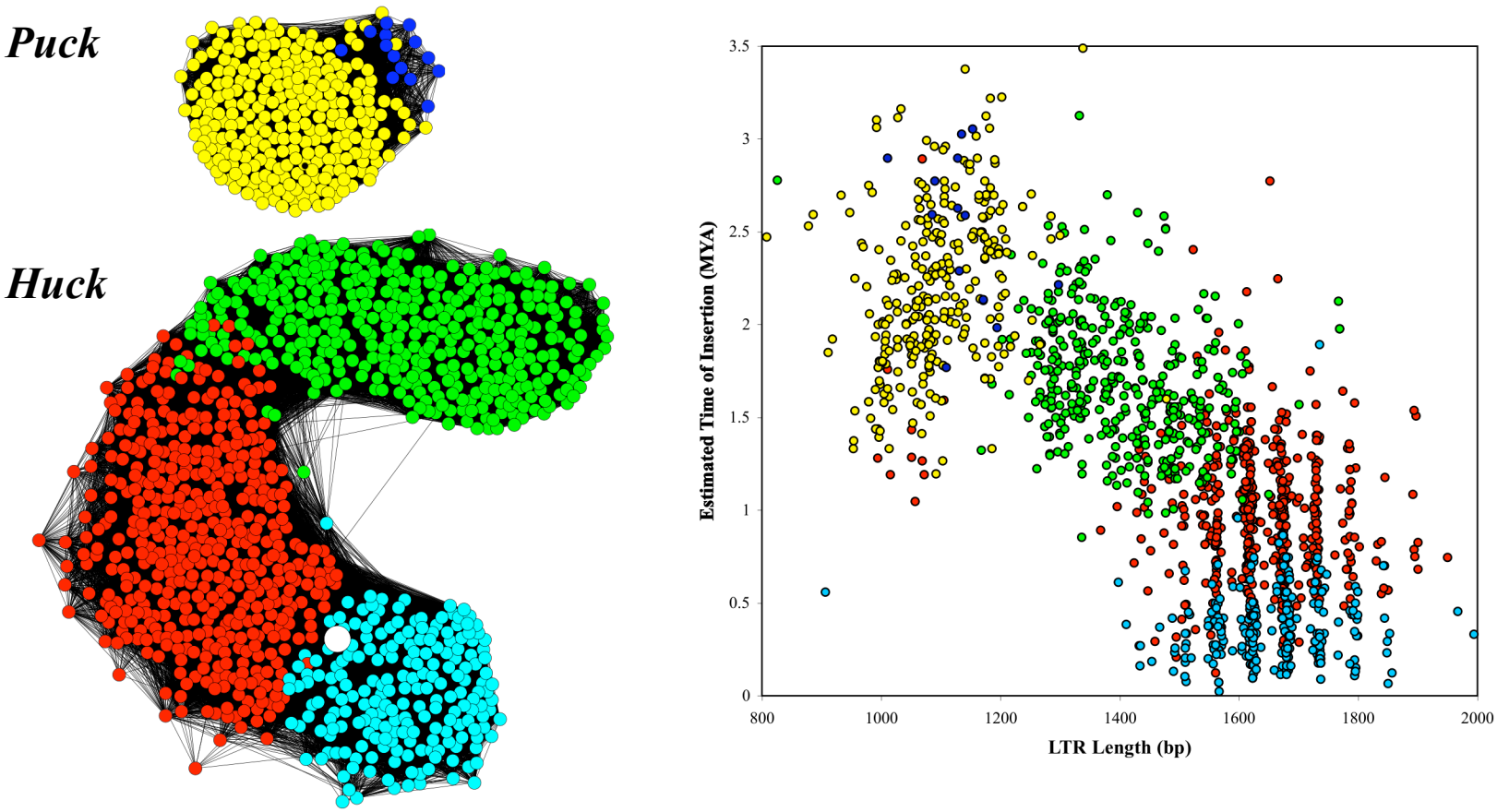


Figure 3.7. MCL clusters of *Huck* and *Puck* correspond to trends in LTR length and age in the two families. Colored nodes in the network of LTRs indicate the three MCL clusters within the *Huck* family and the two MCL clusters within the *Puck* family. The sequence from the original description of the *Huck* family [46] is indicated at the large white node in the *Huck* network. The age of insertion and length of the 5' LTR are illustrated in the graph on the right.

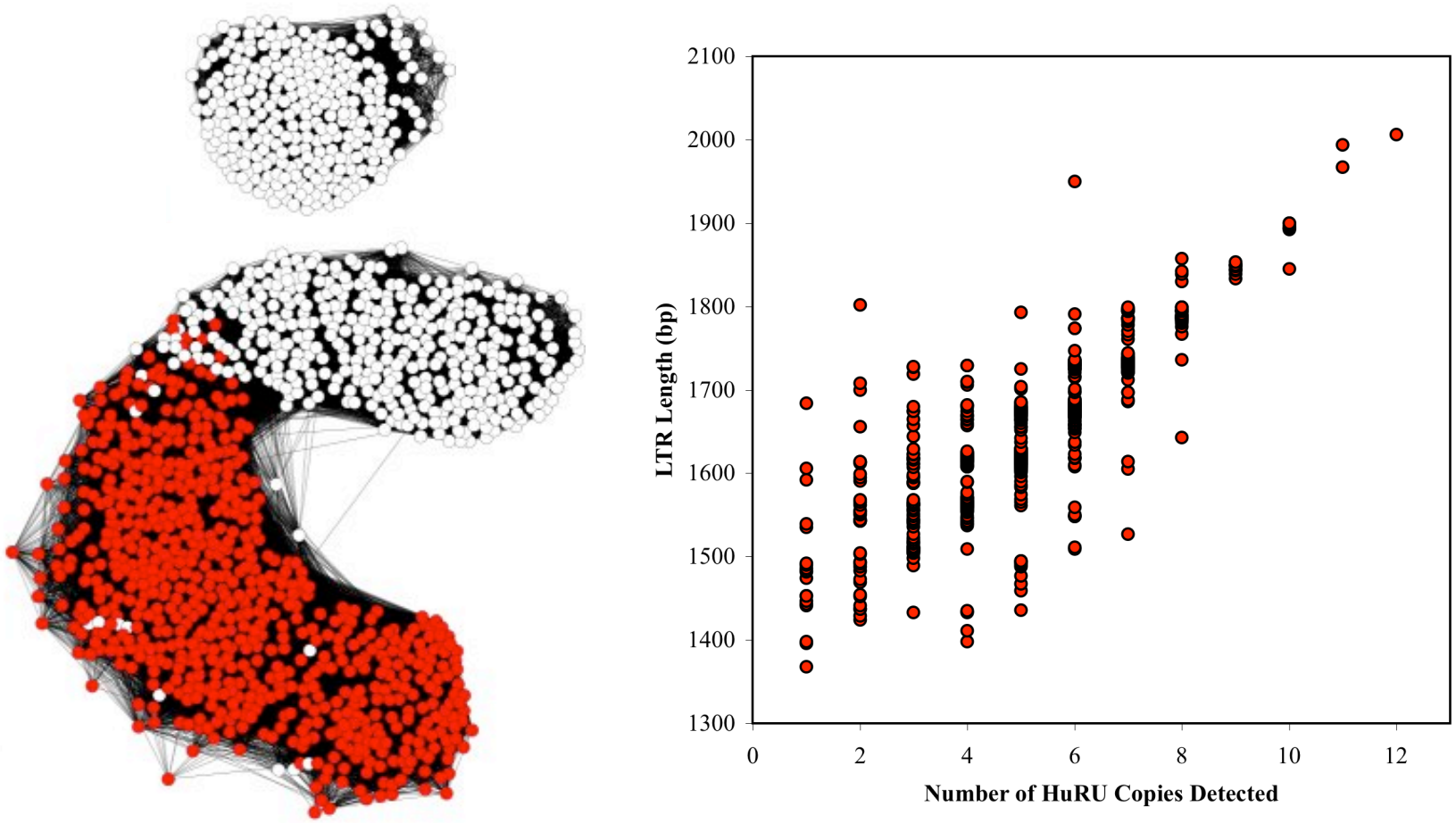


Figure 3.8. The 57 bp *Huck* Repeat Unit is largely responsible for increase in LTR length in the *Huck* family. The red nodes on the network of the *Huck* and *Puck* represent the nodes with a detectable occurrence of the *Huck* Repeat Unit in the LTR retrotransposons as detected by vmatch. The plot on the right represents the relationship between the number of copies of HuRU detected and the length of the LTR for these nodes.

REFERENCES

1. Bennetzen JL (2000) Transposable element contributions to plant gene and genome evolution. Plant Mol Biol 42: 251-269.
2. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, et al. (2009) The B73 maize genome: complexity, diversity and dynamics Science 326: 1112-1115.
3. Dooner HK, He L (2008) Maize genome structure variation: interplay between retrotransposon polymorphisms and genic recombination. Plant Cell 20: 249-258.
4. Xing J, Zhang Y, Han K, Salem AH, Sen SK, et al. (2009) Mobile elements create structural variation: Analysis of a complete human genome. Genome Res 19: 1516-1526.
5. Wright SI, Le QH, Schoen DJ, Bureau TE (2001) Population dynamics of an *Ac*-like transposable element in self- and cross-pollinating arabidopsis. Genetics 158: 1279-1288.
6. Lockton S, Ross-Ibarra J, Gaut BS (2008) Demography and weak selection drive patterns of transposable element diversity in natural populations of *Arabidopsis lyrata*. Proc Natl Acad Sci U S A 105: 13965-13970.
7. Kidwell MG, Lisch D (1997) Transposable elements as sources of variation in animals and plants. Proc Natl Acad Sci U S A 94: 7704-7711.
8. Gogvadze E, Buzdin A (2009) Retroelements and their impact on genome evolution and functioning. Cell Mol Life Sci.
9. Weil CF (2009) Too many ends: aberrant transposition. Genes Dev 23: 1032-1036.
10. Britten RJ, Kohne DE (1968) Repeated sequences in DNA. Science 161: 529-&.
11. Bairoch A (1991) PROSITE: a dictionary of sites and patterns in proteins. Nucleic Acids Res 19 Suppl: 2241-2245.
12. Attwood TK, Beck ME, Bleasby AJ, Parry-Smith DJ (1994) PRINTS--a database of protein motif fingerprints. Nucleic Acids Res 22: 3590-3596.
13. Barker WC, Pfeiffer F, George DG (1996) Superfamily classification in PIR-International Protein Sequence Database. Methods Enzymol 266: 59-71.
14. Corpet F, Gouzy J, Kahn D (1998) The ProDom database of protein domain families. Nucleic Acids Res 26: 323-326.
15. Krause A, Nicodeme P, Bornberg-Bauer E, Rehmsmeier M, Vingron M (1999) WWW access to the SYSTERS protein sequence cluster set. Bioinformatics 15: 262-263.
16. Ponting CP, Schultz J, Milpetz F, Bork P (1999) SMART: identification and annotation of domains from signalling and extracellular protein sequences. Nucleic Acids Res 27: 229-232.
17. Huang H, Xiao C, Wu CH (2000) ProClass protein family database. Nucleic Acids Res 28: 273-276.
18. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, et al. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res 29: 37-40.
19. Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, et al. (2001) TIGRFAMs: a protein family resource for the functional identification of proteins. Nucleic Acids Res 29: 41-43.
20. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, et al. (2003) PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 13: 2129-2141.
21. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, et al. (2004) The Pfam protein families database. Nucleic Acids Res 32: D138-141.

22. Hulo N, Bairoch A, Bulliard V, Cerutti L, Cuche BA, et al. (2008) The 20 years of PROSITE. Nucleic Acids Res 36: D245-249.
23. Dongen SMv (2000) Graph Clustering by Flow Simulation. Amsterdam: University of Utrecht. 169 p.
24. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30: 1575-1584.
25. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, et al. (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8: 973-982.
26. Agarwal P, States DJ (1994) The Repeat Pattern Toolkit (RPT): analyzing the structure and evolution of the *C. elegans* genome. Proc Int Conf Intell Syst Mol Biol 2: 1-9.
27. Bao Z, Eddy SR (2002) Automated *de novo* identification of repeat sequence families in sequenced genomes. Genome Res 12: 1269-1276.
28. Pevzner PA, Tang H, Tesler G (2004) *De novo* repeat classification and fragment assembly. Genome Res 14: 1786-1796.
29. Kim S, Lee J (2006) BAG: a graph theoretic sequence clustering algorithm. Int J Data Min Bioinform 1: 178-200.
30. Rho M, Choi JH, Kim S, Lynch M, Tang H (2007) *De novo* identification of LTR retrotransposons in eukaryotic genomes. BMC Genomics 8: 90.
31. Chen Y, Zhou F, Li G, Xu Y (2009) MUST: a system for identification of miniature inverted-repeat transposable elements and applications to *Anabaena variabilis* and *Haloquadratum walsbyi*. Gene 436: 1-7.
32. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498-2504.
33. Hu Z, Snitkin ES, DeLisi C (2008) VisANT: an integrative framework for networks in systems biology. Brief Bioinform 9: 317-325.
34. Frey BJ, Dueck D (2007) Clustering by passing messages between data points. Science 315: 972-976.
35. Leone M, Sumedha, Weigt M (2007) Clustering by soft-constraint affinity propagation: applications to gene-expression data. Bioinformatics 23: 2708-2715.
36. Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5: 101-113.
37. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, et al. (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome Res 12: 1611-1618.
38. Foundation FS (2010) http://www.gnu.org/prep/standards/.
39. Estill JC, Bennetzen JL (2009) The DAWGPAWS pipeline for the annotation of genes and transposable elements in plant genomes. Plant Methods 5: 8.
40. Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, et al. (2009) Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. PLoS Genet 5: e1000732.
41. McCarthy EM, McDonald JF (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics 19: 362-367.
42. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. Nature Genetics 20: 43-45.

43. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-4680.
44. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586-1591.
45. Ma J, Bennetzen JL (2004) Rapid recent growth and divergence of rice nuclear genomes. Proc Natl Acad Sci U S A 101: 12404-12410.
46. SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, et al. (1996) Nested retrotransposons in the intergenic regions of the maize genome. Science 274: 765-768.
47. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797.
48. Stamatakis A, Ludwig T, Meier H (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21: 456-463.
49. SanMiguel P, Vitte C (2008) The LTR-retrotransposons of maize. In: Bennetzen JL, Hake SC, editors. Handbook of Maize: Genetics and Genomics Springer.
50. Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35: W265-268.
51. DeBarry JD, Liu R, Bennetzen JL (2008) 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: 235.
52. Sanz-Alferez S, SanMiguel P, Jin YK, Springer PS, Bennetzen JL (2003) Structure and evolution of the *Cinful* retrotransposon family of maize. Genome 46: 745-752.
53. Arcot SS, Shaikh TH, Kim J, Bennett L, Alegria-Hartman M, et al. (1995) Sequence diversity and chromosomal distribution of "young" *Alu* repeats. Gene 163: 273-278.
54. Arcot SS, Adamson AW, Lamerdin JE, Kanagy B, Deininger PL, et al. (1996) *Alu* fossil relics--distribution and insertion polymorphism. Genome Res 6: 1084-1092.
55. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.
56. Brohee S, van Helden J (2006) Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics 7: 488.
57. Vlasblom J, Wodak SJ (2009) Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. BMC Bioinformatics 10: 99.
58. Kijima TE, Innan H (2010) On the estimation of the insertion time of LTR retrotransposable elements. Mol Biol Evol 27: 896-904.
59. Chaparro C, Guyot R, Zuccolo A, Piegu B, Panaud O (2007) RetrOryza: a database of the rice LTR-retrotransposons. Nucleic Acids Res 35: D66-70.

CHAPTER 4

AFFINITY PROPAGATION CLUSTERING EFFICIENTLY AND EFFECTIVELY DEFINES REPRESENATIVE EXEMPLARS FROM LARGE MOLECULAR SEQUENCE DATABASES OF TRANSPOSABLE ELEMENTS[1]

[1] Estill, J.C., and J.L. Bennetzen. To be submitted to *BMC Bioinformatics*.

Abstract

A common need in the curation of large molecular sequence databases is the generation of representative subsets of the entire database of sequences. These needs occur for protein sequence databases as well as databases of transposable elements. One common solution to this problem is to group sequences into multiple alignments and provide a consensus sequence for this alignment. This sequence represents a putative ancestral sequence for this group and can generally be used to represent the derived sequences within this group, or the multiple alignment can be used to generate a profile hidden Markov model of this multiple alignment. Another approach is to generate a database of representative exemplars from the overall dataset. This differs from consensus-based methods in that extant representatives are used as exemplars of the overall dataset. The affinity propagation based method for exemplar assignment uses a graph theory based approach to find representative nodes from a large network of similarity among groups of objects. We apply this approach to find representative sequences from a large database of maize LTR retrotransposons, and discuss how this approach can be used to generate a canonical database suitable for repeat identification and masking of transposable elements.

Introduction

A commonly required outcome in working with large sequence datasets is the generation of a non-redundant dataset that adequately represents all other members of the set. Historically, this has been achieved by defining a consensus sequence for all members of groups within the larger dataset, or by providing extant representatives that can serve as representative exemplars of their group. For example, the PFAM seed sequences [1] represent a database of exemplar sequences or other members of a PFAM group. The exemplar can be used to generate multiple

alignments and profile hidden Markov models (pHMMs) of the multiple alignments [2] for pHMM-based searches of the database. These exemplars may also be used to assign newly annotated protein genes to families using a simple local alignment program such as BLAST.

The identification of non-redundant representative sets is a particularly important need in the study of transposable elements (TEs). TEs are mobile genetic elements that are ubiquitous among eukaryotic genomes [3]. These elements can have profound impacts on host genome architecture [4] and gene evolution [5]. In plants, these elements are largely responsible for the great variation in genome size [6] and, in many grass species such as maize, they can comprise over 85% of the genome [7]. TEs are are categorized into the Class I retrotransposons that propagate through the use of an RNA intermediate and the Class II DNA transposons that replicate via a direct cut and paste of the DNA molecule [8]. These larger taxonomic categories are variously divided into subsets using both structural and phylogenetic criteria [9]. The recently-sequenced genome of B73 maize [7] has permitted comprehensive discovery of the LTR retrotransposons in a complex angiosperm genome, and they have been found to contribute over 75% of the nuclear DNA [10]. Four LTR retrotransposon families individually contribute more DNA to the maize genome than the size of the entire Arabidopsis genome, although the majority of elements occur at low to moderate copy numbers [10]. The full set of annotated LTR retrotransposon elements in maize is one of the largest datasets of TEs in any genome sequenced to date, and as such represents a useful case study for the generation of nonredundant datasets from the larger comprehensive databases of elements.

Many of the existing tools used to annotate the repetitive fraction of the genome make use of canonical databases that are representative of the full TE diversity of the genome. The tools include RepeatMasker [11], Censor [12] and TENest [13] which process local alignment

queries against databasea of characterized repeat sequences. In general, the goals of these tools are to both "mask" the genome of TE DNA to facilitate annotation of other genome features and to assign these masked regions to individual families. These searches are intrinsically limited by the quality and comprehensiveness of the canonical database used for querying. The canonical TE databases much be free of chimeric annotations to assure high quality taxonomic assignment of the newly masked TE regions, and they should ideally be nonredundant to allow searches to proceed in a manageable amount of time.

A number of existing TE databases have attempted to generate non-redundant databases using *ad hoc* methods. These include the TREP database of Triticeae repeats which makes the use of manual methods to generate non-redundant datasets [14] as well as the RepBase [15] database which attempts to curate non-redundant sets of TEs. The MIPs REcat database (http://mips.gsf.de/proj/plant/webapp/recat/) generates non-redundant sets using a threshold value approach that removes identical repeats at a 98% identity limit and takes the longest sequence as a representative. However, this 98% value is arbitrary, and the use of the longest sequences as a representative could select the sequences that are most likely to be contaminated with heterologous insertions or chimeric assemblies.

The affinity propagation algorithm (AP) has been developed to find representative subsets from a larger dataset of objects [16]. This program takes as its input a similarity matrix for the objects in the dataset, as well as a vector of preferences establishing the probability that each individual object would be selected as a representative. The output of the program is the association of each object to a cluster set, and the assignment of a representative object for each cluster. This algorithm does not require *a priori* knowledge of the number of clusters desired, and can thus provide the number of objects needed to represent the overall dataset as an outcome

of the algorithm. For large datasets of over 1000 objects, AP can define these clusters with a lower error rate than alternative clustering algorithms such as a vertix substitution heuristic [17,18]. Furthermore, AP can discover these clusters in computational times that are orders of magnitude faster than existing approaches [18]. For example, AP can cluster a dataset of over 17,000 Netflix movies in 2 hours versus the 10 days needed for alternative approach that yields an inferior result [18]. The AP algorithm has already been applied to a broad domain of data clustering scenarios, including finding representative faces from large image collections [19], locating the optimal placement of transportation infrastructure [16] and the clustering of gene expression data [20].

Here, we propose using affinity propagation-based clustering to generate non-redundant exemplar databases of transposable elements. We will specifically use this method to generate a non-redundant dataset of LTR retrotransposons in maize suitable for use for TE annotation and masking with the RepeatMasker program [11]. We will compare the results from repeat masking with affinity propagation-based clusters to randomly selected sequences and to MCL clustering-based sequences.

Materials and Methods

LTR Retrotransposon Discovery in the B73 Maize Genome

The protocols used to annotate the LTR retrotransposon in maize have been described elsewhere [10]. Generally the DAWGPAWS suite of scripts (http://dawgpaws.sourceforge.net) was used to generate a dataset of full length LTR retrotransposons for the maize genome as annotated by the program LTR_STRUC [21]. These results were further structurally characterized for biological features using the *cnv_ltrstruc2ann.pl* program of the RepMiner

package (Chapter 3), and these results were loaded into an MS Access database to facilitate further analysis. These annotated sequences were then sorted into families using the program RepMiner, as described previously [10].

Generating the Similarity Matrix

The 5' LTR sequence for each LTR retrotransposon was exported from the MS Access database and used to generate a FASTA format file. This FASTA file was formatted with formatdb to generate a query database, and the input file was searched against this database to generate an "all-by-all" similarity matrix. This search was conducted using NCBI BLASTn version 2.2.23 with an e-value threshold $e < 1x10^{-5}$ and all significant hits were returned. The resulting BLAST report was converted to a similarity matrix using the *cnv_blast2sim.pl* program from RepMiner (http://sourceforge.net/projects/repminer/), with the tiled bitscore of the high scoring segment pairs (HSPs) as the metric of similarity.

Generating Representative Sequence Databases

Sets of representative sequences were generated under three protocols: 1) random sequence selection, 2) selection of MCL clustering attractors, and 2) selection of affinity propagation exemplars. For the randomly selected sequence set, sequences were randomly chosen without replacement from the larger MS access database, and these sequence selections were used to generate FASTA file sets for repeat masking that contained sequence sets in size intervals of 10 for sets less than 100 sequences and for size intervals of 100 for sets with more than 100 sequences. This was used to generate a range of sequence sets for repeat masking that ranged from 10 randomly selected sequences to 1300 randomly selected sequences (Table 4.1).

As an alternative to AP clustering, sets of representative sequences of various sizes were generated from version 1.007 of the MCL clustering program [22] by selecting variable values

for the inflation parameter [22]. The inflation parameter values that were used ranged from 1 to 5, with larger values generating more clusters in the resulting dataset. The central attractor sequence for each cluster set was used as the representative sequence for each cluster. The inflation values used to generate clusters and the resulting numbers of sequences selected are indicated in Table 5.1.

Affinity propagation-based clustering used the sparse matrix implementation of AP functions implemented in MATLAB (v 7.5.0.338) that are available from the Brendan Frey laboratory website [23]. With AP clustering, the number of clusters and exemplars returned from a dataset is a result of the choice of the vector of availabilities [24]. This availability represents the "willingness" of each sequence to serve as a representative, and each sequence has its own individual availability value. Different availabilities can be set for each sequence such that some sequences are more likely to be selected than others, or all sequences can be assigned the same availability. These availability values are represented in the same units as the data and can range from negative infinity to infinity. Higher availability values will generate more clusters and exemplars.

For the use of AP to generate a representative set of sequences, the same availability value for all sequences in the dataset was employed. The set of availability values used to generate sets of sequence for repeat masking was generally derived from the distribution of tiled HSP bitscores in the similarity matrix. The median value of the bitscore was generally used. Selected values in standard deviation steps above this median value and below this median value were also employed, with the goal of generating a range of exemplar sequence counts that were within the same range as the number of clusters selected by the MCL program. The availability

values that were used and the numbers of sequences selected by these values are indicated in Table 4.1.

Repeat Masking with Representative Sequence Sets

The ability of the representative subset of sequences to represent the overall database of TEs was assessed by masking the entire dataset of sequences using the individual representative subsets. This masking was done with RepeatMasker (version 3.2.6) using the WuBLAST engine [11]. The general efficacy of the masking was assessed as the count of the masked residues divided by the total number of residues in the full dataset of over 123 million bases. The masking reports generated by RepeatMasker were not used for this purpose because they contain spatial overlaps that would lead to an overestimation of the masking efficacy. Instead, the masked bases were directly counted in the masked FASTA files using custom PERL scripts. These results are summarized graphically in Figure 4. as the ratio of total sequences masked in the masked ratio column of Table 4.1. The count of the number of sequences that were not even partially masked in the dataset was recorded, as was the time required to complete each RepeatMasker job.

Visualizing Exemplars on the Network of Maize LTR Retrotransposons

An "All-By-All" BLAST matrix of the 5' LTRs of the full database of LTR retrotransposons was visualized as a network using RepMiner (Chapter 3). The 463 sequences selected by affinity propagation clustering resulting from an availability of A = -19,196 were then mapped onto this network to allow for a visualization of the distribution of affinity propagation-based exemplars in the entire database. This visualization is shown in Figure 4.2.

Results

The full results for the Repeat Masking using the different types of representative sequences are shown in Table 4.1 and the percent of the families masked are illustrated in Figure 4.1. These results show that, for small numbers of sequences in the masking database (< 400), the exemplars for AP clearly outperformed equivalently sized databases of randomly selected sequences or MCL selected sequences. The smallest database that Affinity Propagation would select is 140 sequences, and this masking database was able to mask nearly 95% of the bases in the comprehensive dataset. Only 20 sequences were not masked at all in this masking result. This masking database of 140 sequences could mask the comprehensive dataset in a little over five hours of compute time making it one of the most efficient sets used. However, since there are over 400 families of LTR retrotransposons in the maize genome and only 140 sequences in the masking database, the database would not correctly assign family identity even though it would recognize these sequences as LTR retrotransposons.

Increasing the affinity propagation-based masking database size to 6,745 sequences allows 99.75% of the bases to be masked. However this result is still not comprehensive, with 11 LTR retrotransposons remaining completely unmasked. This moderate increase in masking percent comes at the expense of a 70 times increase in computational time required, to over two weeks of required masking time.

The sequences that were not masked by the affinity propagation-based exemplars were generally single copy families with a small size and lacking internal protein coding domains. Interestingly, the MCL-selected sequences did a better job in generating a more taxonomically comprehensive coverage of the transposable elements. When selecting 391 sequences for a masking database, the MCL masking completely missed only 8 sequences. However, this came

at the cost of overall masking efficacy since the sequences in this dataset only identified 77% of the overall dataset of transposable elements. It would thus appear that the MCL algorithm is optimizing at least partial coverage of the largest number of sequences, while not maximizing the total number of bases that are masked.

Discussion

It is possible that an exhaustive exact search for representative sequences could generate a masking database that outperforms the results for affinity propagation-based clustering. However, the computational time required for such an exact search would be prohibitive, and the user would be required to have *a priori* knowledge of an effective database size to set as the target search size. The affinity propagation-based approach is a heuristic method that quickly identifies a highly quality dataset suitable for repeat masking without any *a priori* knowledge required [24]. The number of exemplars that adequately represent a database are an outcome of the affinity propagation-based approach, in addition to the database created. Therefore, the numbers of exemplars created from a dataset of transposable elements is a result that can be compared among genomes as a measure of comparative TE diversity. Newly sequenced genomes without a prior study of TE diversity could thus quickly be compared to the diversity of previously studies genomes. This makes affinity propagation similar to unsupervised approaches such as MCL for diversity discovery in that the number of clusters is an outcome of the clustering method and not a required input.

An advantage of affinity propagation over other methods of unsupervised clustering is that the units used to generate preference values are the same as the values in the similarity matrix. For example, if the similarity values are bitscores, than the preference values are

bitscores. This allows preference values to be anchored in observable parameters. Values used for preferences can therefore be extracted from the distribution of empirical bitscore values that exist in the similarity matrix. For example, one could choose to always use the median value from the set of bitscores to serve as the preference parameter. The same references points in the distribution of values in the similarity matrices could be used across genomes and allow for reasonably comparable results. This is an advantage over MCL clustering, which makes use of an "inflation" value that is difficult to interpret in terms of the current understanding of the biology of a given system being clustered.

The vector of preference values used by affinity propagation also allows for a fine tuned selection of individual preferences for each sequence. While all sequences can be assigned the same preference value, as was done in this manuscript, individual sequences could be assigned a unique preference value based on *a priori* knowledge of the taxonomy of the element. For example, MCL clustering results could be used as input to set the preference values to the median similarity score for a given MCL cluster or a named taxonomic group. Also, since sequences can be assigned unique preference values, sequences that the user does not want to be available for selection as an exemplar can be assigned a preference value of negative infinity. This could be used to avoid selection of known chimeric sequences, while still assigning them to an appropriate cluster. Alternatively, sequences that the user wishes to have a high probability of being selected can have a proportionally higher preference value assigned to them. For example, if the user is extending an already established database of representative sequences, the currently identified exemplars could be assigned a high preference value that would insure that they would be selected as exemplars in a new round of exemplar discovery. Furthermore, if one wanted to generate a phylogeny of exemplar TEs across genomes based on a particular protein coding

region, it is also possible to set higher preferences values for sequences containing these coding regions than for sequences that lack coding regions.

The strengths of using affinity propagation clustering to define representative databases for repeat masking are directly demonstrated in this study. However, one disadvantage of this approach is that it may not yield taxonomically comprehensive databases when selecting for small databases. For example, the database of 140 exemplar sequences does an excellent job masking the comprehensive dataset (Table 4.1), but it cannot fully represent the taxonomy of LTR retrotransposons in a maize genome that includes over 400 families. It is therefore useful to consider taxonomic representation when generating a database to use for repeat identification and not just repeat masking. In generating an exemplar database for use in annotating the maize genome, we selected the affinity propagation-based exemplar set that included all families of LTR retrotransposons. While this increases the required time to mask the database for a minimal increase in the proportion of the database that is masked, it ensures that masked elements are properly assigned to the correct family.

Table 4.1 Masking results for representative sequence datasets. Parameter values indicate the number of sequences selected (n), the MCL inflation parameter (i), or the vector of availabilities for affinity propagation (A). The numbers of representative sequences in the masking database are indicated, and the computational time required to run the masking job on the sequence database is indicated in hours. The number of unmasked sequences is the count of sequences that were not at least partially masked.

Method	Parameter Values	Representative Sequences	Time (Hours)	Masked Ratio	Unmasked Sequences
Random Selection	$n = 10$	10	2.5	0.5984	1066
	$n = 20$	20	3.0	0.7995	764
	$n = 30$	30	3.8	0.7984	789
	$n = 40$	40	4.5	0.8159	588
	$n = 50$	50	4.8	0.7894	484
	$n = 60$	60	5.6	0.8580	445
	$n = 70$	70	6.2	0.8490	617
	$n = 80$	80	6.9	0.8705	464
	$n = 90$	90	7.3	0.8749	339
	$n = 100$	100	7.7	0.9094	378
	$n = 200$	200	13.3	0.9337	203
	$n = 300$	300	20.4	0.9470	241
	$n = 400$	400	26.0	0.9523	141
	$n = 500$	500	32.5	0.9598	145
	$n = 600$	600	38.3	0.9630	105
	$n = 700$	700	45.2	0.9649	93
	$n = 800$	800	49.4	0.9663	93
	$n = 900$	900	55.3	0.9665	62
	$n = 1000$	1000	61.9	0.9683	79
	$n = 1100$	1100	70.6	0.9677	86
	$n = 1200$	1200	75.6	0.9709	78
MCL	$i = 1.1$	141	5.6	0.4826	2047
	$i = 1.2$	279	16.2	0.7123	50
	$i = 1.3$	391	19.6	0.7698	8
	$i = 1.4$	449	21.3	0.9413	8
	$i = 1.5$	481	22.0	0.9421	6
	$i = 1.6$	503	22.8	0.9460	6
	$i = 1.7$	520	23.1	0.9483	6
	$i = 1.8$	532	23.9	0.9517	0
	$i = 1.9$	548	24.3	0.9554	0
	$i = 2.0$	566	25.2	0.9618	0
	$i = 3.0$	688	31.5	0.9691	0
	$i = 4.0$	765	37.8	0.9703	0
	$i = 5.0$	832	44.6	0.9724	0
Affinity Propagation	$A = -4{,}312{,}406$	141	5.3	0.9458	20
	$A = -2{,}156{,}001$	140	5.6	0.9463	20
	$A = -1{,}293{,}439$	141	5.7	0.9399	21
	$A = -862{,}158$	144	6.1	0.9478	20
	$A = -430{,}877$	160	7.5	0.9565	20
	$A = -172{,}110$	180	9.6	0.9637	21
	$A = -128{,}980$	196	10.8	0.9653	22
	$A = -85{,}852$	226	12.5	0.9677	22
	$A = -42{,}724$	304	17.6	0.9716	22
	$A = -38{,}797$	323	19.0	0.9722	22
	$A = -19{,}196$	463	27.4	0.9751	25
	$A = -9{,}396$	692	41.4	0.9794	23
	$A = -5{,}476$	888	52.0	0.9812	21
	$A = -3{,}516$	1028	60.6	0.9821	23
	$A = -1{,}556$	1212	70.5	0.9839	19
	$A = 71$	1454	83.6	0.9857	19
	$A = 404$	1617	86.8	0.9875	13
	$A = 1383$	1990	100.2	0.9893	12
	$A = 2364$	2526	119.8	0.9920	13
	$A = 4324$	4281	213.8	0.9960	11
	$A = 6284$	6745	387.0	0.9975	11

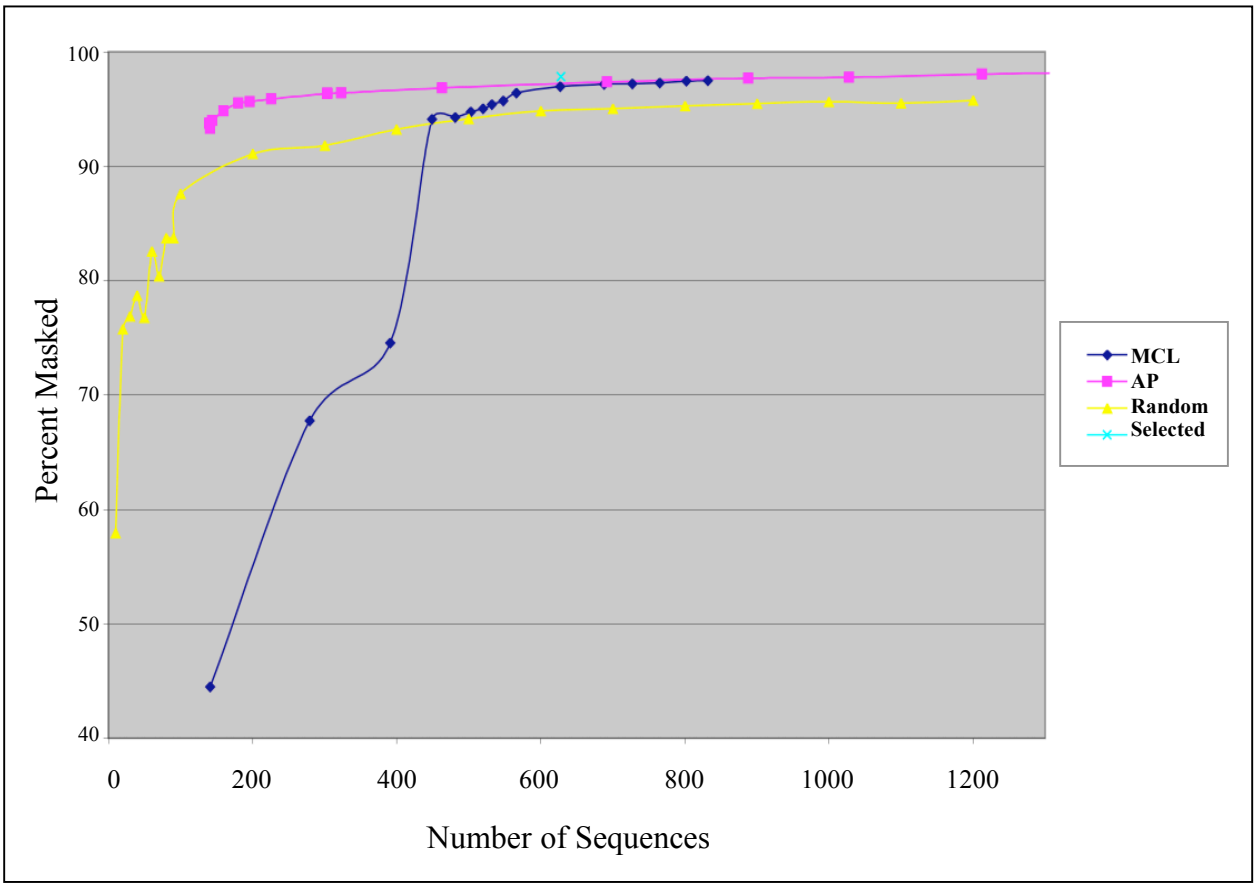


Figure 4.1 The masking efficacy of the different representative sequence sets presented as the percent of total bases masked in the dataset of all identified LTR retrotransposons. The results for MCL clustering, affinity propagation clustering (AP) and randomly selected sequences are shown.

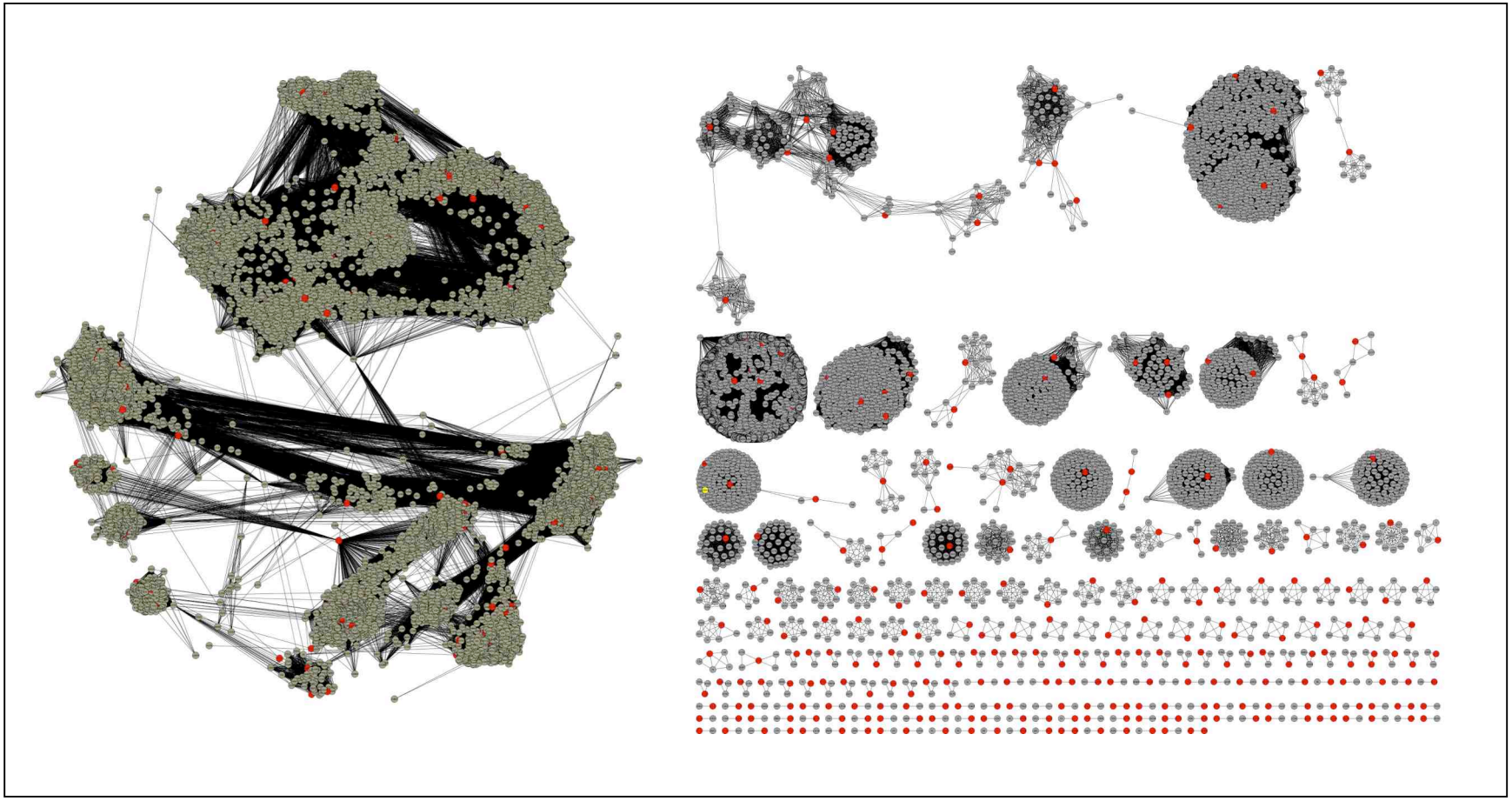


Figure 4.2. Distribution of exemplar sequences across the network of maize LTR retrotransposons. The visualization of the network of maize LTR retrotransposons was generated with the RepMiner program [25]. Each node represents a full length LTR retrotransposon in the dataset, and edges represent significant BLAST matches among sequences. Nodes colored in red indicate sequences that were selected as exemplars from the overall network of LTR retrotransposons in the dataset. Single copy families are not shown in this visualization.

References

1. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, et al. (2006) Pfam: clans, web tools and services. Nucleic Acids Res 34: D247-251.
2. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14: 755-763.
3. Pritham EJ (2009) Transposable elements and factors influencing their succœss in eukaryotes. J Hered 100: 648-655.
4. Weil CF (2009) Too many ends: aberrant transposition. Genes Dev 23: 1032-1036.
5. Gogvadze E, Buzdin A (2009) Retroelements and their impact on genome evolution and functioning. Cell Mol Life Sci.
6. Bennetzen JL, Ma JX, Devos K (2005) Mechanisms of recent genome size variation in flowering plants. Annals of Botany 95: 127-132.
7. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, et al. (2009) The B73 maize genome: complexity, diversity and dynamics Science 326: 1112-1115.
8. Finnegan DJ (1989) Eukaryotic transposable elements and genome evolution. Trends in Genetics 5: 103-107.
9. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, et al. (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8: 973-982.
10. Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, et al. (2009) Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. PLoS Genet 5: e1000732.
11. Smit A, Hubley R, Green P (1996-2004) RepeatMasker Open-3.0.
12. Jurka J, Klonowski P, Dagman V, Pelton P (1996) CENSOR--a program for identification and elimination of repetitive elements from DNA sequences. Comput Chem 20: 119-121.
13. Kronmiller BA, Wise RP (2008) TEnest: Automated chronological annotation and visualization of nested plant transposable elements. Plant Physiol 146: 45-59.
14. Wicker T (2009) TREP, the Triticeae Repeat Sequence Database.
15. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110: 462-467.
16. Frey BJ, Dueck D (2007) Clustering by passing messages between data points. Science 315: 972-976.
17. Brusco MJ, Kohn HF (2008) Comment on "clustering by passing messages between data points". Science 319: -.
18. Frey BJ, Dueck D (2008) Response to comment on "clustering by passing messages between data points". Science 319: -.
19. Dueck D, Frey B. Non-metric affinity propagation for unsupervised image categorization; 2007. IEEE. pp. 1-8.
20. Leone M, Sumedha, Weigt M (2007) Clustering by soft-constraint affinity propagation: applications to gene-expression data. Bioinformatics 23: 2708-2715.
21. McCarthy EM, McDonald JF (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics 19: 362-367.
22. Dongen SMv (2000) Graph Clustering by Flow Simulation. Amsterdam: University of Utrecht. 169 p.
23. Frey B http://www.psi.toronto.edu/affinitypropagation/software/apclusterSparse.m.
24. Frey BJ, Dueck D (2005) Mixture modeling by affinity propagation. Advances in Neural Information Processing Systems 18.

25. Estill JC, Baucom RS, Bennetzen JL (2009) RepMiner : http://repminer.sourceforge.net/.

CHAPTER 5

EXCEPTIONAL DIVERSITY, NON-RANDOM DISTRIBUTION AND RAPID EVOLUTION OF RETROELEMENTS IN THE B73 MAIZE GENOME[1]

[1] Baucom, RS*, J.C. Estill*, C. Chapparro, N. Upshaw, A. Jogi, J.M. Deragon, R.P. Westerman, P.J. SanMiguel, and J.L Bennetzen. 2009. PLOS Genetics. 5(11): e1000732
**Authors contributed equally to this work*

Abstract

Recent comprehensive sequence analysis of the maize genome now permits detailed discovery and description of all transposable elements (TEs) in this complex nuclear environment. Reiteratively optimized structural and homology criteria were used in the computer-assisted search for retroelements, TEs that transpose by reverse transcription of an RNA intermediate, with the final results verified by manual inspection. Retroelements were found to occupy the majority (>75%) of the nuclear genome in maize inbred B73. Unprecedented genetic diversity was discovered in the long terminal repeat (LTR) retrotransposon class of retroelements, with >400 families (>350 newly discovered) contributing >31,000 intact elements. The two other classes of retroelements, SINEs (four families) and LINEs (at least 30 families), were observed to contribute 1,991 and ~35,000 copies, respectively, or a combined ~1% of the B73 nuclear genome. With regard to fully intact elements, median copy numbers for all retroelement families in maize was 2 because >250 LTR retrotransposon families contained only one or two intact members that could be detected in the B73 draft sequence. The majority, perhaps all, of the investigated retroelement families exhibited non-random dispersal across the maize genome, with LINEs, SINEs and many low-copy-number LTR retrotransposons exhibiting a bias for accumulation in gene-rich regions. In contrast, most (but not all) medium- and high-copy-number LTR retrotransposons were found to preferentially accumulate in gene-poor regions like pericentromeric heterochromatin, while a few high-copy-number families exhibited the opposite bias. Regions of the genome with the highest LTR retrotransposon density contained the lowest LTR retrotransposon diversity. These results indicate that the maize genome provides a great number of different niches for the survival and

procreation of a great variety of retroelements that have evolved to differentially occupy and exploit this genomic diversity.

Author Summary

Although TEs are a major component of all studied plant genomes, and are the most significant contributors to genome structure and evolution in almost all eukaryotes that have been investigated, their properties and reasons for existence are not well understood in any eukaryotic genome. In order to begin a comprehensive study of TE contributions the structure, function and evolution of both genes and genomes, we first set out to identify all of the TEs in maize and investigated whether there were non-random patterns in their dispersal. We used homology and TE structure criteria in an effort to discover all of the retroelements in the recently sequenced genome from maize inbred B73. We found that the retroelements are incredibly diverse in maize, with many hundreds of different families that show different insertion and/or retention specificities across the maize chromosomes. Most of these element families are present in low copy numbers, and had thus been missed by previous searches that relied on a high-copy-number criterion. Different element families exhibited very different biases for accumulation across the chromosomes, indicating that they can detect and utilize many different chromatin environments.

Introduction

Transposable elements (TEs) were first discovered in maize (*Zea mays*) [1], but have subsequently been found in almost every organism investigated, from archaea and eubacteria to animals, plants, fungi and protists [2]. TEs are dynamic, abundant and diverse components of

higher eukaryotic genomes, where they play key roles in the evolution of genes and genomes. The class I TEs transpose through reverse transcription of a transcribed RNA intermediate, while most class II TEs transpose through a cut-and-paste mechanism that mobilizes the DNA directly. However, there are some class II TEs, for instance *IS91* of bacteria and *Helitrons* in eukaryotes, that are believed to transpose through a rolling-circle DNA replication process that does not involve element excision [3,4].

In most plant species, a particular type of class I element, the long terminal repeat (LTR) retrotransposons, has been observed to be the major TE, accounting for >80% of the nuclear DNA in many angiosperms [5]. The other two types of class I elements, LINEs and SINEs, have also been observed in all flowering plant genomes that have been carefully annotated, but their copy numbers and overall contributions to genome composition have not usually been large. However, in lily (*Lilium speciosum*) and grapevine (*Vitis vinifera*), LINEs appear to be more numerous and/or active than in most plant species investigated [6,7].

A wealth of recent studies has indicated that the class I elements, especially LTR retrotransposons, are primary contributors to the dynamics of genome structure, function and evolution in higher plants. Even within species, the LTR retrotransposon arrangement and copy number can vary dramatically in different haplotypes [8-11]. Some LTR retrotransposons acquire and amplify gene fragments [12,13], and sometimes fuse their coding potential with those of other genes [14], to create "exon shuffled" candidate genes that have the potential to evolve novel genetic functions [15]. Retroelements of all types may also serve as sites for the ectopic recombination events that can cause chromosomal rearrangements: duplications, deletions, inversions and translocations. Retroelement insertions can donate their transcriptional regulatory functions to any adjacent gene, and the prevalence of this process over evolutionary time is

indicated by the many fragments of retroelements and other TEs that are found in current plant gene promoters [16].

In angiosperms, the major factors responsible for the greater than 1000-fold variation in genome size are polyploidy and retroelement amplification [5]. In some lineages, amplification of only one or a few LTR retrotransposon families has been observed to more than double genome size in just a few million years [17,18]. In other organisms, like maize, many different LTR retrotransposon families have amplified in recent times to create a large and complex genome [19].

Despite the abundance, ubiquity and genetic contributions of TEs in plants, no previous investigation has made comprehensive efforts to fully discover or characterize all of the TEs in any angiosperm genome. Even the best annotated plant genomes, those of *Arabidopsis thaliana* and rice (*Oryza sativa*), were initially examined only at a cursory level to find highly repetitive elements and those with homology to previously known TEs. Hence, subsequent studies on these genomes continue to yield new families of TEs of various types. The first exception to this rule has been the draft sequence analysis of the ~2300 Mb maize genome, where a consortium of TE researchers has used several independent approaches in an attempt to discover and describe as many TEs as possible in this complex genome [20].

Even before its nearly full genome analysis, maize was the source of the best-studied TE populations in plants, including the LTR retrotransposons, where detailed analysis of small segments of the genome uncovered a great diversity of elements in different families that are mostly arranged as nested insertions [21]. The maize LTR retrotransposons were classified into 47 families [22], and comparisons between families indicated differences in their times of

transposition [23], their preferential associations with different chromosomal regions [23-25], and their levels of expression [26].

In order to fully describe the contributions of TEs to genome structure and function, one needs to first find and describe all of the TEs in a genome. Given that that average flowering plant genome is ~6500 Mb [27], they are expected to be composed of complex intermixtures and highly variable structures of TEs, so identification and analysis of the complete TE set will be a daunting task. Hence, we know very little about TE abundances and arrangements in anything but unusually tiny plant genomes, like those of Arabidopsis, rice and sorghum. Here, a comprehensive identification and characterization of retroelements is reported for the maize genome from inbred line B73 [20]. Hundreds of new retroelement families were discovered, and dramatic preferences in their distributions, associations and activities were uncovered. These first comprehensive studies open a window onto the true complexity of genome structure and evolution in a moderate-sized angiosperm genome.

Results

Retroelement discovery

In order to find all elements, LTR retrotransposons were sought by a combination of approaches that relied on both structure and homology, as described in Materials and Methods. The structure of an integrated LTR retrotransposon can be simply described as a terminal 5' repeat that starts/ends in TC/GA), followed by a primer binding site that is used for the initiation of reverse transcription (i.e., replication), followed by polycistronic (and sometimes frame-shifted) genes that encode for several proteins necessary for element replication and integration, followed by a polypurine tract that is involved in the switch to second strand DNA synthesis,

followed by the 3' LTR. Searching for these canonical structures employed LTR_STRUC [28], combined with custom Perl scripts. All intact LTR retrotransposons were identified in a set of 16,960 sequenced maize BACs (bacterial artificial chromosomes) [20]. In addition, LTR retrotransposons homologous to known TEs in the maize LTR retrotransposon exemplar database (http://maizetedb.org/) were found by running the RepeatMasker program (vers 3.19) [29] on the assembled B73 genome using default parameters.

The element discovery process yielded 406 unambiguously distinct families of LTR retrotransposons that contained at least one intact member (Table 5.1), with intact being defined as the presence of two LTRs flanked by target site duplications (TSDs). Families were defined by established sequence relatedness criteria [30], and most families were named using the sequence-based criteria developed by San Miguel and coworkers [31]. Of these families, the great majority (363) were found by this structure-based screen and had not been previously described. A few (90) additional full-length LTR retrotransposons were identified that lacked sufficient structural or internal sequence information to allow one to determine their family status, and these are currently given the generic family name "unknown" (see Materials and Methods).

LINEs were detected by their TSDs flanking a block of sequence of appropriate length (5-10 kb for L1-like superfamily member searches and 3-5 kb for RTE-like superfamily member searches), terminated on one end with a simple sequence repeat, usually poly A. Further, these candidates were required to encode at least one LINE-specific protein motif.

SINEs are non-autonomous retroelements that use the enzymatic machinery of autonomous LINEs to retropose (for a review see [32]). SINE discovery was mainly based on the detection of the characteristic internal RNA polymerase III promoter, as described in Materials

and Methods. Prior to this search, only the ZmAU SINE family had been identified in maize [33]. Using a structure-based approach, an additional three SINE families were discovered, and are now named ZmSINE1, ZmSINE2 and ZmSINE3 (Figure 5.1A). All four maize SINE consensus sequences possess an internal RNA polymerase III promoter composed of conserved A and B boxes, suggesting an ancestral relationship to tRNAs. As for the pSINE family in rice and the TS SINE family in tobacco [34,35], ZmAU, ZmSINE1 and ZmSINE2 members ends with a poly(T) stretch of 4 to more than 20 bases, a feature found only in these five plant SINE families [32]. In contrast, ZmSINE3 members end with a poly(A) stretch, a feature found for Brassicaceae SINEs [36] as well as for all other eukaryotic tRNA-related SINEs [32]. Despite this structural difference, ZmSINE2 and ZmSINE3 likely have the same LINE partner as they show strong 3'-end sequence homologies with the maize LINE1-1 consensus sequence (Figure 5.1B). This implies that, in the target-primed reverse transcription process leading to SINE integration by the LINE machinery, the same LINE reverse transcriptase can prime reverse transcription on a poly(A) as well as a poly(U)-ending RNA template.

Retroelement abundance and diversity in B73 maize

Because TEs in maize and other organisms tend to insert into each other, it was possible that other TE sequences inside a retroelement might be misidentified as an intrinsic part of the retroelement. Hence, all of the retroelements identified in maize were carefully compared to the comprehensive databases for other (i.e., class I) TEs in maize [20] to produce a filtered set of retroelement sequences.

The filtered LTR retrotransposon sequences for all 406 families were used with a RepeatMasker approach [29] to find all of the significant homologies in the B73 draft sequence [20]. At the default settings employed, similarity as small as a contiguous perfect match of 24 bp

was identified as a valid homology. With this approach, over 1.1 million LTR retrotransposon fragments were identified in the B73 maize genome, contributing ~1.5 Gb, or about 75% of the ~2.05 Gb of the genome that has been sequenced (Table 5.1; [20]). As expected, the most abundant families were those that had been previously known, like *Huck*, with the four most numerous families each contributing 7-12% of the nuclear DNA. The 20 most numerous LTR retrotransposon families generate ~70% of the sequenced B73 genome (Table 5.2), while the remaining 386 families mostly consist of low-copy-number families with a high diversity but lesser genomic abundance (Figure 5.2 and Table 5.S1).

Many cases were observed of gene fragments inside LTR retrotransposons (Table 5.S2). A total of 425 intact LTR retrotransposons were observed to contain gene fragments, from a minimum of 189 independent gene fragment captures. No case was identified, under the conditions employed, where a single LTR retrotransposon contained inserted fragments from more than one standard nuclear gene. Other classes of TEs in maize are even more active in gene fragment acquisition, including 1194 gene fragment captures by *Helitrons* and 462 by other DNA transposons, including Pack-MULEs [20]. It is not known whether these gene fragments play any role in maize genetic function, for instance in the creation of a new gene or in epigenetic regulation of their donor loci.

Thirty different families (with family members defined as those with >80% sequence identity [30]) of LINEs were detected in the maize genome, with 13 of these not having been previously found and/or identified as separate families (Table 5.1). Approximately 35,000 LINEs (many as fragments of intact elements) were found in the B73 sequence, but this number is certain to be an overestimate caused by the many gaps and incorrect assemblies that are

expected in the current maize genome draft sequence [20]. These LINEs contribute 20 Mb of DNA to the draft genome sequence, or about 1% of the total (Table 5.1).

Overall, SINEs represent around 0.5 Mb and 0.02% of the sequenced portion of the B73 maize genome [20]. The copy numbers are 49, 134 and 23 for the ZmAU, ZmSINE1 and ZmSINE3 families, respectively. ZmSINE2 is the major SINE family, with 1382 members. Based on phylogenetic criteria (Figure 5.S1), the ZmSINE2 family can be further divided into three distinct subfamilies.

A phylogenetic approach was used to study the amplification dynamics of SINEs in maize. The ZmSINE1, ZmSINE2 and ZmSINE3 families contain very young members (Figure 5.S1), close to the family consensus, suggesting very recent transposition activity. Tree topologies for these families are also typical of the "gene founder" model wherein a very small number of "master" elements are active while the vast majority of derived copies have no significant amplification potential [37]. The ZmAu family is mainly composed of more diverged members, suggesting little or no activity in the recent past.

LTR retrotransposon superfamilies and families

In order to look at the behaviors (e.g., insertion specificities or amplification level) of the TEs across a genome, it is essential to determine their relatedness and then use this information to generate families of close relatives. Once families are generated, then family-specific behaviors can be investigated. Transposable elements of all classes tend to vary in relatedness across a spectrum, such that two TEs recently derived by transposition from the same parent element may be 100% identical in sequence, while others with a more ancient relationship can show any degree of further divergence. However, the very rapid removal of DNA from higher plant genomes [38,39], especially from maize [40], by the progressive accumulation of small

deletions indicates that TEs that last shared a common ancestor more than a few million years ago (mya) are usually largely or fully deleted from the genome. Hence, TE families can be defined by an arbitrary but consistent criterion of nucleotide sequence divergence, and a value of 80% identity has been selected by a consortium of researchers in this field [30].

In the maize genome, the classification of LTR retrotransposons into families was a major challenge because of the exceptional complexity that was observed. Nonetheless, similar to the case in the much simpler rice genome [41], all-by-all BLAST analysis of LTRs was sufficient to unambiguously define families by the 80% identity rule. Not all families could be classified in their appropriate superfamily (i.e., *copia* or *gypsy*), usually because of an absence of the genes needed for the definitive gene order criterion or for phylogenetic analysis, and these were dubbed RLX. The individual family identifications were clear, however, and each family was given a unique name. Some of these family designations conflict with previous names [42], but these earlier names were not applied with any specific rule, and thus were certain to be both misleading and temporary. For instance, the LTR retrotransposon collection called CRM [20] was actually found to represent four related, but clearly separate, LTR retrotransposon families that we have now named *CRM1*, *CRM2*, *CRM3/CentA,* and *CRM4*. Our consistent analysis using agreed-upon criteria [30] caused other such splittings of previously lumped families, and also lumped some different named families into single families that fit the 80% identity criterion (e.g., *Cinful* and *Zeon* are actually a single family that has now been named *Cinful-zeon*). The new names, and the names that had previously been applied by unspecified and/or inconsistent homology criteria, are now shown in Table 5.S1.

Dispersal of retroelements across the B73 maize genome

The assembled physical and genetic map of maize inbred B73 [20] allows placement of any class of sequence along that portion of the genome that was sequenced. Overall, LTR retrotransposons are found to be most abundant in pericentromeric heterochromatin and least abundant in the more gene-rich arms on all chromosomes (Figure 5.3). However, different LTR retrotransposons are found to be differentially clustered in such analyses, with the general observation that the *gypsy* superfamily of LTR retrotransposons is concentrated in the pericentromeric heterochromatin while the *copia* superfamily shows a preferential accumulation in the more euchromatic regions of the chromosome arms [20]. Despite this general pattern, individual families show deviations from the rule. For instance, the *gypsy* family *Huck* was found to exhibit a more '*copia*-like' distribution on chromosome 1 (Figure 5.S2). Another *gypsy* family, *Grande*, shows a relatively even distribution across 10 Mb bins of this same chromosome. Hence, there are families that accumulate in a pattern that contrasts with the general behavior of their superfamilies in maize.

A more dramatic correlation between LTR retrotransposon family property and insertion/accumulation pattern was observed by comparing the copy numbers of intact elements in a LTR retrotransposon family with the nature of the sequences within 500 bp (on each side) of the insertion site. Low-copy-number families were found to be most often inserted into the regions in or near genes (or gene fragments), while high-copy-number families were observed to primarily accumulate inside other LTR retrotransposons (Figure 5.4).

LINEs of both RIT and RIL (L1-like) families were found to be fairly evenly distributed across all chromosomes, with a higher abundance in distal regions of the chromosomes (Figure 5.S3). Although maize LINEs have been observed to show a preferential association with genic

regions, especially introns [43], their common occurrence in pericentromeric DNA suggests that many insertions are not in or near genes.

Of the 1991 SINEs discovered, 1174 were found in the introns or UTRs (untranslated regions) of genes and 21 in putative coding exons (data not shown). Only 796 were found in the intergenic space that makes up more than 85% of the sequenced B73 genome [20]. Hence, like SINEs in other species, these small TEs show a very strong preference for association with genes in the maize nuclear genome. In this regard, the general distribution of SINEs across the maize chromosomes (Figure 5.S4) was found to exhibit a pattern quite similar to the gene distribution [20].

Correlated patterns of retroelement distribution

As previously observed in other organisms by numerous scientists studying many different genomes, maize TEs were found to make up a greater quantity of the total DNA in the gene-poor pericentromeric regions than in other parts of the genome (Figure 5.3). However, as mentioned above and observed previously (reviewed in [44]), LINEs, SINEs and some LTR retrotransposon families accumulate preferentially in areas that are near genes.

Figure 5.5 shows the relationship between LTR retrotransposon abundance and LTR retrotransposon family richness across chromosome 1 of maize inbred B73, and this general pattern was found to be the same across all other chromosomes (data not shown; Table 5.S3). Hence, on all maize chromosomes, those regions that have the most total LTR retrotransposons also have the fewest kinds of LTR retrotransposons. This observation echoes the relationship between the number of species and the abundance of individual species in both terrestrial and aquatic environments, but has no precedent that we are aware of in TE studies.

The insertion dates of intact LTR retrotransposons was observed to vary according to the distance from the centromere. Younger elements are enriched in the euchromatic regions whereas older elements are most abundant in the pericentromeric regions (Figure 5.6). An analysis of variance showed that the average insertion date per 1 Mb bin varied according to distance from the centromere ($F = 2.08$; $P < 0.0001$), and this relationship held across most of the chromosomes (Table 5.3).

The average date of LTR retrotransposon insertion for a given family was also observed to correlate with the current perceived copy numbers of the LTR retrotransposon families. As a general pattern, the lower-copy-number elements were more ancient insertions (averaging about 1.2 mya) compared to the highest-copy-number elements (averaging about 0.7 mya) (Figure 5.7). Because most of the higher-copy-number LTR retrotransposons are of the *gypsy* superfamily (Table 5.2), and show an overall pericentromeric accumulation bias [20], one expected the opposite result because of the slower rate of LTR retrotransposon removal in gene-poor (and thus recombination-poor) regions like the pericentromeres [45].

Discussion

Limitations of the dataset and problems this might generate

The landmark sequencing of the very complex and fairly large maize genome was accomplished at a small fraction of the cost of previous clone-by-clone sequencing projects because of the expertise of the researchers involved, because of a low redundancy of initial shotgun sequencing and because of a decision to not finish any regions of the genome that appeared to lack gene candidates [20]. Hence, a very comprehensive TE discovery and masking process was necessary to facilitate finishing that was efficiently targeted on genes. One

disadvantage of this approach, however, is that most sequenced regions are composed of many tiny contiguous sequences (contigs). Our analysis of the current B73 assemblies (data not shown) indicates a median contig size of ~7 kb with ~60% of the assembly occurring in contigs larger then 30 kb. Thus, a structure-based search approach that requires intact elements, like the one employed in this project, will miss any families where the only intact members are fractured by sequence gaps or inaccurate scaffolding of contigs. This is expected to be most problematic for large TEs (like LTR retrotransposons) and for those that only have a few intact members. Hence, our prediction that ~75% of the B73 maize genome is composed of LTR retrotransposons is a minimum estimate.

Also because of the many tiny sequence gaps in the assembly, there will be many occasions when an intact retroelement was identified by RepeatMasking as several fragments of an element. Hence, calculation of the ratio of intact to fragmented LTR retrotransposons is not valid with this dataset. In contrast, this same analysis with the random sampling of fully sequenced and annotated clones known as the GeneTrek approach does allow accurate quantification of the relative abundance of different TE structures. In such a GeneTrek analysis, the ratio of intact to truncated LTR retrotransposons in maize was found to be ~2:1 [40,46], quite different from the ratio of ~1:27 that was calculated (Baucom and Bennetzen, data not shown) as an artifact of this same analysis on the currently fractured B73 assembly [20].

There are also many large sequence gaps, and numerous sequenced BACs with no home in the assembly, for the B73 draft sequence [20]. It is likely that about 90% of the maize nuclear genome is present in the current assembly (~2005 Mb out of ~2300 Mb). From all previous full genome sequences in multicellular eukaryotes that have centromeres, the standard observation has been that the majority of the unsequenced regions are in the gene-poor areas around the

centromeres and in other heterochromatic blocks. Because these gene-poor chromosome segments also tend to be LTR retrotransposon-rich, these results provide a further reason to believe that the B73 maize genome contains more than 75% LTR retrotransposons, with an upper limit of ~85%.

Importantly, however, the overall quantitation of retroelement contributions to the B73 genome is not dramatically biased by the gaps and other intrinsic errors in the current assembly. As shown in Figure S5, most LTR retrotransposons exhibit the same relative abundance when used to mask the current B73 draft assembly as they do when used to mask a shotgun dataset from the same B73 line ($R^2 = 0.99$, $p < 0.0001$). The few exceptions to this observation (e.g., *Ipiki*) are likely to be LTR retrotransposons that are preferentially abundant in that ~10% (e.g., near centromeres?) of the maize genome that is not present in the assembly [20].

Previous maize studies had uncovered primarily the high-copy-number retroelements [21,23], with some exceptions of low-copy-number TE discovery associated with particular mutations [47,48] or carefully sequenced and annotated small segments of the maize genome [46]. All of the LTR retrotransposons found in these earlier studies were also found in this analysis, at the approximate predicted frequencies. The major difference, however, was the large dataset available in the current study, and thus the discovery of hundreds of additional LTR retrotransposon families. Only by this comprehensive analysis on the majority of the maize genome was it possible to determine the exceptional complexity of retroelements in maize, and their different properties of dispersal and divergence.

Diversity and its meaning

Rice, with an ~400 Mb nuclear genome, has 172 identified LTR retrotransposon families that contribute ~97 Mb, distributed across 48% with only a single intact element, 20% with 2

intact elements and 32% with 3 or more intact elements [41]. Maize, in contrast, has 406 identified LTR retrotransposon families, just over twice as many, but they contribute ~1700 Mb of DNA to the maize nuclear genome. These maize elements are distributed across 42% singleton intact elements, 21% with 2 intact elements and 37% of families with 3 or more intact elements. Hence, the >17X greater amount of LTR retrotransposons in maize compared to rice is not primarily caused by a greater number of element families in maize but instead by a much higher copy number of a very small number of superabundant families.

Two of the many misconceptions about TE properties in higher eukaryotes are that they are highly repetitive and are randomly scattered about the genome. In fact, many TE families are present in very low copy numbers. The median family copy number of intact LTR retrotransposon with TSDs in B73 maize was measured to be 2 (mean ~77), with a total of 256 families that contained only one or two intact LTR retrotransposons that were detected. Most LTR retrotransposon families are distributed quite unevenly across the genome, probably an outcome of both differences in insertion preferences and different rates of LTR retrotransposon removal in different chromosomal locations [44-46,49]. The previous observation that LTR retrotransposons show a dramatic bias in whether they insert into LTRs or the internal regions of other LTR retrotransposons [21] was not observed, however, and it now seems likely that the previous conclusion was an artifact of a small sample size.

Studies in rice and other organisms suggest that LTR retrotransposons are more rapidly removed (sometimes by unequal homologous recombination to generate solo LTRs) in regions with high recombination rates, like areas around genes and in the cores of centromeres [45,46]. One example of this analysis was that the ratio of solo LTRs to intact elements was found to be much higher in gene-rich and recombination-rich euchromatic regions than in gene-poor and

recombination-poor pericentromeric regions [44]. Although natural selection should also more rapidly remove individuals from a population that contain retroelements or other TEs detrimentally inserted into coding and gene regulatory regions, this process alone cannot explain the differential retroelement accumulation properties that we observe. For instance, why would LINEs, SINEs and low-copy-number LTR retrotransposons not be depleted in genic regions, while high-copy-number LTR retrotransposons are? A simpler explanation is that different retroelements are directed to preferentially insert in different parts of the genome by the biases of their integrases for association with specific chromatin proteins, as observed with Ty elements in yeast [50].

We have no idea how many types of DNA::protein configurations are actually present in plants, of course, but it is very clear that chromatin consists of more than just hetero- and eu-varieties [51], so sufficient variability should be present to allow a great wealth of different TE insertion specificities. Particularly fascinating are the high-copy-number LTR retrotransposons like *Ji* and *Opie* that preferentially avoid insertion into genes, but primarily insert into heterochromatin near genes, while other high-copy-number elements like *Gyma* avoid inserting into genes or heterochromatin near genes, preferring instead an accumulation into large gene-free heterochromatic blocks [46]. Unlike low-copy-number LTR retrotransposons, which are associated with *de novo* mutations in many plant species, neither class of high-copy-number LTR retrotransposons is associated with a mutation caused by insertion into a gene. Perhaps TE insertion profiles will be a uniquely useful route to uncover and map a broad spectrum of novel chromatin structures.

Retroelement distribution and the origin of plant genome complexity

Genomic complexity is not just a matter of the number of different sequences, but also of the variability in their arrangement and stability. The factors that determine differences in these arrangements, such as differential insertion specificities and differences in retention, are only beginning to be understood. It is already clear, though, that TE insertion and retention biases are the major forces that determine local genome structure in maize and other complex plant genomes. The mechanisms responsible for these biases, and their outcome vis-à-vis gene/genome function and evolution, are only now beginning to be understood.

Viewed from the standpoint of the TE, much of the diversity in TE populations and their arrangement takes on a new and informative light. A previous model proposed that low-copy-number TEs must insert near or into genes so that they have a reasonable chance of expression and activity in subsequent generations, while highly repetitive TEs need to avoid insertions that disrupt genes in most cases because 1000 or 10,000 such insertions would lead to a dead host [44]. Hence, abundant TEs rely on their abundance *per se* to guarantee transmission and the opportunity for activity in future generations. The data for LTR retrotransposon abundance versus copy number shown here agrees with this model, as does the fact that (to date) none of the high-copy-number LTR retrotransposons have been shown to cause a *de novo* mutation, while low-copy-number LTR retrotransposons (e.g., *Bs1*, *Tnt1*, *Tos17*) that make up a relatively small part of their genomes have caused many new mutations [47-49,52]. The analysis of the maize genome suggests that the copy number for this transition is fairly low, 10-100 intact copies per genome (Figure 5.4), for this change in lifestyle. LTR retrotransposon families with copy numbers less than ten were usually found to preferentially accumulate in genic regions, while

most LTR retrotransposon families with copy numbers higher than 100 were found to be enriched in gene-poor regions like pericentromeric heterochromatin.

The insertion preferences of LTR retrotransposons can contribute to their potential for more than just transcriptional activity. Elements that land in recombination-rich regions have a greater chance of inter-element unequal events that can create novel LTR retrotransposons with possible new properties [38]. Insertion into an LTR provides the opportunity to acquire the gene regulatory properties of the target LTR retrotransposon. Moreover, insertion of a LTR retrotransposon into a LTR retrotransposon would usually eliminate the target element as a potential competitor for future amplification.

The observed relationship between LTR retrotransposon family richness and LTR retrotransposon abundance across the maize chromosomes is the most compelling indicator, in this study, of the validity of the conceptualization of TEs as competitor organisms whose world is the nuclear genome. When an environment is highly suitable for proliferation of a category of life, a few highly adapted types of individuals (e.g., species or, in this case, LTR retrotransposons) crowd out all other competitors to create a dense but diversity-poor ecosystem. Other species, here proposed to be the lower-copy-number LTR retrotransposons, disseminate themselves at lower abundances across less productive environments that thus become diversity-rich. Of course, it is not at all clear what aspect(s) of these TE-enriched regions might make them "productive" from a TE perspective. Perhaps it is something as simple as a lower rate of TE removal by ectopic recombination [45]. This view of genomic life provides another angle to investigate TEs, as highly adapted commensals, but in no way suggests that they cannot be utilized when the opportunity arises for a process that benefits the plant host. The occasional creation of new genes by TE capture and shuffling of gene fragments or through fusion of TE

genes (or regulatory regions) with nearby genes falls into this category. What remains constant in these considerations is the long-term evolutionary value of the instability and diversity generated by retroelements and other TEs.

Materials and Methods

Generation of the maize LTR retrotransposon exemplar database

New families of maize LTR retrotransposons were discovered by several iterations of masking and re-investigation. First, 5,075 maize BACs were downloaded on February 22, 2007 from the Washington University maize sequencing project [20] and masked using the RepeatMasker program [29] with a database of previously known maize LTR retrotransposons. Masked regions were removed from the sequence, and LTR_STRUC [28] was used to find new elements. This program identifies LTR retrotransposons based on the presence of LTRs, matching target site duplications (TSDs), and the presence of the canonical TG/CA motif found at the 5' and 3' end of each LTR (although deviations are permitted), and thus is a structure-based screen rather than one that requires sequence homology to a known TE. This process was designed to uncover old and fragmented families of LTR retrotransposons after masking out the younger and previously discovered families [21,22].

Next, 15,708 maize BAC sequence data sets were downloaded March 1, 2008 from the Washington University sequencing project and were first masked at a quality score of '40,' then screened with LTR_STUC. 13,362 LTR retrotransposons were found and, along with the sequences uncovered in the initial screen, placed into families using the RepMiner classification tools (http://repminer.sourceforge.net/) [53]. This process generated ~600 maize LTR retrotransposon exemplar sequences that best describe each of 412 identified families. Each

exemplar was annotated for LTR position, the primer-binding site sequence and the genes involved in the transposition process.

Exemplars were identified as members of either the *copia* or *gypsy* superfamilies based on the position of the reverse transcriptase gene in relation to the integrase gene, and by using a maximum-likelihood gene tree of reverse transcriptase. Both methods of superfamily designation were 100% congruent. Exemplar sequences that did not contain internal coding regions with an identifiable homology to LTR retrotransposon genes were given the 'unknown' superfamily designation. Each exemplar was hand-curated to ensure that exemplars where not chimeric annotations that contained insertions of other LTR retrotransposon sequences. DNA transposons inserted within the LTR retrotransposon exemplars were identified by homology-based searches against the maize TE database (http://maizetedb.org/) and were excluded from the exemplar sequence by masking.

Family nomenclature follows established methodology [30] in which the TE classification can be deduced from the full family name. In this system, family names are given a three character prefix that represents the class, order and superfamily of the individual family. For example, families with the RLG prefix represent LTR retrotransposons that are members of the *gypsy* superfamily while the RLC prefix represents families that are members of the *copia* superfamily. LTR retrotransposons that could not be assigned to the *gypsy* or *copia* superfamilies were assigned the RLX prefix.

Annotation of LTR retrotransposon distribution with RepeatMasker

The B73 maize genome represented as an Accessioned Golden Path (AGP) assembly [20] was downloaded from the Arizona Genomics Institute (http://www2.genome.arizona.edu/genomes/maize). This dataset was investigated for LTR

retrotransposon content using the default settings in RepeatMasker [29] with the curated exemplar library of maize LTR retrotransposons (http://maizetedb.org/).

The RepeatMasker annotation of the maize AGP assembly was uploaded to a custom MySQL relational database to facilitate manipulation and querying of sequence features mapped onto the maize genome assembly. The RepeatMasker output files derived from masking the AGP with the exemplar database were translated to General Feature Format (GFF) style coordinates using the cnv_repmask2gff.pl program [54]. These coordinates were uploaded to a MySQL database using custom Perl scripts. The database served as the query engine to trim overlapping features resulting from the RepeatMasker annotation and provided the framework to query distribution related information. The MySQL database schema and custom Perl scripts used to generate the non-redundant distribution information are available from the authors upon request.

Each of the AGP chromosomes was spatially binned into 10 Mb non-overlapping units and the percent LTR retrotransposon composition within each bin was determined, as was the number of distinct families present within each bin. The strength and direction of the correlation between percent LTR retrotransposon composition and family richness was determined using the Resample program [55] separately for each chromosome.

Identification, classification and location of full-length LTR retrotransposons

The sequence files for the 16,007 BAC assemblies incorporated in the maize AGP were downloaded from GenBank. Full-length LTR retrotransposons were identified by LTR_STRUC and mapped onto these BACs through the use of batch annotation scripts available in the DAWGPAWS annotation package [54]. This process resulted in a database of 35,229 full-length LTR retrotransposons.

The 5' LTR sequences of this dataset of full-length LTR retrotransposons were used to classify the elements into families using at least 80% identity in a BLASTn analysis employing the exemplar database. LTR retrotransposons that were not homologous to families present within the exemplar database (1,979) were removed from analysis, with the exception of the gene capture analysis, explained below. Further, sequences that were 2 standard deviations greater in length than the assigned family's mean length (2,135) were also removed from analysis. These sequences were found to harbor full-length insertions of other LTR retrotransposons and thus do not provide an accurate characterization of the most recently intact elements. The resultant database of full-length LTR retrotransposons consisted of 31,115 individual sequences distributed among 406 distinct families. Six families initially identified on the maize BACs used to create the exemplar database were not found in the current assembly of the AGP, potentially due to the fact that 981 BAC sequences released from the Washington University sequencing effort were not used to assemble the AGP. The location of full-length LTR retrotransposons on the AGP was determined using the data conversion table provided by the Arizona Genomics Institute.

LTR retrotransposon insertion history and specificity

The insertion date of each full-length LTR retrotransposon was determined by estimating the amount of divergence between the 5' and 3' LTRs [23]. Perl programs were used to automate this process; the two LTRs of each mined LTR retrotransposon were first aligned using ClustalW [56], and the genetic divergence between the two was estimated using the baseml module of PAML ([57], vers. 4). The time since insertion of each LTR retrotransposon element was estimated using the substitution rate of 1.3×10^{-8} per site per year [11]. To determine if distance to the centromere explained variation in insertion dates, the GLM procedure of the SAS

statistical package (vers. 9.2) was used to perform an analysis of variance with the square-root transformed average insertion date per bin as the dependent variable and the distance of each bin to the centromere as the independent variable. This analysis was performed separately for each chromosome.

Investigation into the insertion-site specificity of each full-length LTR retrotransposon was conducted by a performing a BLASTn search to four separate databases, namely those containing maize genes [20] and those containing DNA transposable elements, *Helitrons*, and LTR retrotransposons (http://maizetedb.org/). 500 bp of maize sequence flanking the 3' and 5' sides of each element was used as the query in separate nucleotide BLAST analyses, and the results were parsed for at least 80% identity. No annotations >5 bp away from the query sequence were included, because the objective was to determine what type of sequence the LTR retrotransposons inserted into, rather than those sequences that were simply nearby.

LTR retrotransposon capture of host gene fragments

A set of curated genes from the rice genome (RAPDB, vers. 4) was used to search the full-length maize LTR retrotransposons for instances of host gene capture. The full-length LTR retrotransposon dataset was screened for homology to rice genes at an Expect value of e^{-5}. Significant BLAST hits were screened for TE genes, and genes were also removed if annotated as 'rice gene family candidate' and present in high copy number (>20), as they are likely to be undiscovered TE genes. The full-length LTR retrotransposons that were not placed into families based on the 80% identity rule were retained in this analysis as they represented ~20% of the total gene capture events. The annotations of these particular LTR retrotransposons indicated that they exhibit general LTR retrotransposon features, such as target site duplications and a TG/CA

motif at the end of the LTRs, and as such represent LTR retrotransposons of 'unknown' family classification.

Maize shotgun data

Trace files of whole genome shotgun (WGS) DNA sequence reads for maize inbred B73 were obtained from those deposited by the Joint Genome Institute (JGI) to the NCBI Trace Archive (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?). These sequence files were trimmed of low quality bases and vector sequence using Lucy [58]. Organellar sequences were identified by BLAT [59]. Alignments to maize chloroplast and mitochondrial DNA and were removed from further analysis. This filtering resulted in a dataset of 1,028,203 high quality sequence reads totaling 79,6326,632 bp of genomic DNA. These data represent an approximately one-third sample sequence coverage of the B73 genome.

The JGI shotgun data were annotated for LTR retrotransposons using RepeatMasker ([29], vers. 3.19) with the same database and parameter set used to annotate the AGP. Overlapping features from the RepeatMasker output were identified using the same methodology described for LTR retrotransposon annotation of the maize AGP assembly. Significant outliers between the ratios found in the AGP and the ratios found in the JGI shotgun data were identified by performing an outlier analysis in the SAS statistical package (vers. 9.2).

SINE detection

The approach to identify potential SINE families was divided into several steps. The first step was the search for anchors, which were defined as small regions containing SINE features (see below). Following that, a 500 bp region flanking the anchor on each side was extracted. These sequences were used to perform a non-stringent search for direct repeats (likely to be TSDs) that were less than 350 bp apart. The sequences that passed the filter were aligned using

ClustalW [56], alignments were refined using muscle [60] and corrected by hand using Seaview [59].

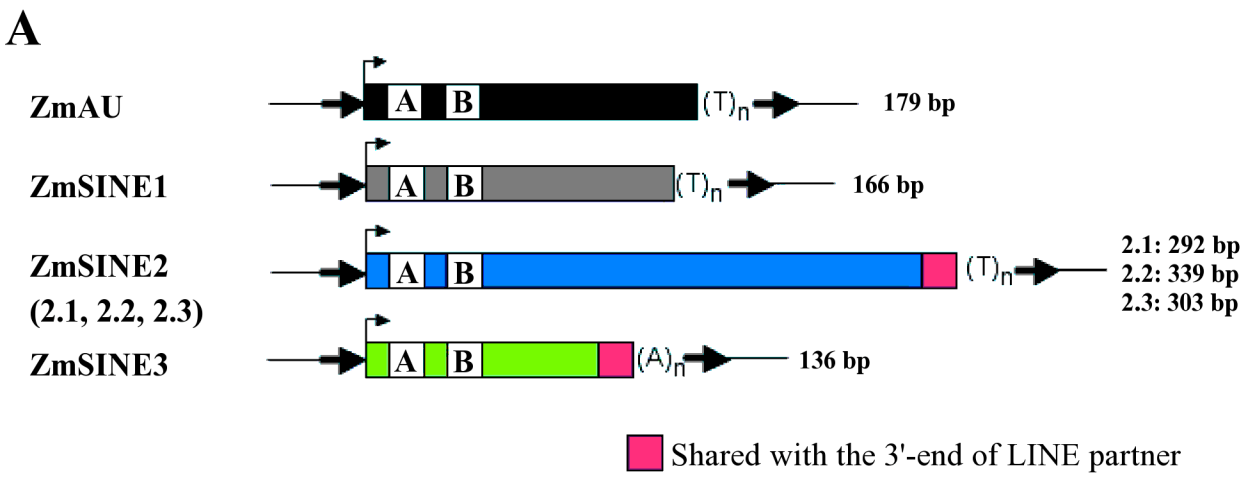
A first approach for SINE identification consisted in developing an hmm model using hmmer (http://hmmer.wustl.edu) for the region harboring the main anchor, which is the internal (tRNA-related) promoter for RNA polymerase III, defined for SINEs as an "A" box (RRYNNRRYGG) around position +14 of the start of the repeat and a B box (GGTTCGANNCC) around position +54 of the start of the repeat. This anchor was designed using known plant SINE elements. This model was then used to search the whole pseudomolecule representing the draft sequence of the B73 maize genome [20]. A second approach consisted in identifying tRNAs using tRNAscan-SE and using those sites described as "Pseudo tRNAs" as anchors. A third approach consisted in using the last 30 bases of maize LINE consensus sequences to screen for homology by BLASTn against the B73 draft genome, and to then use these homologies as anchors. In this case, to make sure that SINEs were distinguished from severely truncated LINEs, these homologies were searched for the presence of internal A and B boxes typical of tRNA-derived SINEs. A search for 5S RNA-derived SINEs was also performed, using as anchor the A/IE/C conserved boxes of the 5S RNA internal polymerase III promoter, without success. SINEs that did not share significant sequence identity (<50%) outside of the common SINE features (internal polymerase III promoter and 3'-terminal end) were classified in distinct families. For SINEs that do have significant homologies (>50%) outside of the common SINE features (>50%), further subfamily classifications were proposed using phylogenetic criteria.

SINE phylogenetic analysis

The SINE sequences were aligned using the ClustalW multiple-alignment program [56] with some manual refinements (i.e., elimination of unnecessary gaps at the beginning and end of the ClustalW alignment). Evolutionary distances were calculated using the Jin-Nei distance method of the Dnadist program (PHYLIP package version 3.573c [61]. The coefficient of variation of the Gamma distribution (to incorporate rate heterogeneity) and the expected transition to transversion ratio (t) were obtained by pre-analyzing the data with the Tree-Puzzle program [62]. Phylogenetic trees were inferred using the Neighbor-Joining (NJ) method (PHYLIP package version 3.573c [61]). Consensus trees were inferred using the Consense program (PHYLIP package). The significance of the various phylogenetic lineages was assessed by bootstrap analyses [63].

Acknowledgements

The authors thank the US National Science Foundation (0607123), l'Agence National de la Recherche (ANR-05-BLAN-0244-02), the CNRS, and the Université de Perpignan for support of this research.



B

```
ekeje/LINE1-1     1 TCTCCTTCTTAA-TATATTT---AAGGCGCAGT----TCCCCTGCGCTTTCGAG
ZmSINE2.1         1 TCTTCTTCTTAA-TACAATGCC-CGGGGGCTGTCTTACCCCCCGC-AGGTCGAG
ZmSINE2.2         1 TCTTCTTCTTAA-TATAATACCGTGGGGGCGGTCTT-TCCCCTAC-CGGCCGAG
ZmSINE2.3         1 TCTC-TTCTTAA-TATAATACCGGGAGGGCGGTCTT-TCCCTCCC-CGGCCGAG
ZmSINE3           1 CCAACTTCTTAAGTGAAAAAGTCAGGGGGACGTCTC-TCCCCT---TGGTCGAG
```

Figure 5.1. Description of the four maize SINE families. (A) Schematic representation of the four consensus maize SINEs. The size of consensus SINE sequences is indicated for each family and subfamily. The position of A and B motifs that constitute the internal (polymerase III) promoter is shown. The 3'-end similarity of ZmSINE2 and ZmSINE3 is also shown. (B) A sequence comparison of the 3'-ends of ZmSINE2.1, ZmSINE2.2, ZmSINE2.3, ZmSINE3 and the putative LINE partner, LINE1-1, is shown. No significant sequence identify (>50%) was detected between other SINE families and other maize LINE consensus sequences.

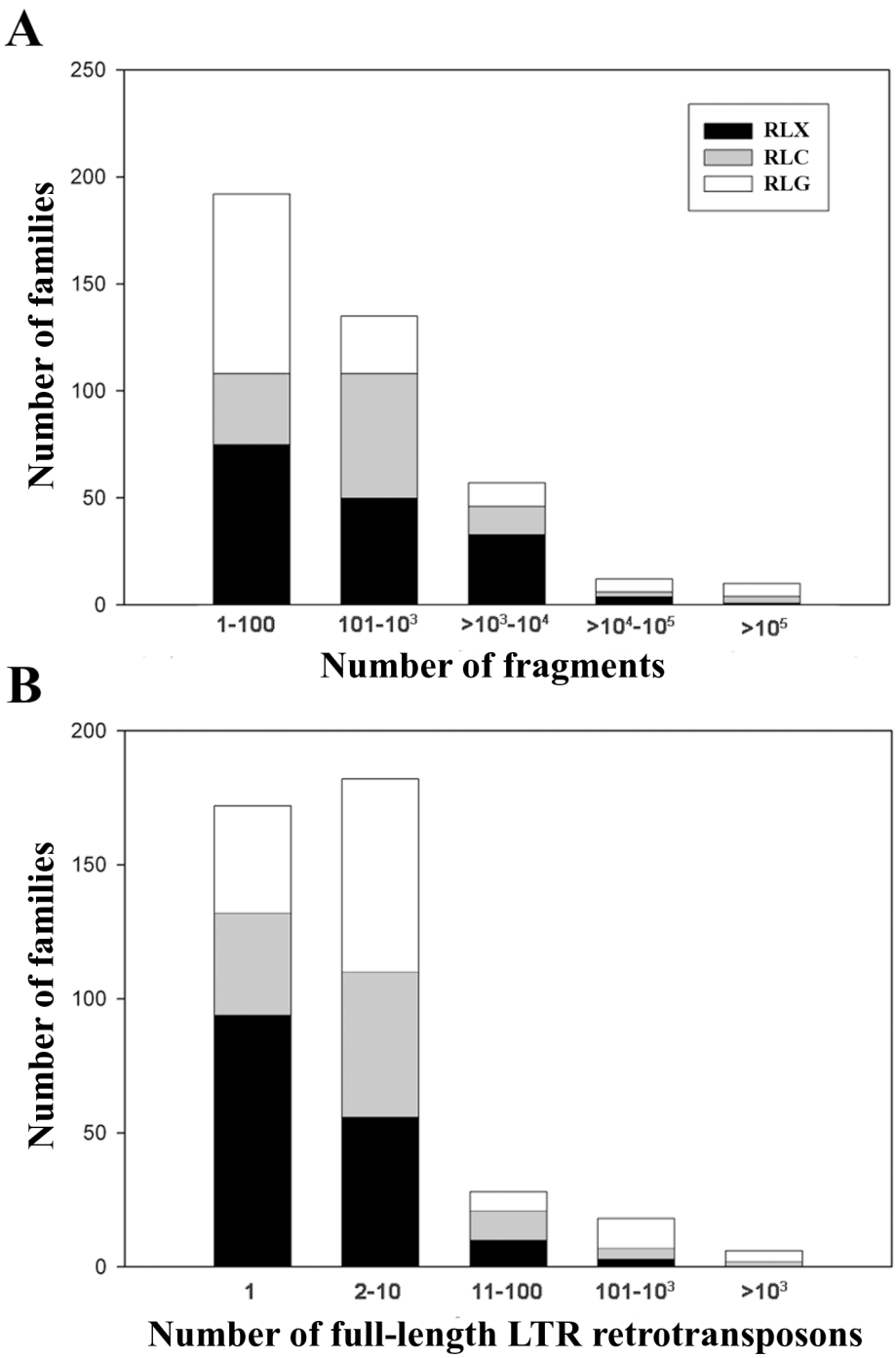


Figure 5.2. Copy number distribution of LTR retrotransposon families in the B73 maize genome. (A) The result of a homology search using the program RepeatMasker (vers. 3.19) with a library of maize LTR retrotransposon exemplars and (B) the result of a combined structure and homology screen that first uncovered the full-length LTR retrotransposons in the genome and then placed them into families, with 80% identity to an element in the exemplar database required for membership in a family.

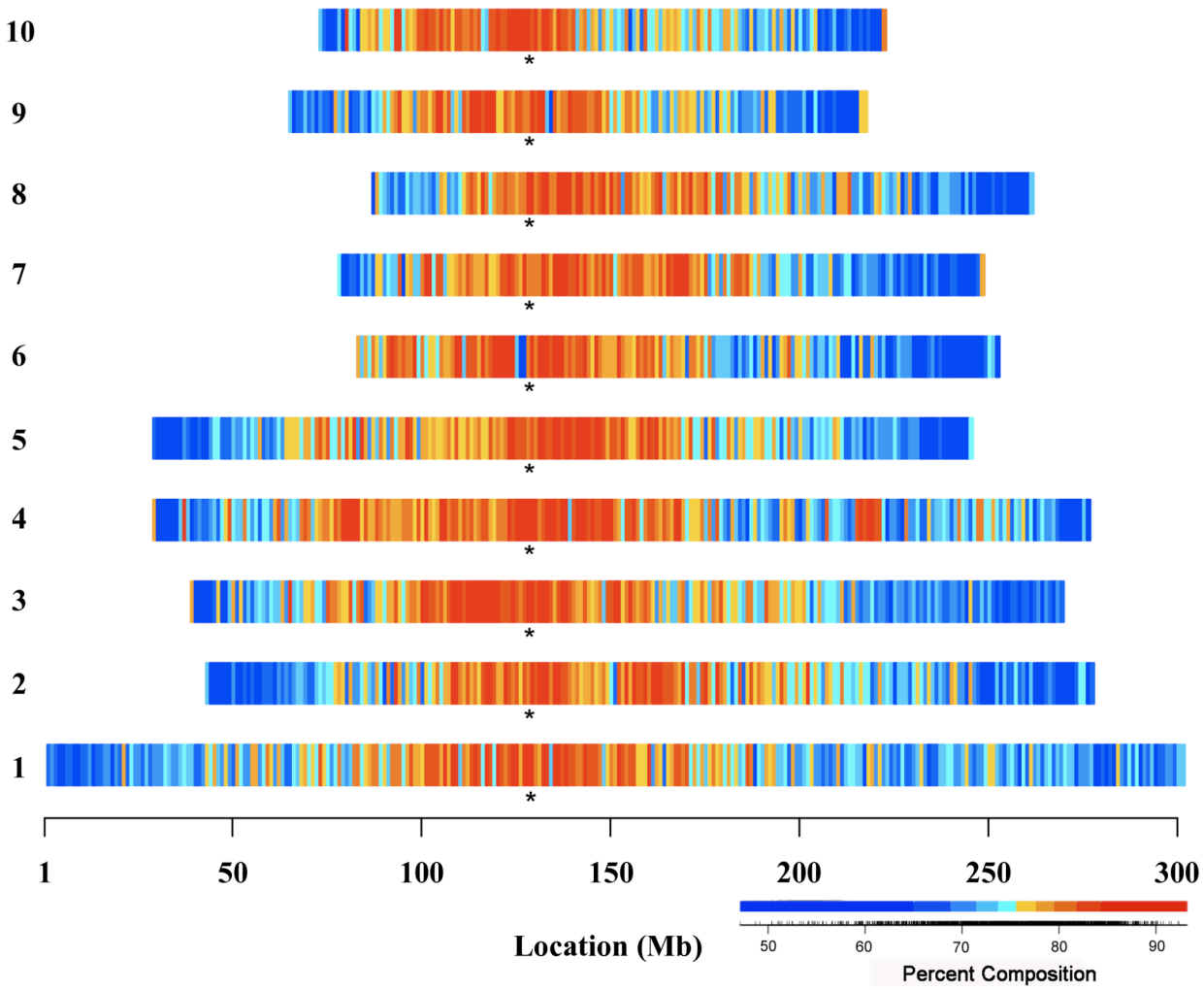


Figure 5.3. The chromosomal distribution of the LTR retrotransposon composition of the B73 maize genome. The RepeatMasker-identified LTR retrotransposons are summarized as percent composition in 1Mb bins along each of the ten chromosomes. The heatmap was derived by classifying the percent composition values into equal interval quantiles. The distribution of these classified values are illustrated as color tiles superimposed under the empirical cumulative distribution of the observed percent composition values. Asterisks indicate approximate centromere positions.

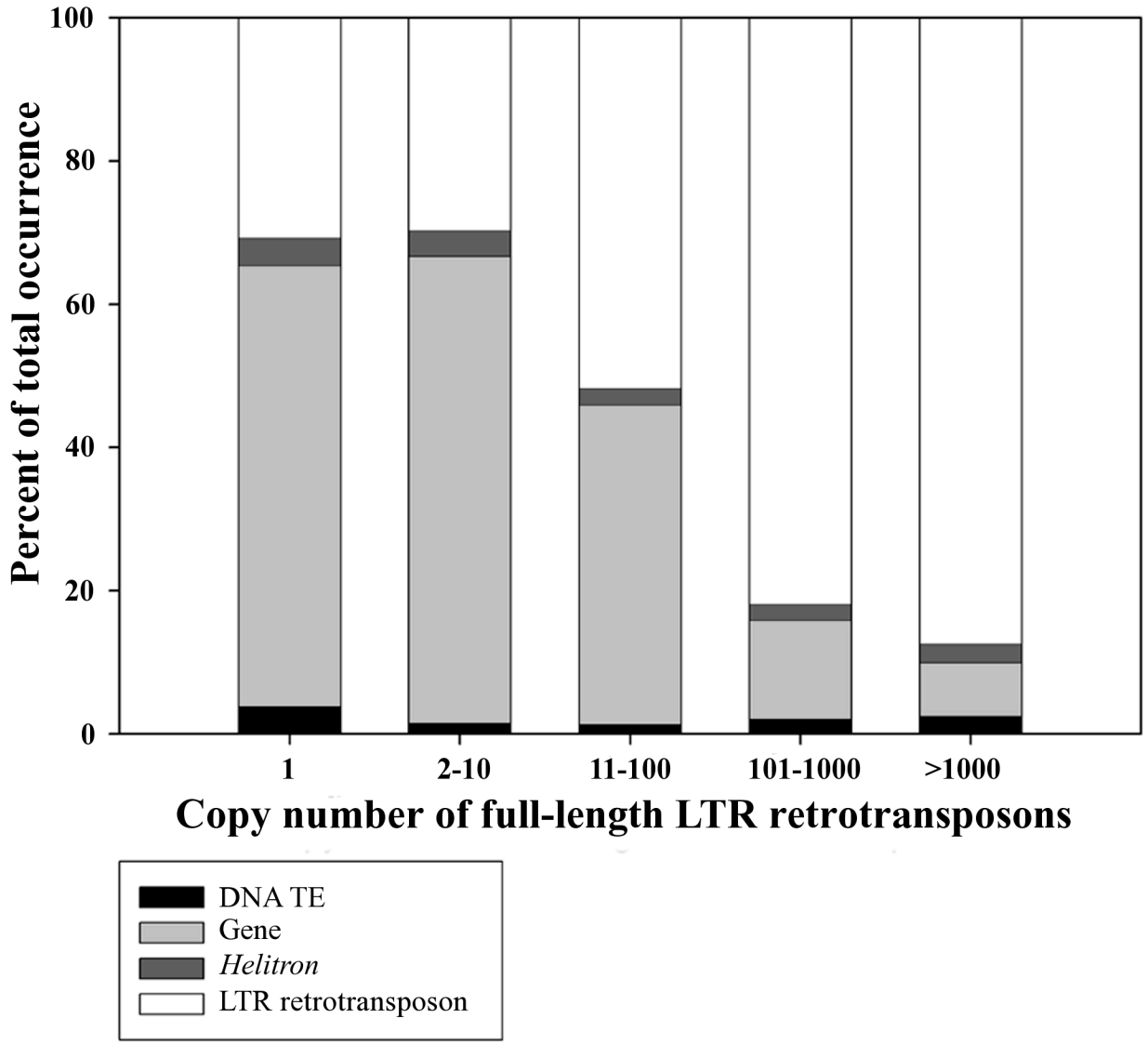


Figure 5.4. The insertion-site preferences of maize LTR retrotransposons. The full-length LTR retrotransposons were placed into bins according to their relative copy number and the results of blast analysis to separate databases of maize genes, cut-and-paste DNA TEs, *Helitrons*, and LTR retrotransposons were summarized according to their copy number classes.

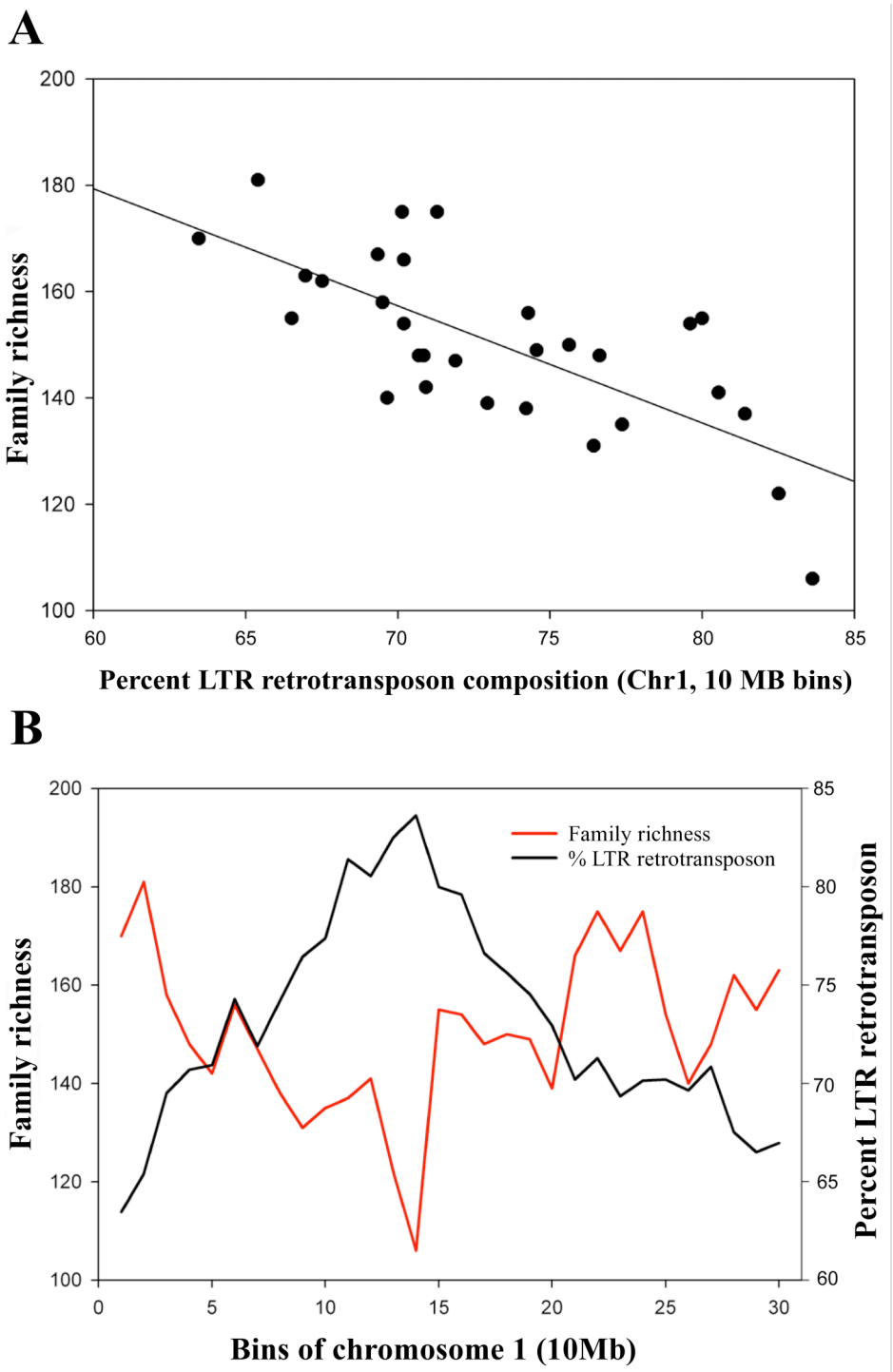


Figure 5.5. Abundance and family richness of LTR retrotransposons found on chromosome 1. (A) The relationship between the % LTR retrotransposon abundance and family richness per 10 Mb bins, and (B) the specific pattern of abundance and richness plotted along the chromosome.

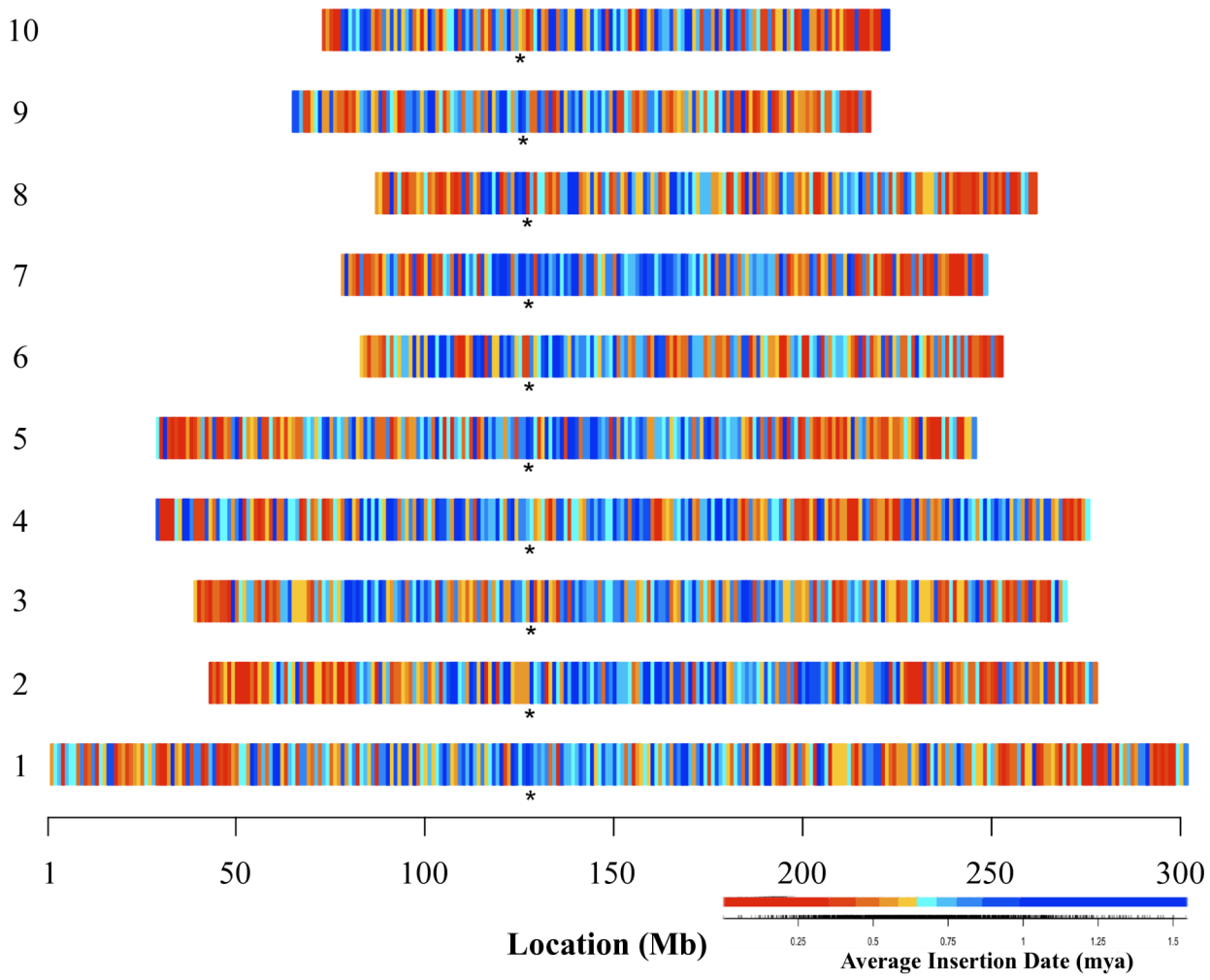


Figure 5.6. The chromosomal distribution of full-length LTR retrotransposon insertion histories. The insertion date of each full-length LTR retrotransposon was determined and these values were averaged for all full length LTR retrotransposons occurring in each 1 Mb bin. The heat map was derived by classifying the average insertion age into equal-interval quantiles. The distribution of these classified ages are illustrated as color tiles superimposed under the empirical cumulative distribution of the average insertion dates for each bin. Asterisks indicate approximate centromere positions.

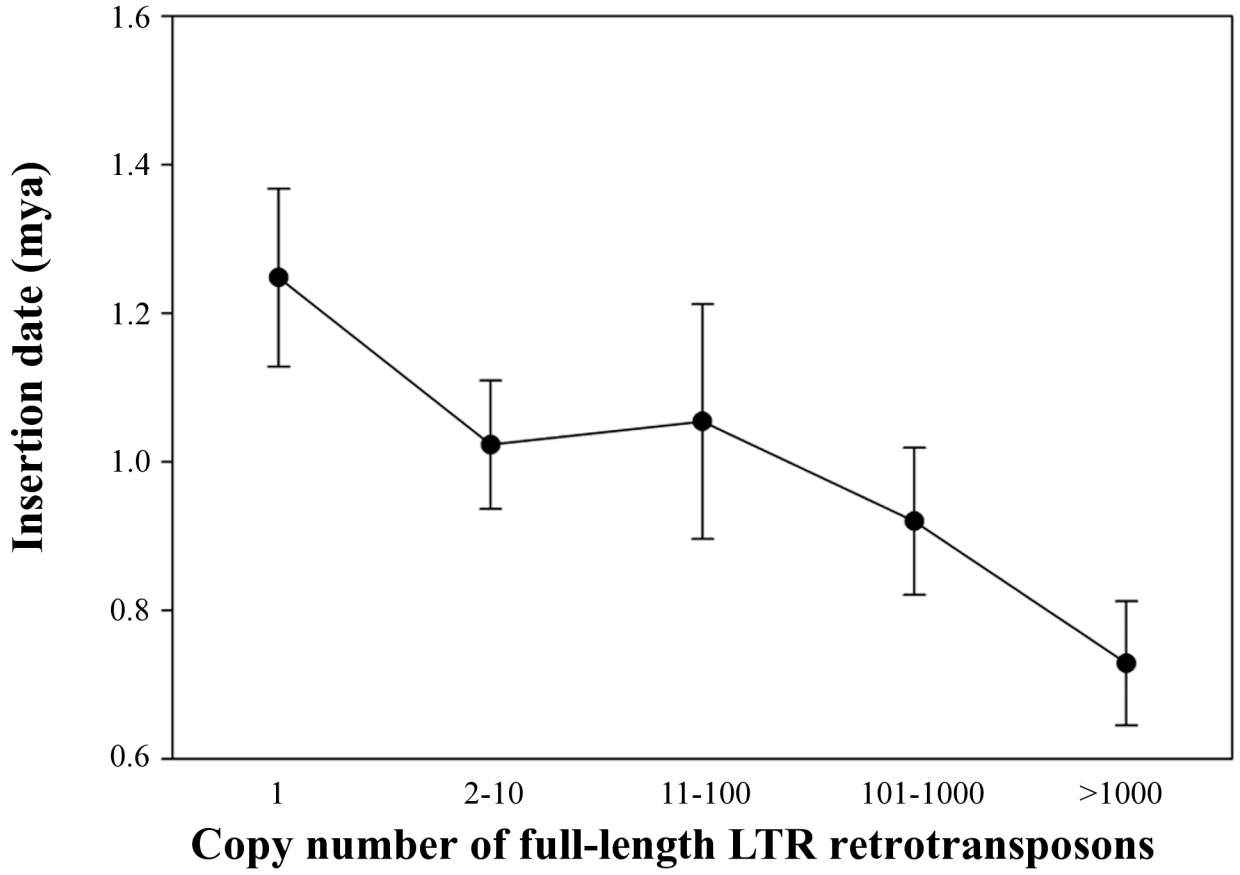


Figure 5.7. The average date of LTR retrotransposon insertion for each of the copy-number classes.

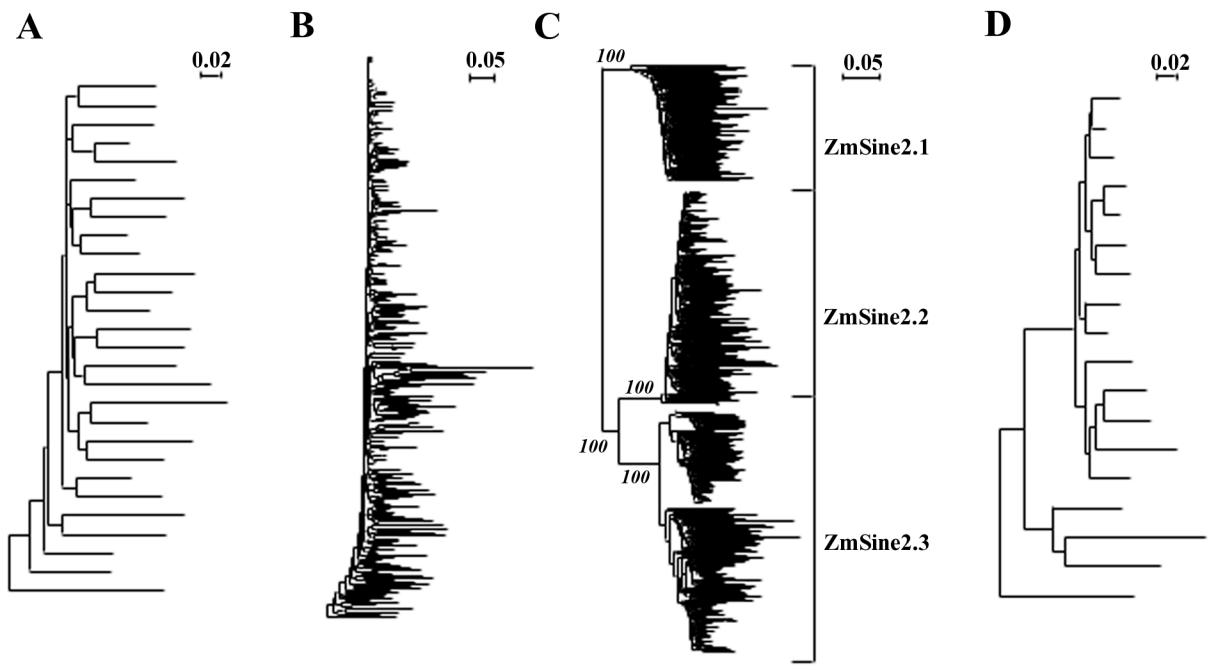


Figure 5.S1. Comparison of maize SINE evolution histories. (A) ZmAU, (B) ZmSINE1, (C) ZmSINE2, and (D) ZmSINE3. All full-length or near full-length elements were analyzed. The phylogenies were obtained using the Neighbor-Joining method. Significant bootstrap values are shown. The nucleotide divergence scale is indicated for each phylogeny.

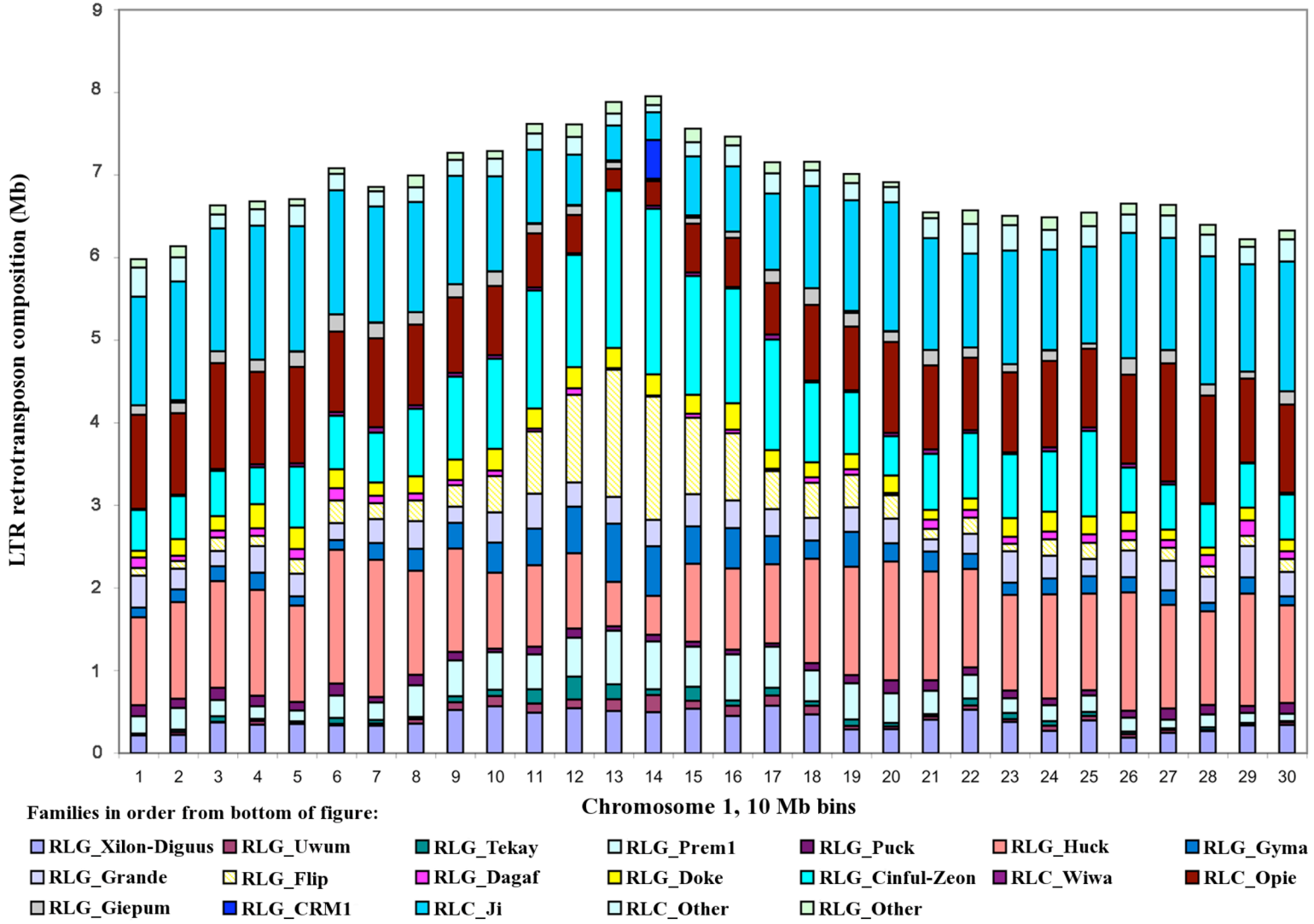


Figure 5.S2. The distribution of LTR retrotransposon family abundance across chromosome 1.

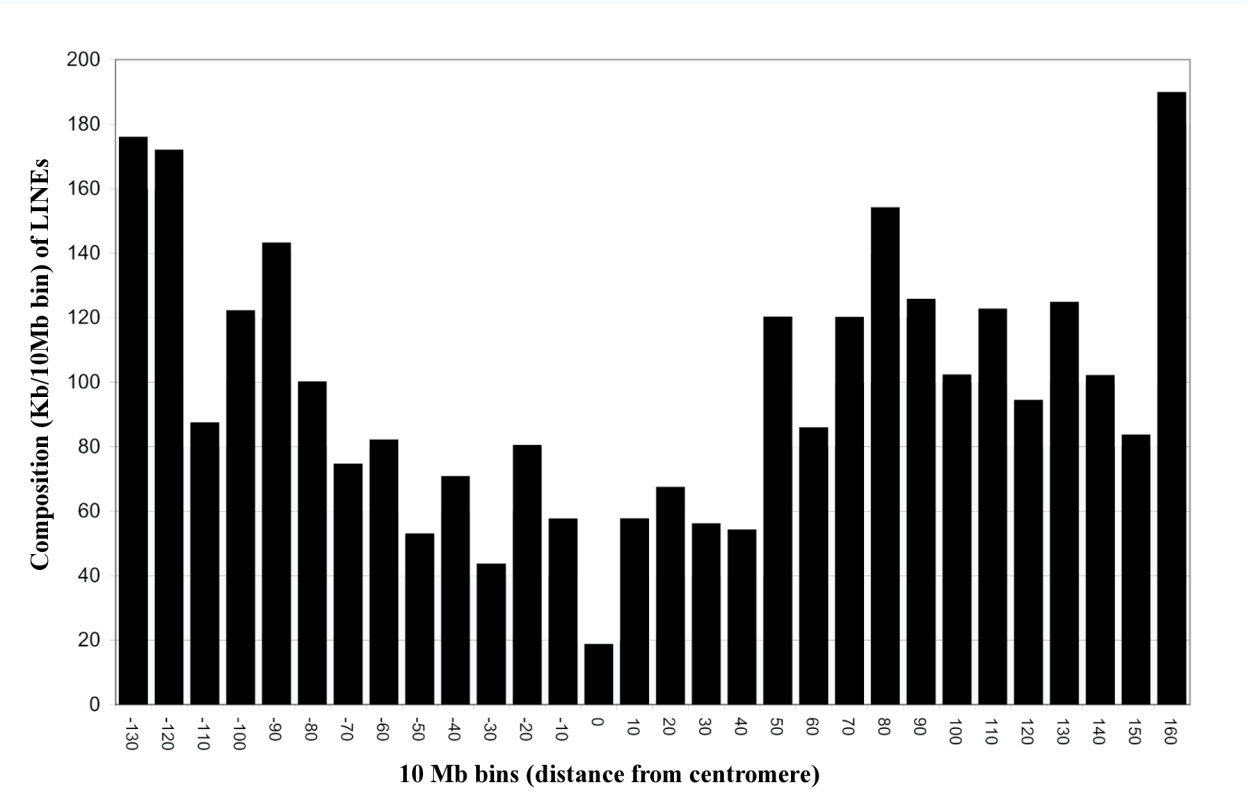


Figure 5.S3. Distribution of LINEs across chromosome 1.

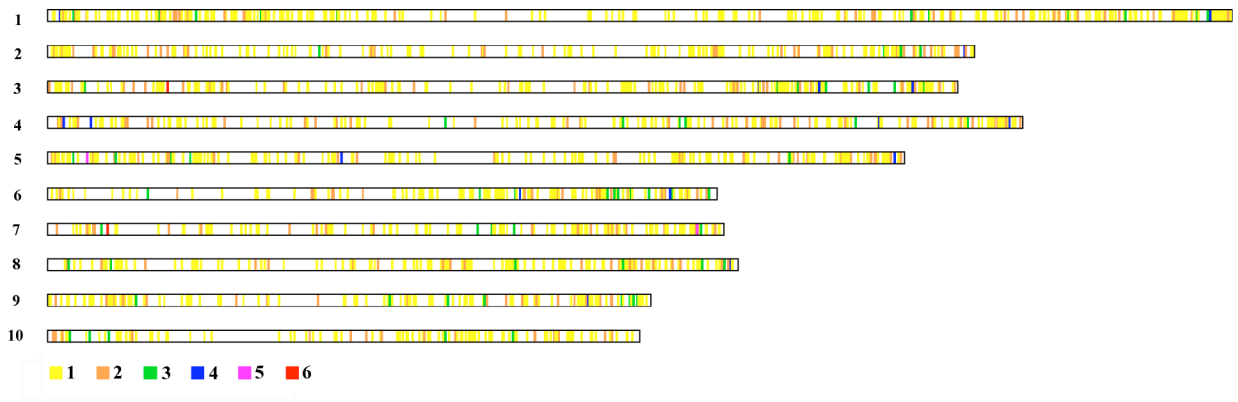


Figure 5.S4. The general distribution of SINEs across the maize chromosomes. Different colors indicate different SINE families, as indicated in the figure.

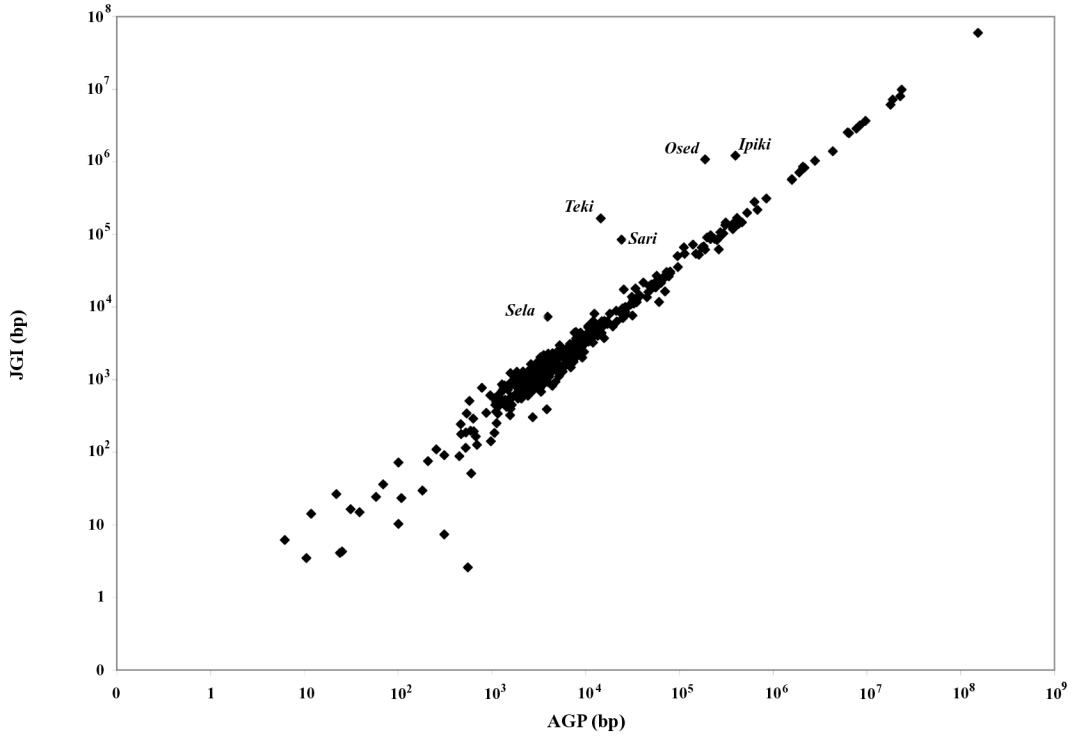


Figure 5.S5. The relationship between the abundance of LTR retrotransposon families found within the AGP compared to their abundance in the sample sequence. Significant outliers are noted on the figure.

Table 5.1. The class I elements within the maize B73 genome.

Superfamily	Number families	Number new families	Number homologous fragments	Mb occupied in the genome	Percent coverage	Number elements containing gene fragments[1]	Number families containing gene fragments
RL *Copia*	109	95	~404,056	484.0	23.7	36	15
RL *Gypsy*	134	117	~476,686	948.3	46.4	168	22
RL Unknown	163	151	~221,635	92.9	4.5	221	44
SINEs	4	3	~1,991	0.5	0.0	n.d.	n.d.
LINEs	31	13	~35,000	20	1.0	n.d.	n.d.
Total	441	379	~1,139,368	1545.7	75.6	425	81

1. n.d. – not determined

Table 5.2. Properties of the top 20 families that comprise ~70% of the maize genome.

Superfamily	Family	Mb in B73, homology search	Count, homology search	Avg. length, homology search	Number of FL elements, structural search	Avg. length, structural search (bp)	Avg. insertion date (mya), FL elements
RLG	*Huck*	233.5	59,208	3,943	3,341	13,407	1.09
RLC	*Ji*	225.8	127,484	1,771	4,093	9,523	0.77
RLG	*Cinful-Zeon*	188.3	82,429	2,284	9,844	8,202	0.60
RLC	*Opie*	178.2	159,512	1,117	3,530	8,888	0.78
RLG	*Flip*	96.3	29,485	3,265	716	14,847	0.86
RLG	*Xilon-Diguus*	83.6	48,297	1,730	197	10,964	0.77
RLG	*Prem1*	77.0	75,605	1,018	1,479	8,958	0.57
RLG	*Gyma*	64.4	39,405	1,635	436	12,797	0.92
RLG	*Grande*	62.3	19,303	3,226	1,338	13,796	0.56
RLG	*Doke*	43.3	19,523	2,217	697	10,630	0.74
RLC	*Giepum*	27.8	28,737	968	186	12,387	0.71
RLX	*Milt*	21.6	16,341	1,319	599	6,308	1.18
RLG	*Puck*	20.7	15,114	1,369	514	9,307	2.17
RLX	*Ruda*	19.2	42,455	451	568	6,485	0.74
RLG	*Tekay*	15.9	15,387	1,031	102	12,102	0.74
RLG	*Uwum*	15.8	13,271	1,191	238	8,495	0.80
RLG	*Dagaf*	15.8	13,991	1,128	185	10,955	0.95
RLX	*Iwik*	8.5	18,024	469	32	13,874	2.29
RLC	*Wiwa*	6.8	4,049	1,675	162	7,935	0.56
RLG	*CRM1*	6.3	3,578	1,761	286	6,918	0.89

FL – Full-length elements

Table 5.3. An analysis of variance showing the relationship between LTR retrotransposon insertion date and the distance to the centromere. Distance to the centromere in 1Mb bins was the dependent variable whereas the square root transformed average insertion date per 1Mb was the independent variable.

Chromosome	df	Type III SS (10^5)	F-value	P-value
1	165	41.45	1.41	0.020
2	143	37.50	0.80	0.886
3	134	30.44	1.43	0.033
4	140	41.01	1.33	0.062
5	107	32.07	1.40	0.044
6	118	28.90	1.36	0.110
7	114	37.99	1.07	0.399
8	126	24.07	0.69	0.945
9	82	23.82	1.74	0.010
10	88	35.71	2.10	0.001

Table 5.S1. Properties of all maize LTR retrotransposons examined in this manuscript.

SF	Family	Mb in B73, homology search	Num. fragments, homology search	Avg. length, homology search	Num. FL, structural search	Avg. length, structural search	Avg. insertion date (mya), FL elements
RLG	*Cinful-Zeon†*	188.308	82429	2284	9844	8202	0.60
RLC	*Ji†*	225.818	127484	1771	4093	9523	0.77
RLC	*Opie*	178.171	159512	1117	3530	8888	0.78
RLG	*Huck*	233.485	59208	3943	3341	13407	1.09
RLG	*Prem1*	76.954	75605	1018	1479	8958	0.57
RLG	*Grande*	62.269	19303	3226	1338	13796	0.56
RLG	*Flip*	96.263	29485	3265	716	14847	0.86
RLG	*Doke*	43.276	19523	2217	697	10630	0.74
RLX	*Milt*	21.552	16341	1319	599	6308	1.18
RLX	*Ruda*	19.152	42455	451	568	6485	0.74
RLG	*Puck†*	20.688	15114	1369	514	9307	2.17
RLG	*Gyma*	64.419	39405	1635	436	12797	0.92
RLG	*CRM1*	6.301	3578	1761	286	6918	0.89
RLG	*CRM4*	5.282	5354	987	249	6004	1.80
RLG	*Uwum*	15.802	13271	1191	238	8495	0.80
RLX	*Name*	0.570	2206	258	212	6024	0.85
RLG	*Xilon-Diguus†*	83.558	48297	1730	197	10964	0.77
RLC	*Giepum*	27.823	28737	968	186	12387	0.71
RLG	*Dagaf*	15.785	13991	1128	185	10955	0.95
RLC	*Wiwa*	6.782	4049	1675	162	7935	0.56
RLC	*Ebel*	4.270	5436	786	150	5301	0.57
RLC	*Raider*	1.972	1090	1809	146	5869	0.73
RLG	*Bosohe*	1.507	1461	1032	106	5182	0.59
RLG	*Tekay*	15.864	15387	1031	102	12102	0.74
RLG	*Ewib*	0.512	794	645	94	2083	1.64
RLG	*Guhis*	2.142	1742	1230	90	7582	1.05
RLX	*Lamyab*	0.181	620	291	88	2551	0.67
RLC	*Agep*	0.254	413	616	67	7636	0.21
RLX	*Baso*	1.117	4401	254	65	12051	2.41
RLC	*Eninu*	1.811	1084	1670	64	7127	0.61
RLX	*Ubow*	4.120	6225	662	57	8935	2.70
RLG	*Bygum*	2.150	3642	590	49	9245	2.44
RLC	*Neha*	2.523	8463	298	39	9610	2.51
RLX	*Iwik*	8.451	18024	469	32	13874	2.29
RLX	*Yraj*	0.796	3530	225	32	6758	0.86
RLX	*CRM3/CentA*	0.727	1445	503	29	5697	0.57
RLC	*Debeh*	0.770	2050	376	27	6711	0.83
RLC	*Stonor*	0.706	1453	486	27	6195	0.36
RLG	*Bogu*	0.492	729	675	27	6733	0.66
RLX	*Yreud*	1.726	5927	291	24	2832	2.76
RLX	*Jeli*	0.079	327	243	22	3097	0.42
RLC	*Fourf*	3.707	4039	918	19	6259	0.62

SF	Family	Mb in B73, homology search	Num. fragments, homology search	Avg. length, homology search	Num. FL, structural search	Avg. length, structural search	Avg. insertion date (mya), FL elements
RLG	*Kubi*	0.321	220	1459	19	5168	0.45
RLX	*Bs1*	0.061	216	285	19	2924	0.63
RLX	*Daju*	0.062	273	226	16	1375	1.07
RLG	*Nopip*	0.079	126	623	14	5191	0.83
RLC	*Wamenu*	0.351	690	509	13	5265	0.26
RLG	*Apil*	0.163	166	985	13	4778	0.71
RLC	*Ahov*	0.158	90	1758	13	5611	0.23
RLC	*Homy*	0.149	164	906	13	5167	0.39
RLC	*Ubat*	0.468	422	1109	11	4740	0.94
RLC	*Ubel*	0.076	195	388	11	4300	0.39
RLX	*CRM2*	1.134	1617	702	10	5864	0.37
RLX	*Avahi*	0.545	3341	163	10	6230	1.58
RLX	*Tuteh*	0.451	1086	415	10	1911	1.63
RLC	*Hesa*	0.244	321	760	10	4772	0.39
RLG	*Bobobo*	0.074	75	987	10	5257	0.07
RLX	*Naseup*	3.123	7202	434	9	12057	2.20
RLC	*Gudyeg*	3.091	3151	981	9	6191	2.00
RLG	*Laiwa*	0.126	211	596	9	4904	0.45
RLX	*Mibaab*	0.024	107	226	9	2572	0.25
RLX	*Odip*	0.238	1267	188	8	3906	0.56
RLG	*Udokup*	0.214	356	600	8	5322	0.38
RLC	*Lafa*	0.135	305	441	8	5015	0.35
RLC	*Fehod*	0.106	124	859	8	5089	0.25
RLX	*Maono*	0.087	242	360	8	2150	1.92
RLG	*Pagof*	0.040	29	1378	8	7523	0.59
RLG	*Sokiit*	0.081	92	878	7	5298	0.70
RLX	*Oweiw*	0.030	65	462	7	2604	0.45
RLC	*Dugiab*	2.809	10178	276	6	5548	3.37
RLC	*Hera*	0.132	230	575	6	5110	0.17
RLC	*Sofi*	0.123	427	288	6	2177	2.40
RLC	*Tata*	0.118	238	495	6	4915	0.93
RLG	*Umojev*	0.069	45	1525	6	5463	0.29
RLC	*Dijap*	3.667	5288	693	5	11505	1.84
RLC	*Anar*	0.669	1010	662	5	5198	0.76
RLG	*Gyte*	0.507	3556	142	5	4224	2.60
RLX	*Sari*	0.241	176	1370	5	2404	4.14
RLG	*Ufonah*	0.122	146	836	5	5456	0.18
RLC	*Kuvi*	0.121	176	686	5	5351	0.85
RLX	*Ukov*	0.098	283	346	5	3278	2.61
RLG	*Guwiot*	0.092	144	637	5	6192	0.92
RLG	*Satulo*	0.087	78	1115	5	5340	0.22
RLG	*Ovikoh*	0.068	48	1426	5	6954	0.36
RLG	*Yfages*	0.068	133	511	5	5220	0.10
RLX	*Halo*	0.030	89	337	5	6817	0.91

SF	Family	Mb in B73, homology search	Num. fragments, homology search	Avg. length, homology search	Num. FL, structural search	Avg. length, structural search	Avg. insertion date (mya), FL elements
RLX	*Mopin*	0.028	103	272	5	4451	0.61
RLX	*Neafu*	0.023	87	270	5	2997	0.07
RLX	*Elalal*	0.014	27	514	5	1845	0.57
RLX	*Dady*	0.012	38	321	5	3866	1.40
RLC	*Victim*	0.708	949	746	4	4849	2.20
RLG	*Udav*	0.520	3977	131	4	9964	0.50
RLX	*Defub*	0.339	865	392	4	10053	3.33
RLC	*Fipi*	0.317	638	497	4	5211	5.09
RLX	*Vedi*	0.293	2444	120	4	3446	0.33
RLX	*Small*	0.250	456	548	4	4037	0.82
RLG	*Amiin*	0.137	95	1445	4	4774	0.31
RLC	*Janoov*	0.135	165	816	4	6606	0.65
RLG	*Nasi*	0.076	79	964	4	5410	1.38
RLG	*Fege*	0.070	110	639	4	5143	1.23
RLG	*Tituer*	0.064	84	759	4	7065	0.64
RLG	*Abiri*	0.063	41	1528	4	5853	0.72
RLG	*Witi*	0.053	40	1330	4	7857	0.07
RLG	*Ojav*	0.053	44	1200	4	5358	0.13
RLG	*Fuved*	0.051	35	1443	4	5310	0.23
RLC	*Giream*	0.050	137	365	4	4618	1.35
RLG	*Soger*	0.040	44	916	4	5597	0.42
RLG	*Uper*	0.037	33	1116	4	5455	0.07
RLG	*Okopam*	0.035	41	864	4	5923	0.67
RLX	*Gufa*	0.033	113	289	4	3523	0.63
RLG	*Lute*	0.033	40	814	4	5570	0.37
RLX	*Ovev*	0.030	130	233	4	8896	2.51
RLC	*Atop*	0.025	45	557	4	5128	0.14
RLC	*Ijiret*	0.024	32	754	4	4795	0.92
RLX	*Gotur*	0.010	68	144	4	2582	1.83
RLX	*Tisy*	0.007	24	271	4	4893	0.33
RLX	*Hiimam*	1.612	6528	247	3	4583	2.75
RLX	*Bene*	0.952	1425	668	3	9383	3.03
RLC	*Donuil*	0.558	1303	428	3	10759	1.67
RLC	*Kake*	0.314	518	606	3	7372	0.22
RLC	*Hopscotch*	0.171	344	497	3	4883	0.11
RLC	*Hani*	0.129	174	742	3	10008	1.24
RLG	*Waepo*	0.114	111	1023	3	5473	1.19
RLG	*Ubid*	0.111	116	959	3	4631	0.07
RLG	*Wuwe*	0.109	609	179	3	6550	0.70
RLC	*Bovo*	0.107	158	674	3	7787	0.34
RLX	*Fajy*	0.097	483	201	3	1932	3.38
RLG	*Olepo*	0.070	63	1109	3	5379	0.77
RLG	*Hooni*	0.069	115	602	3	5415	0.33
RLG	*Weaniv*	0.066	74	895	3	5330	0.09

SF	Family	Mb in B73, homology search	Num. fragments, homology search	Avg. length, homology search	Num. FL, structural search	Avg. length, structural search	Avg. insertion date (mya), FL elements
RLG	*Upus*	0.060	70	861	3	5454	0.42
RLC	*Nowuv*	0.060	130	459	3	5520	0.59
RLX	*Bumy*	0.057	244	233	3	5386	1.63
RLG	*Awuhe*	0.052	50	1044	3	5472	0.95
RLG	*Maro*	0.045	45	994	3	5507	0.77
RLC	*Guvi*	0.040	58	692	3	5808	0.75
RLG	*Magellan*	0.040	31	1294	3	5620	0.25
RLC	*Uhun*	0.038	117	326	3	7376	0.49
RLC	*Sehoad*	0.036	67	531	3	4624	0.22
RLX	*Wiolus*	0.034	97	351	3	5213	1.30
RLX	*Toro*	0.031	122	257	3	3872	1.92
RLG	*Reina*	0.028	29	952	3	5407	0.98
RLX	*Ifab*	0.026	124	208	3	3174	0.13
RLC	*Guwo*	0.025	48	530	3	8758	0.74
RLX	*Rijuep*	0.022	56	397	3	4595	0.14
RLG	*Omoha*	0.022	41	531	3	5394	0.71
RLX	*Teuta*	0.021	136	154	3	2216	1.55
RLC	*Bote*	0.018	112	161	3	6160	0.53
RLX	*Jupek*	0.014	24	570	3	4100	0.35
RLG	*Bobeg*	0.013	49	270	3	3077	0.19
RLX	*Ipiki*	3.945	1454	2713	2	9203	3.10
RLG	*Ywyt*	2.643	3015	876	2	11087	3.00
RLX	*Osed*	1.877	1052	1785	2	2203	6.90
RLX	*Wihov*	0.617	548	1127	2	15043	1.40
RLX	*Leso*	0.602	4380	137	2	1373	3.28
RLG	*Ekoj*	0.501	1728	290	2	3700	0.44
RLC	*Pute*	0.478	982	487	2	5108	1.60
RLC	*Japov*	0.375	713	526	2	7452	1.01
RLG	*Sawujo*	0.347	2395	145	2	19972	3.07
RLX	*Vora*	0.263	2250	117	2	4285	1.62
RLC	*Totu*	0.258	688	375	2	4965	5.77
RLX	*Nabu*	0.215	464	462	2	3245	n.d.
RLG	*Piube*	0.157	114	1379	2	5206	0.04
RLX	*Onub*	0.137	153	893	2	16847	0.25
RLC	*Dolovu*	0.134	138	972	2	5170	1.19
RLX	*Kaise*	0.125	199	628	2	1180	1.17
RLG	*Lyruom*	0.122	261	469	2	6106	1.68
RLC	*Omud*	0.112	250	448	2	5355	1.68
RLC	*Gekog*	0.110	203	543	2	4773	0.39
RLG	*Ulik*	0.109	109	1002	2	5494	0.95
RLC	*Huti*	0.097	208	467	2	4754	1.62
RLG	*Anysaf*	0.097	136	712	2	5201	0.29
RLC	*Nitat*	0.093	164	565	2	5122	0.00
RLG	*Rufefu*	0.092	79	1167	2	5565	0.23

SF	Family	Mb in B73, homology search	Num. fragments, homology search	Avg. length, homology search	Num. FL, structural search	Avg. length, structural search	Avg. insertion date (mya), FL elements
RLG	*Riiryl*	0.091	100	906	2	5504	0.24
RLG	*Oveah*	0.089	115	771	2	5338	0.52
RLC	*Iseb*	0.080	81	984	2	5337	0.00
RLG	*Ojam*	0.077	82	934	2	5353	0.00
RLC	*Bihar*	0.075	220	340	2	4155	1.73
RLG	*Iwim*	0.074	73	1016	2	6637	0.11
RLG	*Lowy*	0.071	77	925	2	5197	0.00
RLC	*Yrer*	0.070	91	769	2	5282	0.28
RLC	*Ehahu*	0.063	62	1021	2	5056	1.42
RLG	*Ovamef*	0.062	95	655	2	5404	1.07
RLG	*Ures*	0.062	96	644	2	5484	0.73
RLG	*Ojokat*	0.058	55	1055	2	5557	0.25
RLC	*Mauky*	0.055	104	532	2	5390	0.94
RLG	*Wemu*	0.052	52	998	2	5609	0.65
RLC	*Seko*	0.049	82	596	2	5247	0.61
RLG	*Naijaj*	0.048	74	642	2	5558	1.21
RLG	*Notu*	0.046	40	1141	2	5380	0.12
RLG	*Aneas*	0.044	40	1103	2	5519	0.51
RLG	*Bomevy*	0.044	56	778	2	5496	0.14
RLG	*Usuf*	0.043	39	1097	2	5230	0.63
RLC	*Tatu*	0.042	104	406	2	20664	0.60
RLG	*Boha*	0.042	50	831	2	5940	0.66
RLG	*Gofi*	0.041	66	624	2	5354	0.15
RLG	*Jaws*	0.041	237	172	2	6074	2.50
RLC	*Tiwewi*	0.039	66	593	2	14210	0.57
RLC	*Seufyt*	0.037	101	368	2	5276	0.00
RLC	*Wawo*	0.036	229	159	2	9479	0.67
RLG	*Ywuv*	0.036	61	591	2	5167	2.31
RLX	*Raga*	0.036	115	311	2	3194	0.00
RLC	*Tufe*	0.033	316	104	2	13464	6.43
RLC	*Rely*	0.033	55	592	2	4706	0.00
RLX	*Kupu*	0.032	88	366	2	10252	2.03
RLX	*Mako*	0.032	194	165	2	3010	0.14
RLG	*Tuku*	0.030	41	729	2	5600	0.67
RLC	*Fuvej*	0.029	55	536	2	3288	0.74
RLX	*Petopi*	0.029	157	187	2	3290	0.00
RLC	*Gilovu*	0.028	82	339	2	4861	0.09
RLC	*Ulyg*	0.028	49	566	2	8767	0.60
RLG	*Weki*	0.027	24	1140	2	5492	0.00
RLG	*Ugano*	0.026	26	985	2	5608	1.40
RLG	*Wyly*	0.024	24	986	2	5545	0.55
RLG	*Yvoj*	0.023	26	871	2	5419	0.22
RLX	*Mafogo*	0.022	82	267	2	2942	0.06
RLC	*Huta*	0.021	65	329	2	4725	2.06

SF	Family	Mb in B73, homology search	Num. fragments, homology search	Avg. length, homology search	Num. FL, structural search	Avg. length, structural search	Avg. insertion date (mya), FL elements
RLG	*Uwuw*	0.020	27	746	2	5494	0.65
RLX	*Wiru*	0.019	86	223	2	3542	1.39
RLC	*Ugog*	0.019	57	332	2	4964	2.42
RLG	*Bavav*	0.019	44	426	2	5671	0.00
RLX	*Ruwi*	0.018	85	211	2	3059	0.86
RLX	*Bawigu*	0.017	69	244	2	3324	0.37
RLX	*Fanuab*	0.016	87	187	2	4319	0.30
RLX	*Sido*	0.015	51	302	2	2652	3.51
RLX	*Mafigi*	0.013	75	179	2	3444	1.45
RLG	*Suda*	0.013	22	592	2	5311	0.54
RLX	*Nakuuv*	0.013	40	319	2	7368	0.59
RLX	*Urogor*	0.011	41	265	2	1888	1.10
RLX	*Ewot*	0.011	46	235	2	3171	0.00
RLX	*Juta*	0.009	39	226	2	3138	0.00
RLC	*Erev*	0.008	31	252	2	5429	0.58
RLX	*Mijuw*	0.006	12	462	2	5529	0.35
RLC	*Nuhan*	4.640	17550	264	1	14219	6.41
RLC	*Machiavelli*	3.638	3451	1054	1	5701	2.12
RLX	*Loukuv*	2.933	13645	215	1	9909	3.36
RLX	*Vegu*	2.728	2113	1291	1	6245	4.75
RLX	*Hutu*	2.368	10561	224	1	20168	3.13
RLG	*Lata*	1.880	3996	470	1	9068	4.57
RLX	*Naiba*	1.394	6872	203	1	21781	0.92
RLX	*Afuv*	0.965	6027	160	1	14195	n.d.
RLC	*Ibulaf*	0.800	813	984	1	4625	1.51
RLX	*Ojah*	0.774	4272	181	1	9869	n.d.
RLC	*Uloh*	0.638	510	1250	1	7334	0.00
RLX	*Uwub*	0.565	5910	96	1	11064	0.72
RLX	*Kawivo*	0.542	3422	158	1	11243	n.d.
RLG	*Pebi*	0.416	928	448	1	11583	0.24
RLX	*Buire*	0.361	3604	100	1	2632	7.75
RLX	*Utar*	0.326	2728	119	1	1925	n.d.
RLX	*Afeke*	0.303	2984	102	1	1857	n.d.
RLX	*Panen*	0.265	1549	171	1	2925	n.d.
RLX	*Ugymos*	0.264	1221	216	1	5243	n.d.
RLX	*Pibo*	0.258	1488	173	1	2318	n.d.
RLX	*Lyna*	0.249	454	549	1	7435	0.21
RLX	*Habu*	0.232	2865	81	1	1554	4.88
RLC	*Depuw*	0.195	161	1211	1	5471	0.74
RLC	*Labe*	0.171	332	516	1	3005	2.49
RLX	*Kahoba*	0.157	1295	121	1	2635	n.d.
RLX	*Ywely*	0.149	1222	122	1	3130	n.d.
RLG	*Ahoru*	0.148	385	384	1	5702	0.00
RLX	*Teki*	0.145	139	1041	1	1514	n.d.

SF	Family	Mb in B73, homology search	Num. fragments, homology search	Avg. length, homology search	Num. FL, structural search	Avg. length, structural search	Avg. insertion date (mya), FL elements
RLC	*Lusi*	0.136	307	442	1	4993	0.62
RLX	*Ujinas*	0.125	333	374	1	7136	0.65
RLC	*Ajipe*	0.122	495	246	1	2543	2.56
RLX	*Vusu*	0.113	400	284	1	7071	n.d.
RLC	*Afad*	0.107	217	493	1	5166	0.75
RLC	*Urum*	0.103	212	487	1	5202	0.09
RLC	*Jelat*	0.098	461	213	1	1602	n.d.
RLG	*Vufi*	0.092	66	1396	1	5418	0.13
RLC	*Hago*	0.088	103	854	1	5511	0.73
RLX	*Ilyl*	0.087	577	151	1	1408	n.d.
RLC	*Owiit*	0.084	130	647	1	5100	0.00
RLC	*Tiwe*	0.084	189	444	1	4812	0.15
RLC	*Uwaf*	0.081	199	407	1	4903	0.00
RLG	*Gati*	0.080	103	776	1	5617	0.71
RLG	*Mufeub*	0.080	241	331	1	5131	0.19
RLC	*Pifo*	0.076	115	662	1	5241	0.33
RLG	*Ikal*	0.069	53	1310	1	5481	0.27
RLG	*Ewog*	0.067	107	625	1	5436	1.65
RLC	*Ydut*	0.066	83	794	1	4666	0.86
RLG	*Mywur*	0.063	89	704	1	5230	0.68
RLC	*Vuijon*	0.062	116	538	1	9200	0.91
RLC	*Volo*	0.059	122	481	1	5199	0.37
RLC	*Vodida*	0.058	105	554	1	4857	0.00
RLG	*Kase*	0.058	57	1017	1	5304	1.51
RLC	*Ogiv*	0.057	356	160	1	5212	2.12
RLG	*Ewiut*	0.057	62	919	1	5268	0.00
RLX	*Dala*	0.053	386	137	1	1688	3.89
RLC	*Niki*	0.051	102	502	1	5151	0.50
RLX	*Joemon*	0.051	439	116	1	2194	n.d.
RLX	*Etug*	0.051	171	296	1	3425	2.16
RLC	*Ajajog*	0.050	64	782	1	5469	n.d.
RLG	*Nobe*	0.050	49	1021	1	5463	0.24
RLG	*Gylu*	0.048	53	902	1	5240	0.13
RLX	*Soefes*	0.047	191	245	1	2977	2.34
RLX	*Rulo*	0.046	334	139	1	3353	n.d.
RLG	*Oguod*	0.046	55	842	1	5467	0.38
RLG	*Nakovu*	0.045	30	1507	1	5448	0.11
RLX	*Anim*	0.045	138	324	1	1659	n.d.
RLC	*Ruhi*	0.045	125	357	1	4710	0.38
RLG	*Epohi*	0.043	86	504	1	5106	3.55
RLG	*Ijaat*	0.042	45	940	1	5544	1.79
RLC	*Finaij*	0.042	76	554	1	4787	1.10
RLG	*Ytub*	0.041	54	752	1	5516	0.27
RLX	*Arar*	0.039	75	524	1	4148	0.52

SF	Family	Mb in B73, homology search	Num. fragments, homology search	Avg. length, homology search	Num. FL, structural search	Avg. length, structural search	Avg. insertion date (mya), FL elements
RLX	*Sela*	0.039	301	130	1	1729	1.02
RLC	*Ytar*	0.038	81	471	1	4844	0.18
RLC	*Ajit*	0.037	69	533	1	4989	1.98
RLC	*Naasuj*	0.033	49	677	1	4756	0.00
RLG	*Ruugu*	0.033	107	309	1	5273	0.77
RLG	*Loba*	0.033	64	509	1	3478	0.00
RLX	*Hoda*	0.032	147	217	1	2955	1.15
RLG	*Uwew*	0.032	50	631	1	5281	0.27
RLG	*Hute*	0.031	72	435	1	3857	0.42
RLX	*Demo*	0.031	185	168	1	1742	1.60
RLG	*Sowu*	0.031	46	677	1	3888	0.00
RLG	*Uvet*	0.029	31	941	1	5352	0.00
RLX	*Vafim*	0.029	184	157	1	3391	0.30
RLG	*Bowuow*	0.028	59	479	1	5400	0.39
RLX	*Yemi*	0.028	107	263	1	9765	0.39
RLX	*Jakek*	0.027	202	136	1	1775	n.d.
RLG	*Lywy*	0.027	42	643	1	5275	0.81
RLX	*Ilofaw*	0.026	109	242	1	3305	0.56
RLC	*Atej*	0.026	100	260	1	5407	0.38
RLX	*Nisow*	0.026	124	209	1	3635	1.13
RLG	*Sywu*	0.026	40	640	1	5397	2.57
RLX	*Kinosi*	0.025	125	199	1	3550	0.11
RLX	*Alaw*	0.025	126	196	1	1858	2.52
RLG	*Usif*	0.023	53	440	1	4004	2.98
RLG	*Lise*	0.023	49	475	1	5350	0.00
RLG	*Labu*	0.023	42	553	1	5474	0.33
RLG	*Moorud*	0.023	37	613	1	5469	0.99
RLX	*Beby*	0.022	163	138	1	3094	0.18
RLX	*Neteut*	0.022	76	292	1	2858	4.45
RLX	*Emuh*	0.022	167	133	1	1281	n.d.
RLX	*Fate*	0.022	96	230	1	3415	0.11
RLX	*Ulon*	0.022	82	269	1	3607	n.d.
RLG	*Dabe*	0.022	30	723	1	5444	2.84
RLC	*Votaed*	0.021	137	157	1	16653	1.85
RLG	*Rowi*	0.021	28	736	1	5494	0.12
RLX	*Eguh*	0.020	12	1635	1	1457	n.d.
RLX	*Pope*	0.019	95	200	1	2452	n.d.
RLX	*Saahol*	0.018	114	162	1	3686	1.58
RLG	*Dadeir*	0.018	34	542	1	5516	1.26
RLG	*Rimaar*	0.017	31	544	1	5489	0.60
RLX	*Epom*	0.016	100	163	1	2206	n.d.
RLC	*Taname*	0.016	45	359	1	4652	0.37
RLX	*Mulaf*	0.016	93	172	1	3288	0.00
RLG	*Seuwe*	0.016	28	572	1	5623	1.37

SF	Family	Mb in B73, homology search	Num. fragments, homology search	Avg. length, homology search	Num. FL, structural search	Avg. length, structural search	Avg. insertion date (mya), FL elements
RLC	*Nene*	0.016	29	541	1	5309	2.31
RLG	*Boja*	0.016	17	919	1	5463	1.60
RLG	*Kise*	0.015	29	511	1	5454	2.59
RLC	*Ubep*	0.015	33	449	1	4748	0.00
RLC	*Ypel*	0.015	31	471	1	1823	n.d.
RLX	*Nana*	0.014	40	360	1	2763	0.95
RLX	*Uvis*	0.013	55	242	1	5145	0.43
RLX	*Vufe*	0.013	36	355	1	5475	0.00
RLX	*Wuywu*	0.013	22	577	1	1881	0.08
RLG	*Ivuk*	0.012	24	498	1	5290	2.33
RLX	*Kahowu*	0.012	50	239	1	3628	0.00
RLX	*Vuna*	0.012	34	339	1	17890	n.d.
RLX	*Mewu*	0.011	143	80	1	1759	n.d.
RLC	*Onal*	0.011	37	306	1	4826	0.00
RLG	*Taro*	0.011	26	428	1	5607	0.88
RLX	*Poarow*	0.011	32	347	1	2913	1.70
RLX	*Niypo*	0.011	30	369	1	1826	0.69
RLX	*Bori*	0.011	29	367	1	2725	n.d.
RLX	*Wawu*	0.010	40	243	1	1472	1.65
RLX	*Uluil*	0.007	21	331	1	1361	n.d.
RLX	*Okur*	0.007	28	240	1	17410	0.55
RLC	*Muekeh*	0.006	17	371	1	1379	2.03
RLX	*Jube*	0.006	14	431	1	1341	7.85
RLX	*Luteja*	0.006	3	1974	1	1364	n.d.
RLX	*Fosu*	0.006	12	480	1	2146	n.d.
RLX	*Muusi*	0.005	14	384	1	1703	n.d.
RLC	*Waneer*	0.005	11	488	1	3900	n.d.
RLX	*Ewigyw*	0.005	14	376	1	20063	0.70
RLX	*Tojena*	0.005	26	181	1	1306	0.00
RLX	*Vaofen*	0.005	9	517	1	2784	2.01
RLX	*Nyjuvy*	0.005	14	323	1	2416	0.00
RLX	*Ovys*	0.004	6	640	1	6471	n.d.
RLX	*Okoj*	0.003	9	345	1	1557	n.d.
RLX	*Oviil*	0.003	9	345	1	8696	n.d.
RLX	*Lenu*	0.003	17	150	1	1715	0.08
RLX	*Miva*	0.002	1	2205	1	1881	n.d.
RLX	*Owume*	0.002	3	694	1	1221	5.31
RLX	*Epiil*	0.002	6	303	1	4669	0.20
RLC	*Hiri*	0.001	13	83	1	4542	n.d.
RLX	*Ohag*	0.001	6	168	1	11090	n.d.
RLX	*Fara*	0.001	10	101	1	5310	n.d.
RLX	*Wugaab*	0.001	6	115	1	5422	n.d.
RLX	*Gate*	0.001	8	71	1	4601	n.d.
RLX	*Etin*	0.000	8	47	1	4375	n.d.

SF	Family	Mb in B73, homology search	Num. fragments, homology search	Avg. length, homology search	Num. FL, structural search	Avg. length, structural search	Avg. insertion date (mya), FL elements
RLX	*Ajeb*	0.000	4	86	1	4353	n.d.
RLX	*Ewuof*	0.000	3	114	1	5050	n.d.
RLX	*Efaw*	0.000	4	78	1	3870	n.d.
RLX	*Adun*	0.000	5	51	1	2783	n.d.
RLX	*Guafa*	0.000	3	80	1	3598	n.d.
RLX	*Biwa*	0.000	3	73	1	14520	n.d.
RLX	*Guali*	0.000	2	59	1	6076	n.d.
RLX	*Bosovu*	0.000	3	34	1	3053	n.d.
RLX	*Doba*	0.000	1	93	1	4005	n.d.
RLX	*Tywo*	0.000	1	62	1	3101	n.d.
RLX	*Beboso*	0.000	1	61	1	2563	n.d.
RLX	*Baha*	0.000	1	29	1	2478	n.d.

Table 5.S2. Gene capture events uncovered in the full-length LTR retrotransposons mined from the B73 genome.

SF	Family[1]	Fragment(s) aquired	Copy number in B73	bp
RLC	*Bote*	FAD-binding family	1	138
RLC	*Debeh*	HAT dimerisation family	1	166
RLC	*Ebel*	Rice gene family candidate, 0006430	1	117
RLC	*Ebel*	Rice gene family candidate, 0011514	1	106
RLC	*Ebel*	Rice gene family candidate, 0012045	2	166
RLC	*Ebel*	WD40 repeat family	2	564
RLC	*Erev*	Peptidase aspartic, catalytic family	2	406
RLC	*Fourf*	Peptidase S10, serine carboxypeptidase family	17	1853
RLC	*Gudyeg*	Peptidase S10, serine carboxypeptidase family	7	742
RLC	*Homy*	ABC transporter related family	12	3565
RLC	*Iseb*	Typical LEA gene family	2	275
RLC	*Ji*	AAA ATPase, core family	2	220
RLC	*Ji*	Chaperonin Cpn60/TCP-1 family	1	75
RLC	*Ji*	Coatomer, epsilon subunit family	1	162
RLC	*Ji*	Cystathionine beta-synthase, core family	2	172
RLC	*Ji*	Heat shock protein 70 family	2	350
RLC	*Ji*	Inorganic H+ pyrophosphatase family	2	442
RLC	*Ji*	Pentatricopeptide repeat family	2	204
RLC	*Ji*	Plant peroxidase family	1	80
RLC	*Ji*	Protein kinase, core family	2	203
RLC	*Ji*	Rice gene family candidate, 0010468	1	68
RLC	*Ji*	Zinc finger, TTF-type family	1	1458
RLC	*Opie*	Peptidase S14, ClpP family	2	246
RLC	*Opie*	Regulator of chromosome condensation/beta-lactamase-inhibitor protein II family	1	211
RLC	*Sofi*	Rice gene family candidate, 0006675	2	148
RLC	*Ubel*	Ribulose-phosphate binding barrel family	4	812
RLG	*Anysaf*	Rice gene family candidate, 0008142	1	53
RLG	*Bogu*	Peptidase aspartic, catalytic family	5	1233
RLG	*Bosohe*	Peptidase aspartic, catalytic family	20	6340
RLG	*Bygum*	Homeobox family	1	98
RLG	*Cinful-Zeon*	AAA ATPase, core family	1	153
RLG	*Cinful-Zeon*	Caffeic acid O-methyltransferase family	24	1668
RLG	*Cinful-Zeon*	DNA polymerase III, subunits gamma and tau family	1	108
RLG	*Cinful-Zeon*	Glyceraldehyde 3-phosphate dehydrogenase family	2	168
RLG	*Cinful-Zeon*	Ribosomal protein S4E family	1	55
RLG	*Cinful-Zeon*	Rice gene family candidate, 0008745	1	158
RLG	*Cinful-Zeon*	Rice gene family candidate, 0011056	1	68
RLG	*Ewib*	Helicase, C-terminal family	5	461
RLG	*Ewiut*	Peptidase aspartic, catalytic family	1	93
RLG	*Flip*	Putative von Willebrand factor, type A family	2	106

SF	Family[1]	Fragment(s) aquired	Copy number in B73	bp
RLG	*Flip*	Short organization (SHO) gene family	1	142
RLG	*Gofi*	Chromo family	2	232
RLG	*Huck*	Armadillo-like helical family	1	71
RLG	*Huck*	Caffeic acid O-methyltransferase family	14	841
RLG	*Huck*	Peptidase aspartic, catalytic family	1	1147
RLG	*Huck*	Rice gene family candidate, 0011056	1	74
RLG	*Iwim*	Peptidase aspartic, catalytic family	2	140
RLG	*Jaws*	Peptidase aspartic, catalytic family	2	132
RLG	*Laiwa*	Peptidase aspartic, catalytic family	9	8780
RLG	*Maro*	Chromo family	3	199
RLG	*Mufeub*	Peptidase aspartic, catalytic family	1	113
RLG	*Mywur*	Peptidase aspartic, catalytic family	1	74
RLG	*Oguod*	Chromo family	1	62
RLG	*Ojokat*	Chromo family	2	156
RLG	*Prem1*	AAA ATPase, core family	1	61
RLG	*Prem1*	Protein phosphatase 2C-related family	1	93
RLG	*Prem1*	Rice gene family candidate, 0006743	1	107
RLG	*Prem1*	Rice gene family candidate, 0011056	7	476
RLG	*Prem1*	Rice gene family candidate, 0011109	5	411
RLG	*Puck*	Homeobox family	12	2523
RLG	*Puck*	Putative Eggshell protein family	4	207
RLG	*Puck*	Rice gene family candidate, 0007576	1	258
RLG	*Sowu*	Chromo family	1	65
RLG	*Suda*	Peptidase aspartic, catalytic family	1	69
RLG	*Ubid*	Chromo family	2	150
RLG	*Udav*	Peptidase aspartic, catalytic family	2	124
RLG	*Ugano*	Chromo family	2	118
RLG	*Uwew*	Peptidase aspartic, catalytic family	1	92
RLG	*Vufi*	Chromo family	1	70
RLG	*Weaniv*	Peptidase aspartic, catalytic family	3	267
RLG	*Wuwe*	Chromo family	2	118
RLG	*Ytub*	Chromo family	1	94
RLX	*Bawigu*	Rice gene family candidate, 0011592	2	1102
RLX	*Beby*	C2 calcium-dependent membrane targeting family	1	104
RLX	*Bs1*	Aldehyde dehydrogenase family	4	1170
RLX	*Bs1*	Mitochodrial transcription termination factor-related family	5	405
RLX	*Bs1*	Pentatricopeptide repeat family	2	1002
RLX	*Bs1*	Transcription factor jumonji/aspartyl beta-hydroxylase family	1	92
RLX	*Defub*	Peptidase M16, C-terminal family	1	127
RLX	*Gotur*	Cytochrome b5 family	1	169
RLX	*Gufa*	Heat shock protein 70 family	2	296
RLX	*Ifab*	Ribose 5-phosphate isomerase family	3	235

SF	Family[1]	Fragment(s) aquired	Copy number in B73	bp
RLX	*Jeli*	Pentatricopeptide repeat family	1	402
RLX	*Jeli*	Protein of unknown function DUF248, methyltransferase putative family	5	2430
RLX	*Jeli*	Rice gene family candidate, 0010431	5	800
RLX	*Jeli*	Rice gene family candidate, 0011246	1	115
RLX	*Jupek*	Peptidase S9A, prolyl oligopeptidase family	3	712
RLX	*Kinosi*	Protein prenyltransferase family	1	145
RLX	*Lamyab*	Peptidase aspartic, catalytic family	13	9474
RLX	*Mafigi*	Raffinose synthase family	2	202
RLX	*Mako*	Heat shock protein Hsp90 family	2	480
RLX	*Mibaab*	AAA ATPase, core family	1	952
RLX	*Mibaab*	Copper amine oxidase family	1	57
RLX	*Mibaab*	Pathogenesis-related transcriptional factor and ERF family	2	362
RLX	*Mibaab*	Rice gene family candidate, 0007820	1	53
RLX	*Mibaab*	Thiolase family	4	460
RLX	*Mopin*	Glucose/ribitol dehydrogenase family	3	261
RLX	*Mulaf*	Xanthine/uracil/vitamin C permease family	1	394
RLX	*Name*	Snf7 family	1	94
RLX	*Neafu*	Enoyl-CoA hydratase/isomerase family	2	519
RLX	*Neafu*	Malic oxidoreductase family	3	411
RLX	*Nyjuvy*	Rice gene family candidate, 0009691	1	95
RLX	*Oweiw*	AAA ATPase, core family	6	780
RLX	*Oweiw*	Elongation factor G, III and V family	1	119
RLX	*Petopi*	Galactokinase family	2	428
RLX	*Raga*	Rice gene family candidate, 0007733	2	236
RLX	*Ruwi*	Cytochrome P450 family	1	103
RLX	*Ruwi*	Rice gene family candidate, 0011119	1	65
RLX	*Saahol*	Protein prenyltransferase family	1	145
RLX	*Toro*	Tubulin family	3	186
RLX	*Tuteh*	Plant peroxidase family	1	72
RLX	*Tuteh*	Rice gene family candidate, 0006969	1	177
RLX	*Tuteh*	Zinc finger, RING-type family	1	104
RLX	*Vafim*	Cinnamyl alcoholdehydrogenase family	1	181
RLX	*Vedi*	Histone H3 family	2	390
RLX	*Vedi*	Lecithin:cholesterol acyltransferase family	2	572
RLX	*Wiolus*	Armadillo-like helical family	3	725
RLX	*Wiru*	Rice gene family candidate, 0011592	1	324
RLX	*Yemi*	Transcription factor jumonji/aspartyl beta-hydroxylase family	1	668
RLX	*unk*	Aminotransferase, class I and II family	1	87
RLX	*unk*	ATP-dependent DNA helicase RecQ family	1	110
RLX	*unk*	Caffeic acid O-methyltransferase family	4	255
RLX	*unk*	Chaperonin clpA/B family	1	182
RLX	*unk*	Chaperonin Cpn10 family	1	163

SF	Family[1]	Fragment(s) aquired	Copy number in B73	bp
RLX	*unk*	Chromo family	2	118
RLX	*unk*	Cinnamyl alcoholdehydrogenase family	1	181
RLX	*unk*	Copper amine oxidase family	1	57
RLX	*unk*	Cystathionine beta-synthase, core family	1	82
RLX	*unk*	Cytochrome P450 family	1	248
RLX	*unk*	Enoyl-CoA hydratase/isomerase family	1	266
RLX	*unk*	Exoribonuclease, phosphorolytic domain 1 family	1	186
RLX	*unk*	F-box family	1	272
RLX	*unk*	Glyoxalase/extradiol ring-cleavage dioxygenase family	1	132
RLX	*unk*	Heat shock protein 70 family	3	248
RLX	*unk*	Homeobox family	4	619
RLX	*unk*	HSP20-like chaperone family	1	100
RLX	*unk*	Laccase family	1	232
RLX	*unk*	Mandelate racemase/muconate lactonizing enzyme family	1	67
RLX	*unk*	Mitochodrial transcription termination factor-related family	1	81
RLX	*unk*	NPH3 family	2	120
RLX	*unk*	Nucleotide-binding, alpha-beta plait family	1	137
RLX	*unk*	Pectate lyase family	1	208
RLX	*unk*	Pentatricopeptide repeat family	1	178
RLX	*unk*	Peptidase aspartic, catalytic family	19	9271
RLX	*unk*	Peptidase S10, serine carboxypeptidase family	4	384
RLX	*unk*	Peptidase S14, ClpP family	1	136
RLX	*unk*	Peptidase S9A, prolyl oligopeptidase family	1	246
RLX	*unk*	Pheophorbide a oxygenase family	1	410
RLX	*unk*	Phytochrome A/B/C/D/E family	1	308
RLX	*unk*	Prephenate dehydratase family	1	208
RLX	*unk*	Protein of unknown function DUF266, plant family	2	258
RLX	*unk*	Protein of unknown function DUF569 family	1	117
RLX	*unk*	Protein of unknown function DUF6, transmembrane family	1	132
RLX	*unk*	Protein of unknown function DUF914, eukaryotic family	2	938
RLX	*unk*	Protein phosphatase 2C-related family	4	363
RLX	*unk*	Protein-L-isoaspartate(D-aspartate) O-methyltransferase family	1	130
RLX	*unk*	Putative Clp, N-terminal family	1	124
RLX	*unk*	Putative Glycine rich family	1	223
RLX	*unk*	Ribosomal protein L14b/L23e family	1	55
RLX	*unk*	Ribosomal protein L4/L1e, archeabacterial like family	2	108
RLX	*unk*	Ribosomal protein S4E family	3	577
RLX	*unk*	Rice gene family candidate, 0006076	1	150
RLX	*unk*	Rice gene family candidate, 0008584	1	354
RLX	*unk*	Rice gene family candidate, 0009691	1	95
RLX	*unk*	Rice gene family candidate, 0010797	1	108
RLX	*unk*	Rice gene family candidate, 0011119	1	65

SF	Family[1]	Fragment(s) aquired	Copy number in B73	bp
RLX	*unk*	RNA polymerase Rpb2, domain 7 family	1	201
RLX	*unk*	SANT, DNA-binding family	1	66
RLX	*unk*	Thiamine pyrophosphate enzyme, N-terminal TPP binding region family	2	362

Table 5.S3. The resampled correlation coefficients describing the relationship between percent LTR retrotransposon abundance and family richness for each chromosome.

Chromosome	r	SE	CI %
1	-0.72	0.10	(-0.86, -0.47)
2	-0.61	0.13	(-0.81, -0.30)
3	-0.74	0.08	(-0.87, -0.56)
4	-0.85	0.05	(-0.94, -0.73)
5	-0.57	0.16	(-0.79, -0.22)
6	-0.58	0.16	(-0.86, -0.25)
7	-0.73	0.12	(-0.92, -0.44)
8	-0.73	0.13	(-0.92, -0.44)
9	-0.71	0.13	(-0.92, -0.45)
10	-0.58	0.26	(-0.86, 0.12)

References

1. McClintock B (1948) Mutable loci in maize. Year Book: Carnegie Institute of Washington pp. 155-169.
2. Berg DE, Howe MM (1989) Mobile DNA. Washington, DC: American Society for Microbiology.
3. Kapitonov VV, Jurka J (2001) Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci USA 98: 8714-8719.
4. Mendiola MV, Delacruz F (1992) *IS91* transposase is related to the rolling-circle-type replication proteins of the pub110 family of plasmids. Nucleic Acids Res 20: 3521-3521.
5. Bennetzen JL (2005) Transposable elements, gene creation and genome rearrangement in flowering plants. Curr Opin Genet Dev 15: 621-627.
6. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, et al. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449: 463-U465.
7. Leeton PRJ, Smyth DR (1993) An abundant LINE-like element amplified in the genome of *Lilium speciosum*. Mol Gen Genet 237: 97-104.
8. Bennetzen JL, Ramakrishna W (2002) Exceptional haplotype variation in maize. Proc Natl Acad Sci USA 99: 9093-9095.
9. Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A (2005) Evolution of DNA sequence nonhomologies among maize inbreds. Plant Cell 17: 343-360.
10. Fu HH, Dooner HK (2002) Intraspecific violation of genetic colinearity and its implications in maize. Proc Natl Acad Sci USA 99: 9573-9578.
11. Ma JX, Bennetzen JL (2004) Rapid recent growth and divergence of rice nuclear genomes. Proc Natl Acad Sci USA 101: 12404-12410.
12. Bureau TE, White SE, Wessler SR (1994) Transduction of a cellular gene by a plant retroelement.. Cell 77: 479-480.
13. Jin YK, Bennetzen JL (1994) Integration and nonrandom mutation of a plasma-membrane proton atpase gene fragment within the *Bs1* retroelement of maize. Plant Cell 6: 1177-1186.
14. Wang W, Zheng HK, Fan CZ, Li J, Shi JJ, et al. (2006) High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell 18: 1791-1802.
15. Gilbert W (1978) Why genes in pieces? Nature 271: 501-501.
16. Wessler SR, Bureau TE, White SE (1995) LTR-retrotransposons and MITEs - important players in the evolution of plant genomes. Curr Opin Genet Dev 5: 814-821.
17. Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. Genome Res 16: 1252-1261.
18. Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, et al. (2006) Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. Genome Res 16: 1262-1269.
19. SanMiguel P, Bennetzen JL (1998) Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Ann Bot-London 82: 37-44.
20. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, et al. (2009) The B73 maize genome: complexity, diversity and dynamics. Science, in review.

21. SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, et al. (1996) Nested retrotransposons in the intergenic regions of the maize genome. Science 274: 765-768.
22. Kronmiller BA, Wise RP (2008) TEnest: Automated chronological annotation and visualization of nested plant transposable elements. Plant Physiol 146: 45-59.
23. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. Nat Genet 20: 43-45.
24. Edwards KJ, Veuskens J, Rawles H, Daly A, Bennetzen JL (1996) Characterization of four dispersed repetitive DNA sequences from *Zea mays* and their use in constructing contiguous DNA fragments using YAC clones. Genome 39: 811-817.
25. Fengler K, Allen SM, Li BL, Rafalski A (2007) Distribution of genes, recombination, and repetitive elements in the maize genome. Crop Sci 47: S83-S95.
26. Meyers BC, Tingley SV, Morgante M (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res 11: 1660-1676.
27. Zonneveld BJM, Leitch IJ, Bennet MD (2005) First nuclear DNA amounts in more than 300 angiosperms. Ann Bot-London 96: 229-244.
28. McCarthy EM, McDonald JF (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics 19: 362-367.
29. Smit AFA, Hubley R, Green P (1996-2004) ReepatMasker Open-3.0. Available online: <http://www.repeatmasker.org>.
30. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, et al. (2007) A unified classification system for eukaryotic transposable elements should reflect their phylogeny. Nat Rev Genet 8: 973-982.
31. SanMiguel PJ, Ramakrishna W, Bennetzen JL, Busso CS, Dubcovsky J (2002) Transposable elements, genes and recombination in a 215-kb contig from wheat chromosome $5A^m$. Funct Integr Genomic 2: 70-80.
32. Kramerov DA, Vassetzky NS (2005) Short retroposons in eukaryotic genomes. Int Rev Cytol Vol 247: 165-221.
33. Yasui Y, Nasuda S, Matsuoka Y, Kawahara T (2001) The Au family, a novel short interspersed element (SINE) from *Aegilops umbellulata*. Theor Appl Genet 102: 463-470.
34. Umeda M, Ohtsubo H, Ohtsubo E (1991) Diversification of the rice *waxy* gene by insertion of mobile DNA elements into introns. Jpn J Genet 66: 569-586.
35. Yoshioka Y, Matsumoto S, Kojima S, Ohshima K, Okada N, et al. (1993) Molecular characterization of a short interspersed repetitive element from tobacco that exhibits sequence homology to specific transfer-RNAs. Proc Natl Acad Sci USA 90: 6562-6566.
36. Deragon JM, Zhang XY (2006) Short interspersed elements (SINEs) in plants: Origin, classification, and use as phylogenetic markers. Syst Biol 55: 949-956.
37. Deininger PL, Batzer MA (2002) Mammalian Retroelements. Genome Res 12: 1455-1465.
38. Devos KM, Brown JKM, Bennetzen JL (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome Resh 12: 1075-1079.
39. Ma JX, Devos KM, Bennetzen JL (2004) Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res 14: 860-869.
40. Vitte C, Bennetzen JL (2006) Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proc Natl Acad Sci USA 103: 17638-17643.

41. Baucom RS, Estill JC, Leebens-Mack J, Bennetzen JL (2009) Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome. Genome Res 19: 243-254.
42. SanMiguel PJ, Vitte C (2009) The LTR-Retrotransposons of Maize. In: Bennetzen JL, Hake S, editors. Handbook of Maize: Genetics and Genomics. New York: Springer. pp. 307-328.
43. Tikhonov AP, SanMiguel PJ, Nakajima Y, Gorenstein NM, Bennetzen JL, et al. (1999) Colinearity and its exceptions in orthologous *adh* regions of maize and sorghum. Proc Natl Acad Sci USA 96: 7409-7414.
44. Bennetzen JL (2000) Transposable element contributions to plant gene and genome evolution. Plant Mol Biol 42: 251-269.
45. Ma JX, Bennetzen JL (2006) Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. Proc Natl Acad Sci USA 103: 383-388.
46. Liu RY, Vitte C, Ma JX, Mahama AA, Dhliwayo T, et al. (2007) A GeneTrek analysis of the maize genome. Proc Natl Acad Sci USA 104: 11844-11849.
47. Johns MA, Mottinger J, Freeling M (1985) A low copy number, *copia*-like transposon in maize. EMBO J 4: 1093-1101.
48. Varagona MJ, Purugganan M, Wessler SR (1992) Alternative splicing induced by insertion of retrotransposons into the maize *waxy* gene. Plant Cell 4: 811-820.
49. Yamazaki M, Tsugawa H, Miyao A, Yano M, Wu J, et al. (2001) The rice retrotransposon *Tos17* prefers low-copy-number sequences as integration targets. Mol Genet Genomics 265: 336-344.
50. Gai XW, Voytas DF (1998) A single amino acid change in the yeast retrotransposon Ty5 abolishes targeting to silent chromatin. Mol Cell 1: 1051-1055.
51. Bennetzen JL (2000) The many hues of plant heterochromatin. Genome Biol 1: 107.101-107.104.
52. Grandbastien MA, Spielmann A, Caboche M (1989) TNT1, a mobile retroviral-like transposable element of tobacco isolated by plant-cell genetics. Nature 337: 376-380.
53. Estill JC, Bennetzen JL (2009) RepMiner: A graph theory approach to the assembly and classification and assembly of the repetitive fraction of genomic sequence data. Available online: http://repminer.sourceforge.net/.
54. Estill JC, Bennetzen JL (2009) The DAWGPAWS pipeline for the annotation of genes and transposable elements in plant genomes. Plant Methods 5.
55. Howell DC (2000) Resampling Procedures. Available online: http://www.uvm.edu/~dhowell/StatPages/Resampling/Resampling.html.
56. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL-W - Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-4680.
57. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586-1591.
58. Chou H, Holmes M (2001) DNA sequence quality trimming and vector removal. Bioinformatics 17: 1093-1104.
59. Galtier N, Gouy M, Gautier C (1996) SEAVIEW and PHYLO_WIN: Two graphic tools for sequence alignment and molecular phylogeny. Comput Appl Biosci 12: 543-548.
60. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797.

61. Felsenstein J (1989) PHYLIP-phylogeny inference package. Cladistics 5: 164-166.
62. Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18: 502-504.
63. Hedges SB (1992) The number of replications needed for accurate estimation of the bootstrap-p value in phylogenetic studies. Mol Biol Evol 9: 366-369.

CHAPTER 6

CONCLUSIONS

Tools for Genome Cartology

A major outcome of my dissertation research is a suite of software tools that facilitate the annotation of genome sequences, and the classification of the genes and transposable elements (TEs) discovered in the annotation process. Software has also been developed for the study of patterns in the distribution of these sequence features when they are mapped to genomic contigs. The application of these tools to a taxonomy and genome cartology of LTR retrotransposons in maize has illustrated the utility of these programs. Future development of this software suite will seek to improve interoperability among the programs, and increase the scalability of these tools so they can more easily be used across multiple genomes.

The DAWGPAWS programs described in chapter 2 are currently being used to annotate genomic contigs in multiple plant genomes, including wheat, Amborella, cotton (C. Grover, pers. comm.) and sunflower (E. Staton, pers. comm.). The recent addition of support for BLAT alignments [1] and SNAP gene annotation [2] should facilitate using DAWGPAWS in novel genomes for which trained gene annotation models have not been developed. All of the DAWGPAWS programs have also all been extended to produce GFF3 output (http://www.sequenceontology.org/gff3.shtml) in addition to the GFF2 output described in chapter 2. The GFF3 output format captures the hierarchical structure of features and sub-features existing in genome annotations. For example, exons can now be designated as subunits of gene features, and long terminal repeats can be recognized as parts of LTR retrotransposons.

I've also submitted additional terms and relationships to the formal sequence ontology used for genome annotation (http://www.sequenceontology.org/) to better support representation of TEs and mathematically defined repeats. These additions will further facilitate interaction of DAWGPAWS with existing tools in the generic model organism (GMOD) suite of genome annotation and visualization programs (http://gmod.org), and this expands the potential user base for DAWPAWS. Many of these contributions also generally provide better support for TEs in all GMOD programs and other in sequence ontology compliant genomic databases. I will continue to increase my contributions to TE annotation components of the GMOD suite of software.

The RepMiner visualization techniques described in chapter 3 have already proved useful in a map of the relationships among the LTR retrotransposons in rice [3]. Initial visualizations of LTR retrotransposon annotations from the grape and poplar genomes indicate that the RepMiner tool could be quite useful for visualizations comparing diversity across genomes (Figure 6.1). When combined with the affinity propagation-based techniques for exemplar discovery described in chapter 4, this suite of tools can directly display the relationships of LTR retrotransposons among multiple genomes by selecting representative extant sequences for comparisons among genomes (Figure 6.2). These sequences can be further analyzed to produce representative phylogenies for the study of LTR retrotransposon evolution across genomes, and will yield insight into the rates and processes of horizontal transfer of TEs.

The application of the techniques described here for classification and visualization of sequences is not limited to LTR retrotransposons. These techniques have already been applied to visualizations of MITEs and the autonomous DNA transposons that putatively promote their mobilization in rice (S. Wessler and G. Yang, pers. comm.). Other classes of TEs could be

visualized and classified using these techniques. Furthermore, any general class of DNA sequence features could benefit from RepMiner-based classification and visualizations.

The Utility of Exemplar Approaches in Bioinformatics

This dissertation represents the first time that the affinity propagation approach has been applied to DNA sequence data, and it has been shown to be a powerful algorithm for building representative sequence databases for repeat masking. This clustering and exemplar discovery method could also provide an important framework for breaking larger problems in informatics into smaller and more tractable subsets. For example, large datasets of sequences requiring a multiple alignment could be distilled into smaller exemplar sets that could be aligned in a much faster time than attempts at alignment of the entire set. In this dissertation, the problem of curating a massive database of TE sequences was reduced to the more manageable task of curating a smaller representative subset of those sequences. Similarly, the entire suite of computationally predicted genes from a newly sequenced genome could be distilled to a representative subset of genes, and knowledgeable biologists could curate these gene models. These curated gene models could then be used to more precisely train automated annotation programs, and could thus increase the accuracy of all computationally-generated gene models. Such divide and conquer techniques for genome annotation will become increasingly necessary as the rate of genome sequencing and assembly surpasses the ability of human curation to keep pace with the data generation.

A major advantage of the affinity propagation approach over other methods of clustering data is that the number of sequences required to represent the larger database is an outcome of the approach and not a required input. This makes this technique particularly useful to novel

genomes in which the number of TE families is not known beforehand. Furthermore, the use of a vector representing the availability of each individual sequence in the dataset provides for finer control of clustering and exemplar discovery than any existing unsupervised method for clustering and classification. For the purposes of finding sequences for repeat masking, I have used a single value for all sequences in the dataset, but finer control could easily allow for expansion of an exemplar sequence database that keeps all existing exemplars in the database, while providing new exemplar sequences when necessary to increase the scope of the database.

Future Developments in Genome Cartology

The results for maize LTR retrotransposons indicates that taking an ecologically informed approach to studying TE distribution is a powerful tool for describing extant patterns of sequence features mapped in the genomic landscape. It is likely that this approach could inform genome evolution studies for other types of sequence features mapped in genomes. I will continue the development of the genome cartology tools used for this dissertation into a more comprehensive suite of statistical and visualization approaches for summarizing genomic diversity. This development will seek to further integrate ecological metrics of diversity and apply statistical tools from spatial econometrics to the study of genome evolution.

LTR Retrotransposons in Maize

The annotation of the genome sequence of maize inbred B73 revealed 31,115 full-length LTR retrotransposons distributed among 406 distinct families. Over 350 of these families were newly described, and most of these new families were in very low copy number in the B73 genome. This illustrates that most of the diversity of these TEs are present in the low-copy-

number elements. This also indicates that structure-based discovery methods are important supplements to repeat-based annotation of transposable elements. For example, discovery of elements by the Recon [4] repeat-based annotation software requires the presence of more than five copies in an assembly. Over 250 families of LTR retrotransposons are present in one to two full-length copies in the B73 genome, and these families would have been missed by this copy-number criterion. In contrast, structure-based discovery methods can find any element for which there is at least a single intact copy in the assembled data set.

The sequencing of the maize genome sought to maximize gene discovery by focusing secondary sequencing efforts and assembly improvement on non-repetitive portions of the genome [5]. The result is a genome assembly that well represents the genic content of maize but has a disproportionately high number of gaps in repeat rich/gene poor regions. Full-length elements existing in regions that contain gaps could not be annotated by the structural criteria used in this dissertation when the sequence assembly is fragmented. Thus, the reported number of maize LTR retrotransposon families and the copy number of full-length elements discovered in this dissertation is an underestimate of the actual number present in B73 maize. Furthermore, as discussed in chapter 5, the centromeric and heterochromatin-rich regions are not proportionately represented in the assembly. The genomic contributions of centromeric repeat (CRM) elements are therefore underrepresented in this dataset and it is possible that there are additional families of CRM and CRM-like elements that were not discovered in this analysis. Furthermore, expanding the annotation of transposable elements to other maize inbreds is likely to discover hundreds of new families that are present in the maize population but do not exist in an intact form in the B73 inbred. As new copies of these elements are discovered with the

resequencing of additional maize genomes, the exemplar database can be expanded to incorporate these new discoveries using the strategy outlined above.

Although the assembly does set some limitations to the comprehensive discovery of all LTR retrotransposons in B73 maize, the data at hand do allow for a broad description of LTR retrotransposon biology and the discovery of spatial patterns in the distribution of these elements. Many of the descriptions for LTR retrotransposons described in this dissertation are consistent with previous analyses of the maize genome. The majority of intact LTR retrotransposons were found to have inserted within the last two million years (Figure 3.4). Relative hotspots of LTR retrotransposon accumulation exist in pericentromeric heterochromatin (Figure 5.3). The families of LTR retrotransposons are nonrandomly distributed in the maize genome (Figure 5.S2) and many families show accumulation patterns that are opposite of the general trend of accumulation for all LTR retrotransposons. These family-level trends would have been missed by lower resolution taxonomies that lumped all LTR retrotransposons together or that simply described spatial trends in *Copia* and *Gypsy* distribution.

It is also likely that the higher taxonomic resolution afforded by the TE classification approach used here will allow for more precise descriptions of the insertion site preferences and the specific genomic context targeted by individual families. These individual family-based studies could be used to shed light on current and past epigenetic states for individual genomic regions. Clustering of the similarities in the ensembles of TE communities will also further enlighten broad-scale studies in genome evolution by highlighting regions with similar genomic "ecosystems". Affinity propagation-based approaches can be used to define exemplar TE ecosystems from this broader survey, and can highlight individual genomic regions for additional research. Studies across genomes would also allow for the discovery of general trends that could

allow isolated sequence contigs to be assigned to genome "ecosystem" types based on the flora of transposable elements that they contain. These studies will likely enhance our view of chromatin beyond the binary euchromatin/heterochromatin nomenclature that is a historical contribution from chromosome staining experiments [6,7]. Transposable elements have persisted in genomes for millions of years, and have likely segregated their distributions based upon genetic and epigenetic information that we are only now beginning to discover and appreciate.

An unexpected outcome of this dissertation was the negative relationship that was discovered between diversity and coverage in the distribution of maize LTR retrotransposons. Multiple studies have undertaken the study of the distribution of individual families or groups, as described in chapter 1. This dissertation represents one of the first assessments of intragenomic spatial patterns in the diversity of the entire ensemble of a group of elements. Future work will be required to distinguish between directed insertion processes, differential removal processes and natural selection (acting on the TEs and/or the host genome) in structuring this pattern. This future work will leverage the tools developed here to assess these spatial patterns that exist for multiple classes of TEs within the maize genome as well as an assessment of these patterns of diversity in other eukaryotic genomes.

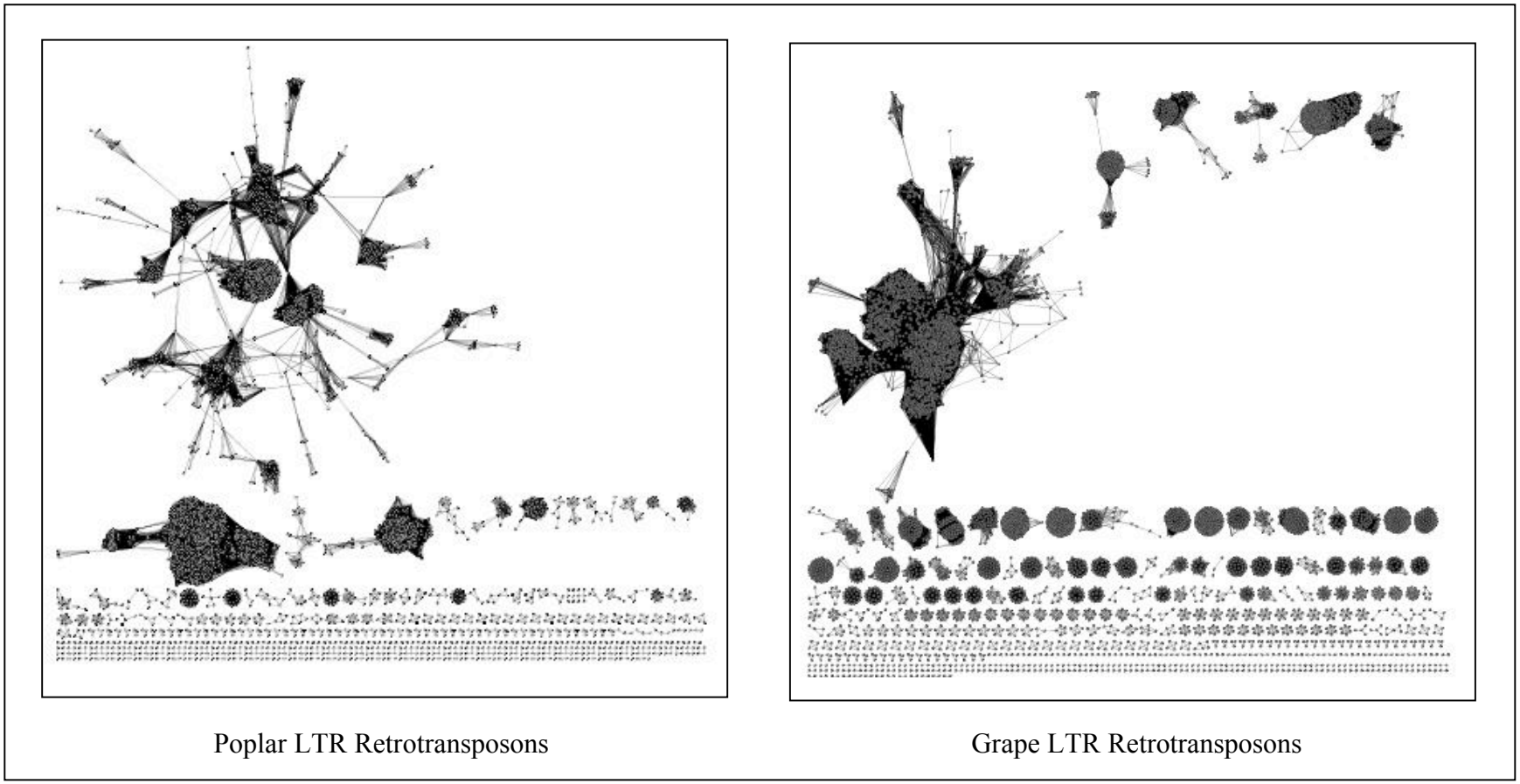


Figure 6.1 Application of the RepMiner approach to a visualization of the diversity of LTR retrotransposons within poplar and grape genomes. LTR retrotransposons were annotated using LTR_Finder [8] with default settings, employing published sequence data available for poplar [9] and grape [10]. The 5' LTRs were clustered with BLASTn ($e < 1x10^{-10}$) and visualized in Cytoscape [11] using RepMiner.

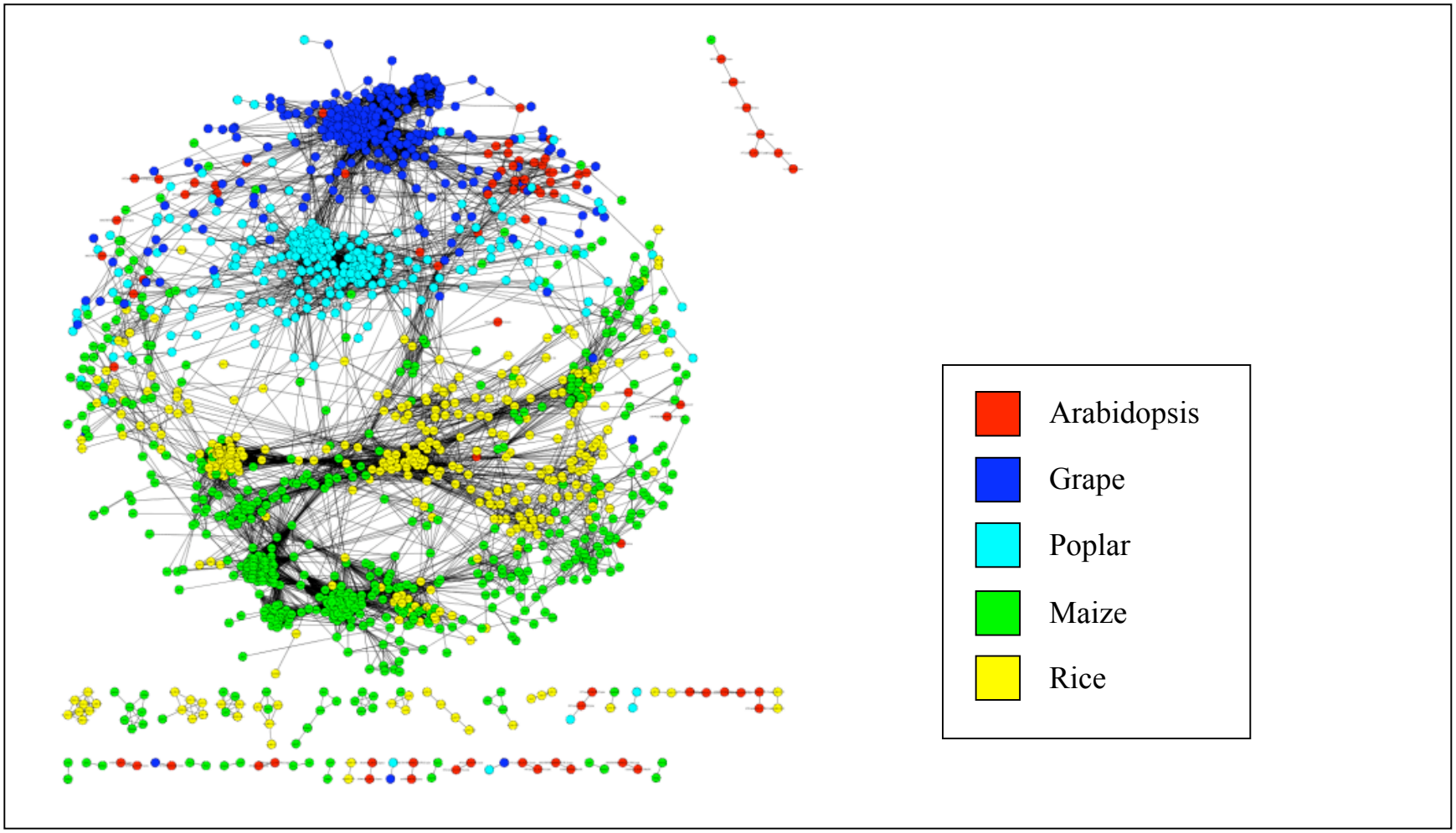


Figure 6.2. Visualization of the relationship among representative exemplar sequences of LTR retrotransposons across multiple plant genomes. LTR retrotransposons were annotated using LTR_Finder [8] using publicly available sequence data for the five host genomes. Representative sequences were identified using affinity propagation-based clustering [12]. The full length sequences of these exemplars were clustered with an "all-by-all" BLASTn search ($e < 1 \times 10^{-5}$) and were visualized in Cytoscape [11] using the RepMiner program.

References

1. Kent WJ (2002) BLAT--the BLAST-like alignment tool. Genome Res 12: 656-664.
2. Korf I (2004) Gene finding in novel genomes. BMC Bioinformatics 5: 59.
3. Baucom RS, Estill JC, Leebens-Mack J, Bennetzen JL (2009) Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome. Genome Res 19: 243-254.
4. Bao Z, Eddy SR (2002) Automated *de novo* identification of repeat sequence families in sequenced genomes. Genome Res 12: 1269-1276.
5. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, et al. (2009) The B73 maize genome: complexity, diversity and dynamics Science 326: 1112-1115.
6. Heitz E (1928) Das heterochromatin der moose. I Jahrb wiss Bot 69: 762–818.
7. Huisinga KL, Brower-Toland B, Elgin SC (2006) The contradictory definitions of heterochromatin: transcription and silencing. Chromosoma 115: 110-122.
8. Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35: W265-268.
9. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, et al. (2006) The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science 313: 1596-1604.
10. Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, et al. (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS One 2: e1326.
11. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498-2504.
12. Frey BJ, Dueck D (2007) Clustering by passing messages between data points. Science 315: 972-976.