INTRINSIC AND EXTRINSIC EVOLUTION OF *HELITRONS* IN FLOWERING PLANT GENOMES

by

LIXING YANG

(Under the Direction of Jeffrey L. Bennetzen)

ABSTRACT

Helitrons are a recently discovered class of eukaryotic transposable elements that are believed to transpose by a rolling circle mechanism. Because *Helitrons* frequently acquire and fuse fragments of multiple genes, they may be major contributors to the creation of novel genes by exon shuffling. This dissertation provides a comprehensive study of *Helitrons* in flowering plant genomes including their identification in several different genomes, their structural features, and their roles in genome evolution.

Helitrons can be difficult to identify because they have few and tiny conserved structures. We developed a new approach to effectively identify *Helitrons* (tested in *Arabidopsis* and nematodes) and further refined and demonstrated this approach on a few completely sequenced plant genomes (*Medicago*, rice and sorghum). We discovered a large number of new elements and new element families, and also a few new *Helitron* characteristics. We identified initiation and termination bypass events that led to new 5′ or 3′ ends that created newly active families and subfamilies of *Helitrons*. We found that *Helitrons* preferentially insert into AT-rich regions, that they prefer to insert near other *Helitrons*, and that the predicted hairpins near their 3′ ends would have high predicted melting temperatures.

Maize *Helitrons* are known to acquire gene fragments frequently. With the completion of the maize genome sequencing project this year, we were able to perform a large-scale search for *Helitrons* in the maize genome. We discovered 1930 intact elements in the maize genome, and were able to predict more than 20,000 total elements that account for just over 2% of the sequence assembly. We found 1194 intact *Helitrons* that contain fragments of regular nuclear genes, from 840 independent acquisition events. A total of 4% of the captured gene fragments appeared to be under negative selection and another 4% under positive selection. The results also indicated that gene fragments acquired in the same orientation as *Helitron* genes persist longer than gene fragments acquired in the antisense orientation.

Finally, we identified candidates for active elements from the rice and maize genomes by transposon display of sibling plants and of plants derived from tissue culture. The results suggest that, in most rice and maize lines, *Helitron* activity is non-existent or quite low.

INDEX WORDS: exon shuffling, gene fragment acquisition, genome evolution, insertion specificity, transposable elements, transposition

INTRINSIC AND EXTRINSIC EVOLUTION OF *HELITRONS* IN FLOWERING PLANT GENOMES

by

LIXING YANG

B.S., Biology, Fudan University, P. R. China 2004

M.S., Statistics, University of Georgia, 2009

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

2009

INTRINSIC AND EXTRINSIC EVOLUTION OF *HELITRONS* IN FLOWERING PLANT GENOMES

by

LIXING YANG

Major Professor: Jeffrey L. Bennetzen

Committee: Kelly R. Dawe
Jessica C. Kissinger
Richard B. Meagher
Susan R. Wessler

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
Dec 2009

DEDICATION

To my father Yuliang Yang and mother Yinglan He, for your only son being at the other end of the earth for over 5 years.

ACKNOWLEDGEMENTS

I would like to thank my major professor, Jeff Bennetzen, for giving me the opportunity to work with him. I couldn't have achieved so much without his guidance, support, encouragement, and patience. I have learned from him not only the topic in my research, but also how to think, how to be a great scientist and how to succeed in my career. I would like to thank my committee members, Kelly Dawe, Jessica Kissinger, Rich Meagher and Sue Wessler, for their help and valuable advice.

During my doctoral studies, many people from the genetics department and the plant biology department have helped me in many ways. I thank Jessica Kissinger and Jonathan Arnold for their support in the early stages of my studies. I thank Daniel Promislow and Jim Leebens-Mack for valuable discussions and their insights. I thank Ning Jiang from Michigan State University for discussions and encouragement. I am grateful for consistent help and support from Janice, Linda, Ana, Darlene, David and other staff members in the genetics department.

I would like to thank all current and previous members of the Bennetzen lab: Regina Baucom, Srinivasa Chaluvadi, Jeremy DeBarry, Taoran Dong, Matt Estep, Jamie Estill, Ansuya Jogi, Fang Lu, Liang Feng, Wenxiang Gao, Jennifer Hawkins, Jianxin Ma, Eugene McCarthy, Ervin Nagy, Antonio Costa deOliveira, Ryan Percifield, Ana Pontaroli, Shavannor Smith, Clementine Vitte, Qin Yao, Hao Wang and Qihui Zhu, for their generous help valuable

discussions, and for providing a great lab environment to work in. I want to thank Qihui Zhu in particular for her help and valuable discussions on my studies.

I also thank Sue Wessler for allowing me to use her lab facilities for transposon display experiments and both current and previous members from the Wessler lab and the Dawe lab for providing materials and helping me on experiments, including Jim Burnette, Tianle Chen, Eunyoung Cho, Nathan Hancock, Yujun Han, Xuexian Li, Jinhua Shi, Guojun Yang, Yaowu Yuan, Feng Zhang and Han Zhang.

Finally, I would like to thank my parents, my wife Jia Xu and all other friends for their consistent support and love.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS v

CHAPTER 1

1 INTRODUCTION AND LITERATURE REVIEW 1

Introduction 1

Transposable Elements 2

Helitron Discovery, Structure and Diversity 7

Rolling Circle Mechanism of Transposition 8

Gene Fragment Acquisitions 10

Approaches to Identify *Helitrons* 11

Outline of the Dissertation 13

Reference 15

2 STRUCTURE-BASED DISCOVERY AND DESCRIPTION OF PLANT AND ANIMAL *HELITRONS* 22

Abstract 23

Introduction 24

Results 26

Discussion 31

Methods 36

Acknowledgements 40

References 41

Figure Legends 52

3 DISTRIBUTION, DIVERSITY, EVOLUTION AND SURVIVAL OF *HELITRONS* IN THE MAIZE GENOME 62

Abstract 63

Introduction 65

Results and Discussion 68

Materials and Methods 84

Acknowledgement 88

Reference 89

Figure Legends 94

4 CONCLUSIONS 103

References 111

APPENDIX *HELITRON* ACTIVITY TESTS 113

Introduction 113

Materials and Methods 115

Results and Discussion 117

Acknowledgement 121

References 122

Figure Legends 129

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Introduction

With the recent advances in DNA sequencing technology, sequence data have been accumulating at an unprecedented rate. The next challenge is to understand the meaning of the sequences and to compare two or more genomes to address interesting evolutionary questions, such as how genes evolve or how genomes evolve.

Plant genome sizes have huge variation, a phenomenon known as the "C value paradox" (Thomas 1971), wherein it is noted that genome size does not correlate with organismal complexity. For example, *Arabidopsis thaliana* has a genome size of 115MB (The *Arabidopsis* Genome Initiative 2000), *Oryza sativa* ssp. *japonica* cultivar Nipponbare ~400MB (Goff et al. 2002), and *Zea mays* ~2500MB (Schnable et al. 2009). The unsequenced flowering plant species have even larger variation. *Fragaria viridis* (strawberry) has been proposed to have the smallest genome in flowering plants (98MB) (Antonius and Ahokas 1996), while *Fritillaria assyriaca* has the largest genome yet identified (~124,000MB) (Bennett and Smith 1991). Although genome size variation is huge, the gene content (number of genes) is relatively quite conserved. For the

dicot *Arabidopsis*, the gene number is about 30,000. For the sequenced monocots rice and maize, the gene numbers are about 30,000 and 35,000, respectively. The major factor responsible for genome size variation is TE content (Bennetzen 2005; Bennetzen et al. 2005). Ancient and recent polyploidy are also important (Blanc and Wolfe 2004; Paterson et al. 2004; Swigonová et al. 2004).

Some plants have been predicted to have very large numbers of genes, which has subsequently been found to be an artifact of poor genome annotation. For example, the protein-encoding gene number for the genome of rice, *Oryza sativa* L. ssp. *japonica*, dropped from 61,668 in the initial annotation (Goff et al. 2002) to 30,192 in the most recent published annotation (Tanaka et al. 2008). The mis-annotation comes from two major sources. Firstly, transposable element (TE) genes and gene fragments that are frequently acquired by certain classes of TEs are often annotated as regular nuclear genes. Secondly, truncated genes are commonly annotated as genes. In the recently released plant genome of sorghum (Paterson et al. 2009), it was found that the gene content was over-estimated because many truncated pseudogenes were annotated as real genes (Bennetzen, Wang, Yang and Zhu, unpublished results).

Transposable Elements

TEs are segments of DNA that can duplicate and insert themselves into new chromosomal locations. They were first identified by Barbara McClintock (McClintock 1951a). She was awarded a Nobel Prize in 1983 for the discovery and characterization of TEs.

TEs have been found in almost all well-studied eukaryotes and are a major component of most plant genomes (Berg and Howe 1989; Finnegan 1989; Feschotte et al. 2002). It has been estimated that 20% of the *Arabidopsis* genome, 40% of the rice genome, and 85% of the maize genome are composed of TEs (Initiative 2000; Goff et al. 2002; Liu and Bennetzen 2008; Schnable et al. 2009). Transposable elements are classified into two major classes: Class I RNA elements and Class II DNA elements, based on their transposition mechanisms. RNA elements transpose by a "copy and paste" mechanism while most DNA elements (except *Helitrons*) are believed to transpose by a "cut and paste" mechanism.

Class I elements transpose through an element-encoded RNA intermediate. They can be further divided into two groups based on their transposition mechanism and structure: LTR retrotransposons and non-LTR retrotransposons. LTR retrotransposons have long terminal repeats (LTRs) and contain *gag* (encoding a capside-like protein) and *pol* (encoding protease, reverse transcriptase, RNase H and integrase) genes (Jin and Bennetzen 1989; SanMiguel et al. 1998). Non-LTR retrotransposons are either long interspersed nuclear elements (LINEs) or short interspersed nuclear elements (SINEs). LINEs encode *ORF1* (*gag*-like protein), *EN* (endonuclease) and *RT* (reverse transcriptase). SINEs do not encode any protein, but their 3′ ends are homologous to LINEs, and they apparently use LINE-encoded functions for their transposition (Ogiwara et al. 1999).

Most DNA transposons are characterized by terminal inverted repeats (TIRs) and cause target site duplications (TSDs) upon insertion. They usually encode transposase, which interacts with the ends of the element to specifically initiate excision (cut-and-paste reaction).

Autonomous elements have open reading frames (ORFs) that encode all of the protein products that are required for their transposition, while non-autonomous elements are usually derived from internal deletion(s) of autonomous elements and depend on the proteins encoded by autonomous elements to transpose.

TEs have been considered to be largely "junk DNA" (Ohno 1972) or "selfish DNA" for decades (Doolittle and Sapienza 1980; Orgel and Crick 1980). Why do TEs predominate in most flowering plant genomes? What kind of effects do they have on gene evolution, genome structure and genome evolution? McClintock proposed that TEs were part of a global stress response that could potentially restructure genomes (McClintock 1984). More and more TEs with acquired host functions have been discovered. This is commonly referred to as TE domestication. Since DNA TEs excise during transposition, such a process can cause chromosomal breakage, because TEs generate single strand or double strand breaks. TEs are usually repetitive. So they can pair ectopically in otherwise non-homologous regions within or between chromosomes. This can induce ectopic chromosome recombination, leading to such rearrangements as inversions, deletions, duplications and translocations (Lonnig and Saedler 2002). Although most such rearrangements will be deleterious, especially when heterozygous (for inversions and translocations), such rearrangements may sometimes increase host fitness (Caceres et al. 1999; Dunham et al. 2002; Babcock et al. 2003).

TEs can serve as mutagens when they insert into a new location. TEs inserted in promoters can influence transcription or transcriptional regulation (McClintock 1967; van de Lagemaat et al. 2003; Han et al. 2004). TEs inserted in introns can influence RNA splicing (Lev-Maor et al. 2003)

or even serve as an intron (Wessler et al. 1987) or an exon (Bejerano et al. 2006). They can also provide poly-A signals (Harendza and Johnson 1990; Krane and Hardison 1990) or can be converted into the functional equivalent of telomeres (Pardue and DeBaryshe 2003). The presence of TEs can change the chromatin status. Inactive TEs are often methylated (Fedoroff and Chandler 1994) and form heterochromatin (Lippman et al. 2004). Changes of TE transcriptional activity may change the activity of adjacent gene (Kashkush et al. 2003).

Some TEs in plants have been found to have a high capacity for shuffling exons. *Bs1* was the first reported retroelement that contained sequences similar to a portion of a normal host gene, a plasma membrane proton ATPase (Bureau et al. 1994; Jin and Bennetzen 1994; Palmgren 1994). *L1* elements were found to be able to acquire additional sequences other than the elements themselves, thus shuffling exons in the human genome (Moran et al. 1999). Novel sequence acquisition is a relatively common phenomenon in DNA elements, especially in the *Mutator* system. A *Mu*-Related Sequence A (MRS-A) was shown to have a portion of a *Mu*-unrelated gene with internal homology to *Mu1* and *Mu2* inverted termini (Talbert and Chandler 1988). In addition, Pack-MULEs have been very well studied in the rice genome. Over 3000 were reported to contain fragments derived from more than 1000 cellular genes, and at least 5% appear to be expressed (Jiang et al. 2004). Some of the gene fragments captured by Pack-MULEs were found to be under negative selection (Hanada et al. 2009).

Hence, because of their effects on genome size, their ability to mobilize genes and gene fragments, and their other contributions to genome rearrangement and possibly gene creation, it

is crucial to study TEs in order to better understand the structure and evolutionary history of genomes.

Active TEs are particularly useful for the study of transposition mechanisms or insertion preference, and can be used for mutagenesis by transposon tagging. There are a number of TEs that have been found in an active state, including the DNA element *Ac*/*Ds*, *En*/*Spm*, *Mutator*, *Cy*, *Dt*, *Uq* and *Mrh* in maize (McClintock 1946; McClintock 1951b; Peterson 1953; Peterson 1965; Robertson 1978; Peterson 1986; Peterson and Salamini 1986; Schnable and Peterson 1986; Cormack et al. 1988). A few LTR retrotransposons (Hirochika 1993; Lucas et al. 1995; Hirochika et al. 1996a; Hirochika et al. 1996b; Hirochika et al. 2000) were found to be active in *Arabodopsis*, rice or tobacco tissue culture cells and in the Arabidopsis *ddm1* mutant. The DNA transposon called the *P* element in *Drosophila*, the DNA transposon *Tc1* in *C. elegans* and the retroelement *L1* in human were found to be active (Rubin et al. 1982; Emmons and Yesner 1984; Dombroski et al. 1991). A reconstructed DNA transposon, *Sleeping Beauty* from fish, was found to be active when transformed into human cells (Ivics et al. 1997). Both DNA transposons and retrotransposons can be reactivated in pollen (Slotkin et al. 2009) by DNA methylation changes. In addition, a MITE called *mPing* was found to be active in rice tissue culture (Jiang et al. 2003) and was also active in transgenic yeast (Yang et al. 2006) and transgenic *Arabidopsis* (Yang et al. 2007). Once active versions were found, detailed analysis of its transposition processes were addressed (Yang et al. 2009).

Helitron Discovery, Structure and Diversity

Helitrons are a new class of elements discovered by computational analysis of repeats in the *Arabidopsis* genome (Kapitonov and Jurka 2001). Four elements encoding Rep/helicase and RPA (replication protein A)-like proteins were identified. Similar elements were also found in the *C. elegans* and rice genomes (Kapitonov and Jurka 2001). Deletion derivatives (nonautonomous elements) were also identified in those genomes. *Helitrons* were found to constitute about 2% each of the *Arabidopsis*, *C. elegans* and rice genomes. Several *Arabidopsis* insertion elements that had previously been detected prior to the discovery and description of *Helitrons* were then identified as *Helitrons*, including Aie (Doutriaux et al. 1998), AthE1 (Surzycki 1999), *Basho* (Le et al. 2000) and ATREP (Kapitonov and Jurka 1999). A previously identified *Drosophila* repeat, *DINE-1,* was found to be a *Helitron* as well, although having a slightly different structure (Yang and Barbash 2008). Since their original discovery in *Arabidopsis*, *Helitrons* have been found in rice, maize, barley, wheat, alfalfa, morning glory, worm, fungi, moss, fruit fly, fish, urchin and bat (Lal et al. 2003; Poulter et al. 2003; Arkhipova and Meselson 2005; Brunner et al. 2005b; Gupta et al. 2005; Lai et al. 2005; Morgante et al. 2005; Wang and Dooner 2006; Xu and Messing 2006; Zhou et al. 2006; Choi et al. 2007; Pritham and Feschotte 2007; Jameson et al. 2008; Sweredoski et al. 2008; Yang and Barbash 2008).

Helitrons are characterized by a 5′ TC terminus and a 3′ CTRR terminus that includes a predicted small hairpin structure near the 3′ CTRR end (Kapitonov and Jurka 2001). They are found to be preferentially inserted into the host dinucleotide AT. Some elements encode

Rep/helicase and RPA-like proteins that may be involved in the transposition process. The elements that encode Rep/helicase are considered putative autonomous elements. *Helitrons* in fish often encode an apurinic-apyrimidinic-like endonuclease (Poulter et al. 2003; Zhou et al. 2006).

Two maize *Helitrons* were identified as mutational insertions into genes (Lal et al. 2003; Gupta et al. 2005). Another maize *Helitron* was found to excise from its insertion site (Li and Dooner 2009), an unexpected result given the presumed transpositional mechanism for *Helitrons* (see below). These results suggest that *Helitrons* in maize have been active recently.

Rolling Circle Mechanism of Transposition

Certain plasmids (pT181, pC194, pLS1, pE194, pWV01, pBAA1, pUB110), bacterial transposons (IS*91*, IS*801*, IS*1294*), and viruses (e.g., geminivirus, nanovirus and circovirus) are known to replicate through a rolling circle mechanism (Stenger et al. 1991; Mankertz et al. 1997; Mahillon and Chandler 1998; Richter et al. 1998; Timchenko et al. 1999; Gutierrez 2000; Khan 2000; Tavakoli et al. 2000; Mankertz and Hillenbrand 2001; Khan 2005). Geminivirus-related DNA (GRD) in *Nicotiana tabacum* (tobacco) has similar terminal sequences to *Helitrons* (i.e. CTAG at one end) (Ashby et al. 1997; Murad et al. 2004). It is proposed that geminiviruses evolved from plant rolling circle transposons, perhaps *Helitrons*. *Helitrons* are believed to share a common ancestor with prokaryotic rolling circle transposons (Kapitonov and Jurka 2001).

Helitrons do not encode any known DNA transposase. However, the Rep-helicase in *Helitrons* contains a motif that is conserved in rolling circle plasmids and DNA viruses

(Mendiola and de la Cruz 1992; Koonin and Ilyina 1993; Kapitonov and Jurka 2001). The rolling circle mechanism proposes that the 5′ TC is the transposition initiation site and 3′ CTRR is the transposition termination site (Kapitonov and Jurka 2001). The rolling circle replication starts from a site-specific nicking by a Rep-protein (Mendiola et al. 1994; Castellano et al. 1999), then one strand is elongated by a complex of host replication proteins such as DNA helicase, DNA polymerase and single-strand DNA binding protein. Subsequently, Rep catalyzes a strand transfer reaction (transposon insertion) to insert single stranded DNA and the insertion is repaired to a double stranded state. Such a rolling circle model explains why there would be no target site duplications created by transposition.

Rep is the transposase equivalent in the rolling circle model. It is a sequence-specific DNA binding protein. It also has nicking, ligation, and topoisomerase activity (Pant et al. 2001; Campos-Olivas et al. 2002). RPA is a sequence-nonspecific single strand DNA binding protein and can interact with Rep (Singh et al. 2006). It binds and stabilizes single-stranded DNA during DNA replication, transcription, recombination and DNA repair processes (Binz et al. 2004). It has been also proposed that the 3′ end of *Helitrons* may be the transposition initiation site because certain geminiviruses initiate their replication at the end that has a predicted DNA hairpin (Castellano et al. 1999), as observed at the 3′ ends of all *Helitrons*.

The rolling circle mechanism predicts that *Helitrons* transpose through a “copy and paste” manner. They would not excise from the original location. However, a recent study (Li and Dooner 2009) discovered a few *Helitron* excision events in maize somatic cells with TA repeats left as footprints. Hence, the exact mechanism for *Helitron* transposition is still unknown.

Gene Fragment Acquisitions

Helitrons have been found to often capture fragments from multiple genes of different chromosomal origin. In the first such observation, a maize *Helitron* inserted into *sh2-7527* was found to have captured a sorting nexin gene (Lal et al. 2003). Since then, a number of maize and other plant *Helitrons* were found to have acquired one or more gene fragments. A *Helitron* inserted into the *ba1-ref* allele was found to have captured 4 gene fragments (Gupta et al. 2005). A family of non-autonomous *Helitron* elements in maize were found to have acquired as many as 7 different gene fragments from different chromosomal locations (Brunner et al. 2005b) and *Helitrons* from the maize *bz* region were found to have captured gene fragments (Lai et al. 2005; Wang and Dooner 2006). Another maize *Helitron* was found to have acquired part of a cytidine deaminase gene (Xu and Messing 2006). Four exons of a cytochrome P450 monooxygenase gene containing a putative ORF was found to have been captured by a maize *Helitron* and the putative ORF is transcribed (Jameson et al. 2008). Transcripts of *Helitron*-acquired gene fragments were found to be processed on some occasions into a single chimeric transcript (Lal et al. 2003; Morgante et al. 2005), and sometimes alternatively spliced (Brunner et al. 2005b). In the exon shuffling model (Gilbert 1987), the intron's role was to serve as a permissively inprecise boundary for exons, thereby increasing the rate at which exons can reassort as independent elements and make complex proteins out of dissimilar domains encoded by genes on different chromosomes. *Helitrons* provide one such way to create proteins by capturing and fusing exons from different genes on different chromosomes. *Helitrons* can not only capture gene fragments,

but also promoters. A *xanA* gene promoter was found to be captured by an *Aspergillus nidulans Helitron*, which could lead to the recruitment of new genes into specific regulatory circuits (Cultrone et al. 2007). Another interesting aspect of *Helitron* gene acquisition is that the gene fragments acquired are usually in the same orientation (Kapitonov and Jurka 2007). Morgante proposed a "transduplication" model (Morgante 2006) that non-autonomous *Helitrons* can carry different gene fragments and insert into a new location. However, the exact molecular basis and mechanism of gene acquisition remain unknown.

Since maize *Helitrons* have been found to frequently acquire more than one gene fragment, it is of great interest to study maize *Helitrons* and the evolutionary role of the gene fragments captured by those elements. It was estimated that 4000 gene fragments in the maize genome are *Helitron* related (Morgante et al. 2005).

Approaches to Identify *Helitrons*

The identification of *Helitrons* is difficult because of their few and tiny structural features. The approaches used up to now can be divided into six categories. One approach is to search for Rep/helicase homology (Kapitonov and Jurka 2001; Poulter et al. 2003; Arkhipova and Meselson 2005; Hood 2005; Zhou et al. 2006; Pritham and Feschotte 2007), and this will find putative autonomous *Helitrons*. The second approach is to identify *Helitrons* as *de novo* insertions (Lal et al. 2003; Gupta et al. 2005; Choi et al. 2007). Insertions that mutated the *sh2* gene and the *ba1* gene in maize both turned out to be *Helitrons*. Another spontaneous mutation in morning glory was also caused by a *Helitron* insertion (Choi et al. 2007). The third approach is to search for

similarity to known *Helitrons* (Kapitonov and Jurka 2001; Choi et al. 2007; Cultrone et al. 2007; Hollister and Gaut 2007; Pritham and Feschotte 2007; Jameson et al. 2008). After finding putative autonomous elements, this method can be used to identify non-autonomous elements. The fourth approach involves characterizing identified repeats (Kapitonov and Jurka 2001; Yang and Barbash 2008). There are a few programs that can find repetitive sequences in a given genome, such as RECON (Bao and Eddy 2002) and Spectral Repeat Finder (Sharma et al. 2004). This approach can only detect *Helitron* families that have copy numbers above a certain arbitrarily chosen threshold. The fifth approach is to search for violations of micro-colinearity (Brunner et al. 2005b; Lai et al. 2005; Morgante et al. 2005; Wang and Dooner 2006; Xu and Messing 2006). Intraspecies comparisons have uncovered a large number of non-colinearities between maize inbred lines (Fu and Dooner 2002; Ramakrishna et al. 2002; Song and Messing 2003; Brunner et al. 2005a). Most differences are due to retroelements and DNA transposons, including *Helitrons*. The sixth approach is to utilize computer programs that facilitate the identification of *Helitrons* based on their structural features, which will be discussed in detail below.

Since both ends of *Helitrons* are quite conserved, a few computational approaches have utilized this feature to identify *Helitrons* in a given genome. Consensus sequences were built for *Arabidopsis Helitrons* (Tempel et al. 2007) and maize *Helitrons* (Du et al. 2008), *Helitron* elements were predicted by searching for conserved ends of *Helitrons*. A profile HMM (Hidden Markov Model) approach was also used to identify rice *Helitrons,* by detecting strong motif patterns for conserved groups of *Helitrons* (Sweredoski et al. 2008). The above approaches were

designed to detect a specific family/subfamily of *Helitron* elements in a specific genome. Another approach used a more general way to detect *Helitron* ends by searching for the hairpin structure and 3′ CTRR sites (Zuccolo pers. comm.). Both approaches designed for identifying maize *Helitrons* (Du et al. 2008 and Zuccolo pers. comm.) predict *Helitron* elements in a similar manner after predicting ends. If two ends were found in the correct orientation and within a certain distance (i.e. 50kb and 30kb), it would be predicted as a *Helitron* element.

Outline of the Dissertation

This dissertation uses the genomic plant sequence data currently available for several plant species (1) to develop a new approach to efficiently identify *Helitrons* and to investigate the properties and evolutionary effects of identified *Helitrons*, (2) to perform a detailed study of maize *Helitrons* to investigate the properties and evolutionary effects of *Helitrons* and *Helitron*-acquired gene fragments, and (3) to screen for *Helitron* activity in rice and maize.

In Chapter 2, I describe a structure-based approach that we developed to allow one to identify intact *Helitrons* efficiently without any previous knowledge of *Helitrons* in that genome. We analyzed several plant and animal genomes, leading to the discovery of hundreds of new *Helitrons*. Analysis of these *Helitrons* has uncovered mechanisms of their evolution, including end creation and sequence acquisition. Preferential accumulation in gene-poor regions and target site specificities were also identified. This study was published as Yang, L. and Bennetzen, J.L. 2009, Structure-based discovery and description of plant and animal *Helitrons*, *Proc Natl Acad*

Sci USA **106:** 12832-12837. *PNAS* allows the first author to use his/her published paper as part of his/her dissertation without request for specific permission.

In Chapter 3, I describe the use of the previously-developed structure-based approach, combined with homology searches, to identify *Helitrons* in the genome of maize inbred B73. A total of 1930 intact *Helitrons* from eight families and >20,000 *Helitron* fragments were identified, accounting for ~2.2% of the B73 genome. We found that more than 60% of maize *Helitrons* have captured fragments of nuclear genes. We showed that *Helitrons* have a strong accumulation bias for certain chromatin structure and that some gene fragments acquired by *Helitrons* do not appear to be evolving in a neutral fashion, but instead have some beneficial or detrimental effect on the elements and/or the host. This study was published as Yang, L. and Bennetzen, J.L. 2009, Distribution, Diversity, Evolution and Survival of *Helitrons* in the Maize Genome, *Proc Natl Acad Sci USA* 106: 19922-19927.

Chapter 4 provides a discussion of all of the results of my Ph.D. studies.

In Appendix, I describe the use of the transposon display technique to test the activity of *Helitrons* in rice and maize. I tested sibling plants and plants derived from tissue cultures in rice and maize, and also recombinant inbred lines of maize. We found *Helitrons* in the lines that were tested to be quite stable, no proven activity was found.

Reference

Antonius K, Ahokas H. 1996. Flow-cytometric determination of polyploidy level in spontaneous clones of strawberries. *Hereditas* **124**: 285-299.

Arkhipova IR, Meselson M. 2005. Diverse DNA transposons in rotifers of the class Bdelloidea. *Proc Natl Acad Sci USA* **102**: 11781-11786.

Ashby MK, Warry A, Bejarano ER, Khashoggi A, Burrell M, Lichtenstein CP. 1997. Analysis of multiple copies of geminiviral DNA in the genome of four closely related *Nicotiana* species suggest a unique integration event. *Plant Mol Biol* **35**: 313-321.

Babcock M, Pavlicek A, Spiteri E, Kashork CD, Ioshikhes I, Shaffer L, Jurka J, Morrow BE. 2003. Shuffling of genes within low-copy repeats on 22q11 (LCR22) by *Alu*-mediated recombination events during evolution. *Genome Res* **13**: 2519-2532.

Bao Z, Eddy SR. 2002. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res* **12**: 1269-1276.

Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ, Haussler D. 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441**: 87-90.

Bennett MD, Smith JB. 1991. Nuclear DNA amounts in angiosperms. *Philos Trans R Soc Lond B Biol Sci* **334**: 309-345.

Bennetzen JL. 2005. Transposable elements, gene creation and genome rearrangement in flowering plants. *Curr Opin Genet Dev* **15**: 621-627.

Bennetzen JL, Ma J, Devos KM. 2005. Mechanisms of recent genome size variation in flowering plants. *Ann Bot* **95**: 127-132.

Berg DE, Howe MM. 1989. *Mobile DNA*.

Binz SK, Sheehan AM, Wold MS. 2004. Replication protein A phosphorylation and the cellular response to DNA damage. *DNA repair* **3**: 1015-1024.

Blanc G, Wolfe KH. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16**(7): 1667-1678.

Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A. 2005a. Evolution of DNA sequence nonhomologies among maize Inbreds. *Plant Cell* **17**: 343.

Brunner S, Pea G, Rafalski A. 2005b. Origins, genetic organization and transcription of a family of non-autonomous helitron elements in maize. *Plant J* **43**: 799-810.

Bureau TE, White SE, Wessler SR. 1994. Transduction of a cellular gene by a plant retroelement. *Cell* **77**: 479-480.

Caceres M, Ranz JM, Barbadilla A, Long M, Ruiz A. 1999. Generation of a widespread *Drosophila* inversion by a transposable element. *Science* **285**: 415-418.

Campos-Olivas R, Louis JM, Cleot D, Gronenborn B, Gronenborn AM. 2002. The structure of a replication initiator unites diverse aspects of nucleic acid metabolism. *Proc Natl Acad Sci USA* **99**: 10310-10315.

Castellano MM, Sanz-Burgos AP, Gutierrez C. 1999. Initiation of DNA replication in a eukaryotic rolling-circle replicon: identification of multiple DNA-protein complexes at the geminivirus origin. *J Mol Biol* **290**: 639-652.

Choi JD, Hoshino A, Park KI, Park IS, Iida S. 2007. Spontaneous mutations caused by a *Helitron* transposon, *Hel-It1*, in morning glory, *Ipomoea tricolor*. *Plant J* **49**: 924-934.

Cormack J, Cox D, Peterson P. 1988. Presence of the transposable element *Uq* in maize breeding material. *Crop Science* **28**(6): 941-944.

Cultrone A, Dominguez YR, Drevet C, Scazzocchio C, Fernandez-Martin R. 2007. The tightly regulated promoter of the *xanA* gene of *Aspergillus nidulans* is included in a helitron. *Mol Microbiol* **63**: 1577-1587.

Dombroski BA, Mathias SL, Nanthakumar E, Scott AF, Kazazian Jr HH. 1991. Isolation of an active human transposable element. *Science* **254**(5039): 1805-1808.

Doolittle WF, Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**: 601-603.

Doutriaux MP, Couteau F, Bergounioux C, White C. 1998. Isolation and characterisation of the RAD51 and DMC1 homologs from *Arabidopsis thaliana*. *Mol Gen Genet* **257**: 283-291.

Du C, Caronna J, He L, Dooner HK. 2008. Computational prediction and molecular confirmation of *Helitron* transposons in the maize genome. *BMC Genomics* **9**: 51.

Dunham MJ, Badrane H, Ferea T, Adams J, Brown PO, Rosenzweig F, Botstein D. 2002. Characteristic genome rearrangements in experimental evolution of *Saccharomyce scerevisiae*. *Proc Nat Acad Sci USA* **99**: 16144-16149.

Emmons SW, Yesner L. 1984. High-frequency excision of transposable element *Tc1* in the nematode Caenorhabditis elegans is limited to somatic cells. *Cell* **36**(3): 599-605.

Fedoroff NV, Chandler V. 1994. Inactivation of maize transposable elements. *Mobile DNA II*: 349-385.

Feschotte C, Jiang N, Wessler SR. 2002. Plant transposable elements: where genetics meets genomics. *Nat Rev Genet* **3**: 329-341.

Finnegan DJ. 1989. Eukaryotic transposable elements and genome evolution. *Trends Genet* **5**: 103-107.

Fu H, Dooner HK. 2002. Intraspecific violation of genetic colinearity and its implications in maize. *Proc Natl Acad Sci USA* **99**: 9573-9578.

Gilbert W. 1987. The exon theory of genes. *Cold Spring Harb Symp Quant Biol* **52**: 901-905.

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**: 92-100.

Gupta S, Gallavotti A, Stryker GA, Schmidt RJ, Lal SK. 2005. A novel class of *Helitron*-related transposable elements in maize contain portions of multiple pseudogenes. *Plant Mol Biol* **57**: 115-127.

Gutierrez C. 2000. DNA replication and cell cycle in plants: learning from geminiviruses. *EMBO J* **19**: 792-799.

Han JS, Szak ST, Boeke JD. 2004. Transcriptional disruption by the *L1* retrotransposon and

implications for mammalian transcriptomes. *Nature* **429**: 268-274.

Hanada K, Vallejo V, Nobuta K, Slotkin RK, Lisch D, Meyers BC, Shiu SH, Jiang N. 2009. The functional role of Pack-MULEs in rice inferred from purifying selection and expression profile. *Plant Cell* **21**: 25-38.

Harendza CJ, Johnson LF. 1990. Polyadenylylation signal of the mouse thymidylate synthase gene was created by insertion of an L1 repetitive element downstream of the open reading frame. *Proc Nat Acad Sci USA* **87**: 2531-2535.

Hirochika H. 1993. Activation of tobacco retrotransposons during tissue culture. *EMBO J* **12**: 2521-2528.

Hirochika H, Okamoto H, Kakutani T. 2000. Silencing of retrotransposons in Arabidopsis and reactivation by the *ddm1* mutation. *Plant Cell* **12**: 357-369.

Hirochika H, Otsuki H, Yoshikawa M, Otsuki Y, Sugimoto K, Takeda S. 1996a. Autonomous transposition of the tobacco retrotransposon *Tto1* in rice. *Plant Cell* **8**: 725-734.

Hirochika H, Sugimoto K, Otsuki Y, Tsugawa H, Kanda M. 1996b. Retrotransposons of rice involved in mutations induced by tissue culture. *Proc Nat Acad Sci USA* **93**: 7783-7788.

Hollister JD, Gaut BS. 2007. Population and evolutionary dynamics of *Helitron* transposable elements in *Arabidopsis thaliana*. *Mol Biol Evol* **24**: 2515-2524.

Hood ME. 2005. Repetitive DNA in the automictic fungus *Microbotryum violaceum*. *Genetica* **124**: 1-10.

The *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796-815.

Ivics Z, Hackett PB, Plasterk RH, Izsvak Z. 1997. Molecular reconstruction of *Sleeping Beauty*, a *Tc1*-like transposon from fish, and its transposition in human cells. *Cell* **91**: 501-510.

Jameson N, Georgelis N, Fouladbash E, Martens S, Hannah LC, Lal S. 2008. *Helitron* mediated amplification of cytochrome P450 monooxygenase gene in maize. *Plant Mol Biol* **67**: 295-304.

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. 2004. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431**: 569-573.

Jiang N, Bao Z, Zhang X, Hirochika H, Eddy SR, McCouch SR, Wessler SR. 2003. An active DNA transposon family in rice. *Nature* **421**: 163-167.

Jin YK, Bennetzen JL. 1989. Structure and coding properties of *Bs1*, a maize retrovirus-like transposon. *Proc Nat Acad Sci USA* **86**: 6235-6239.

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.

Kapitonov VV, Jurka J. 1999. Molecular paleontology of transposable elements from *Arabidopsis thaliana*. *Genetica* **107**: 27-37.

Kapitonov VV, Jurka J. 2001. Rolling-circle transposons in eukaryotes. *Proc Natl Acad Sci USA* **98**: 8714-8719.

Kapitonov VV, Jurka J. 2007. *Helitrons* on a roll: eukaryotic rolling-circle transposons. *Trends Genet* **23**: 521-529.

Kashkush K, Feldman M, Levy AA. 2003. Transcriptional activation of retrotransposons alters

the expression of adjacent genes in wheat. *Nature Genet* **33**: 102-106.

Khan S. 2005. Plasmid rolling-circle replication: highlights of two decades of research. *Plasmid* **53**: 126-136.

Khan SA. 2000. Plasmid rolling-circle replication: recent developments. *Mol Microbiol* **37**: 477-484.

Koonin EV, Ilyina TV. 1993. Computer-assisted dissection of rolling circle DNA replication. *Biosystems* **30**: 241-268.

Krane DE, Hardison RC. 1990. Short interspersed repeats in rabbit DNA can provide functional polyadenylation signals. *Mol Biol Evol* **7**: 1-8.

Lai J, Li Y, Messing J, Dooner HK. 2005. Gene movement by *Helitron* transposons contributes to the haplotype variability of maize. *Proc Natl Acad Sci USA* **102**: 9068-9073.

Lal SK, Giroux MJ, Brendel V, Vallejos CE, Hannah LC. 2003. The maize genome contains a *Helitron* insertion. *Plant Cell* **15**: 381-391.

Le QH, Wright S, Yu Z, Bureau T. 2000. Transposon diversity in *Arabidopsis thaliana*. *Proc Natl Acad Sci USA* **97**: 7376-7381.

Lev-Maor G, Sorek R, Shomron N, Ast G. 2003. The birth of an alternatively spliced exon: 3'splice-site selection in *Alu* exons. *Science* **300**: 1288-1291.

Li Y, Dooner HK. 2009. Excision of *Helitron* Transposons in Maize. *Genetics*.

Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, McCombie WR, Lavine K, Mittal V, May B, Kasschau KD. 2004. Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430**: 471-476.

Liu R, Bennetzen JL. 2008. Enchilada redux: how complete is your genome sequence? *New Phytol* **179**: 249-250.

Lonnig WE, Saedler H. 2002. Chromosome rearrangements and transposable elements. *Annu Rev Genet* **36**: 389-410.

Lucas H, Feuerbach F, Kunert K, Grandbastien MA, Caboche M. 1995. RNA-mediated transposition of the tobacco retrotransposon Tnt1 in *Arabidopsis thaliana*. *EMBO J* **14**: 2364-2373.

Mahillon J, Chandler M. 1998. Insertion sequences. *Microbiol Mol Biol Rev* **62**: 725-774.

Mankertz A, Hillenbrand B. 2001. Replication of porcine circovirus type 1 requires two proteins encoded by the viral rep gene. *Virology* **279**: 429-438.

Mankertz A, Persson F, Mankertz J, Blaess G, Buhk HJ. 1997. Mapping and characterization of the origin of DNA replication of porcine circovirus. *J Virol* **71**(3): 2562.

McClintock B. 1946. Maize genetics. *Carnegie Inst Washington Year Book* **45**: 176-186.

McClintock B. 1951a. Chromosome organization and genic expression. *Cold Spring Harbor Symp Quant Biol* **16**: 13-47.

McClintock B. 1951b. Mutable loci in maize. *Carnegie Inst Washington Year Book* **50**: 174-181.

McClintock B. 1967. Regulation of pattern of gene expression by controlling elements in maize. *Carnegie Inst Washington Year Book* **65**: 568-576.

McClintock B. 1984. The significance of responses of the genome to challenge. *Science* **226**(4676): 792-801.

Mendiola MV, Bernales I, de la Cruz F. 1994. Differential roles of the transposon termini in IS*91* transposition. *Proc Natl Acad Sci USA* **91**: 1922-1926.

Mendiola MV, de la Cruz F. 1992. IS91 transposase is related to the rolling-circle-type replication proteins of the pUB110 family of plasmids. *Nucleic Acids Res* **20**: 3521.

Moran JV, DeBerardinis RJ, Kazazian Jr HH. 1999. Exon shuffling by L1 retrotransposition. *Science* **283**: 1530-1534.

Morgante M. 2006. Plant genome organisation and diversity: the year of the junk! *Curr Opin Biotechnol* **17**: 168-173.

Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A. 2005. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat Genet* **37**: 997-1002.

Murad L, Bielawski JP, Matyasek R, Kovarik A, Nichols RA, Leitch AR, Lichtenstein CP. 2004. The origin and evolution of geminivirus-related DNA sequences in *Nicotiana*. *Heredity* **92**: 352-358.

Ogiwara I, Miya M, Ohshima K, Okada N. 1999. Retropositional parasitism of SINEs on LINEs: identification of SINEs and LINEs in elasmobranchs. *Mol Bio Evol* **16**: 1238-1250.

Ohno S. 1972. So much" junk" DNA in our genome. In *Brookhaven Symp Biol*, Vol 23, pp. 366-370, Gordon and Breach, New York.

Orgel LE, Crick FHC. 1980. Selfish DNA: the ultimate parasite. *Nature* **284**: 604-607.

Palmgren MG. 1994. Capturing of host DNA by a plant retroelement: *Bs1* encodes plasma membrane H+-ATPase domains. *Plant Mol Biol* **25**: 137-140.

Pant V, Gupta D, Choudhury NR, Malathi VG, Varma A, Mukherjee SK. 2001. Molecular characterization of the Rep protein of the blackgram isolate of Indian mungbean yellow mosaic virus. *J Gen Virol* **82**: 2559-2567.

Pardue ML, DeBaryshe PG. 2003. Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. *Annu Rev Genet* **37**: 485-511.

Paterson A, Bowers J, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A et al. 2009. The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**: 551-556.

Paterson A, Bowers J, Chapman B. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Nat Acad Sci USA* **101**: 9903-9908.

Peterson PA. 1953. A mutable pale green locus in maize. *Genetics* **38**(1): 682–683.

Peterson PA. 1965. A relationship between the *Spm* and *En* control systems in maize. *The American Naturalist* **99**(908): 391-398.

Peterson PA. 1986. Mobile elements in maize. *Plant Breed Rev* **4**(3): 81-122.

Peterson PA, Salamini F. 1986. A search for active mobile elements in the Iowa Stiff Stalk Synthetic maize population and some derivatives. *Maydica* **31**: 163-172.

Poulter RTM, Goodwin TJD, Butler MI. 2003. Vertebrate helentrons and other novel *Helitrons*. *Gene* **313**: 201-212.

Pritham EJ, Feschotte C. 2007. Massive amplification of rolling-circle transposons in the lineage

of the bat *Myotis lucifugus*. *Proc Natl Acad Sci USA* **104**: 1895-1900.

Ramakrishna W, Emberton J, Ogden M, SanMiguel P, Bennetzen JL. 2002. Structural analysis of the maize *Rp1* complex reveals numerous sites and unexpected mechanisms of local rearrangement. *Plant Cell* **14**: 3213-3223.

Richter GY, Bjoklof K, Romantschuk M, Mills D. 1998. Insertion specificity and trans-activation of IS801. *Mol Gen Genet* **260**: 381-387.

Robertson DS. 1978. Characterization of a mutator system in maize. *Mutation Research* **51**(1): 21-28.

Rubin GM, Kidwell MG, Bingham PM. 1982. The molecular basis of PM hybrid dysgenesis: the nature of induced mutations. *Cell* **29**(3): 987-994.

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. 1998. The paleontology of intergene retrotransposons of maize. *Nat Genet* **20**: 43-45.

Schnable PS, Peterson PA. 1986. Distribution of genetically active *Cy* transposable elements among diverse maize lines. *Maydica* **31**: 59-81.

Schnable PS Ware D Fulton RS Stein JC Wei F Pasternak S Liang C Zhang J Fulton L Graves TA et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**: 1112-1115.

Sharma D, Issac B, Raghava GPS, Ramaswamy R. 2004. Spectral Repeat Finder (SRF): identification of repetitive sequences using Fourier transformation. *Bioinformatics* **20**: 1405-1412.

Singh DK, Islam MN, Choudhury NR, Karjee S, Mukherjee SK. 2006. The 32 kDa subunit of replication protein A (RPA) participates in the DNA replication of Mung bean yellow mosaic India virus (MYMIV) by interacting with the viral Rep protein. *Nucleic Acids Res* **00**: 1-16.

Slotkin RK, Vaughn M, Borges F, Tanurdzic M, Becker JD, Feijo JA, Martienssen RA. 2009. Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell* **136**: 461-472.

Song R, Messing J. 2003. Gene expression of a gene family in maize based on noncollinear haplotypes. *Proc Natl Acad Sci USA* **100**: 9055-9060.

Stenger DC, Revington GN, Stevenson MC, Bisaro DM. 1991. Replicational release of geminivirus genomes from tandemly repeated copies: evidence for rolling-circle replication of a plant viral DNA. *Proc Natl Acad Sci USA* **88**: 8029-8033.

Surzycki SA. 1999. Characterization of repetitive DNA elements in *Arabidopsis*. *J Mol Evol* **48**: 684-691.

Sweredoski M, DeRose-Wilson L, Gaut BS. 2008. A comparative computational analysis of nonautonomous *Helitron* elements between maize and rice. *BMC Genomics* **9**: 467.

Swigonová Z, Lai J, Ma J, Ramakrishna W, Llaca V, Bennetzen JL, Messing J. 2004. Close split of sorghum and maize genome progenitors. *Genome Res* **14**: 1916-1923.

Talbert LE, Chandler VL. 1988. Characterization of a highly conserved sequence related to mutator transposable elements in maize. *Mol Biol Evol* **5**: 519-529.

Tanaka T, Antonio B, Kikuchi S, Matsumoto T, Nagamura Y, Numa H, Sakai H, Wu J, Itoh T,

Sasaki T. 2008. The rice annotation project database (RAP-DB): 2008 update. *Nucleic Acids Res* **36**: 1028-1033.

Tavakoli N, Comanducci A, Dodd HM, Lett MC, Albiger B, Bennett P. 2000. IS1294, a DNA element that transposes by RC transposition. *Plasmid* **44**: 66-84.

Tempel S, Nicolas J, El Amrani A, Couee I. 2007. Model-based identification of Helitrons results in a new classification of their families in *Arabidopsis thaliana*. *Gene* **403**: 18-28.

Thomas CA. 1971. The genetic organization of chromosomes. *Annual Rev Genetics* **5**: 237-256.

Timchenko T, de Kouchkovsky F, Katul L, David C, Vetten HJ, Gronenborn B. 1999. A single rep protein initiates replication of multiple genome components of faba bean necrotic yellows virus, a single-stranded DNA virus of plants. *J Virol* **73**: 10173-10182.

van de Lagemaat LN, Landry JR, Mager DL, Medstrand P. 2003. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet* **19**: 530-536.

Wang Q, Dooner HK. 2006. Remarkable variation in maize genome structure inferred from haplotype diversity at the *bz* locus. *Proc Natl Acad Sci USA* **103**: 17644-17649.

Wessler SR, Baran G, Varagona M. 1987. The maize transposable element *Ds* is spliced from RNA. *Science* **237**(4817): 916-918.

Xu J, Messing J. 2006. Maize haplotype with a *helitron*-amplified cytidine deaminase gene copy. *BMC Genet* **7**: 52.

Yang G, Nagel DH, Feschotte C, Hancock CN, Wessler SR. 2009. Tuned for Transposition: Molecular Determinants Underlying the Hyperactivity of a *Stowaway* MITE. *Science* **325**(5946): 1391-1394.

Yang G, Weil CF, Wessler SR. 2006. A rice *Tc1*/*mariner*-like element transposes in yeast. *Plant Cell* **18**: 2469-2478.

Yang G, Zhang F, Hancock CN, Wessler SR. 2007. Transposition of the rice miniature inverted repeat transposable element *mPing* in *Arabidopsis thaliana*. *Proc Nat Acad Sci USA* **104**: 10962-10967.

Yang HP, Barbash DA. 2008. Abundant and species-specific *DINE-1* transposable elements in 12 *Drosophila* genomes. *Genome Biol* **9**: R39.

Zhou Q, Froschauer A, Schultheis C, Schmidt C, Bienert GP, Wenning M, Dettai A, Volff JN. 2006. Helitron transposons on the sex chromosomes of the platyfish *Xiphophorus maculatus* and their evolution in animal genomes. *Zebrafish* **3**: 39-52.

CHAPTER 2

STRUCTURE-BASED DISCOVERY AND DESCRIPTION OF PLANT AND ANIMAL *HELITRONS* [1]

[1] Yang, L. and Bennetzen, J.L. 2009. Structure-based discovery and description of plant and animal *Helitrons*. *Proc Natl Acad Sci USA* **106:** 12832-12837. Reprinted here with permission of publisher.

Abstract

Helitrons are recently discovered eukaryotic transposons that are predicted to amplify by a rolling-circle mechanism. They are present in most plant and animal species investigated, but were previously overlooked partly because they lack terminal repeats and do not create target site duplications. *Helitrons* are particularly abundant in flowering plants, where they frequently acquire, and sometimes express, one or more gene fragments. A structure-based search protocol was developed to find *Helitrons*, and was used to analyze several plant and animal genomes, leading to the discovery of hundreds of new *Helitrons*. Analysis of these *Helitrons* has uncovered mechanisms of element evolution, including end creation and sequence acquisition. Preferential accumulation in gene-poor regions and target site specificities were also identified. Overall, these studies provide insights into the transposition and evolution of *Helitrons*, and their contributions to evolved gene content and genome structure.

Keywords:

gene fragment acquisition, *Helitron*, insertion specificity, transposition

Introduction

Helitrons are a new class of transposable elements (TEs) that were initially discovered by repeat-based computational analysis of the genome of the model plant *Arabidopsis thaliana* (Kapitonov and Jurka 2001). Their structural homologies to genes encoding Rep/helicase-like and RPA (replication protein A)-like proteins suggest that they transpose by a rolling circle mechanism, although this conclusion is not yet supported by experimental evidence. Because gemini viruses, some plasmids and some bacterial transposons are known to replicate by a rolling circle mechanism (Stenger et al. 1991; Mendiola et al. 1994), it is possible that *Helitrons* share a very ancient common ancestor with these other mobile sequences (Murad et al. 2004). Several *Arabidopsis* insertion elements that had been detected prior to the discovery and description of *Helitrons* were eventually found to be *Helitrons*, including Aie (Doutriaux et al. 1998), AthE1 (Surzycki 1999), *Basho* (Le et al. 2000) and ATREP (Kapitonov and Jurka 1999).

Helitrons are characterized by a 5′ TC terminus and a 3′ CTRR terminus that includes a predicted small hairpin structure near the 3′ CTRR end (Figure 2.1A). They are found to be preferentially inserted into the dinucleotide AT. Some elements encode Rep/helicase-like and RPA-like proteins that may be involved in the transposition process. The elements that encode Rep/helicase are considered putative autonomous elements.

Since their original discovery in *Arabidopsis*, *Helitrons* have been found in six additional flowering plant species, and also in moss (Rensing et al. 2008), fungi (Poulter et al. 2003; Hood 2005), the worm *Caenorhabditis elegans* (Kapitonov and Jurka 2001), sea urchin (Poulter et al.

2003; Zhou et al. 2006), fish (Poulter et al. 2003), and bats (Pritham and Feschotte 2007). They have been found to often capture gene fragments, sometimes fragments from multiple genes that normally reside in unlinked chromosomal locations (Lal et al. 2003; Gupta et al. 2005; Lai et al. 2005; Wang and Dooner 2006; Xu and Messing 2006). Although the mechanism of gene fragment acquisition has not been determined, it apparently occurs at the DNA level because both introns and exons are found within the acquired DNA. Some of these collections of *Helitron*-acquired gene fragments can be found as chimeric transcripts (Lal et al. 2003; Morgante et al. 2005). In a few detected cases, acquired introns are spliced, sometimes alternatively, and the junctions between fragments can be processed as crude *de novo* introns (Lal et al. 2003; Brunner et al. 2005). This process is quite comparable to the model proposed by Gilbert (Gilbert 1987) to explain the origin of introns. However, Gilbert's "exon shuffling", the fusion of the first short templates for single peptide domains to create the potential to synthesize complex proteins out of dissimilar subunits, was proposed as a mechanism that existed primarily in the early days of life on earth, more than a billion years ago. In maize, Morgante et al. 2005 calculated that there are more than 4000 gene fragment acquisitions within *Helitrons* in a single maize inbred, suggesting that exon shuffling is a very active process right now in at least some flowering plant genomes.

In plants, gene fragment acquisition by TEs is not a process unique to *Helitrons*. *Bs1* of maize was the first reported retrotransposon that contained sequences similar to a portion of a normal host gene, a plasma membrane proton ATPase (Bureau et al. 1994; Jin and Bennetzen 1994; Palmgren 1994), and many additional cases of similar phenomena have now been found

(Wang et al. 2006). Novel sequence acquisition is also observed for DNA elements, especially in the *Mutator* system (Talbert and Chandler 1988; Jiang et al. 2004). In the rice genome, over 3000 *Mutator* elements were reported to contain fragments derived from more than 1000 cellular genes and at least 5% appear to be expressed (Jiang et al. 2004). Hence, some plant species appear to have a manic rate of genic sequence rearrangement, with the potential to create a vast array of novel genes and genetic functions (Bennetzen 2005).

Results

A structure-based approach for *Helitron* identification

The identification of *Helitrons* is difficult because of their few and tiny structural features. The approaches used up to now can be divided into five categories. One approach is to search for Rep/helicase or RPA-like protein homology (Kapitonov and Jurka 2001; Poulter et al. 2003; Arkhipova and Meselson 2005; Pritham and Feschotte 2007). The second approach is to identify *Helitrons* as *de novo* insertions (Lal et al. 2003; Gupta et al. 2005; Choi et al. 2007). Insertions that mutated the *sh2* gene and the *ba1* gene in maize both turned out to be *Helitron* insertions. Another spontaneous mutation in morning glory was also caused by *Helitron* insertion. The third approach is to search for similarity to known *Helitrons* or known *Helitron* ends (Kapitonov and Jurka 2001; Choi et al. 2007; Pritham and Feschotte 2007; Tempel et al. 2007; Du et al. 2008; Sweredoski et al. 2008). The fourth approach involves characterizing identified repeats (Kapitonov and Jurka 2001). There are a few programs that can find repetitive sequences in a given genome, such as RECON (Bao and Eddy 2002) and Spectral Repeat Finder (Sharma et al.

2004). However, high levels of sequence diversity make it difficult to precisely define boundaries of *Helitron* elements found by such programs. This approach can only detect *Helitron* families that have abundances above a certain arbitrarily chosen copy number. The fifth approach is to search for violations of micro-colinearity between genomes (Brunner et al. 2005; Lai et al. 2005; Morgante et al. 2005; Xu and Messing 2006). This method requires well-studied colinear regions from different haplotypes in order to define *Helitron* boundaries precisely, and is very time and labor consuming.

In order to circumvent the limitations of previous approaches, a computer program called HelSearch was developed using the tiny structural features of *Helitrons*, and a requirement for at least two identical 3′ ends between separate elements, to find *Helitrons* in any genomic sequence (see Methods). *A. thaliana*, *Medicago truncatula*, *Oryza sativa*, *Sorghum bicolor*, *C. elegans* and *Drosophila melanogaster* genomes were screened for *Helitrons* using this program. In addition, as a negative control, the genomic sequence of *A. thaliana* was pooled and randomly reconstructed, and then searched with the same procedure. No *Helitron* was identified by HelSearch in the random sequence (Table 2.1).

In two small genomes, those of *A. thaliana* and *C. elegans*, *Helitrons* have been well characterized (Kapitonov and Jurka 2001). A total of 281 intact elements from 10 *Helitron* families were identified in *Arabidopsis* by HelSearch, including 2 new families, and 281 intact elements from 4 families were found in the *C. elegans* genome (no new family) (see Table 2.2 for details). Families were defined as containing elements that shared the same intact 3′ end, and subfamilies as those with same intact 5′ ends and the same intact 3′ end. The *Helitrons* in *A.*

thaliana and *C. elegans* make up 1.3% and 2.3% of their nuclear genomes, respectively, a number consistent with previous calculations (Kapitonov and Jurka 2001). Two *Helitron* families in *Arabidopsis* were found to have acquired gene fragments, compared to 1 family with gene fragments reported in previous studies (Hollister and Gaut 2007).

Few *Helitrons* have been reported in *Medicago*, rice and sorghum. HelSearch identified 230, 651 and 608 intact elements from 10, 23 and 11 families, respectively, in these three species. Only 4 families in *Medicago* and 3 in rice were previously known. *Helitrons* compose about 1.3%, 2.1% and 3.0% of the *Medicago*, rice and sorghum genomes, respectively. Three families in *Medicago*, 7 families in rice and 2 families in sorghum were found to have acquired gene fragments. *Helitrons* seem to acquire gene fragments of all types, as assessed by GO category analysis (Table 2.3). Putative autonomous elements were found in all of the above-mentioned five genomes. No intact *Helitron* was found in the *D. melanogaster* genome. Table 2.1 provides a summary of these results.

A minimum total number of elements was estimated by searching for conserved 3′ ends. In *Arabidopsis*, *C. elegans*, *Medicago*, rice and sorghum genomes, there are at least 1200, 600, 1300, 7000 and 5000 *Helitrons*, respectively.

Although some elements from different genomes belong to the same family, it does not necessarily mean they are the most related elements, because families are defined by a shared 3′ end sequence. For instance, *Medicago Helitrons* exhibit only 2 independent gene acquisition events, both quite ancient. Due to changes in the 3′ end, they have evolved into what we now call different families.

Because our approach can only identify elements that have at least 2 copies in the genome, a linear regression analysis was performed to estimate the number of intact single copy elements that are likely to have been missed (Figure 2.2). By this approach, single copy families of *Helitrons* in *A. thaliana*, *O. sativa*, *M. truncatula* and *S. bicolor* were estimated to number approximately 19, 162, 9 and 271, respectively. Hence, HelSearch is predicted to find the great majority of elements in each of these genomes, but to identify less than half of the different element families. It should be noted, however, that these low-copy-number elements have been missed by all previous searches.

Helitron distribution along plant and animal chromosomes

The identified *Helitrons* in the *Arabidopsis*, rice and *C. elegans* genomes were mapped along the sequenced genomes in these species. The results indicate that *Helitrons* in *Arabidopsis* are enriched in gene-poor pericentromeric regions (Figure 2.3A), showing a similar pattern to that previously seen for LTR retrotransposons and opposite to the DNA transposons that have been found to be preferentially associated with gene-rich regions in those genomes (Feng et al. 2002; Wright et al. 2003). Similarly, *Helitrons* in the *C. elegans* genome were found to be most abundant in the gene-poor terminal regions of each chromosome (Figure 2.3B). In *C. elegans*, retrotransposons are enriched in the gene rich chromosome center and DNA transposons are relatively more abundant toward the chromosome termini, as now observed for *Helitrons* (Waterston 1998; Duret et al. 2000; Surzycki and Belknap 2000). Rice, on the other hand (Figure 2.3C), exhibited a less ordered pattern of *Helitron* distribution, with some pericentromeric regions rich in these elements and others less so.

Helitron structures and specificities

The hairpins of *Helitrons* have higher predicted melting temperatures than are found in similar hairpins that are not associated with *Helitrons* (Figure 2.4). For reasons unknown, the two eudicot plants (*Arabidopsis* and *Medicago*) and *C. elegans* demonstrated a much higher and uniform range of predicted *Helitron* melting temperatures (Tm) than did the two monocots (rice and sorghum).

Fifty base pairs upstream and downstream of all intact *Helitron* insertions were used to screen for insertion specificities beyond the flanking AT dinucleotide. The results (Figure 2.5A-E) indicate that insertion regions are relatively A/T rich. The region from the insertion sites to about 12 bp downstream shows an 82% A plus T abundance. As a control, AT dinucleotide sites were chosen randomly from each genome, and 50 bp both upstream and downstream were scored for A/T composition. A Chi-square test comparing A/T content to *Helitron* insertion sites at each position was performed and the results (Figure 2.5F) demonstrated that most of the positions from the insertion site to 12 bp downstream and 3 bp upstream were significantly more A/T rich than would have been expected by chance.

New end creation and sequence acquisition

Analysis of the complete set of *Helitrons* in *Arabidopsis* indicated that these elements can acquire new sequences by recognizing either a new 3′ termination site or a new 5′ start site. Figure 2.6A shows an example of a *Helitron* family with a new 3′ end. Because of their lesser homology to each other (84-90% pairwise identity over the entire elements [Fig 2.6A and data not shown]), *Helitrons* in Sequences 7 to 10 are proposed to have the ancestral structure.

Helitrons represented by Sequences 1 to 6 are expected to be recently derived, as they exhibit 90-98% similarity. This proposition is supported by neighbor joining tree analysis with 1000 bootstrap replicates (Fig. 2.6B). If one element had a mutation in the hairpin, noted as the first star from the left, that damaged the rolling-circle stop signal, then a transpositional replication event might not stop at the original termination site. However, by chance, another potential stop site will sometimes be present somewhere downstream. A *Helitron* thus acquires another stop and a few hundred base pairs of sequence to complete the transposition event, giving rise to the *Helitrons* in Sequences 1 to 6.

In Figure 2.6C, Sequences 1 through 5 appear to represent *Helitrons* that are now recognizing a new 5′ start site without any mutation to the normal 5′ start sites. That 1-5 are recently derived is supported by the fact that *Helitrons* in Sequence 1-5 have 96-98% similarity, while the *Helitrons* represented by Sequences 6-11 exhibit 88-96% similarity.

Hence, it appears that *Helitrons* can acquire flanking sequences by recognizing a new 5′ start site or a new 3′ termination site. We observed similar end sequence acquisitions for *Helitron* families in other species, including *Hup* in rice and *Hip* in sorghum (data not shown).

Discussion

Efficient discovery of *Helitron* transposable elements

HelSearch proved superior to any of the five previous methods of identifying *Helitrons*. In the *A. thaliana* genome, where *Helitrons* are fairly well studied, new *Helitron* families were still identified. For genomes where *Helitrons* have been identified but are not very well characterized

such as *Medicago* and rice, the new approach uncovered a large number of families that were not previously detected.

The HelSearch approach yielded 3 false positives in the rice genome, all LTR retrotransposons having the same terminal structures as *Helitrons*. Ten previously identified *Helitrons* were missed in the *Arabidopsis* genome, 20 in the *C. elegans* genome and 2 in the rice genome. All of the false negatives from the *Arabidopsis* and rice genomes are putative autonomous elements, and thus are easily detected by a homology-based search for *Helitron*-encoded proteins. Those missed from the *C. elegans* genome have an unusual *Helitron* 3′ end structure (namely, no gap between the predicted hairpin and the conserved 3′ end sequences). In most of the previous *Helitron* discovery procedures, many intact elements were missed because they have low copy numbers, have accumulated many mutations including big insertions/deletions, or are unusually large in size (8-20 kb). It can be difficult to build multiple sequence alignments because of these *Helitron* properties. The sensitivity of the structure-based approach (correctly identified / (correctly identified + false negatives)) is 93%, and the specificity (correctly identified / (correctly identified + false positives)) is 99%.

There are several reasons why all other techniques for *Helitron* discovery have been less sensitive and/or more error-prone. A Rep/helicase protein-based search yields a large number of false negatives, because the majority of *Helitrons* are non-autonomous elements. A similarity-based search will not identify any new families, and will thus work poorly in newly studied genomes. Such programs (Du et al. 2008; Sweredoski et al. 2008) only capture variations within the known families and are likely to provide incorrect annotation on nested *Helitrons*. A repeat-

based search requires extensive manual curation to identify *Helitron* families, an overwhelming task in large genomes with substantial DNA repetition such as sorghum or maize, and misses the great majority of families because they have a low copy number.

Based on the overall sensitivity and specificity, the structure-based approach to identify *Helitron* elements is quite successful and especially useful to identify *Helitron* elements in a newly characterized genome. However, because at least two copies are needed to make an alignment, single copy *Helitrons* will be missed. Finally, the HelSearch program does not identify *Helitron* fragments for families that contain no intact copies. As with most informatic approaches, a full set of tools are best utilized to provide a comprehensive discovery process. HelSearch, accompanied by homology searches to Rep/Helicase and known *Helitrons,* would be an especially comprehensive strategy.

Distributions of *Helitrons* within and between genomes

In all three genomes where these elements could be comprehensively mapped, those with both near-complete sequence descriptions and a large number of *Helitrons*, it was observed that *Helitrons* preferentially accumulate in gene-poor regions. It is not clear whether this is caused by insertion specificities, possibly slower rates of DNA removal in gene-poor heterochromatic regions, and/or selection against *Helitron* insertions in gene-rich regions. These questions can be best answered in a genome with highly active *Helitrons*, so that *de novo* insertions could be analyzed prior to the action of natural selection or sequence degradation and removal.

When compared to genome size, it is apparent that different eukaryotic genomes can accumulate very different numbers of *Helitrons*. In this study, the genomes of sorghum (~750

Mb), *Medicago* (~250 Mb), rice (~400 Mb), *Arabidopsis* (~120 Mb), *Drosophila* (~150 Mb) and *C. elegans* (~100 Mb) were predicted to contain a respective minimum of 22, 2, 8, 2, 0 and 2 Mb of *Helitrons*. None of these elements appear to be major contributors to genome size in any species studied. Moreover, their absolute quantities do not correlate with genome size, indicating that the host characteristics that allow different levels and rates of LTR retrotransposon amplification and removal (Vitte and Bennetzen 2006) do not act on *Helitrons* in an absolutely parallel manner.

Helitron properties

Although all of the eukaryotic genomes investigated had a wealth of possible *Helitron* ends, as indicated by the presence of the short terminal consensus sequences and a nearby 3′ hairpin, it was found that *Helitrons* tended to have hairpins with a high predicted Tm. In the rolling circle transposition model (Kapitonov and Jurka 2001), helicase unwinds the double stranded DNA and the hairpin serves as a stop signal to terminate the transposition event. A high melting temperature may allow a *Helitron* hairpin to serve as a particularly powerful stop signal for transcription and/or rolling-circle replication. Roles for the non-*Helitron* short hairpins deserve to be investigated, although it should be noted that the randomly reconstructed *Arabidopsis* sequence also yielded many such short hairpins (about 1/3 as many as seen in the real *Arabidopsis* genome). Thus, some may have no function but be an unavoidable outcome of other issues of sequence arrangement. It should be noted, however, that HelSearch's requirement for identification of two of the same 3′ ends to prove an element was a *Helitron* will mean that some single copy intact *Helitrons* are missed. Hence, it is possible that all high Tm hairpins of the

approximate size found in *Helitrons* are actually associated with *Helitrons*. It is also true, from the results with rice and sorghum, that some fairly low Tm hairpins can function in *Helitrons*, but these may often be newly created elements ends that have not yet been fully selected for a high Tm.

All transposable elements exhibit a degree of insertion specificity, some for specific sequences but more commonly (at least in eukaryotes) for a specific set of chromatin-associated proteins, like silencers or RNA polymerase subunits. Because of its mechanism of rolling-circle transposition, it is possible that the A/T-rich insertion specificity observed for *Helitrons* would facilitate helicase unwinding associated with the next rounds of transposition. Interestingly, the AT-richness bias is mostly for the region 3′ to the insertion site. This should give *Helitrons* a bias not only for the region of their insertion but also for their orientation.

Element evolution and the acquisition of new element sequences

One of the great mysteries of *Helitron* function is how they acquire new internal sequences. When these sequences are parts of genes, the acquisition has an increased potential for the creation of a new gene. Studies in maize suggested that gene fragments are sequentially acquired, perhaps during transposition, and can occur at both ends (Brunner et al. 2005). The process observed for both 3′ and 5′ end sequence acquisition in this study suggested the possibility that *Helitrons* may skip the original end to thereby acquire new sequences. Of course, this may not be the only way for *Helitrons* to acquire new sequences. For instance, an integrase like that seen in integrons might initiate a site-specific recombination event that would lead to some gene capture

events (Hall and Collis 1995), although an integrase of this type has not yet been reported to be encoded within any *Helitrons*.

The rolling circle transposition model proposes that the hairpin serves as a stop signal during *Helitron* transposition. With the presence of other *Helitrons* in a genome, and other *Helitron*-end like sequences, it is likely that an end-like sequences could be acquired from many genomic locations, including from nearby *Helitrons* to create chimeric elements.

Future prospects

For *Helitrons*, the issues of transposition mechanism, gene acquisition processes and the fates of acquire gene fragments remain unresolved. Future research will need to approach these issues, and can be done best in species that have a high likelihood of containing active *Helitrons*. At this time, maize is a particularly strong candidate for an optimal species for these studies because of the presence of recently created mutations (Lal et al. 2003; Gupta et al. 2005) and the abundance of haplotype variation associated with *Helitron* presence/absence (Morgante et al. 2005). The use of HelSearch on the maize genome is underway (Yang and Bennetzen, unpub. res.) and is yielding a great wealth of new *Helitrons* for functional, structural and evolutionary analysis.

Methods

Structure-based *Helitron* identification: *A. thaliana* genomic sequence build 6 (TAIR6), *M. truncatula* sequence version 1.0, *O. sativa* ssp. *japonica* cultivar Nipponbare sequence version 4.0, *S. bicolor* sequence Sbi1 assembly, *C. elegans* sequence build WS144, and *D. melanogaster*

sequence build 4.1 were downloaded to screen for *Helitrons*. A random sequence assembly was also generated with the same genome size and GC content as the *A. thaliana* genome by randomly rearranging each nucleotide into a full genome pseudomolecule.

The program "HelSearch" was designed to search for CTRRT in genomic sequence first. To narrow the results, insertion site T was included in this search. The proposed *Helitron* end (helend) structure is composed of a minimum of 6 hairpin pairs (2 mismatches allowed) upstream of the CTRR, a 2-4 base pair loop, and 5-8 base pairs between the hairpin and CTRR. The program was developed in PERL to search for these features in a given genome. Identified candidate helends were grouped together by their hairpin structure. Flanking sequences were obtained for each helend, and multiple alignments by CLUSTALW were performed for those sequences within each group. These alignments were inspected manually to define boundaries of *Helitron* elements. Two or more sequences with a clear 5′ end at the TC dinucleotide and clear 3′ boundary were defined as *Helitron* elements (Figure 2.1B). BLASTX screening of the NCBI non-redundant protein database was used to find gene fragments acquired by *Helitrons*, with a required expect value of *e*-10, or *e*-5 if a homology was found in a different species (self hits, hypothetical proteins and tranposase proteins were ignored).

The HelSearch program can be downloaded at http://lxyang.myweb.uga.edu/helsearch1.0.tar.gz and http://sourceforge.net/projects/helsearch/. All *Helitron* sequences described in this article can be downloaded at http://lxyang.myweb.uga.edu/All_Helitrons_AT_CE_MT_OS_SB.tar.gz.

Helitron family and subfamily assignment: There is no current knowledge involving *Helitron* cis or trans activation, so it is not possible to categorize *Helitrons* as non-autonomous or autonomous members that respond to the same transposition function. So, a family classification was instead assigned as a description of general ancestry, wherein sequences with the most similar 3′ ends (30 bp with at least 80% identity) were classified as members of the same family and sequences with the most similar 5′ ends (30 bp with at least 80% identity) were classified as members of the same subfamily. A short word starting with "H" was assigned as the name for each *Helitron* family. The same family name, when used for elements from different species, indicates that the shared name describes elements with the same 3′ end (80% identity over 30 bp).

Helitron properties: Predicted melting temperatures of hairpins were calculated by the melt program in the UNAFold 3.3 software package (Markham and Zuker 2005). The program was run on Windows. Parameters were set as follows: DNA molecule, sodium concentration 1, magnesium concentration 0. Flanking sequences (50 base pairs both upstream and downstream) of all intact *Helitrons* insertion sites were used to calculate base composition, with PICTOGRAM.

Helitron abundance: The HelSearch program identifies only intact elements (i.e., those with both a conserved 3′ end and a conserved 5′ end). In order to find the genome contribution of all *Helitron* elements, both intact and fragmented, in a given genome, a BLAST search against an entire genome was performed using all intact elements. Hits with at least 100 bp of 80% identity were counted to calculate genome contribution. The total number of 3′ ends (30 bp with at least

80% identity to 3′ ends of all intact elements) was used to estimate the minimum total number of elements in a given genome.

Acknowledgements

We thank Dr. Renyi Liu and Dr. Clementine Vitte for advice and training on issues regarding the discovery of transposable elements, gene fragments and genes in plant genome sequence data. This research was supported by National Science Foundation grant DBI 0607123.

References

Arkhipova, I.R. and Meselson, M. 2005. Diverse DNA transposons in rotifers of the class Bdelloidea. *Proc Natl Acad Sci USA* **102:** 11781-11786.

Bao, Z. and Eddy, S.R. 2002. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res* **12:** 1269-1276.

Bennetzen, J.L. 2005. Transposable elements, gene creation and genome rearrangement in flowering plants. *Curr Opin Genet Dev* **15:** 621-627.

Brunner, S., Pea, G., and Rafalski, A. 2005. Origins, genetic organization and transcription of a family of non-autonomous helitron elements in maize. *Plant J* **43:** 799-810.

Bureau, T.E., White, S.E., and Wessler, S.R. 1994. Transduction of a cellular gene by a plant retroelement. *Cell* **77:** 479-480.

Choi, J.D., Hoshino, A., Park, K.I., Park, I.S., and Iida, S. 2007. Spontaneous mutations caused by a *Helitron* transposon, *Hel-It1*, in morning glory, *Ipomoea tricolor*. *Plant J* **49:** 924-934.

Doutriaux, M.P., Couteau, F., Bergounioux, C., and White, C. 1998. Isolation and characterisation of the RAD51 and DMC1 homologs from *Arabidopsis thaliana*. *Mol Gen Genet* **257:** 283-291.

Du, C., Caronna, J., He, L., and Dooner, H.K. 2008. Computational prediction and molecular confirmation of *Helitron* transposons in the maize genome. *BMC Genomics* **9:** 51.

Duret, L., Marais, G., and Biemont, C. 2000. Transposons but not retrotransposons are located preferentially in regions of high recombination rate in *Caenorhabditis elegans*. *Genetics* **156:** 1661-1669.

Feng, Q., Zhang, Y., Hao, P., Wang, S., Fu, G., Huang, Y., Li, Y., Zhu, J., Liu, Y., and Hu, X. 2002. Sequence and analysis of rice chromosome 4. *Nature* **420:** 316-320.

Gilbert, W. 1987. The exon theory of genes. *Cold Spring Harb Symp Quant Biol* **52:** 901-905.

Gupta, S., Gallavotti, A., Stryker, G.A., Schmidt, R.J., and Lal, S.K. 2005. A novel class of *Helitron*-related transposable elements in maize contain portions of multiple pseudogenes. *Plant Mol Biol* **57:** 115-127.

Hall, R.M. and Collis, C.M. 1995. Mobile gene cassettes and integrons: capture and spread of genes by site-specific recombination. *Mol Micro* **15:** 593-600.

Hollister, J.D. and Gaut, B.S. 2007. Population and evolutionary dynamics of *Helitron* transposable elements in *Arabidopsis thaliana*. *Mol Biol Evol* **24:** 2515-2524.

Hood, M.E. 2005. Repetitive DNA in the automictic fungus *Microbotryum violaceum*. *Genetica* **124:** 1-10.

Jiang, N., Bao, Z., Zhang, X., Eddy, S.R., and Wessler, S.R. 2004. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431:** 569-573.

Jin, Y.K. and Bennetzen, J.L. 1994. Integration and nonrandom mutation of a plasma membrane proton ATPase gene fragment within the *Bs1* retroelement of maize. *Plant Cell* **6:** 1177-1186.

Kapitonov, V.V. and Jurka, J. 1999. Molecular paleontology of transposable elements from *Arabidopsis thaliana*. *Genetica* **107:** 27-37.

Kapitonov, V.V. and Jurka, J. 2001. Rolling-circle transposons in eukaryotes. *Proc Natl Acad Sci USA* **98:** 8714-8719.

Lai, J., Li, Y., Messing, J., and Dooner, H.K. 2005. Gene movement by *Helitron* transposons contributes to the haplotype variability of maize. *Proc Natl Acad Sci USA* **102:** 9068-9073.

Lal, S.K., Giroux, M.J., Brendel, V., Vallejos, C.E., and Hannah, L.C. 2003. The maize genome contains a *Helitron* insertion. *Plant Cell* **15:** 381-391.

Le, Q.H., Wright, S., Yu, Z., and Bureau, T. 2000. Transposon diversity in *Arabidopsis thaliana*. *Proc Natl Acad Sci USA* **97:** 7376-7381.

Markham, N.R. and Zuker, M. 2005. DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res***:** 577-581.

Mendiola, M.V., Bernales, I., and de la Cruz, F. 1994. Differential roles of the transposon termini in IS*91* transposition. *Proc Natl Acad Sci USA* **91:** 1922-1926.

Morgante, M., Brunner, S., Pea, G., Fengler, K., Zuccolo, A., and Rafalski, A. 2005. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat Genet* **37:** 997-1002.

Murad, L., Bielawski, J.P., Matyasek, R., Kovarik, A., Nichols, R.A., Leitch, A.R., and Lichtenstein, C.P. 2004. The origin and evolution of geminivirus-related DNA sequences in *Nicotiana*. *Heredity* **92:** 352-358.

Palmgren, M.G. 1994. Capturing of host DNA by a plant retroelement: *Bs1* encodes plasma membrane H+-ATPase domains. *Plant Mol Biol* **25:** 137-140.

Poulter, R.T.M., Goodwin, T.J.D., and Butler, M.I. 2003. Vertebrate helentrons and other novel *Helitrons*. *Gene* **313:** 201-212.

Pritham, E.J. and Feschotte, C. 2007. Massive amplification of rolling-circle transposons in the lineage of the bat *Myotis lucifugus*. *Proc Natl Acad Sci USA* **104:** 1895-1900.

Rensing, S.A., Lang, D., Zimmer, A.D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T., Perroud, P.F., Lindquist, E.A., and Kamisugi, Y. 2008. The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science* **319:** 64-69.

Sharma, D., Issac, B., Raghava, G.P.S., and Ramaswamy, R. 2004. Spectral Repeat Finder (SRF): identification of repetitive sequences using Fourier transformation. *Bioinformatics* **20:** 1405-1412.

Stenger, D.C., Revington, G.N., Stevenson, M.C., and Bisaro, D.M. 1991. Replicational release of geminivirus genomes from tandemly repeated copies: evidence for rolling-circle replication of a plant viral DNA. *Proc Natl Acad Sci USA* **88:** 8029-8033.

Surzycki, S.A. 1999. Characterization of repetitive DNA elements in *Arabidopsis*. *J Mol Evol* **48:** 684-691.

Surzycki, S.A. and Belknap, W.R. 2000. Repetitive-DNA elements are similarly distributed on *Caenorhabditis elegans* autosomes. *Proc Natl Acad Sci USA* **97:** 245-249.

Sweredoski, M., DeRose-Wilson, L., and Gaut, B. 2008. A comparative computational analysis of nonautonomous *Helitron* elements between maize and rice. *BMC Genomics* **9:** 467.

Talbert, L.E. and Chandler, V.L. 1988. Characterization of a highly conserved sequence related to mutator transposable elements in maize. *Mol Biol Evol* **5:** 519-529.

Tempel, S., Nicolas, J., El Amrani, A., and Couee, I. 2007. Model-based identification of Helitrons results in a new classification of their families in *Arabidopsis thaliana*. *Gene* **403:** 18-28.

Vitte, C. and Bennetzen, J.L. 2006. Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. *Proc Natl Acad Sci USA* **103:** 17638-17643.

Wang, Q. and Dooner, H.K. 2006. Remarkable variation in maize genome structure inferred from haplotype diversity at the *bz* locus. *Proc Natl Acad Sci USA* **103:** 17644-17649.

Wang, W., Zheng, H., Fan, C., Li, J., Shi, J., Cai, Z., Zhang, G., Liu, D., Zhang, J., and Vang, S. 2006. High rate of chimeric gene origination by retroposition in plant genomes. *Plant Cell* **18:** 1791-1802.

Waterston, R. 1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. The C. elegans Sequencing Consortium. *Science* **282:** 2012-2018.

Wright, S.I., Agrawal, N., and Bureau, T.E. 2003. Effects of recombination rate and gene density on transposable element distributions in *Arabidopsis thaliana*. *Genome Res* **13:** 1897-1903.

Xu, J. and Messing, J. 2006. Maize haplotype with a *helitron*-amplified cytidine deaminase gene copy. *BMC Genet* **7:** 52.

Zhou, Q., Froschauer, A., Schultheis, C., Schmidt, C., Bienert, G.P., Wenning, M., Dettai, A., and Volff, J.N. 2006. Helitron transposons on the sex chromosomes of the platyfish *Xiphophorus maculatus* and their evolution in animal genomes. *Zebrafish* **3:** 39-52.

Table 1: Summary of *Helitron* discovery and description in six multicellular eukaryotes.

Organism	Common Name	Genome Screened	No. of Intact Elements	No. of Families	No. of New Families	Putative Autonomous Elements	No. of Families with Acquired Gene Fragment(s)	Total No. of Elements	Genome Abundance
Random Sequence (ATH)	Negative control	115 MB	0	0	0	-	0	-	0
Arabidopsis thaliana	Mustard weed	115 MB	281	10	2	+	2	1242	1.30%
Medicago truncatula	Barrel Medic	243 MB	230	10	6	+	3	1386	1.29%
Oryza sativa ssp. japonica	Rice	389 MB	651	23	18	+	7	6947	2.09%
Sorgum bicolor	Sorghum	748 MB	608	11	11	+	2	4875	3.00%
Caenorhabditis elegans	Worm	100 MB	281	4	0	+	0	600	2.30%
Drosophila melanogaster	Fruit Fly	154 MB	0	0	0	-	0	-	0

Table 2.2: Summery of *Helitron* families

(A) *Helitron* families in *A. thaliana*

Family Name	Number of Intact Element	Size Range (bp)	Original Name(s) of Known *Helitron(s)*	Putative Autonomous Elements	Gene Fragment(s)
Hug	195	360-2734	ATREP1,ATREP2,ATREP3,ATREP5,ATREP11,AtREP21,HELITRONY1A		+
Hali	27	775-1709	ATREP6,ATREP9		
Hari	22	642-1500	ATREP10		
Hane	15	1346-2111	AtREP20	+	
Hani	11	606-655	ATREP13		
Hup	3	2106-2544	ATREP4		
Hame	3	1340-1628	ATREP12		
Hip	2	1189-1219			
Hake	2	3389-3395			+
Haki	1	817-903	ATREP7		
Total	281	360-3395			

(B) *Helitron* families in *C. elegans*

Family Name	Number of Intact Element	Size Range (bp)	Original Name(s) of Known *Helitron(s)*	Putative Autonomous Elements	Gene Fragment(s)
Hike	206	182-4802	HELITRON2_CE,HELITRONY2_CE,HELITRONY3_CE,HELITRONY4_CE	+	
Hibe	38	866-8965	HELITRON1_CE,HELITRONY1_CE	+	
Hino	30	1313-1605	NDNAX3_CE		
Hite	7	1709-2164	NDNAX1_CE		
Total	281	182-8965			

(C) *Helitron* families in *M. truncatula*

Family Name	Number of Intact Element	Size Range	Original Name(s) of Known *Helitron(s)*	Putative Autonomous Elements	Gene Fragment(s)
Hip	99	1281-5357	HELITRON1N_MT	+	+
Hem	53	375-3249			
Heno	33	2445-5717	HELMET	+	+
Heto	33	3055-5643			+
Hefo	3	8279-8394	HELMET2	+	
Heke	2	450-472			
Hebo	2	817			
Hecu	2	1910-1913			
Heku	2	3043			
Hecho	1	14261	HELITRON1MT	+	
Total	230	375-14261			

(D) *Helitron* families in *O. sativa* ssp. *japonica* cultivar Nipponbare

Family Name	Number of Intact Element	Size Range (bp)	Original Name(s) of Known *Helitron(s)*	Putative Autonomous Elements	Gene Fragment(s)
Hup	280	230-15654		+	+
Hip	196	242-4984			+
Hair	116	180-2577			+
Hole	18	2357-4705			
Hoy	10	350-10152	HELITRON7_OS	+	
Hocku	6	1058-6263			+
Hova	5	3776-13092	HELITRON2_OS, HELITRON3_OS	+	
Holie	4	705-1911			
Horia	4	2635-3198			
Howi	3	1896-1955			
Hota	2	2085-2133			+
Hoku	2	1429-1695			
Honu	2	210-219			
Hok	2	187-189			
Hani	2	1140-1143			
Hoju	2	2180-2233			
Holda	2	15063-15147	HELITRON1_OS, HELITRON5_OS,HELITRON8_OS	+	
Holku	2	1891-1977			
Homno	2	3532-3790			+
Holta	2	1172-1422			
Holvo	2	6117-6186			+
Howa	2	1503-1506			
Hofo	2	911-912			
Total	668	180-15147			

(E) *Helitron* families in *S. bicolor*

Family Name	Number of Intact Element	Size range	Original Name(s) of Known *Helitron(s)*	Putative Autonomous Elements	Gene Fragment(s)
Hip	289	289-21899		+	+
Hair	153	238-532			
Hole	119	1248-4499			+
Hok	22	356-3005			
Hoy	9	986-11062			
Huyo	6	268-384			
Hubi	2	148-149			
Husi	2	3423-3426			
Huca	2	172-174			
Huphi	2	1718-1903			
Huga	2	2435-2568			
Total	608	148-21899			

Table 2.3: Gene fragments acquired by *Helitrons*

Name	Size	Gene fragments	Expect value	Identity (%)
Hug_AT2_1	2182	putative protein	e-35	94
		protein in rice	e-14	46
Hake_AT1_1	3389	transferase	e-24	93
Hug_AT4_75	747	hypothetical protein (yeast)	e-14	51
Hup_OS3_1	3410	RNA polymerase beta chain	e-42	84
Hota_OS1_1	2083	unknown	e-40	94
Hocku_OS1_1	6261	FAD-binding domain-containing protein-like	e-79	81
		unknown	e-16	90
Hup_OS1_3	2467	calmodulin	e-17	80
Hup_OS1_56	2830	RNA polymerase beta subunit	e-9	51
Hup_OS1_65	2359	calmodulin	e-8	43
Hup_OS1_157	2687	ARM repeat fold domain	e-41	80
Hair_OS8_1	2372	unknown	e-10	57
Hair_OS9_1	1927	unknown	e-30	89
		unknown	e-30	87
Hair_OS10_1	373	unknown	e-10	73
Hip_OS5_1	4313	unknown	e-11	43
Hip_OS1_1	1404	root cap protein 1-like	e-31	82
Hip_OS7_1	4908	unknown	e-22	78
		unknown	e-11	53
Hip_OS7_4	4140	unknown	e-36	92
		unknown	e-18	97
Homno_OS1_1	3530	unknown	e-71	96
Holvo_OS1_1	6184	Putative RING zinc finger protein	e-35	65
Hip_MT1_1	5016	Nucleic acid-binding, OB-fold	e-14	85
Hip_MT1_13	4364	Nucleic acid-binding, OB-fold	e-18	91
Hip_MT1_43	3913	Nucleic acid-binding, OB-fold	e-10	50
Hip_MT2_19	3768	Polynucleotidyl transferase, Ribonuclease H fold	e-23	58
Heno_MT1_1	4140	Nucleic acid-binding, OB-fold	e-111	93
Heto_MT1_1	3202	Nucleic acid-binding, OB-fold	e-51	68
Heto_MT1_5	3187	Nucleic acid-binding, OB-fold	e-52	69
Heto_MT1_15	5428	Nucleic acid-binding, OB-fold	e-29	59
Hip_MT3_1	4228	Nucleic acid-binding, OB-fold	e-10	86
Hole_SB1_19	2261	putative glucosyl transferase	e-147	87
Hole_SB1_23	1610	glycerol-3-phosphate dehydrogenase	e-17	82
Hole_SB1_28	3451	putative proteasome 26S non-ATPase subunit 2	e-10	40

Hole_SB1_40	2610	AP-3 complex delta subunit-like protein	0	68
Hole_SB1_46	4125	tetratricopeptide repeat (TPR)-containing protein	e-19	50
Hole_SB1_53	2795	Putative ubiquitin carboxyl terminal hydrolase	e-13	81
Hole_SB1_55	2936	putative RNA-binding protein	e-39	93
Hole_SB1_57	3159	hypothetical protein in rice	e-15	58
Hole_SB1_59	2699	DNA directed RNA polymerase	e-13	72
Hole_SB1_68	4499	UMP synthase	e-97	96
Hip_SB5_1	4072	protein phosphatase 2C-related	e-122	68
Hip_SB2_43	1763	F-box domain containing protein	e-21	68
Hip_SB2_44	1721	hypothetical protein in rice	e-56	77
Hip_SB2_54	1675	unknown in maize	e-29	76
Hip_SB2_56	2493	unknown in maize	e-19	90
Hip_SB3_1	4400	unknown in maize	e-19	83
Hip_SB3_4	5982	unknown in maize	e-12	52
Hip_SB3_9	2408	cycloartenol synthase	e-79	86

Figure Legends

Figure 2.1: (A) *Helitron* structure: 5′ TC and 3′ CTRR termini (shown in upper case), a predicted small hairpin structure near the 3′ end, and insertion into the target dinucleotide AT (shown in lower case). (B) The end sequences of eight *Helitrons* (grey rectangles) from *A. thaliana.* 5′ start with TC, 3′ end with CTRR, insert between AT dinucluetide (red rectangle), and flanking sequences are shown. Sequences within the insertion site align very well and sequences outside of the insertion cannot be aligned because they are insertions at different genomic locations.

Figure 2.2: Distribution of *Helitron* copy numbers in different subfamilies. (A) X axis: *Helitron* copy number in each subfamily; Y axis: number of subfamilies with this element copy number. *Helitron* copy number results are shown for *A. thaliana*, *C. elegans*, *O. sativa*, *M. truncatula* and *S. bicolor* in this Figure. (B) X axis: *Helitron* copy numbers in each subfamily (2-5 copies shown); Y axis: log transformed number (base 10) of the number of subfamilies with this copy number. Linear regression lines are also shown and used to estimate the number of single copy elements.

Figure 2.3: Distribution of *Helitrons* along the chromosomes of *A. thaliana* (A), *C. elegans* (B) and *O. sativa* (C). The X axis shows the chromosome locations and the Y axis indicates the copy numbers of *Helitrons* (both intact and truncated elements) per 100kb. Narrow lines in (A) and (C) on the X axis indicate the positions of centromeres.

Figure 2.4: Melting temperature distributions of predicted hairpins in the five studied genomes. Blue denotes predicted hairpins across the entire genome, and red denotes predicted hairpins in

identified *Helitron*s. The X axis indicates predicted melting temperature (0 - 100℃) while the Y axis shows the frequency of predicted hairpins with that predicted Tm.

Figure 2.5: Insertion site base composition. Left side of *Helitrons*: 20 base pairs upstream of *Helitron* insertion sites. Right side of *Helitrons*: 20 base pairs downstream of *Helitron* insertion sites. Black lines in (A) – (E) indicate average GC content of the genome. (A) *Helitrons* of *A. thaliana*. (B) *Helitrons* of *C. elegans*. (C) *Helitrons* of *O. sativa*. (D) *Helitrons* of *M. truncatula*. (E) *Helitrons* of *S. bicolor*. (F) Chi-square tests of GC content for 50 base pairs upstream and downstream of *Helitron* insertion sites on each position compared to a random AT flanking site from the same genome. The *P* value is indicated by color, with black representing the most significant and yellow the least significant associations.

Figure 2.6: A mechanism for *Helitron* sequence acquisition and new end creation. (A) *Helitrons* in *Arabidopsis* pick up a new 3′ end. Stars indicate sequence divergence in the helend. A pink rectangle encloses the 3′ CTRR of *Helitrons*, while blue underlining shows the predicted *Helitron* hairpins. (B) Neighbour joining tree for sequences in (A), with 1000 bootstrap replicates. (C) *Helitrons* in *Arabidopsis* acquire new 5′ termini. Pink underlining shows the 5′ TC of these *Helitrons*.

Figure 2.1

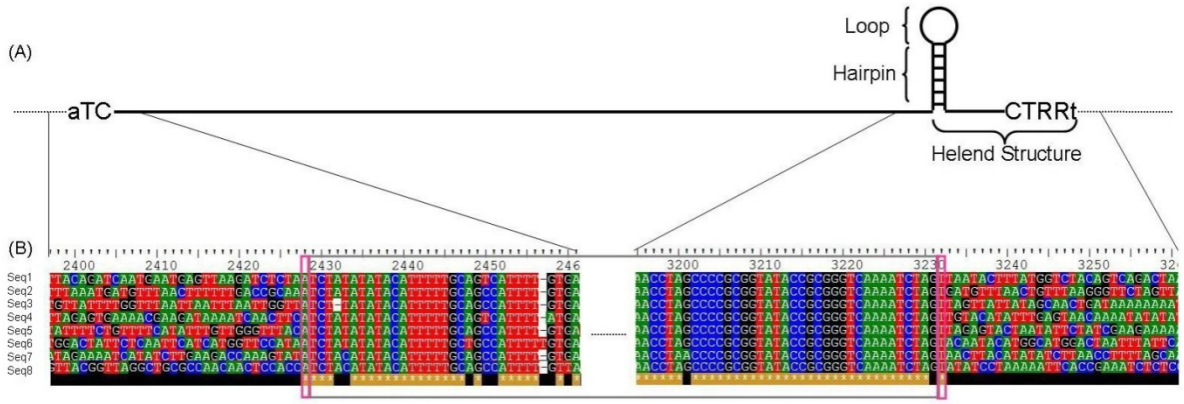
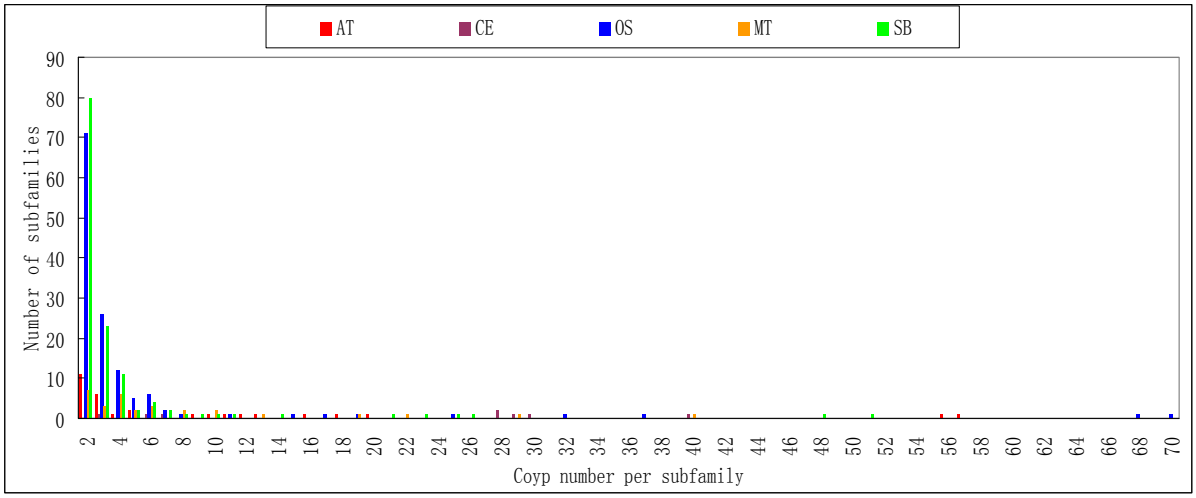
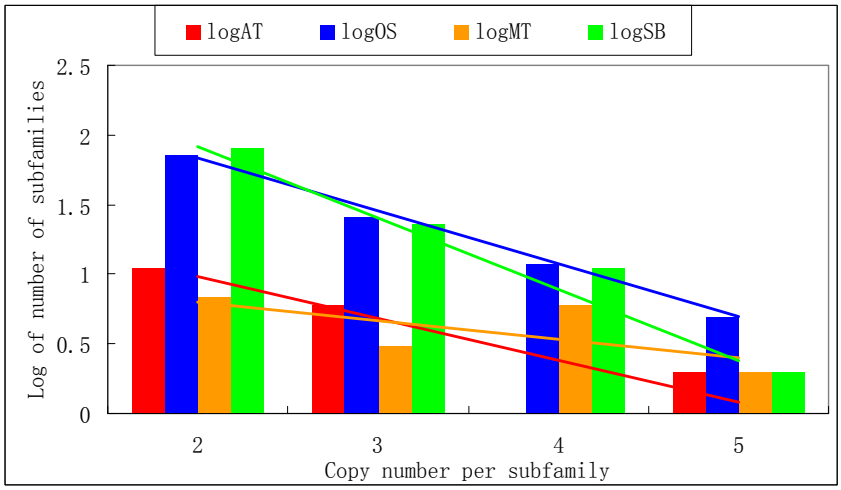

(A)
Loop
Hairpin
aTC
CTRRt
Helend Structure
(B)
2400
2410
2420
2430
2440
2450
3200
3210
3220
3240
3250
Seq1
Seq2
Seq3
Seq4
Seq5
Seq6
Seq7
Seq8

Figure 2.2

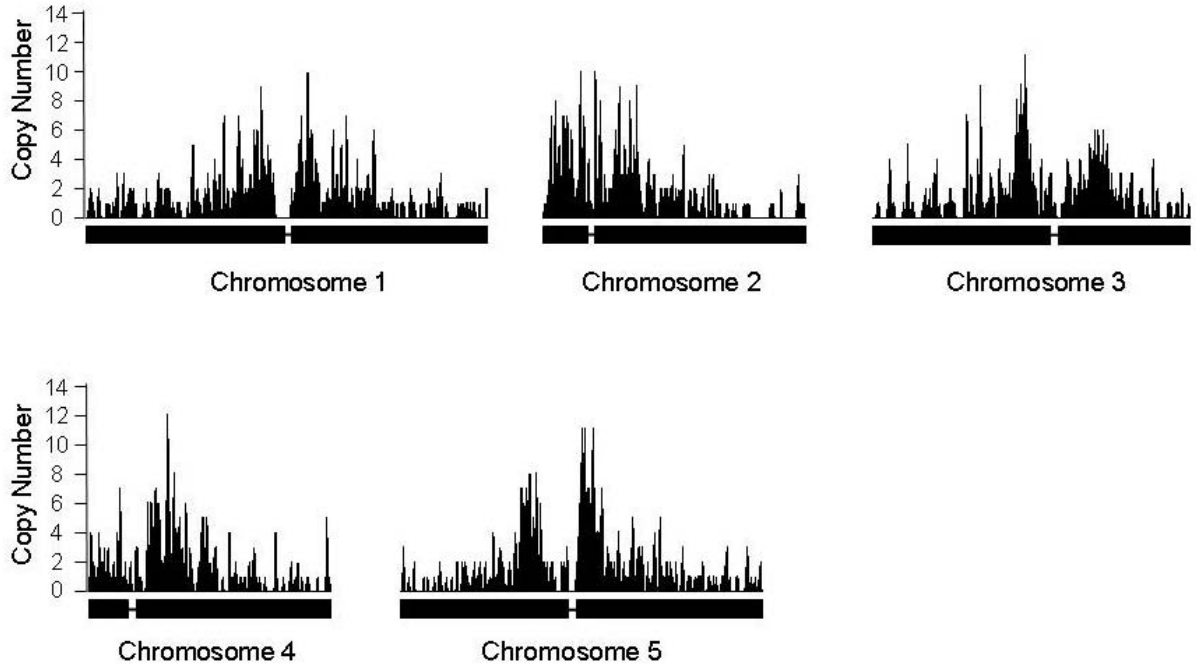

AT
CE
OS
MT
SB
Number of subfamilies
Coyp number per subfamily

(A)

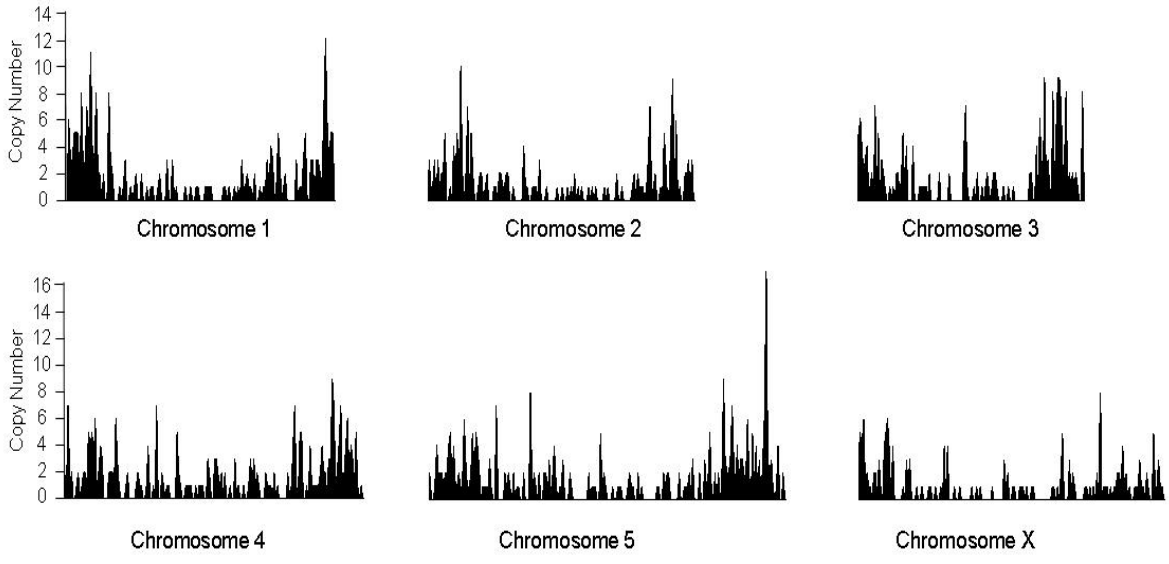

logAT
logOS
logMT
logSB
Log of number of subfamilies
Copy number per subfamily

(B)

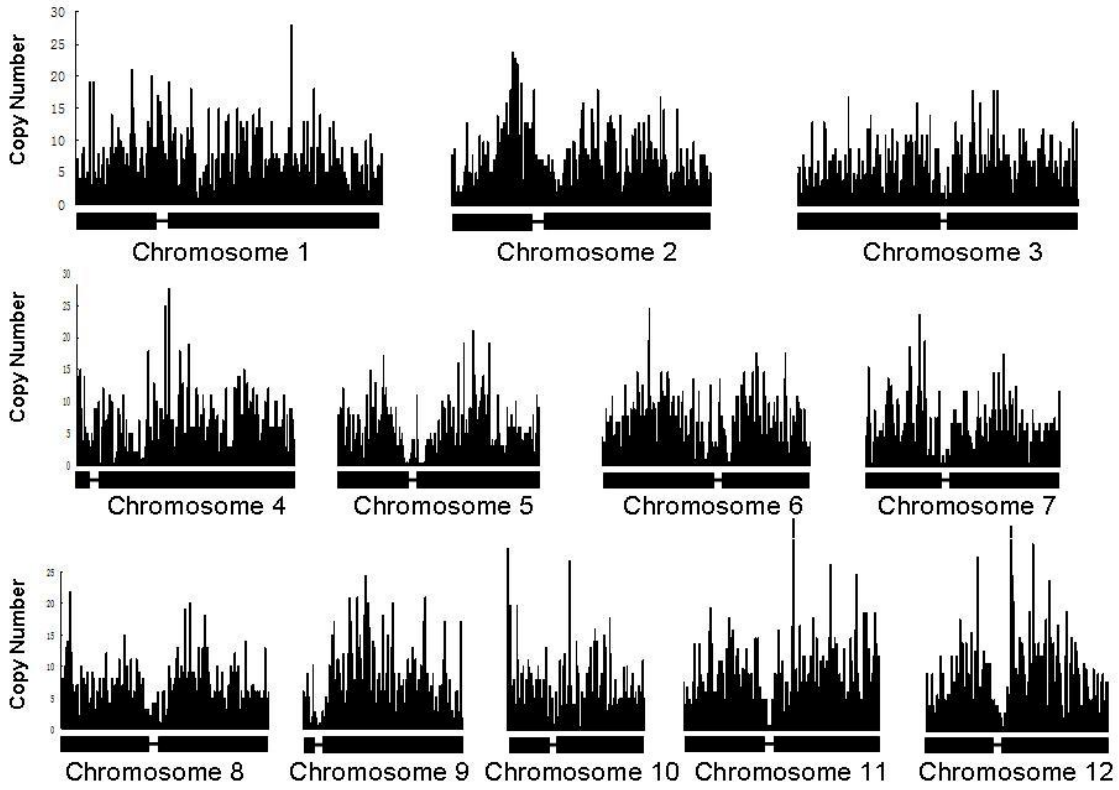
Figure 2.3:


Copy Number
14
12
10
8
6
4
2
0
Chromosome 1
Chromosome 2
Chromosome 3
Chromosome 4
Chromosome 5

(A)


Copy Number
14
12
10
8
6
4
2
0
Chromosome 1
Chromosome 2
Chromosome 3
16
Chromosome 4
Chromosome 5
Chromosome X

(B)


Copy Number
30
25
20
15
10
5
0
Chromosome 1
Chromosome 2
Chromosome 3
Chromosome 4
Chromosome 5
Chromosome 6
Chromosome 7
Chromosome 8
Chromosome 9
Chromosome 10
Chromosome 11
Chromosome 12

(C)

Figure 2.4

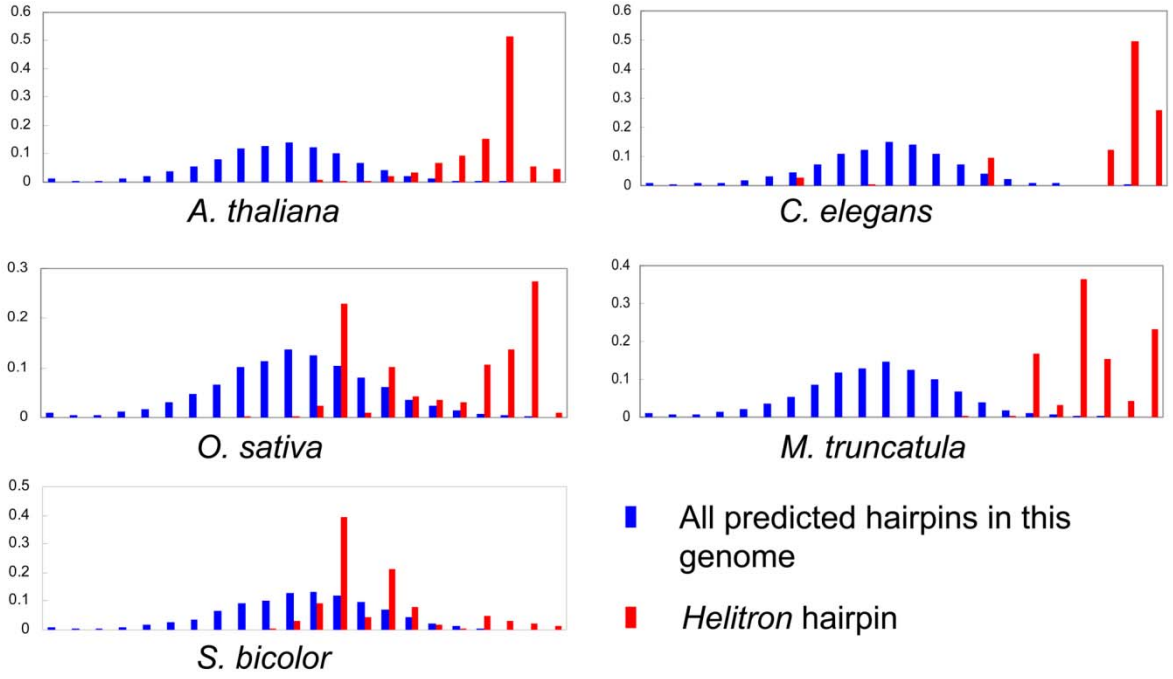

0.6
0.5
0.4
0.3
0.2
0.1
0
A. thaliana
C. elegans
0.3
0.2
0.1
O. sativa
0.4
M. truncatula
0.5
S. bicolor
All predicted hairpins in this genome
Helitron hairpin

Figure 2.5

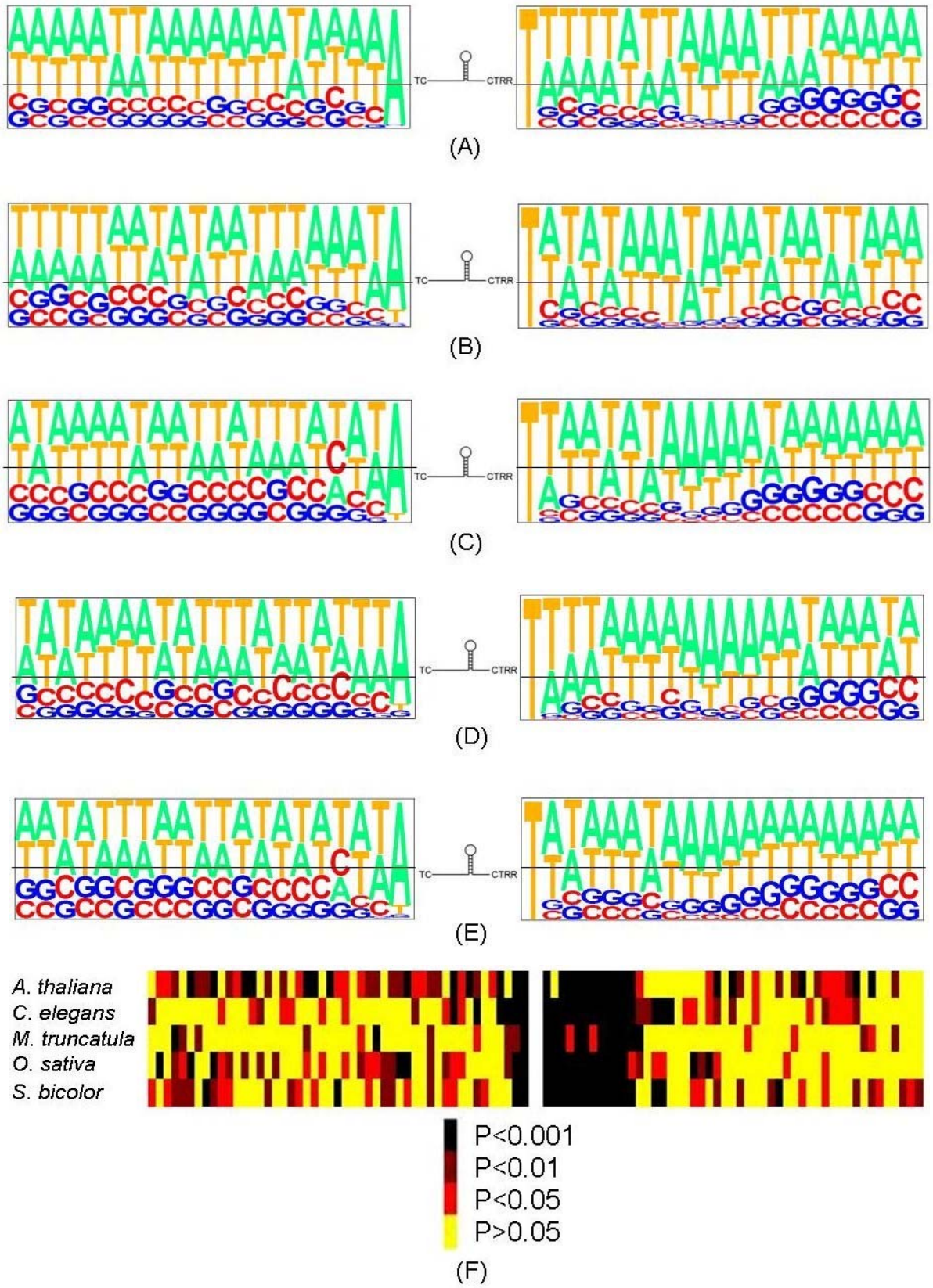
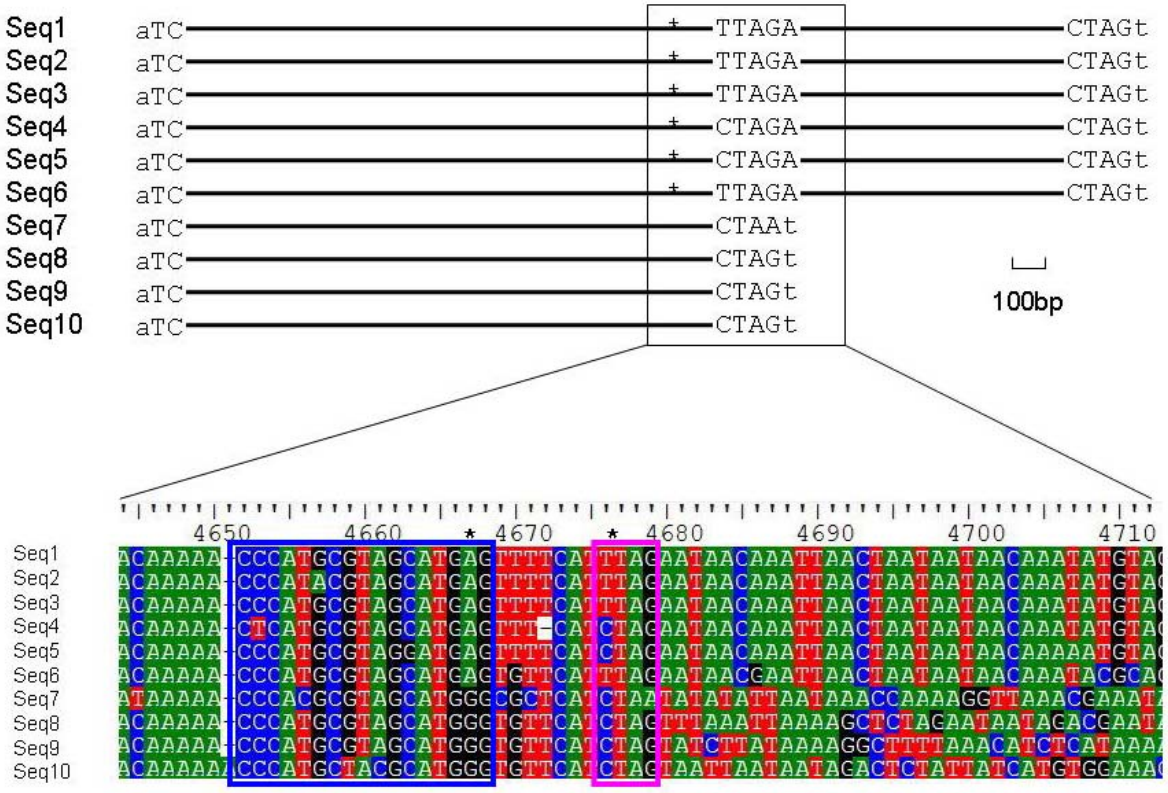
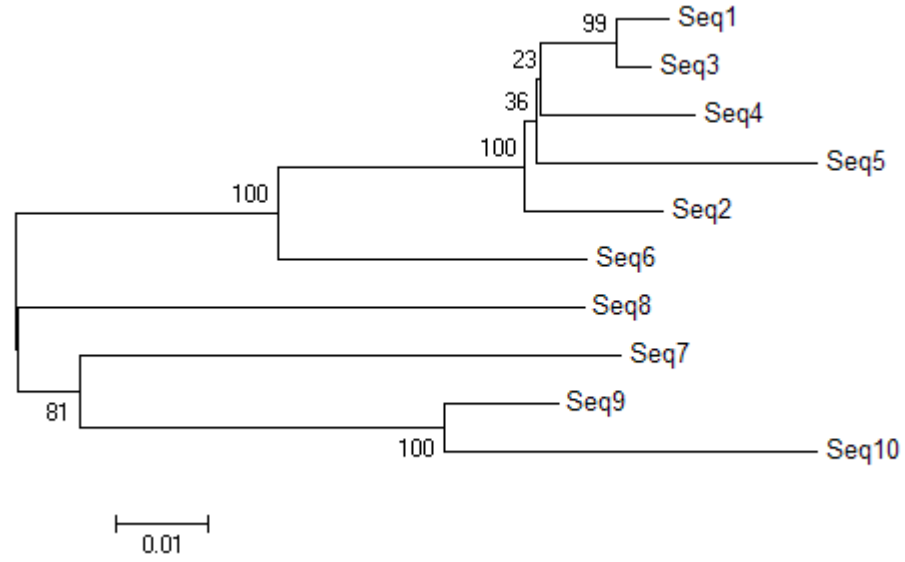

TC
CTRR
(A)
(B)
(C)
(D)
(E)
A. thaliana
C. elegans
M. truncatula
O. sativa
S. bicolor
P<0.001
P<0.01
P<0.05
P>0.05
(F)

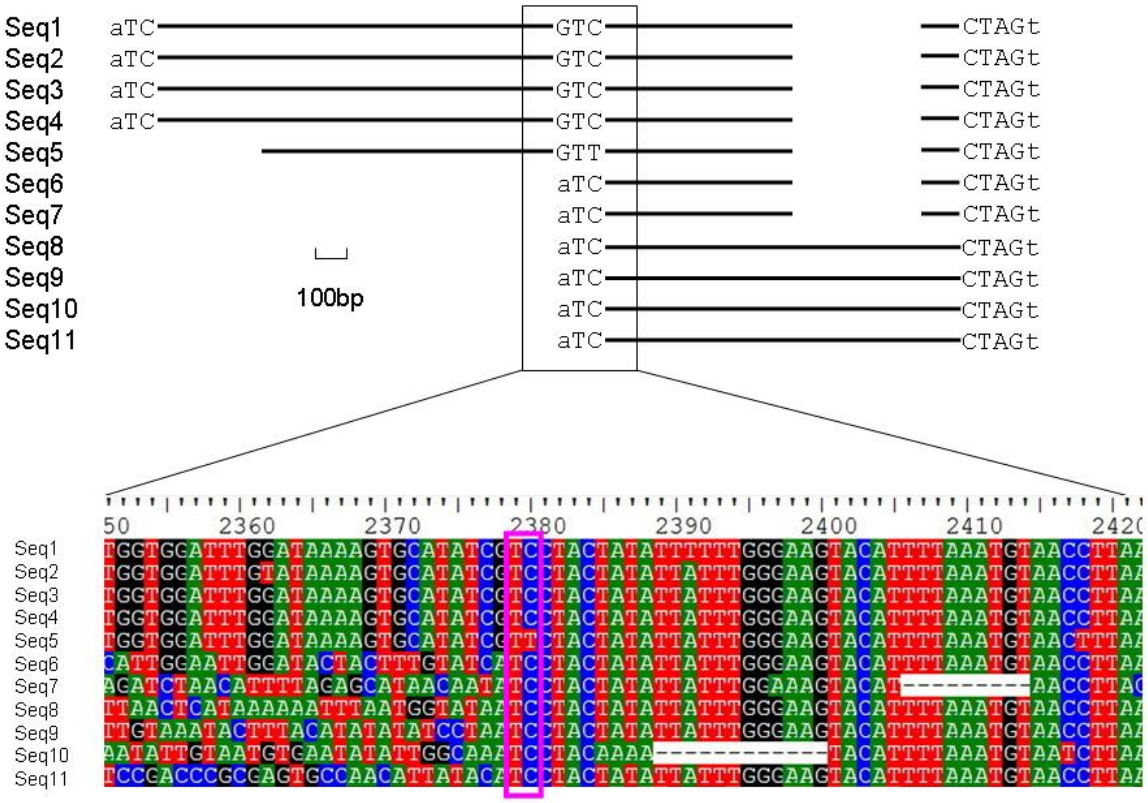
Figure 2.6


Seq1 aTC * TTAGA CTAGt
Seq2 aTC * TTAGA CTAGt
Seq3 aTC * TTAGA CTAGt
Seq4 aTC * CTAGA CTAGt
Seq5 aTC * CTAGA CTAGt
Seq6 aTC * TTAGA CTAGt
Seq7 aTC CTAAt
Seq8 aTC CTAGt
Seq9 aTC CTAGt
Seq10 aTC CTAGt
100bp
4650
4660
4670
4680
4690
4700
4710
Seq1
Seq2
Seq3
Seq4
Seq5
Seq6
Seq7
Seq8
Seq9
Seq10

(A)


99
Seq1
23
Seq3
36
Seq4
100
Seq5
100
Seq2
Seq6
Seq8
Seq7
81
Seq9
100
Seq10
0.01

(B)


Seq1
Seq2
Seq3
Seq4
Seq5
Seq6
Seq7
Seq8
Seq9
Seq10
Seq11
aTC
GTC
CTAGt
GTT
100bp
50
2360
2370
2380
2390
2400
2410

(C)

CHAPTER 3

DISTRIBUTION, DIVERSITY, EVOLUTION AND SURVIVAL OF *HELITRONS* IN THE MAIZE GENOME [2]

[2] Yang, L. and Bennetzen, J.L. 2009. Distribution, Diversity, Evolution and Survival of *Helitrons* in the Maize Genome. *Proc Natl Acad Sci USA* 106: 19922-19927. Reprinted here with permission of publisher.

Abstract

Homology and structure-based approaches were used to identify *Helitrons* in the genome of maize inbred B73. A total of 1930 intact *Helitrons* from eight families (62 subfamilies) and >20,000 *Helitron* fragments were identified, accounting for ~2.2% of the B73 genome. Transposition of at least one of these families is ongoing, but the most prominent burst of amplification activity was ~250,000 years ago. Sixty percent of maize *Helitrons* were found to have captured fragments of nuclear genes (~840 different fragment acquisitions, with tens of thousands of predicted gene fragments inside *Helitrons* within the B73 assembly). Most acquired gene fragments are undergoing random drift, but 4 percent were calculated to be under purifying selection, while another 4 percent exhibit apparent adaptive selection, suggesting beneficial effects for the host or for *Helitron* transposition/retention. Gene fragment capture is frequent in some *Helitron* subfamilies, with as many as ten unlinked genes providing DNA inserts within a single element. Gene fragment acquisition appears to positively influence element survival and/or ability of the *Helitron* to acquire additional gene fragments. *Helitrons* with gene fragment captures in the antisense orientation have a lesser chance of survival. *Helitron* distribution in maize exhibits severe biases, including preferential accumulation in relatively gene-rich regions. Insertions, however, are not usually found inside genes. Rather, *Helitrons* preferentially insert near (but not into) other *Helitrons*. This biased accumulation is not due to a preference for cis or nearby transposition, suggesting a specific association between *Helitron* integration functions and unknown chromatin characteristics that specifically mark *Helitrons*.

Keywords:

exon shuffling, gene fragment acquisition, genome evolution, insertion specificity, transposable elements

Introduction

Helitrons are a class of transposable elements (TEs) that were initially discovered by computational analysis of repetitive nuclear DNA in the model plant *Arabidopsis thaliana* (Kapitonov and Jurka 2001). Subsequent studies have shown that *Helitrons* are broadly distributed in eukaryotes, including in all studied plants (Yang and Bennetzen 2009). They are characterized by a 5′ TC terminus and a 3′ CTRR terminus accompanied by a predicted small hairpin structure near the 3′ end. *Helitrons* preferentially insert into the dinucleotide AT, but they do not generate target site duplications. Some elements encode Rep/helicase-like and/or RPA-like proteins that are believed to be involved in the transposition process. The bacterial IS*91* element also encodes a Rep/helicase protein, and is known to transpose via a rolling circle process (Mendiola and de la Cruz 1992; Mendiola et al. 1994), so it is expected that *Helitrons* use the same mechanism for amplification and insertion within eukaryotic genomes.

Plant *Helitrons* often capture gene fragments. Sometimes, fragments from multiple genes that normally reside in unlinked chromosomal locations are found inside individual elements (Lal et al. 2003; Brunner et al. 2005; Gupta et al. 2005; Lai et al. 2005; Morgante et al. 2005; Wang and Dooner 2006; Xu and Messing 2006; Jameson et al. 2008; Sweredoski et al. 2008). This gene fragment acquisition is known to occur at the DNA level because contiguous introns and exons are found within the acquired DNA. The captured fragments are usually small, but more than one case has been observed where a full gene has been captured by a *Helitron* in maize (Lal et al. 2003; Jameson et al. 2008). Although the mechanism of gene fragment

acquisition is not known, it usually (~88% of the time in maize (Kapitonov and Jurka 2007), based on a small number of *Helitrons*) involves capture of fragments in the same transcriptional orientation as the *Helitron* gene encoding Rep/helicase. Some of these *Helitrons* produce chimeric transcripts where acquired introns are spliced, sometimes alternatively, and the junctions between fragments are occasionally processed as crude *de novo* introns (Lal et al. 2003; Brunner et al. 2005). Hence, the fusion and expression of different gene fragments that are catalyzed by *Helitrons* have many of the properties of exon shuffling, a mechanism proposed for the creation of the first multi-domain genes (Gilbert 1987). In maize, more than 4000 gene fragment acquisitions were predicted to be within *Helitrons* in a single maize inbred (Morgante et al. 2005), and dozens more have been seen in other plant species (Yang and Bennetzen 2009), suggesting that the creation of novel gene candidates might be a very active process in plants.

In all of the plant and animal species investigated, *Helitrons* have been found to show significant biases in structure and distribution. The elements in nematodes, rice and *Arabidopsis* accumulate primarily in gene-poor regions, although this bias is less dramatic in the rice genome (Yang and Bennetzen 2009). The predicted hairpin loops near the 3′ end of all *Helitrons* exhibit a very high predicted melting temperature (Tm), but less so in the grasses rice and sorghum than in the eudicots *Arabidopsis* and *Medicago truncatula* (Yang and Bennetzen 2009).

Genomic copy numbers and diversity in *Helitrons* are quite variable. For instance, the moss *Physcomitrella patens* has only one identified family consisting of eight very similar members (suggesting very recent transposition) (Rensing et al. 2008). In dramatic contrast, vesper bats contain highly abundant *Helitrons* (for instance, more than 100,000 per haploid genome in

Myotis lucifugus), but they appear to be missing from all other orders of placental mammals that have been investigated (Pritham and Feschotte 2007). In limited analyses in flowering plants (angiosperms), four relatively small genomes have been found to each contain several hundred elements from at least ten families per species, with some of these families (defined by a unique 3′ end structure) found in most of these taxa (Yang and Bennetzen 2009).

The recent full genome sequence analysis of the nuclear genome in maize inbred B73 has provided the opportunity for comprehensive discovery of *Helitrons* in this important plant species (Schnable et al. 2009). *Helitrons* in maize have been very recently active, as suggested by their association with mutations in *sh2* and *ba1* genes (Lal et al. 2003; Gupta et al. 2005) and their numerous presence/absence polymorphisms across maize haplotypes (Lai et al. 2005; Morgante et al. 2005; Wang and Dooner 2006; Xu and Messing 2006). Hence, maize should be an ideal organism for the further study of *Helitron* evolution and function. We employed a structure-based approach (Yang and Bennetzen 2009) to find *Helitrons* in maize, uncovering about 2000 intact elements and many thousands of *Helitron* fragments that together comprise over 2% of the genome (Schnable et al. 2009). This article describes the discovery and analysis process for these *Helitrons* in maize, and the properties of the identified elements. The results indicate unexpected specificities in both action and evolution, and emphasize the exceptional role that *Helitrons* play in the rearrangement and enrichment of eukaryotic genomes.

Results and Discussion

Helitron identification and classification

The BAC-by-BAC (bacterial artificial chromosome) sequence data for the maize nuclear genome in inbred B73, covering ~90% of that genome, was analyzed one BAC at a time, encompassing ~2.05 Gb of analyzed DNA (Schnable et al. 2009). Initially, a structure-based approach was used to search for *Helitrons* in this sequence. The program employed, HelSearch (Yang and Bennetzen 2009), initially identifies only intact elements (with at least two intact copies, so ends can be verified). A total of 1923 intact elements were found in this manner. Each element was inspected manually, and all were confirmed *Helitrons,* so the false positive rate in this analysis was zero. Seven intact *Helitrons* identified by earlier studies (Brunner et al. 2005; Gupta et al. 2005; Lai et al. 2005; Morgante et al. 2005) were not found in this structure-based search because only one intact copy was present in the assembled sequences. When combined, the resulting 1930 intact *Helitrons* account for 6.2 Mb of nuclear DNA in the B73 reference sequence (Schnable et al. 2009).

Intact *Helitrons* were classified into 8 families based on the presence of a unique most-terminal 3′ end, with two elements requiring at least 80% identity to be considered members of the same family (Yang and Bennetzen 2009). Members of a family that had very different 5′ ends were classified into subfamilies, with sequences sharing the most 5′-terminal 30 bp (at least 80% identity) classified as members of the same subfamily. This definition of families and subfamilies is based on two primary criteria: namely, the >80% homology rule for family designation that a consortium of plant TE researchers has concluded is the appropriate threshold for distinguishing

families (Wicker et al. 2007) and the fact that the ends of TEs of all types are the primary sites where activating factors (e.g., transposases) determine the specificity of the unique family each unique activating protein will mobilize and/or integrate. Of these 8 families (62 subfamilies), 5 were previously unknown in maize (Table 3.1). The family we named *Hip* is the most abundant, with 1897 elements that make up 98% of the total intact elements, including the previously identified *Helitrons* inserted into *sh2-7527* and *ba1-ref*, and those initially named *HelA*, *McC_bz1_GHI*, *B73_9002_NOPQ*, *B73_14578* and *Hel-BSSS53-z1C1* (Lal et al. 2003; Brunner et al. 2005; Gupta et al. 2005; Lai et al. 2005; Morgante et al. 2005; Xu and Messing 2006). We compared our subfamilies with the results of Du et al. (Du et al. 2009). Eight of our subfamilies were not found in the Du et al. database, while 35 of our identified subfamilies were unnamed in their analysis and only found in their medium quality or high quality output of Helitronfinder. Four of the subfamilies in Du et al. (Du et al. 2009) were not found in our analysis.

Like previously investigated genomes (Yang and Bennetzen 2009), maize has some unique *Helitron* families and some shared families. *Hip*, the largest family in maize, is also the most numerous family in sorghum and the second most abundant in rice. *Hip* is the largest family in *Medicago,* while it is a low-copy-number family in *Arabidopsis*. However, the dominance of *Hip* in maize far exceeds the relative dominance of any *Helitron* family in any other investigated angiosperm genome, where the most abundant intact *Helitron* of any single family makes up only 42% to 70% of the total intact *Helitrons*. All of the other maize *Helitron* families that were identified have not yet been found in other genomes. Sorghum and rice share some families, such as *Hair*, *Hole*, *Hoke* and *Hoy* (Yang and Bennetzen 2009), but no intact elements of these

families were found in maize. *Hole*, *Hoke* and *Hoy* are low-copy-number families, while *Hair* is a high-copy-number family, having over 100 intact elements in both sorghum and rice. The 5′ and/or 3′ ends of some of these families from rice and sorghum do exist in the maize genome, but intact elements were not identified, so we could not guarantee that the maize end homologies were evidence of true *Helitrons*. Although some elements from different genomes belong to the same family, this does not necessarily mean that they are the most related elements, because families are defined by a shared 3′ end sequence. Due to dramatic changes in the 3′ end (for instance, new end sequence acquisition) (Yang and Bennetzen 2009), some *Helitrons* have evolved very recently into what we now call different families. These shared *Helitron* families in rice, sorghum and maize are not *per se* proof of a common ancestry in an ancient grass progenitor, and they do not contain highly similar internal regions suggestive of very recent horizontal transfer. The results suggest a highly conserved set of 3′ end sequences that are needed for *Helitron* function, but repeated convergent evolution to these sequences is also possible. Further analysis across a broader range of grass species will help resolve this issue.

Counting the total number of 5′ and 3′ ends for these 8 families in the B73 maize dataset yielded a respective ~22,000 and ~21,000 apparent ends. This provides an estimate of the total number of *Helitrons*, intact plus fragmented, of >22,000. These results suggest a very high ratio of fragmented to intact elements, but this is at least partly an artifact of the short length of contiguous sequences (contigs) in the current version of the B73 reference sequence (Schnable et al. 2009). Any of the numerous errors in order or orientation of contigs within a BAC sequence

scaffold could lead to the misinterpretation of an intact *Helitron* as two or more fragments of *Helitrons.*

Because TEs of different types can insert into each other, it was often not clear if the DNA inside an identified *Helitron* was *Helitron*-specific or representative of an unrelated TE. Hence, maize TE databases for long terminal repeat (LTR) retrotransposons, and cut-and-paste DNA transposons were obtained (Schnable et al. 2009). These databases were used to remove all of these types of TEs from the maize *Helitron* database. The resultant *Helitron* database was used in a BLAST-based search of the entire maize genomic sequence. Homologies with at least 100 contiguous base pairs of at least 80% identity were identified. In addition, a BLASTX search for *Helitron*-specific helicases was performed with a minimum Expect value of *e*-10. The results from both BLAST searches were combined, and redundancy was removed. From these analyses, *Helitrons* were found to contribute at least 2.2% of the maize nuclear genome in line B73, a total of 45.5 Mb in the 2045 Mb of the genome that has been sequenced (Schnable et al. 2009). Because the HelSearch analysis requires two intact elements in a single subfamily before the existence of a *Helitron* can be confirmed, many *Helitron* subfamilies or families with only a single intact member or no intact members might be missed. Hence, linear regression analysis was performed to estimate the number of intact single-copy elements that are likely to have been missed (Fig. 3.1) (Yang and Bennetzen 2009). By this approach, it can be estimated that there are ~20 intact single-copy subfamilies of *Helitrons* in the maize genome that we did not identify, and provides an interesting comparison to the 62 *Helitron* subfamilies discovered with more than one intact member. Hence, as in previous angiosperm investigations (Yang and Bennetzen 2009), it is

clear that a significant amount of *Helitron* diversity is missed by this and all previous studies, but that the great majority of *Helitron* genomic contributions have been identified in maize by discovery of the multicopy element families and subfamilies.

As observed in other species (Yang and Bennetzen 2009), the 3′ hairpins of maize *Helitrons* average a much higher predicted melting temperatures than that predicted for the full complement of similar hairpins in the maize genome sequence (Fig. 3.2). Unlike the full set of predicted hairpins, the maize *Helitron* hairpins do not exhibit a normal distribution, suggesting a strong selection for a high complementarity and GC content in these components of *Helitron* structure.

Helitron divergence

LTR retrotransposon insertion dates can be estimated by the degree of divergence of their two LTRs (SanMiguel et al. 1998; Ma and Bennetzen 2004; Ma et al. 2004) because they should be 100% identical at the time of insertion (SanMiguel et al. 1998). A similar approach was used to estimate the amplification time of intact *Helitrons,* relying not on conserved end homology but on the relatedness of different copies. By an all-by-all BLAST search of all intact *Helitrons,* the most related elements were identified. These are most likely to be derived from each other or from a shared common ancestor at a more recent time than from *Helitrons* that are less similar in sequence (Lerat et al. 2003; Hollister and Gaut 2007). An intact *Helitron* was aligned with its 2nd best hit (i.e., not with itself) by CLUSTALW. The degree of divergence was determined using the baseml module of PAML (v4.2) and used to calculate amplification dates. Figure 3.3A shows the predicted amplification dates for these intact *Helitrons*. The vast majority (99%) of the

intact *Helitrons* are predicted to have amplified within the last 6 million years, with an amplification peak at 0.25 million years ago (MYA). About 3% of intact maize *Helitrons* are 100% identical, suggesting very recent duplication and insertion. For comparison, the predicted amplification times of intact *Helitrons* in rice, sorghum and *Arabidopsis* were calculated in the same manner (Fig. 3.3A). Rice and sorghum both gave ~3% intact *Helitrons* with 100% identity and an amplification peak about 0.25 MYA, nearly identical to what was observed for maize despite their divergence from common ancestors many more than 11 MYA (Swigonova et al. 2004). In contrast, less than 1% of intact *Helitrons* in *Arabidopsis* are 100% identical, and the overall *Helitron* population exhibited a major amplification between 1 and 2 MYA (Fig. 3.3A).

Because transition mutation frequencies are quite variable across genomic regions, and especially frequent in TEs compared to coding exons (SanMiguel et al. 1998), it seemed appropriate to further pursue this analysis by ignoring transitions and only counting transversions. When divergence and amplification dates were recalculated in this manner, similar results were obtained (Fig. 3.3B).

In order to carefully characterize the nature of *Helitron* divergence in recent times, one hundred intact *Helitrons* were randomly selected for further analysis. Each was aligned to its 2nd best hit by CLUSTALW. Nucleotide changes and indels were counted (Table 3.2). Over 6000 single nucleotide changes were found. Small indels (1-5 bp) were found to be more frequent than larger indels, a result previously observed as an outcome of DNA removal processes in angiosperms (Ma and Bennetzen 2004). The ratio of transition to transversion mutations was

observed to be 3.05 to 1, suggesting that most *Helitrons* are heavily cytosine-methylated in the maize genome (SanMiguel et al. 1998).

Gene fragment acquisition

A BLASTX search against the NCBI non-redundant protein database was performed for each intact element. Over 60% of the intact elements (1194 elements) were found to have acquired one or more gene fragments. See Fig. 3.4 for examples. The *Helitron* with the most captured gene fragments (a total of ten) is a truncated element (the 5′ boundary could not be precisely identified, so this element is not mentioned in Fig. 3.5).

The numbers of gene fragment(s) acquired per element is shown in Fig. 3.5. Intact maize *Helitrons* were further classified into 498 "exemplars", where each exemplar had a unique internal sequence that was more than 20% different at the nucleotide level from any other *Helitron* internal regions. There are 104 *Helitron* exemplars that have not captured any gene fragments and 152 exemplars that have acquired one gene fragment. Two exemplars represent unrelated element groups that have each taken up as many as 7 different gene fragments. If these gene fragments acquisitions were independent events, one expects that the frequencies of exemplars with one, two, three, or more captured fragments would follow a Poisson distribution. The parameter λ was estimated to be 1.81 (overall mean). Given that there have been 840 different acquisition events, the predicted frequencies of exemplars in each category are shown in Fig. 3.5. Note that the observed exemplars that have acquired 2 or 3 gene fragments are fewer than predicted while those that have acquired 5 or 6 gene fragments are more frequent than predicted by a random acquisition model. A goodness of fit test yielded a *P* value of <0.0001,

suggesting that the capture events are not independent. These data suggest that acquisition of one gene fragment facilitates the capture of additional gene fragments or that there is superior survival for fragment-containing *Helitrons*. Alternatively, or in addition, some maize *Helitrons* may have an unusual propensity (of unknown molecular basis) for gene fragment acquisition that leads to an over-representation of *Helitrons* that contain multiple gene fragment captures.

There are over 700 intact elements that have acquired a phosphatase 2C-like gene fragment, mostly in the *Hip1* subfamily and its derivative subfamilies, *Hip26* and *Hip31*. *Hip1* is the most numerous (over 1000 intact elements) and most active subfamily in B73 maize. It is likely that the maize phosphatase 2C-like gene fragment capture was a single event that has been amplified by *Helitron* transposition. Subfamily *Hip26* was created by acquisition of a new 5′ end by a *Hip1* subfamily member, and the 5′ ends of *Hip1* and *Hip31* have experienced a number of small sequence changes that have led to their classification as different subfamilies. Sorghum *Helitron Hip_SB5_1* also has acquired a phosphotase 2C-like gene fragment (Yang and Bennetzen 2009), but a different portion of the gene from a different member of the phosphotase 2C-like gene family. This coincidence may be pure chance, or may reflect either some exceptional lability in this class of gene that promotes its acquisition/retention or, more likely, that capture and/or retention of a phosphotase 2C-like gene fragment has some selective advantage for a *Helitron* or for its host genome. As expected for a predicted independent origin, *Hip_SB5_1* in sorghum does not evidence any similarity to maize *Helitrons* at the nucleotide level, except at the 3′ end.

It is thought that the gene fragments captured by *Helitrons* have a strong bias in the orientation of their acquisition and/or retention (Kapitonov and Jurka 2007). When investigated in the available dataset from the B73 genome of maize, 282 exemplars were found to have gene fragments that are exclusively in the same orientation as the *Helitron* Rep/helicase gene. There are 20 exemplars that have gene fragments in the opposite orientation, and another 92 exemplars with multiple gene fragments that are in opposing orientations in individual elements. Hence, by this method of analysis, acquisition and/or retention of gene fragments within *Helitrons* is strongly biased (282/20 = 14.1 to 1 ratio) towards a conserved orientation that is compatible with *Helitron* promoter-driven expression.

To test whether *Helitrons* preferentially acquire gene fragments in the same orientation, or if gene fragments are acquired randomly, gene fragment orientation was plotted against element amplification time (Fig. 3.6). The *Helitrons* that had acquired gene fragments in the sense orientation had a broad amplification time distribution, but the majority of elements containing a gene fragment in the opposite orientation were of very recent origin. A Mann-Whitney test for significance in the difference in the respective median and mean ages of sense (0.71 MYA and 1.38 MYA) and antisense (0.38 MYA and 1.14 MYA) gene fragments gave a *P* value of 0.089. These results suggest that gene fragments acquired in the antisense orientation are more rapidly removed by selection. However, the effect is not overpowering, suggesting that *Helitrons* may also preferentially acquire gene fragments in the same orientation, so these combined biases account for the current ~14:1 ratio of sense to antisense inserts.

Helitrons in the maize B73 genome were estimated to have acquired ~4000 gene fragments in a previous study (Morgante et al. 2005). However, because this study was based on a presence/absence comparison between inbreds B73 and Mo17, many shared *Helitrons* and gene fragment captures could have been missed. Hence, the total number of gene fragments acquired by *Helitrons* was estimated from (a) the number of observed gene fragments inside intact *Helitrons* (2152) and (b) the ratio of intact *Helitrons* to total *Helitrons* (>22,000 to 1930 or ~11.4 to 1) detected in the assembled B73 reference genome (Schnable et al. 2009). If there is no bias for or against the truncation and sequence decay of *Helitrons* with respect to acquired gene fragments, then more than 24,000 gene fragments inside *Helitrons* are estimated in the B73 genome. Of course, enumeration of these captured fragments is likely to yield an overestimation because of the fragmented status of the B73 reference sequence assembly (Schnable et al. 2009).

Evolution of gene fragments

A simple way to determine whether the fragments inside *Helitrons* perform an important function is to look for biases either for or against synonymous mutations. Previous studies have reported such selection on acquired DNA fragments inside other TEs (Jiang et al. 2004; Juretic et al. 2005; Hanada et al. 2009). However, these investigations can be misleading if the authors compared the acquired gene fragment to its presumed ortholog in the host genome. Any perceived evidence of purifying selection, for instance, might actually be an outcome of sequence divergence between the allele of the host gene chosen as the ortholog and the actual allele that was the origin of the acquired gene fragment. To search for evidence of natural selection on gene fragments acquired by *Helitrons*, 44 *Helitrons* with independent gene

fragments acquisition were randomly chosen. Because some elements had acquired more than one gene fragment, this study included a total of 85 different gene fragments. Seven gene fragments were too short for informative analysis. Nucleotide sequences of gene fragments from the same acquisition event in different *Helitrons* were aligned and used to calculate the nonsynonymous and synonymous substitution dN/dS (ω) ratio between acquired gene fragments. A *P* value of 0.05 with Bonferroni corrections was chosen as the significance level. As a result, 3 gene fragments (4% of the total) exhibited significant evidence of purifying selection. For the abundant phosphatase 2C-like gene fragment acquired by *Helitrons*, 50 elements were randomly chosen and analyzed. The results indicated that these gene fragment are under strong negative selection ($P<0.0001$).

To investigate how inaccurate this analysis would have been if we compared gene fragments inside *Helitrons* to the host gene from which they were derived, a BLASTN search against the B73 maize gene set (Schnable et al. 2009) was first performed to identify the genes that contributed gene fragments to *Helitrons*. Seven gene fragments were not attributable to any corresponding host gene. For the 71 other inserts in *Helitrons,* captured gene fragments within the exemplar were always most homologous to only one maize gene. This suggests that none of these acquisitions within an exemplar group were independent captures from different genes in the same gene family. With the best candidate donor gene now identified, it was simple to calculate ω for the combined dataset of *Helitron* gene fragments and their corresponding host genes. From this analysis, 9 gene fragments appeared to be under significant purifying selection (*P* value smaller than 0.05 with the Bonferroni correction). Hence, this type of comparison to

host genes is likely to provide a major overestimate of purifying selection on gene fragments inside TEs.

To determine whether any of the gene fragments were under adaptive selection, M1 (neutral) and M2 (positive selection) models were built by PAML, and chi square tests were performed on those two models. Three gene fragments (4% of the total) yielded *P* values for adaptive selection that were smaller than 0.05 with a Bonferroni correction (Supplementary Table S4).

Transcriptional activity

Screening maize EST databases (>99.5% identical to a full length EST sequence) indicated that at least 9% of the identified maize *Helitron* sequences are expressed in at least one tissue). Comparison to an sRNA database (Johnson et al. 2007) indicated that ~90% of the identified *Helitrons* had at least one small RNA match, suggesting that maize *Helitrons* are subject to significant levels of epigenetic suppression.

Insertion preferences

Sequences flanking intact *Helitron* insertion sites were obtained in order to assess any possible insertion preference. Figure 3.7A and 3.7B shows the analysis suggesting that *Helitrons* insert preferentially into AT-rich maize DNA, as reported for *Helitrons* in other plant and animal genomes (Yang and Bennetzen 2009). As seen previously, the last 3 bp upstream and 8-10 bp downstream of the insertion exhibit an extreme AT richness, reflecting an insertion orientation bias (Yang and Bennetzen 2009).

Helitron distributions on each chromosome of the B73 reference maize genome (Schnable et al. 2009) were determined for both intact and fragmented elements. Fragmented

elements were considered valid if they contained at least 100 bp of contiguous >80% identity to a known intact element. In contrast to *Helitrons* in other plant genomes (Yang and Bennetzen 2009), maize *Helitrons* were found to be most abundant in gene-rich regions of the chromosomes (Supplementary Fig. 3.7C). The reasons for this dramatic difference are not known, but there are several possibilities.

It should be remembered that the *Helitron* distribution in a genome is the result of the balance between TE insertion and DNA removal (Bennetzen 2005; Bennetzen et al. 2005; Bennetzen 2007). The great haplotype variability in maize (Fu and Dooner 2002; Wang and Dooner 2006) indicates that TEs are completely removed from a region in less than two million years (Ma and Bennetzen 2004; Ma et al. 2004). Moreover, DNA removal (at least by unequal homologous recombination) appears to be more rapid in gene-rich regions (which have a higher level of meiotic recombination) than in pericentromeric regions (Ma and Bennetzen 2004; Ma et al. 2004; Ma and Bennetzen 2006). Hence, if maize *Helitrons* are mostly younger and more active than those in other studied species (Lal et al. 2003; Gupta et al. 2005), one might see a greater level of accumulation in gene-rich regions because there has been less time for their removal. To test this hypothesis, non-parametric Mann-Whitney tests were performed on *Helitrons* amplification times, to see if the maize *Helitron* amplification times (Fig. 3.3) are significantly different in other genomes. The intact maize *Helitrons* were compared to those in *Arabidopsis*, rice and sorghum. On average, maize *Helitrons* were found to be significantly younger than *Arabidopsis Helitrons* (P=0.017), but were determined to be significantly older than *Helitrons* in rice and sorghum (P<0.0001). However, if transversions alone are employed in

the calculations, maize *Helitrons* amplification dates were not significantly different from those in *Arabidopsis*, but were still significantly older than those in rice and sorghum (P <0.0001). T test approximations yielded the same conclusions. This analysis indicates that maize *Helitrons* are not enriched in gene-rich regions because of a shorter average duration in regions of the genome that show a high rate of DNA removal.

A second model proposes that the gene-rich regions of maize are structurally much more like the gene-poor regions of the sorghum, rice, *Arabidopsis* and nematode genomes than they are like the gene-rich regions of these small genomes. Because the gene-poor regions of these four small genomes are composed of large blocks of repetitive DNA intermixed with rare genes, they may have the same chromatin conformations as the gene-rich regions of the maize genome, which are composed of large blocks of repetitive DNA intermixed with genes. If chromatin composition determines the targeting of *Helitrons*, as it does for Ty elements in yeast (Brady et al. 2008), then it is likely that *Helitrons* in both maize and the four smaller genomes are all inserting in the same preferred chromatin types. In maize, we know that some families of LTR retrotransposons exhibit a bias toward insertion into heterochromatin near genes, while others exhibit a bias for insertion into heterochromatin that is not near genes (Liu et al. 2007; Baucom et al. 2009; Schnable et al. 2009). *Helitrons* in maize appear to share the behavior of those elements with the gene-associated heterochromatin bias.

Sequences encompassing 500 bp upstream and 500 bp downstream of each intact *Helitron* were retrieved and annotated to search for additional insertion site characteristics. Using the annotations provided for the B73 draft sequence (Schnable et al. 2009), it was observed that 6%

(109) of *Helitrons* are inserted into or near genes, 34% (658) are inserted into or near LTR retrotransposons, 23% (436) are inserted into or near non-*Helitron* DNA transposons, 26% (497) are inserted into (117 elements) or near (380 elements) *Helitrons* and the other 11% were on small fragments that did not allow definitive determination of the nature of the target site. Given that the *Helitrons* in B73 make up only ~2% of the genome, a random *Helitron* insertion model predicts only 64 intact *Helitrons* inserted into or within 500 bp range of another *Helitron,* not the 497 that were observed. This ~7.8-fold bias for accumulation near *Helitrons* is more dramatic than the approximately 2-fold bias for accumulation near other DNA transposons (that contribute ~10% of the B73 genome (Schnable et al. 2009)), but both are significant (P <0.0001).

One possible explanation for the clustering of *Helitrons* might be that they primarily transpose to nearby sites. However, it was observed that only 31 *Helitrons* were inserted into or near their most related *Helitron* in the B73 genome, while a random model predicts that there would have been 46 such cases. Hence, *Helitrons* in maize significantly (P=0.01) avoid inserting into or near their parental element, thus rejecting the short transposition distance model. For 117 *Helitrons* inserted into another *Helitron*, 84 were observed to be in the same orientation as the target *Helitron*, while 33 were in the opposite orientation. For 380 *Helitrons* inserted near another *Helitron*, 235 were in the same orientation, while 145 were in the opposite orientation (73 head to head and 72 tail to tail). If the insertion orientation had no bias, the ratio of insertion in the same direction and in the opposite direction would be 1:1. Chi-square tests gave P values of <0.0001 for non-randomness of both *Helitron*-internal and nearby insertions. None of these pairs of inserted *Helitrons* were present as more than one paired copy, so they are not co-

integrants, and none are present as the same element in tandem, as observed in vesper bats (Pritham and Feschotte 2007).

A total of 226 *Helitrons* were observed to be inserted into or near a *Helitron* of the same subfamily, while 229 are inserted into or near a *Helitron* of a different subfamily. If there was no bias for accumulation relative to the specific *Helitron* subfamily properties, one would expect 115 *Helitrons* inserted into or near a *Helitron* of the same subfamily. Hence, there is a significant (P <0.0001) bias for *Helitrons* to accumulate near *Helitrons* with similar terminal sequences, although they are usually different members of that subfamily.

Taken together, all of these characteristics of *Helitron* distribution indicate a strong bias for insertion into regions of the genome that contain *Helitrons* of the same family and subfamily. It is more difficult to explain these results as an outcome of biases in DNA removal, because there is no precedent for a DNA removal process that is somehow more active at removing DNA that is more weakly related (i.e., a more distant *Helitron* family) to a nearby *Helitron* than it is at removing DNA that is more strongly related (i.e., the same *Helitron* subfamily). A simpler model proposes that *Helitrons* in the genome exist in a unique chromatin state (perhaps with associated proteins involved in rolling circle amplification) that attracts other *Helitrons*, and attracts those most aggressively that have the same coevolved association between their structure and the particular rolling circle amplification enzymes encoded by that *Helitron* subfamily.

Materials and Methods

***Helitron* identification and abundance.** The structure-based approach to *Helitron* discovery that was employed has been described previously (Yang and Bennetzen 2009). Maize genomic sequences were downloaded from http://www2.genome.arizona.edu/genomes/maize (Schnable et al. 2009). After the structural search, a second search was employed on the residual sequence data to find any previously identified *Helitrons* that may have been missed. A BLAST search of all intact maize *Helitrons* was performed against a comprehensive maize TE database containing LTR retrotransposons and cut-and-paste DNA transposons (Schnable et al. 2009). All identified homologies were manually inspected. Non-*Helitron* TE fragments were removed from the *Helitron* database (replaced by "N"s). In order to find the genome contribution of all *Helitron* elements, both intact and fragmented, a BLAST search of the entire genome was performed against all intact elements with non-*Helitron* TE fragments removed. Homologies with at least 80% identity and at least 100 bp in contiguous length were counted. A BLASTX search for *Helitron*-specific Rep/helicase genes was also performed against the B73 draft sequence (Schnable et al. 2009) with a maximum Expect value of *e*-10. *Helitron*-specific Rep/helicases were retrieved from Repbase (Jurka et al. 2005). The results from both BLAST searches were combined to calculate genome contribution. The total number of 5′ and 3′ ends (30 bp with at least 80% identity to intact elements) was used to estimate the minimum total number of elements in the genome.

Helitron family, subfamily and exemplar assignment. Sequences with the most similar 3′ ends (30 bp with at least 80% identity) were classified as members of the same family and sequences with the most similar 5′ ends (30 bp with at least 80% identity) were classified as members of the same subfamily. A short word starting with "H" was assigned as the name for each *Helitron* family, and a number follows the family name to denote the subfamily. Exemplars were defined as those *Helitrons* with unique internal sequences (less than 80% identity to any other exemplar). These classifications, combined, yielded 498 exemplars in 62 subfamilies within 8 families.

Helitron divergence and estimation of Helitron amplification dates. An all-by-all BLAST search was performed with all intact B73 maize *Helitrons*. Each intact *Helitron* was aligned with its 2nd best hit by CLUSTALW and the corresponding divergence was calculated by BaseML module of PAML (v4.2) (Yang 2007). The amplification time was calculated from the formula $T = k/2r$ (k=divergence), using the substitution rate, r, of 1.3×10^{-8} per site per year for rice, sorghum and maize (Ma and Bennetzen 2004) and 1.05×10^{-8} per site per year for *Arabidopsis* (DeRose-Wilson and Gaut 2007). Non parametric Mann-Whitney tests were performed to see if maize *Helitrons* are significantly older than *Helitrons* in other species. The NPAR1WAY procedure of SAS (v 9.1) was used to calculate P values.

Helitron gene fragment acquisition. A BLASTX search of the NCBI non-redundant protein database (as of Oct 3 2008) was performed for each intact *Helitron*. Gene fragments were identified if a homology was detected with a maximum Expect value of *e*-10, or *e*-5 if the homology was from a species other than maize. TE-related proteins were excluded.

Gene fragment evolution. Forty four *Helitron* exemplars that had acquired gene fragment(s) were randomly chosen. PERL programs were used to facilitate the alignment of nucleotide sequences of gene fragments within the same exemplar type, with the initial alignment based on the predicted amino acid sequences. Alignments were manually inspected. Sequences following a stop codon were removed. Sequences <50 bp also were not evaluated. The codeml module of PAML (v4.2) (Yang 2007) was used to calculate dN/dS (ω) ratios to infer selection on gene fragments. Model M0 with ω fixed at 1 and model M0 with ω estimated from the data were built. Twice the differences of log likelihood of the above 2 models were calculated and chi-square tests with degree of freedom 1 were performed to assess whether ω was significantly different from 1. BLASTN searches against the maize gene set (Schnable et al. 2009) were performed to identify the host genes that the gene fragments in *Helitrons* had been acquired from. Model M1 (neutral) and model M2 (positive selection) were built. Twice the differences in the log likelihood of the 2 models above were calculated and chi-square tests with degree of freedom 2 were performed to assess whether significant adaptive selection was underway. All PAML analyses were run multiple times to verify convergence.

Helitron insertion preferences. Flanking sequences (50 bp both upstream and downstream) of all intact *Helitrons* insertion sites were used to calculate base composition. Base composition was calculated by PICTOGRAM. Chi-square tests of GC content for 50 bp upstream and downstream of *Helitron* insertion sites on each position compared to random AT flanking sites were performed. Flanking sequences (500 bp both upstream and downstream) of all intact *Helitrons* insertion sites were searched by BLAST against a comprehensive maize TE database (Schnable

et al. 2009) with an Expect value cutoff of *e*-10. The insertion sites for intact *Helitrons* were compared to maize gene annotation as well. A prediction of the number of intact *Helitrons* inserted into or within 500 bp of another *Helitron* for a random insertion process was calculated by employing the number of intact *Helitrons* [1,930] × (total length of *Helitrons* in maize genome [45.5 Mb] + (500+500) × total number of *Helitrons* [22,000]) / genome size [2,045 Mb]. The predicted number of insertions into or near the same subfamily by a random insertion process was calculated given the total number of insertions into or near *Helitrons* multiplied by the frequency of each subfamily. The predicted number of insertions into or near the most closely related *Helitron* if there was not bias for insertion was calculated using the total number of insertion into the same subfamily multiplied by the frequency of the most related elements. Chi-square tests were performed by comparing the observed and expected numbers with 1 degree of freedom to assess whether the insertions were significantly biased or not.

Acknowledgement

We thank R. Baucom, J. Estill, J. Leebens-Mack, H. Wang, and Q. Zhu for assistance with the APOLLO and PAML programs; three anonymous reviewers for their suggestions to improve this manuscript; D. Promislow for his advice on data analysis; and C. Du and H. Dooner for assistance with comparisons of our databases. This research was supported by NSF grant DBI-0607123.

Reference

Baucom, R.S., Estill, J.C., Chapparo, C., Deragon, J., Westerman, R.P., SanMiguel, P.J., and Bennetzen, J.L. 2009. Retroelement diversity, distribution and evolution in the B73 maize genome. *PLoS Genet* **5:** e1000732.

Bennetzen, J.L. 2005. Transposable elements, gene creation and genome rearrangement in flowering plants. *Curr Opin Genet Dev* **15:** 621-627.

Bennetzen, J.L. 2007. Patterns in grass genome evolution. *Curr Opin Plant Biol* **10:** 176-181.

Bennetzen, J.L., Ma, J., and Devos, K.M. 2005. Mechanisms of recent genome size variation in flowering plants. *Ann Bot* **95:** 127-132.

Brady, T.L., Schmidt, C.L., and Voytas, D.F. 2008. Targeting integration of the Saccharomyces Ty5 retrotransposon. *Methods Mol Biol* **435:** 153-163.

Brunner, S., Pea, G., and Rafalski, A. 2005. Origins, genetic organization and transcription of a family of non-autonomous helitron elements in maize. *Plant J* **43:** 799-810.

DeRose-Wilson, L.J. and Gaut, B.S. 2007. Transcription-related mutations and GC content drive variation in nucleotide substitution rates across the genomes of *Arabidopsis thaliana* and *Arabidopsis lyrata*. *BMC Evol Biol* **7:** 66.

Du, C., Fefelova, N., Caronna, J., He, L., and Dooner, H.K. 2009. The polychromatic *Helitron* landscape of the maize genome. *Proc Nat Acad Sci USA***:** In press.

Fu, H. and Dooner, H.K. 2002. Intraspecific violation of genetic colinearity and its implications in maize. *Proc Natl Acad Sci USA* **99:** 9573-9578.

Gilbert, W. 1987. The exon theory of genes. *Cold Spring Harb Symp Quant Biol* **52:** 901-905.

Gupta, S., Gallavotti, A., Stryker, G.A., Schmidt, R.J., and Lal, S.K. 2005. A novel class of *Helitron*-related transposable elements in maize contain portions of multiple pseudogenes. *Plant Mol Biol* **57:** 115-127.

Hanada, K., Vallejo, V., Nobuta, K., Slotkin, R.K., Lisch, D., Meyers, B.C., Shiu, S.H., and Jiang, N. 2009. The functional role of Pack-MULEs in rice inferred from purifying selection and expression profile. *Plant Cell* **21:** 25-38.

Hollister, J.D. and Gaut, B.S. 2007. Population and evolutionary dynamics of *Helitron* transposable elements in *Arabidopsis thaliana*. *Mol Biol Evol* **24:** 2515-2524.

Jameson, N., Georgelis, N., Fouladbash, E., Martens, S., Hannah, L.C., and Lal, S. 2008. *Helitron* mediated amplification of cytochrome P450 monooxygenase gene in maize. *Plant Mol Biol* **67:** 295-304.

Jiang, N., Bao, Z., Zhang, X., Eddy, S.R., and Wessler, S.R. 2004. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431:** 569-573.

Johnson, C., Bowman, L., Adai, A., Vance, V., and Sundaresan, V. 2007. CSRDB: a small RNA integrated database and browser resource for cereals. *Nucleic Acids Res* **35:** D829-D833.

Juretic, N., Hoen, D.R., Huynh, M.L., Harrison, P.M., and Bureau, T.E. 2005. The evolutionary fate of MULE-mediated duplications of host gene fragments in rice. *Genome Res* **15:** 1292-1297.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110:** 462-467.

Kapitonov, V.V. and Jurka, J. 2001. Rolling-circle transposons in eukaryotes. *Proc Natl Acad Sci USA* **98:** 8714-8719.

Kapitonov, V.V. and Jurka, J. 2007. *Helitrons* on a roll: eukaryotic rolling-circle transposons. *Trends Genet* **23:** 521-529.

Lai, J., Li, Y., Messing, J., and Dooner, H.K. 2005. Gene movement by *Helitron* transposons contributes to the haplotype variability of maize. *Proc Natl Acad Sci USA* **102:** 9068-9073.

Lal, S.K., Giroux, M.J., Brendel, V., Vallejos, C.E., and Hannah, L.C. 2003. The maize genome contains a *Helitron* insertion. *Plant Cell* **15:** 381-391.

Lerat, E., Rizzon, C., and Biemont, C. 2003. Sequence divergence within transposable element families in the *Drosophila melanogaster* genome. *Genome Res* **13:** 1889-1896.

Liu, R., Vitte, C., Ma, J., Mahama, A.A., Dhliwayo, T., Lee, M., and Bennetzen, J.L. 2007. A GeneTrek analysis of the maize genome. *Proc Natl Acad Sci USA* **104:** 11844-11849.

Ma, J. and Bennetzen, J.L. 2004. Rapid recent growth and divergence of rice nuclear genomes. *Proc Natl Acad Sci USA* **101:** 12404-12410.

Ma, J. and Bennetzen, J.L. 2006. Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. *Proc Natl Acad Sci USA* **103:** 383-388.

Ma, J., Devos, K.M., and Bennetzen, J.L. 2004. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res* **14:** 860-869.

Mendiola, M.V., Bernales, I., and de la Cruz, F. 1994. Differential roles of the transposon termini in IS*91* transposition. *Proc Natl Acad Sci USA* **91:** 1922-1926.

Mendiola, M.V. and de la Cruz, F. 1992. IS91 transposase is related to the rolling-circle-type replication proteins of the pUB110 family of plasmids. *Nucleic Acids Res* **20:** 3521.

Morgante, M., Brunner, S., Pea, G., Fengler, K., Zuccolo, A., and Rafalski, A. 2005. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat Genet* **37:** 997-1002.

Pritham, E.J. and Feschotte, C. 2007. Massive amplification of rolling-circle transposons in the lineage of the bat *Myotis lucifugus*. *Proc Natl Acad Sci USA* **104:** 1895-1900.

Rensing, S.A., Lang, D., Zimmer, A.D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T., Perroud, P.F., Lindquist, E.A., and Kamisugi, Y. 2008. The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science* **319:** 64-69.

SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., and Bennetzen, J.L. 1998. The paleontology of intergene retrotransposons of maize. *Nat Genet* **20:** 43-45.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., and Liang, C. 2009. The B73 maize genome: complexity, diversity and dynamics. *Science* **326:** 1112-1115.

Sweredoski, M., DeRose-Wilson, L., and Gaut, B.S. 2008. A comparative computational analysis of nonautonomous *Helitron* elements between maize and rice. *BMC Genomics* **9:** 467.

Swigonova, Z., Lai, J., Ma, J., Ramakrishna, W., Llaca, V., Bennetzen, J.L., and Messing, J. 2004. Close split of sorghum and maize genome progenitors. *Genome Res* **14:** 1916-1923.

Wang, Q. and Dooner, H.K. 2006. Remarkable variation in maize genome structure inferred from haplotype diversity at the *bz* locus. *Proc Natl Acad Sci USA* **103:** 17644-17649.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., and Panaud, O. 2007. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8:** 973-982.

Xu, J. and Messing, J. 2006. Maize haplotype with a *helitron*-amplified cytidine deaminase gene copy. *BMC Genet* **7:** 52.

Yang, L. and Bennetzen, J.L. 2009. Structure-based discovery and description of plant and animal *Helitrons*. *Proc Natl Acad Sci USA* **106:** 12832-12837.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24:** 1586-1591.

Table 3.1: Maize *Helitron* families

Family Name	Intact *Helitron* Copy #	Size Range (bp)	Putative Autonomous Elements	Gene Fragment(s)
Hip	1897	202-35925	+	+
Hide	15	2061-6564	-	+
Hill	5	330-356	-	-
Hink	4	2712-3705	-	+
Hint	3	1548-1559	-	+
Hire	2	2980-2990	-	+
Hit	2	962-975	-	-
Hitch	2	908-991	-	-
Total	1930	202-35925		

Table 3.2: Types of divergence observed in a random selection of 100 intact *Helitrons*

Single base changes	Indels (bp)						
	1	2-5	6-10	11-20	21-100	101-500	>500
6208	350	130	29	21	17	3	0

Figure Legends

Fig. 3.1: Distribution of *Helitron* copy numbers in different subfamilies. (A) X axis, *Helitron* copy number in each subfamily; Y axis, number of subfamilies with this element copy number. (B) X axis, *Helitron* copy numbers in each subfamily (2-5 copies shown); Y axis, log transformed number (base 10) of the number of subfamilies with this copy number. Linear regression lines are also shown and used to estimate the number of single copy subfamilies that have been missed by this analysis.

Fig. 3.2: Melting temperature distributions of predicted hairpins. Blue denotes predicted hairpins across the entire genome, and red denotes predicted hairpins in intact *Helitrons*. The X axis indicates predicted melting temperature while the Y axis shows the frequency of predicted hairpins with that predicted Tm. Predicted melting temperatures of hairpins were calculated by the melt program in the UNAFold 3.3 software package. Parameters were set as follows: DNA molecule, sodium concentration 1, magnesium concentration 0.

Fig. 3.3: (A) *Helitron* amplification dates for maize, rice, sorghum and *Arabidopsis*. Solid circles denote the mean amplification date and solid triangles denote the median amplification date. (B) *Helitron* amplification dates for maize, rice, sorghum and *Arabidopsis*. Only transversions were counted. Solid circles denote the mean amplification date and solid triangles denote the median amplification date.

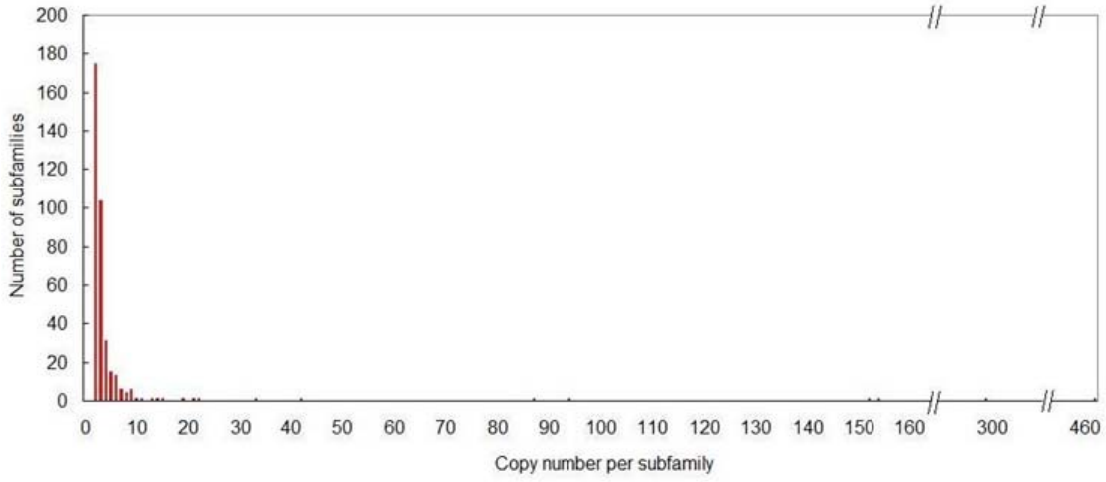
Fig. 3.4: Examples of maize *Helitrons* (*Hip1_120* and *Hide1_2*) that have captured gene fragments. Each colored box denotes an exon of an acquired gene fragment. Exons that come from the same gene are shown in the same color.

Fig. 3.5: Distribution of the number of gene fragments captured per *Helitron* exemplar. X axis indicates the number of gene fragments captured per exemplar. Y axis indicates the number of exemplar types that have acquired this many gene fragments. The dark boxes indicate actual *Helitron* exemplar numbers while the dotted boxes indicate the predicted exemplar numbers in each class if gene fragment acquisitions were random and independent.

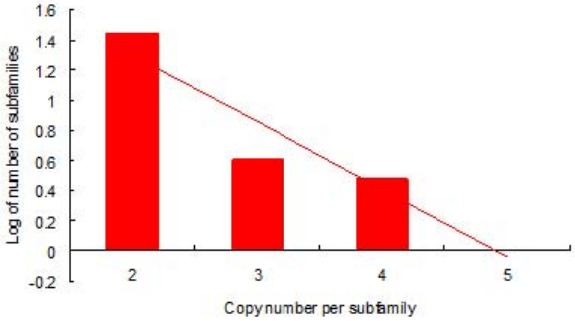
Fig. 3.6: The ages of *Helitrons* compared to the orientations of acquired gene fragments. X axis indicates the amplification dates. Y axis indicates three categories: gene fragment orientation in the sense transcriptional orientation (S) as the *Helitron* genes, gene fragments orientation opposite/antisense to that of the *Helitron* (A), and at least 2 different gene fragments acquired by the same *Helitron*, with at least one in each possible orientation (S&A).

Fig. 3.7: (A) Insertion site base composition. Left side of *Helitrons*, 20 bp upstream of *Helitron* insertion sites. Right side of *Helitrons*, 20 bp downstream of *Helitron* insertion sites. (B) Chi-square tests of GC content for 50 bp upstream and downstream of *Helitron* insertion sites on each position compared to random AT flanking sites. The *P* value is indicated by color, with black representing the most significant and yellow the least significant associations. (C) Distribution of *Helitrons* along the chromosomes. The X axis shows the chromosome locations and the Y axis indicates the copy numbers of *Helitrons* (both intact and truncated elements) per 500 kb. Narrow lines on the X axis indicate the positions of centromeres.

Figure 3.1:


Number of subfamilies
200
180
160
140
120
100
80
60
40
20
0
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 300 460
Copy number per subfamily

(A)


Log of number of subfamilies
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
2 3 4 5
Copy number per subfamily

(B)

Fig. 3.2:

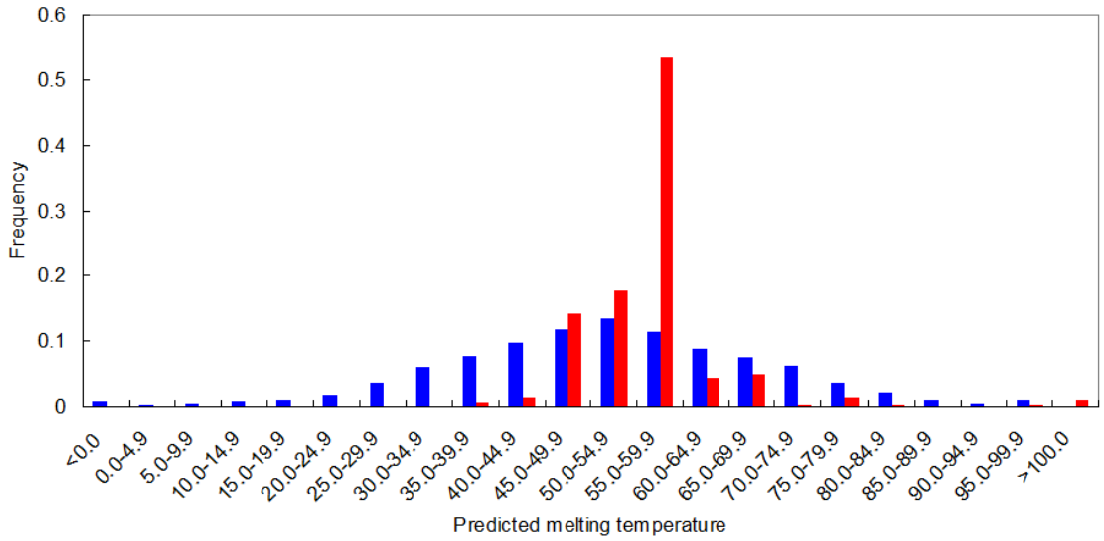
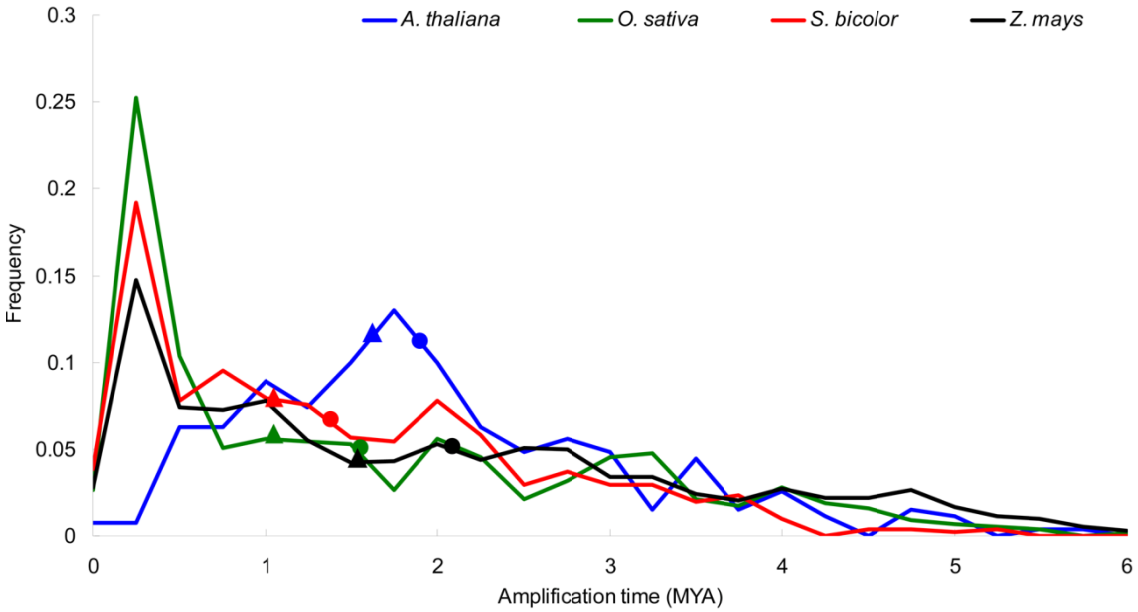
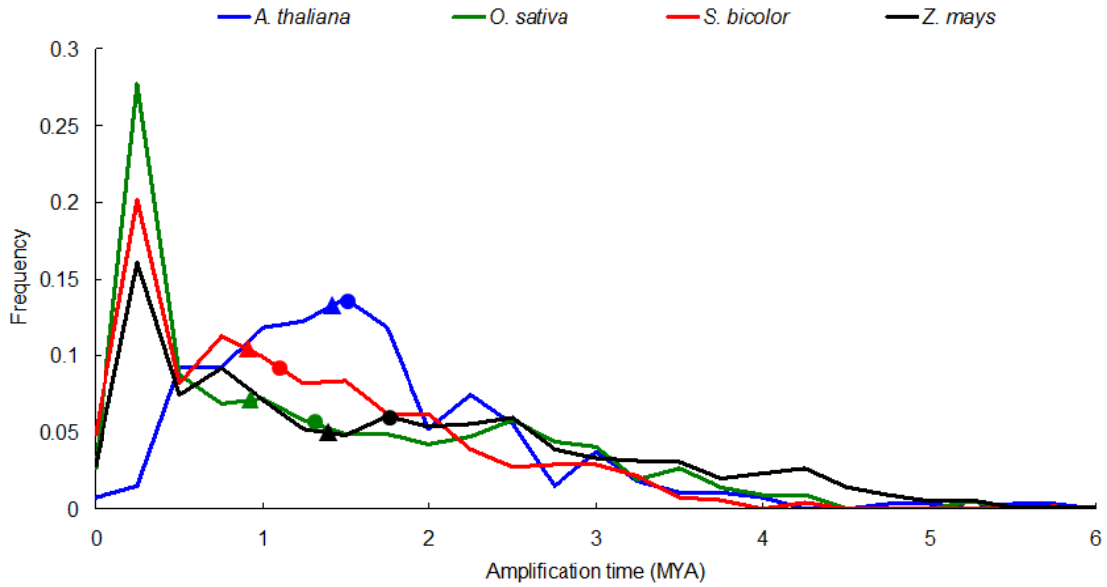

0.6
0.5
0.4
0.3
0.2
0.1
0
Frequency
<0.0
0.0-4.9
5.0-9.9
10.0-14.9
15.0-19.9
20.0-24.9
25.0-29.9
30.0-34.9
35.0-39.9
40.0-44.9
45.0-49.9
50.0-54.9
55.0-59.9
60.0-64.9
65.0-69.9
70.0-74.9
75.0-79.9
80.0-84.9
85.0-89.9
90.0-94.9
95.0-99.9
>100.0
Predicted melting temperature

Fig. 3.3:


A. thaliana
O. sativa
S. bicolor
Z. mays
Frequency
0.3
0.25
0.2
0.15
0.1
0.05
0
0
1
2
3
4
5
6
Amplification time (MYA)

(A)


A. thaliana
O. sativa
S. bicolor
Z. mays
Frequency
0.3
0.25
0.2
0.15
0.1
0.05
0
0
1
2
3
4
5
6
Amplification time (MYA)

(B)

Fig. 3.4:

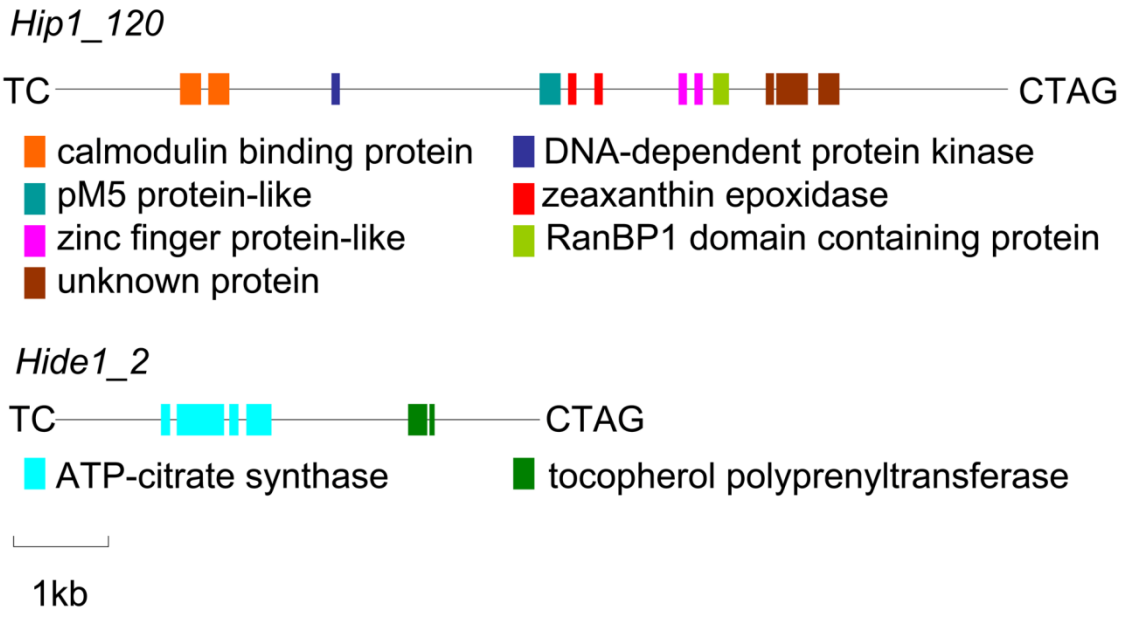

Hip1_120
TC
CTAG
calmodulin binding protein
DNA-dependent protein kinase
pM5 protein-like
zeaxanthin epoxidase
zinc finger protein-like
RanBP1 domain containing protein
unknown protein
Hide1_2
TC
CTAG
ATP-citrate synthase
tocopherol polyprenyltransferase
1kb

Fig. 3.5:

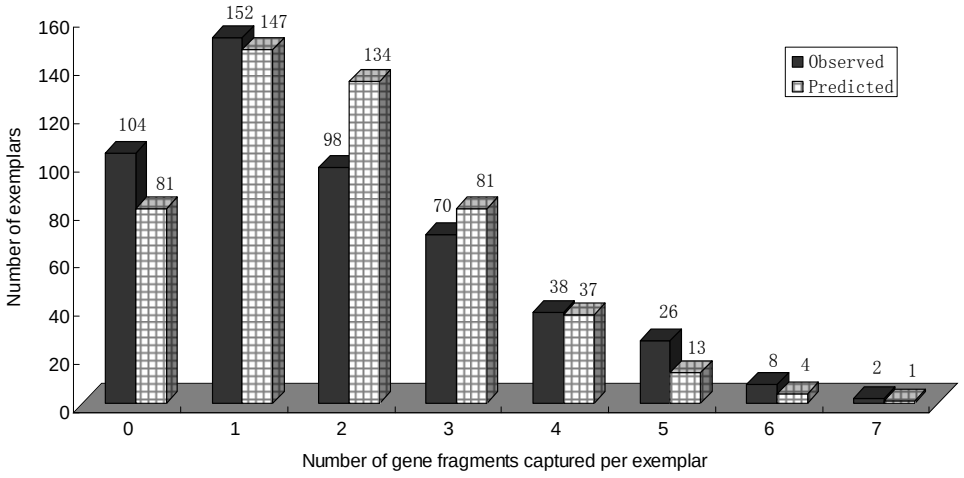

Observed
Predicted
Number of exemplars
Number of gene fragments captured per exemplar
0
20
40
60
80
100
120
140
160
0
1
2
3
4
5
6
7
104
81
152
147
98
134
70
81
38
37
26
13
8
4
2
1

Fig. 3.6

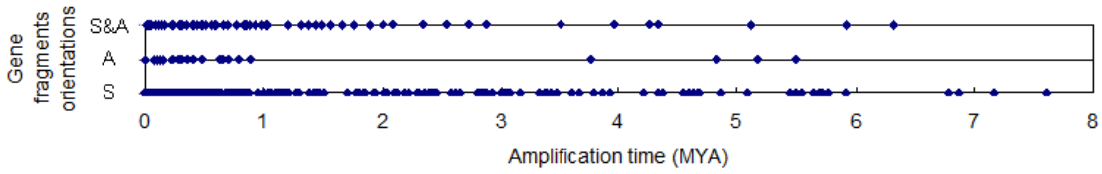
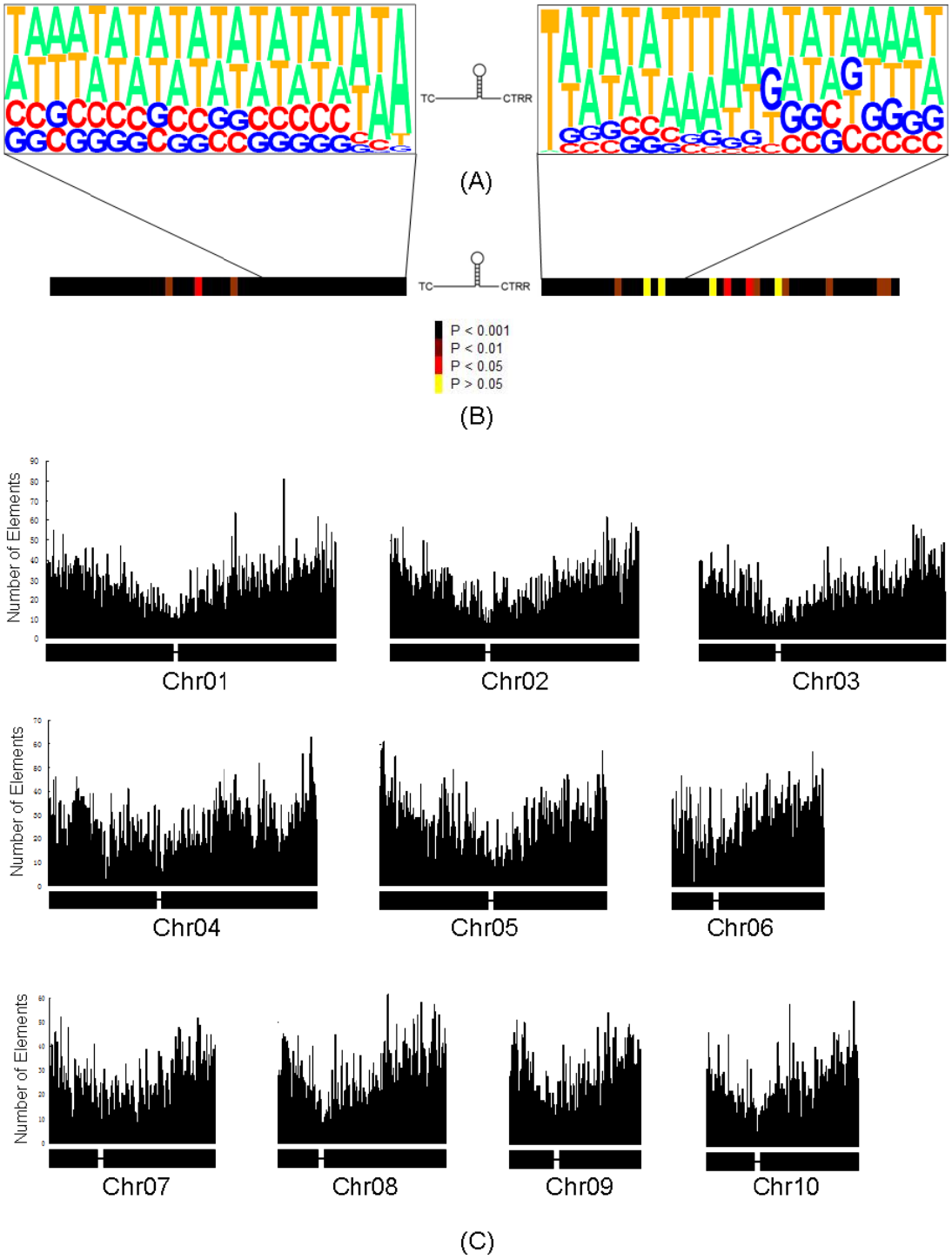

Gene fragments orientations
S&A
A
S
0
1
2
3
4
5
6
7
8
Amplification time (MYA)

Fig. 3.7:


TC
CTRR
(A)
TC
CTRR
P < 0.001
P < 0.01
P < 0.05
P > 0.05
(B)
Number of Elements
Chr01
Chr02
Chr03
Number of Elements
Chr04
Chr05
Chr06
Number of Elements
Chr07
Chr08
Chr09
Chr10
(C)

CHAPTER 4

CONCLUSIONS

TEs are defined by their mobility and ability to increase their copy number in a genome. Because of their abundance and various effects on host genomes, it is important that we understand the diversity of TEs and the dynamics of TE activity.

The results of our studies show that *Helitrons* contribute 1-3% of the *Arabidopsis*, *Medicago*, rice, sorghum and maize genomes (see Chapters 2 and 3). Lacking large structural features, *Helitrons* are difficult to identify in a genome. We developed a structure-based search algorithm that also takes advantage of the repetitive nature of most *Helitron*s to confirm *Helitron* identity (see Chapter 2). The program "HelSearch" can identify intact *Helitrons* with as few as 2 copies present in a given genome without any previous knowledge about the genome except the genomic sequences. We tested HelSearch on a few genomes (*Arabidopsis*, *C. elegans*, *Medicago*, rice and sorghum). Its overall sensitivity and specificity were found to be superior to other methods previously used in identifying *Helitrons*. *Helitrons* were identified in all of the flowering plant genomes that we investigated. We then performed a comprehensive search for *Helitrons* in the maize genome with a combination of our structure-based approach and a

homology-based search (see Chapter 3). Tens of thousands of elements were identified. We found that *Hip* is the most abundant *Helitron* family in the flowering plants investigated, and is especially abundant in maize.

HelSearch can only identify *Helitron* elements with at least 2 copies in the genome because multiple sequence alignments with more than one sequence are needed to identify the boundaries of elements. Single copy elements will not be found. It needs to be noted that those elements are often missed by other methods, although they can be detected if they retain similarity to at least one *Helitron*-specific gene like Rep/helicase or by detecting vacant sites when comparing different genomes. We were able to estimate the number of single copy elements that we failed to identify in a genome (see Chapters 2 and 3). Single copy or low-copy-number elements are predicted to represent the majority of TE diversity in a genome, indicating that very few families/subfamilies are amplified into high copy numbers, as with *Hip* in maize. This characteristics is also observed for LTR retrotransposons in maize (Baucom et al. 2009a; Baucom et al. 2009b). Also, since multiple sequence alignments are needed to identify the boundaries of elements, old elements that have accumulated too many mutations (especially large indels) are also difficult to identify by HelSearch. As a result, some large elements seen by others were not identified by HelSearch in some genomes (see Chapter 2).

The comprehensive element discoveries in the investigated genomes made it possible to assess additional structural features of *Helitrons* (see Chapters 2 and 3). We found that the predicted hairpins just upstream of *Helitron* 3′ ends would, if they form *in vivo*, have much higher melting temperatures than similar-size predicted hairpins in every one of these genomes.

It is believed that these 3′ hairpins are involved in *Helitron* transposition, perhaps by serving as termination signals for rolling circle replication (Kapitonov and Jurka 2001).

We found that *Helitrons* preferentially insert into AT-rich regions, with an apparent orientation bias for a higher AT content in the sequences 3′ to the insertion. *Helitrons* were found to accumulate preferentially in gene poor regions of rice, *Arabidopsis* and *C. elegans*, but maize *Helitrons* were found to be enriched in the euchromatic regions on the chromosome arms relative to pericentromeric heterochromatin. We did consider the possibility that the bias we observed is due to different rates of DNA removal in these species rather than different *Helitron* insertion biases in maize compared to the other species. We determined the approximate transposition dates of maize *Helitrons* and found that maize *Helitron* insertions are significantly older than rice and sorghum *Helitron* insertions. Because processes like unequal homologous recombination are known to preferentially remove TEs and other DNAs from relatively gene-rich regions, a greater rate of *Helitron* removal from gene-rich regions of rice and sorghum could be responsible for the differences observed when comparing maize (with older *Helitrons*, on average, and thus a possibly slower rate of element removal) *Helitron* distributions with those seen in sorghum or rice.

It is also possible that the chromatin structure of maize chromosome arms is more similar to the pericentromeric regions in other plant species than to their gene-rich regions because maize has a similar gene density in its gene-rich regions to that observed in the gene-poor regions of rice and sorghum (Ma and Bennetzen 2006; Paterson et al. 2009). Thus, *Helitrons* may have an insertion bias to insert into heterochromatin that is near genes. Work on LTR

retrotransposons has shown that some of these elements preferentially insert into heterochromatin that is near genes, while other families insert preferentially into heterochromatin that is far from genes (Liu et al. 2007).

We investigated maize *Helitron* insertion sites with respect to the nature of nearby sequences (see Chapter 3). We discovered that maize *Helitrons* preferentially insert near DNA transposons, especially near other *Helitrons*. We also found that they preferentially insert near elements of the same family and same subfamily but not near the closest-related elements. Hence, *Helitrons* do not transpose locally. Furthermore, we found that the *Helitrons* that insert inside another *Helitron* are usually in the same orientation as the one they inserted into and that the *Helitrons* that insert near another *Helitron* are usually in the same orientation as well. This clustering suggests that they have a strong insertion bias for some unknown chromatin structure that is associated with *Helitrons*.

Automated gene annotation has always been partially inaccurate. Gene number of the rice genome, *O. sativa*, dropped from 61,668 in initial annotations (Goff et al. 2002) to 30,192 in the most recent annotation (Tanaka et al. 2008). The annotation of randomly-selected maize BACs showed that at least 13% of the identified maize genes are severely truncated gene fragments (i.e. missing >30% of the coding regions) (Liu et al. 2007). Pseudogenization and TE acquisition are found to be two major phenomena that contribute to such gene annotation inaccuracy. After gene duplication, one copy of the duplicated gene may accumulate mutations, and become a pseudogene. Pseudogenized gene fragments often can be resolved by comparing orthologous or homoeologous regions from different species (Ilic et al. 2003). Recent studies have shown that

Pack-MULEs and *Helitrons* frequently acquire gene fragments from host genomes (Jiang et al. 2004; Morgante et al. 2005). Hence, we performed a detailed analysis of gene fragments captured by *Helitrons* in the maize genome to assist maize genome annotation (Schnable et al. 2009) and to have the opportunity to assess the evolutionary impact of *Helitrons* with acquired gene fragments.

After the identification of 1930 intact *Helitrons* and numerous fragments of *Helitrons* in the maize genome, we found that >60% of the intact maize *Helitrons* have acquired one or more gene fragments (see Chapter 3). We then were able to identify over 4000 *Helitron*-related genes in the maize predicted gene set (Schnable et al. 2009). We also observed that most *Helitron*-acquired gene fragments are undergoing random drift, but 4% were under negative selection while another 4% were under positive selection. Of the 6 acquired gene fragments shown to be under significant negative or positive selection, there are 3 that contain more than one exon. For 1 of these cases (a cyclin type B-like gene that is under positive [diversifying] selection), only the exons are conserved, but the single intron is not. For this gene, a 100% EST match from inbred B73 to the acquired gene fragment was also found. The other 2 acquired gene fragments, with comparable levels of sequence change in both introns and exons, may be selected at the RNA level. These results provide support for the possibility of selection acting on expressed gene products. In at least some cases, gene fragment acquisition appears to positively influence element survival and/or the ability of the *Helitron* to acquire additional gene fragments. Taken together, this evidence suggests that *Helitron*-acquired gene fragments sometimes have beneficial effects for the host and/or for *Helitron* transposition/retention. The most abundant *Helitron*

subfamily in maize has acquired a phosphatase 2C-like gene fragment and amplified it to >700 copies in the genome. The phosphatase 2C-like gene fragment displays strong purifying selection. A sorghum *Helitron* also acquired a phosphatase 2C-like gene fragment, but from a different member of that gene family. It is possible that the survival of these independent acquisitions in two different plant species are purely by chance, but it is also possible that phosphatase 2C-like gene acquisition is beneficial to the host and/or to the *Helitrons*.

Now that gene fragments with possible beneficial functions have been identified, one could attempt to test their function experimentally. If they do possess any function, we might be able to observe a phenotype by mutating the gene fragment and then growing the plant under various conditions. The easiest way to knock out gene function would be to transform an RNAi construct in order to suppress the expression of transcripts, but this would need to be done carefully (e.g., at the junction between the acquired gene fragment and the element-specific transcript) to avoid knocking out the endogenous gene from which the fragment was acquired. The gene fragments that we identified as under negative or positive selection were detected at the amino acid level. Thus, the 4% under negative selection and another 4% under positive selection may be under-estimations because there might be gene fragments that act at the RNA level instead of the amino acid level. Once the gene fragment is transcribed, the RNA product may enter the RNAi pathway to regulate the host gene. We show in Chapter 3 that ~90% of maize *Helitrons* have at least one small RNA match. There are other ways to provide hints as to the functions of acquired gene fragments. For instance, transforming a stronger promoter to create over expression of gene

fragments and looking for a phenotype or transforming the *Helitron* with the captured fragments into a different species to observe any phenotypic effects.

An active *Helitron* element has not been identified yet. Thus, the mechanisms of transposition and gene fragment acquisition can only be hypothesized from *Helitron* structural features, encoded proteins and other evidences found at the nucleotide level (Kapitonov and Jurka 2001; Morgante 2006; Sweredoski et al. 2008). In Chapter 2, we proposed a possible mechanism of *Helitrons* acquiring new sequences by bypassing current ends to generate novel 5′ or 3′ ends. In order to shed new light on the molecular basis of *Helitron* action, efforts were made to search for active elements (see Appendix). We identified a few candidates that have estimated transposition within the last 5,000 years (100% identical in sequences), suggesting recent activity in rice and maize. It was shown that, in our collections of rice and maize siblings, tissue culture-derived lines and recombinant inbred lines, *Helitron* activity is very low. We failed to confirm any active *Helitrons* in the above materials. Based on our transposon display results, *Helitron* activity in rice and maize siblings and tissue culture derived plants is lower than 10^{-2}. Exploring more sources of tissue culture cells may help identify active elements since tissue culture is known to be able to activate TEs (Peschke et al. 1987). Genomic DNA can also be pooled (e.g. DNA from 30 tissue culture cell lines be combined into one pool) in a transposon display analysis in order to screen many lines efficiently. It may also be possible to modify the *Helitron* elements molecularly to create activity. The expression of the key protein Rep/helicase might not be strong enough to induce transposition. A stronger promoter can be transformed into maize or rice to induce stronger expression. Furthermore, the *Helitron* ends might have accumulated

mutations that prevent them from transposing efficiently. A consensus sequence, which could provide the ancestor status of the subfamily, might be built. The ends of *Helitron* can be modified based on the consensus sequence to improve the activity, as was done for *Sleeping Beauty* (Ivics et al. 1997).

In the research described in this dissertation, I developed a novel tool to identify *Helitrons* based on their structure. This tool has offered an opportunity to search for *Helitrons* in any genome. Among the advantages of HelSearch is that it does not require any previous knowledge of a genome. Thus, it can be used in a broad range of species (plants, animals, fungi). Such a general approach (aligning and detecting the boundaries of repeats) also has the potential to be utilized in searching for other types of TEs. Our studies of *Helitrons* in a few flowering plant genomes have provided new insights into the actions and evolution of these elements. However, this is only the beginning of attempts to understand the evolutionary roles of *Helitrons*. Further analysis is necessary to provide a clearer view of how *Helitrons* evolve and how they affect host genomes. I see three areas of special interest: (1) To explore more genomes for the existence and characteristics of *Helitrons* to better understand genome evolution; (2) To explore further for active elements or molecularly modify elements to create activity; (3) To identify candidates of functional gene fragments acquired by *Helitrons* and verify their functions experimentally.

References

Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, Deragon J-M, Westerman RP, SanMiguel PJ, Bennetzen JL. 2009a. Exceptional Diversity, Non-Random Distribution, and Rapid Evolution of Retroelements in the B73 Maize Genome. *PLoS Genet* **5**(11): e1000732.

Baucom RS, Estill JC, Leebens-Mack J, Bennetzen JL. 2009b. Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome. *Genome Res* **19**(2): 243-254.

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**: 92-100.

Ilic K, SanMiguel PJ, Bennetzen JL. 2003. A complex history of rearrangement in an orthologous region of the maize, sorghum, and rice genomes. *Proceedings of the National Academy of Sciences* **100**: 12265-12270.

Ivics Z, Hackett PB, Plasterk RH, Izsvak Z. 1997. Molecular reconstruction of *Sleeping Beauty*, a *Tc1*-like transposon from fish, and its transposition in human cells. *Cell* **91**: 501-510.

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. 2004. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431**: 569-573.

Kapitonov VV, Jurka J. 2001. Rolling-circle transposons in eukaryotes. *Proc Natl Acad Sci USA* **98**: 8714-8719.

Liu R, Vitte C, Ma J, Mahama AA, Dhliwayo T, Lee M, Bennetzen JL. 2007. A GeneTrek analysis of the maize genome. *Proc Natl Acad Sci USA* **104**: 11844-11849.

Ma J, Bennetzen JL. 2006. Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. *Proc Natl Acad Sci USA* **103**: 383-388.

Morgante M. 2006. Plant genome organisation and diversity: the year of the junk! *Curr Opin Biotechnol* **17**: 168-173.

Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A. 2005. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat Genet* **37**: 997-1002.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A. 2009. The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**: 551-556.

Peschke V, Phillips R, Gengenbach B. 1987. Discovery of transposable element activity among progeny of tissue culture--derived maize plants. *Science* **238**: 804-907.

Schnable PS Ware D Fulton RS Stein JC Wei F Pasternak S Liang C Zhang J Fulton L Graves TA et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**: 1112-1115.

Sweredoski M, DeRose-Wilson L, Gaut BS. 2008. A comparative computational analysis of nonautonomous *Helitron* elements between maize and rice. *BMC Genomics* **9**: 467.

Tanaka T, Antonio BA, Kikuchi S, Matsumoto T, Nagamura Y, Numa H, Sakai H, Wu J, Itoh T, Sasaki T. 2008. The rice annotation project database (RAP-DB): 2008 update. *Nucleic Acids Res* **36**: D1028-D1033.

Appendix

HELITRON ACTIVITY TESTS

Introduction

Active TEs are essential for the study of transposition mechanisms and are the best tool for investigating insertion preferences and can be used for transposon mutagenesis. There are a number of TEs that have been found to be active, including *Ac* in many maize lines (McClintock 1946; McClintock 1951; Fedoroff et al. 1983) and tissue culture cells (Peschke et al. 1987) and *En*/*Spm* of maize (Peterson 1953; McClintock 1954; Peterson 1965; Peschke and Phillips 1991). Two possible independent origins of *Mu* transposon activity, one called *Mutator* and one dubbed *Cy*, have also been identified in maize (Robertson 1978; Schnable and Peterson 1986). DNA transposons called *P* elements were found to be active in *Drosophila* (Rubin et al. 1982), while DNA transposon *Tc1* is also active in *C. elegans* (Emmons and Yesner 1984). Human retrotransposon *L1* has also been observed to be active (Dombroski et al. 1991).

TEs are usually silenced by DNA methylation associated with an RNAi mechanism (Yoder et al. 1997; Martienssen 1998; Lippman and Martienssen 2004). Otherwise, a genome might become very unstable due to TE activity. Certain changes to cells may break this silencing

process and thereby activate transposons. Tissue culture-derived plants often display some abnormalities (Peschke et al. 1987), including chromosomal breakage and DNA methylation changes, which could lead to the activation of transposons. Active retroelements have been observed in plants only in a few cases, despite their very great contributions to genome composition. In many cases, retrotransposons are active in tissue culture cells. For instance, tobacco retrotransposon *Tto1* has been found to be active in tobacco tissue culture (Hirochika 1993) and was also active in transgenic rice tissue culture (Hirochika et al. 1996a). Retrotransposon *Tnt1* was found to be active in tobacco tissue culture (Lucas et al. 1995) and when transformed into *Arabidopsis* tissue culture cells (Lucas et al. 1995). Rice retrotransposon *Tos17* is active in rice tissue culture (Hirochika et al. 1996b). Silenced retrotransposons can be reactivated in the *ddm1* mutant, which alters chromatin structure and global DNA methylation status (Hirochika et al. 2000). Silenced DNA transposons can also be reactivated in *ddm1* mutant background (Miura et al. 2001). It also has been reported that some DNA transposons and retrotransposons can be reactivated in pollen (Slotkin et al. 2009) by changes in the DNA methylation state. Plants derived from a specific tissue culture line have been shown to have activated MITEs from the rice *mPing* family, and this has been very useful for subsequent characterizations (Jiang et al. 2003; Yang et al. 2009).

There are two major reasons that active TEs are difficult to find. First, probably because TE insertions are often deleterious, TE activity is usually suppressed in an organism, and thus difficult to detect. *P* elements in *Drosophila* have an insertional mutagenesis activity at a 10^{-5} level (Rubin et al. 1982). By comparison, the insertion mutagenesis activity of *Mutator* elements

in maize is between 10^{-3} and 10^{-5} (Bennetzen 1996), while for *Ac*/*Ds* elements in maize it is 10^{-6} (Walbot 1992).

As stated above, TEs can be activated in tissue culture cells. For example, 10 regenerated maize plants out of 301 (3%) contained active *Ac* transposon (Peschke et al. 1987). *Spm* activity was reactivated at a 1% level in lines derived from maize tissue culture cells (Peschke and Phillips 1991). In regenerated rice plants, the *mPing* MITE exhibited an activity level as high as 40 new insertions per plant per generation (Naito et al. 2006). Certain TE families might be too old, and thus have accumulated too many mutations in necessary ORFs or the essential binding sites, thereby having lost the ability to transpose any more. TE activity requires functional proteins (transposase, reverse transcriptase, etc.) being produced. These proteins can bind to the corresponding TE to initiate and progress the transposition event. Thus, changes in TE-encoded proteins and/or changes in essential binding sites would result in lost activity.

Because two mutagenic maize *Helitrons* have been identified as presumably recent events (Lal et al. 2003; Gupta et al. 2005), it seems likely that *Helitrons* in maize have been active in the last few decades. Hence, this seems to be a good organism to search for currently active elements. Here, we report a screen for *Helitron* activity in rice and maize by the transposon display assay.

Materials and Methods

See Table A.1 for a description of the rice lines that have been investigated and Table A.2 for the maize lines that have been investigated. Note that the DNA from the maize tissue culture

lines (HiII 0, HiII 1, HiII 2, HiII 3, 2-64, XL30, XL32) were isolated from the plant produced by crossing to inbred B73 after the other parent (HiII) was selected from callus after tissue culture.

DNA restriction digestions and adaptor ligations

Each DNA sample (300-500ng) was completely digested by *Mse*I enzyme and ligated in 40 µl with standard *Mse*I AFLP adapters, 2 units of *Mse*I, 2 units of T4 DNA ligase, 5 µg BSA and manufacture-supplied buffer at 37°C for 12 hours. Aliquots of the restriction/ligation reactions were visualized on 1% agarose gels stained with ethidium bromide, to check the quality of DNA restriction/ligation, and the remaining reaction mix was diluted 4-fold with ddH_2O.

Pre-selective amplification

PCR amplifications were performed by using a primer complementary to the *Mse*I adapter sequence along with another primer specific to the termini of *Helitrons*. Reactions were performed in 20 µl containing 1 µl of the diluted restriction/ligation reactions, 10 pmol of each primer, 1× Roche PCR buffer, 0.2 mM dNTPs, and 1 unit Roche DNA polymerase. These and subsequent reactions were carried out with a Robocycler Gradient Temperature Cycler (Stratagene). The temperature cycling parameters were as follows: 72°C/2 min; 94°C/3 min; 24 cycles of 94°C/45 sec, 59°C/45 sec, and 72°C/45 sec, and a final cycle of 72°C/5 min. After visualizing aliquots of each PCR on 1% agarose gels stained with ethidium bromide, the remaining volumes were diluted 10-fold with ddH_2O.

Selective amplification

The second round (selective) PCR amplification used the *Mse*I adapter primers and a radio labeled (^{33}P) primer specific to the termini of *Helitrons* (downstream of the pre-selective

primers). Reactions were performed in 10 µl containing 1.5 µl of the diluted pre-selective amplification reactions, 2.5 pmol of each primer, 1× Roche PCR buffer, 0.2 mM dNTPs, and 0.4 unit Roche DNA polymerase. Temperature cycling used a ‘‘touchdown’’ protocol: 94°C/3 min, followed by 94°C/45 sec, 66°C/45 sec, and 72°C/45 sec. In subsequent cycles, the annealing temperature was reduced from 65°C to 59°C in 1°C increments each cycle. Twenty-seven cycles were performed at the 58°C annealing temperature, followed by a final cycle of 72°C for 5 min.

Gel electrophoresis

For detection of radioactive amplification products, 10 µl of loading/denaturing buffer (98% deionized formamide / 10 mM EDTA, pH 8.0 / 0.025% xylene cyanol / 0.025% bromophenol blue) was added to the PCR reactions. Samples were denatured at 95°C for 5 min and placed on ice, and 4 µl of the mixture was immediately loaded on 6% denaturing (7.5 M urea) acrylamide-bisacrylamide (19:1) gels in 1× TBE buffer (89 mM Tris / 89 mM borate / 2 mM EDTA / pH 8.0). After samples were electrophoresed (35 mA constant) for 2.5 h, the gel was transferred to filter paper, dried, and exposed to an x-ray film for 48 h.

See Table A.3 for the primers used for rice *Helitrons*, the *Mse*I adapter and the *Mse*I-specific primer. See Table A.4 for the primers used for maize *Helitrons*.

Results and Discussion

Transposon display is a modified AFLP technique (Casa et al. 2000) that has been extensively used to detect active elements. AFLP technology has the advantage that the number of mobile elements investigated in an experiment can be determined by the investigator through

choice of the specificity of the primers employed. Visualizing dozens of bands on a single gel is considered to be optimal because smaller numbers would require more labor to run more amplifications and gels, while hundreds of bands on a single gel would lead to extensive band overlap that could hide new bands.

Some of the rice *Helitrons* identified are 100% identical (*Hup_OS1*), with a time since transposition estimated to be in less than 5,000 years ago in Chapter 3, and are present only in *japonica* subspecies but not in *indica.* These characteristics all suggest recent activity. Their activity can be further investigated by analysis of the transposon display banding patterns. We first performed transposon display on *mPing* as a positive control. The MITE *mPing* is known to be active in some rice tissue culture cells and in the line Gimbozu EG4 (Jiang et al. 2003; Kikuchi et al. 2003; Nakazaki et al. 2003). Nipponbare has about 50 copies of *mPing* and Gimbozu EG4 has over 1000 copies. We used primers specific to *mPing* to test the banding patterns. The results were consistent with previous studies (Figure A.1A). We used primers specific to rice *Helitron* candidates *Hup_OS1_27* (P1 and P5) and *Hup_OS1_7* (P2 and P9). Both of them have 10-20 copies in the *japonica* genome. We tested *Helitron* activity in Gimbozu EG4 and Nipponbare siblings (Figure A.1B). Banding patterns for *Hup_OS1_7* are shown. No newly transposed element was identified. We also tested *Helitron* activity in rice tissue culture cells (Jiang et al. 2003) (Figure A.1C). Banding patterns for *Hup_OS1_27* are shown. No newly transposed element was identified. The *Helitron* copy numbers in these lines are fairly constant, as indicated by the banding patterns.

Some maize *Helitrons* have previously been predicted to be active within the last 100 years (Lal et al. 2003; Gupta et al. 2005). We identified a few active-element candidates. They have 100% identical copies in the genome and have relatively high copy numbers (>100). Having a high copy number in the genome suggests that they were quite active at some time in the recent past, because older elements would have been removed by illegtimate recombination (Ma et al. 2004). Hence, they are better candidates than low-copy-number elements. We designed 6 primer sets specific to different elements (Table A.4).

For PM11 and PM12, we had no success with PCR. Here, we show the results of *Hip1_14* (Figure A.2A) which has 59 copies in the sequenced B73 genome (Schnable et al. 2009). The banding pattern from B73 confirmed this approximate copy number. The 8 IBM lines showed bands from either B73 or Mo17. The 8 siblings from AA8xB73 showed bands from either AA8 or B73. And no new bands were found in any tissue culture individuals (25-30, 2-64) other than the bands found in HiII and B73 parents. These results all suggest that *Helitron Hip1_14* is not active in the tested lines and tissue cultures.

We also show the results of *Hip1_5* (Figure A.2B), which has >100 copies in the sequenced B73 genome (Schnable et al. 2009). The banding pattern from B73 confirmed this approximate copy number. The 8 IBM lines showed bands from either B73 or Mo17. The 8 siblings from A112xA113 showed bands from either A112 or A113 and the 8 siblings from AA8xB73 showed bands from either AA8 or B73. No new bands were found in any plants derived from separate tissue culture cells (HiII 0, HiII 1, HiII 2, HiII 3, 2-64, XL30, XL32) compared to the HiII and B73 parents. These all suggested that *Helitron Hip1_5* is not active in

any of the tested lines, including the tissue culture-derived lines. Activity was not identified from other maize *Helitron* families listed in Table A.4 (*Hip1-1*, *Hip1-3*, *Hip2-1* and *Hip2-2*) as well (data not shown).

In summary, maize *Helitron* activity is very low (zero or close to zero) in all recombinant inbred lines, tissue culture-derived lines and sibling lines tested in this study. The recombinant inbred line we tested was selfed for 8 generations. The TE activity has to be at least 10^{-2} (i.e., one new insertion per 100 plants per generation) to have a good chance to be observed in 10 plants. We also tested about 10 independent plants derived from tissue culture, including siblings from plants derived from tissue culture. The TE activity also has to be at a 10^{-1} to 10^{-2} level to be detected in these experiments. It is possible that *Helitrons* are active at a level lower than 10^{-2}, and thus were not detected in our assay.

Previous studies indicate that *Helitrons* have been active in maize in the recent 100 years (Lal et al. 2003; Gupta et al. 2005) by identifying mutations in genes as *Helitron* insertions. However, this does not guarantee we can identify activity in just a few generations (8 generations for the maize recombinant inbred lines). Furthermore, the maize lines we used are unlikely to be the same as the ones in which these mutations first arose. Further analysis may uncover *Helitron* activity, perhaps by exploring more tissue culture sources or perhaps by testing lines with mutations that alter chromatin structure or DNA methylation status (Chandler and Walbot 1986; Schwartz and Dennis 1986; Hirochika et al. 2000; Miura et al. 2001).

Acknowledgement

I thank E. Cho, J. Ning, J. Shi, X. Li, H. Zhang and F. Lu for providing the rice and maize DNA. I also thank E. Cho and T. Chen for training in the transposon display technique.

References

Bennetzen JL. 1996. The *Mutator* transposable element system of maize. *Curr Top Microbiol Immunol* **204**: 195-229.

Casa AM, Brouwer C, Nagel A, Wang L, Zhang Q, Kresovich S, Wessler SR. 2000. Inaugural article: the MITE family *Heartbreaker* (*Hbr*): Molecular markers in maize. *Proc Natl Acad Sci USA* **97**: 10083-10089.

Chandler VL, Walbot V. 1986. DNA modification of a maize transposable element correlates with loss of activity. *Proc Nat Acad Sci USA* **83**(6): 1767-1771.

Dombroski BA, Mathias SL, Nanthakumar E, Scott AF, Kazazian Jr HH. 1991. Isolation of an active human transposable element. *Science* **254**(5039): 1805-1808.

Emmons SW, Yesner L. 1984. High-frequency excision of transposable element *Tc1* in the nematode Caenorhabditis elegans is limited to somatic cells. *Cell* **36**(3): 599-605.

Fedoroff N, Wessler S, Shure M. 1983. Isolation of the transposable maize controlling elements *Ac* and *Ds*. *Cell* **35**(1): 235-242.

Gupta S, Gallavotti A, Stryker GA, Schmidt RJ, Lal SK. 2005. A novel class of *Helitron*-related transposable elements in maize contain portions of multiple pseudogenes. *Plant Mol Biol* **57**: 115-127.

Hirochika H. 1993. Activation of tobacco retrotransposons during tissue culture. *EMBO J* **12**: 2521-2528.

Hirochika H, Okamoto H, Kakutani T. 2000. Silencing of retrotransposons in Arabidopsis and reactivation by the *ddm1* mutation. *Plant Cell* **12**: 357-369.

Hirochika H, Otsuki H, Yoshikawa M, Otsuki Y, Sugimoto K, Takeda S. 1996a. Autonomous transposition of the tobacco retrotransposon *Tto1* in rice. *Plant Cell* **8**: 725-734.

Hirochika H, Sugimoto K, Otsuki Y, Tsugawa H, Kanda M. 1996b. Retrotransposons of rice involved in mutations induced by tissue culture. *Proc Nat Acad Sci USA* **93**: 7783-7788.

Jiang N, Bao Z, Zhang X, Hirochika H, Eddy SR, McCouch SR, Wessler SR. 2003. An active DNA transposon family in rice. *Nature* **421**: 163-167.

Kikuchi K, Terauchi K, Wada M, Hirano HY. 2003. The plant MITE *mPing* is mobilized in anther culture. *Nature* **421**: 167-170.

Lal SK, Giroux MJ, Brendel V, Vallejos CE, Hannah LC. 2003. The maize genome contains a *Helitron* insertion. *Plant Cell* **15**: 381-391.

Lippman Z, Martienssen R. 2004. The role of RNA interference in heterochromatic silencing. *Nature* **431**(7006): 364-370.

Lucas H, Feuerbach F, Kunert K, Grandbastien MA, Caboche M. 1995. RNA-mediated transposition of the tobacco retrotransposon Tnt1 in *Arabidopsis thaliana*. *EMBO J* **14**: 2364-2373.

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.

Martienssen R. 1998. Transposons, DNA methylation and gene control. *Trends Genet* **14**(7): 263-264.

McClintock B. 1946. Maize genetics. *Carnegie Inst Washington Year Book* **45**: 176-186.

McClintock B. 1951. Mutable loci in maize. *Carnegie Inst Washington Year Book* **50**: 174-181.

McClintock B. 1954. Mutations in maize and chromosomal aberrations in Neurospora. *Carnegie Inst Washington Year Book* **53**: 254-260.

Miura A, Yonebayashi S, Watanabe K, Toyama T, Shimada H, Kakutani T. 2001. Mobilization of transposons by a mutation abolishing full DNA methylation in Arabidopsis. *Nature* **411**(6834): 212-214.

Naito K, Cho E, Yang G, Campbell MA, Yano K, Okumoto Y, Tanisaka T, Wessler SR. 2006. Dramatic amplification of a rice transposable element during recent domestication. *Proc Nat Acad Sci USA* **103**(47): 17620-17625.

Nakazaki T, Okumoto Y, Horibata A, Yamahira S, Teraishi M, Nishida H, Inoue H, Tanisaka T. 2003. Mobilization of a transposon in the rice genome. *Nature* **421**: 170-172.

Peschke V, Phillips R, Gengenbach B. 1987. Discovery of transposable element activity among progeny of tissue culture--derived maize plants. *Science* **238**: 804-907.

Peschke VM, Phillips RL. 1991. Activation of the maize transposable element *Suppressor-mutator (Spm)* in tissue culture. *Theor Appl Genet* **81**: 90-97.

Peterson PA. 1953. A mutable pale green locus in maize. *Genetics* **38**(1): 682–683.

Peterson PA. 1965. A relationship between the *Spm* and *En* control systems in maize. *The American Naturalist* **99**(908): 391-398.

Robertson DS. 1978. Characterization of a mutator system in maize. *Mutation Research* **51**(1): 21-28.

Rubin GM, Kidwell MG, Bingham PM. 1982. The molecular basis of PM hybrid dysgenesis: the nature of induced mutations. *Cell* **29**(3): 987-994.

Schnable PS, Peterson PA. 1986. Distribution of genetically active *Cy* transposable elements among diverse maize lines. *Maydica* **31**: 59-81.

Schnable PS Ware D Fulton RS Stein JC Wei F Pasternak S Liang C Zhang J Fulton L Graves TA et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**: 1112-1115.

Schwartz D, Dennis E. 1986. Transposase activity of the *Ac* controlling element in maize is regulated by its degree of methylation. *Molecular and General Genetics MGG* **205**(3): 476-482.

Slotkin RK, Vaughn M, Borges F, Tanurdzic M, Becker JD, Feijo JA, Martienssen RA. 2009. Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell* **136**: 461-472.

Walbot V. 1992. Strategies for mutagenesis and gene cloning using transposon tagging and T-DNA insertional mutagenesis. *Annu Rev Plant Physiol Plant Mol Biol* **43**(1): 49-78.

Yang G, Nagel DH, Feschotte C, Hancock CN, Wessler SR. 2009. Tuned for Transposition: Molecular Determinants Underlying the Hyperactivity of a *Stowaway* MITE. *Science* **325**(5946): 1391-1394.

Yoder JA, Walsh CP, Bestor TH. 1997. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet* **13**(8): 335-340.

Table A.1: Rice lines used in this study

Name	Note	Source
Nipponbare	5 siblings	Eunyoung Cho (University of Georgia)
Gimbozu EG4	5 siblings	Eunyoung Cho
N, NC	Nipponbare cells before tissue culture and after tissue culture	Ning Jiang (Michigan State University)
C, OC	C5924 cells before tissue culture and after tissue culture	Ning Jiang

Table A.2: Maize lines used in this study

Name	Note	Source
B73, Mo17	inbred lines	Jinhua Shi (University of Georgia)
IBM24,25,26,28,29,30,31,32	recombinant inbred lines derived from B73 crossed to Mo17	Jinhua Shi
IBM1,23,42,46,48,51,68,86	recombinant inbred lines derived from B73 crossed to Mo17	Jinhua Shi
HiII WT	HiII wild type	Han Zhang (University of Georgia)
HiII 0	HiII tissue culture-derived plants without any transgene	Xuexian Li (University of Georgia)
HiII 1, HiII 2, HiII 3	HiII tissue culture-derived plant with transgene	Han Zhang
2-64, XL32, XL30	HiII tissue culture-derived plants with transgene	Xuexian Li
25,26,27,28,29,30	6 siblings (HiII after tissue culture with transgene, crossed to B73)	Han Zhang
A112, A113	HiII plants derived from tissue culture, with transgene	Fang Lu (University of Georgia)
AA1,2,3,4,5,6,7,8	8 siblings from A112xA113	Fang Lu
AB1,2,3,4,5,6,7,8	8 siblings from AA8xB73	Fang Lu

Table A.3: PCR primers for rice *Helitrons*

Name	Sequence	Element	Note	Annealing Temperature (°C)
P1	GCGCGGGCTACCTTCCTA	*Hup_OS1_27*	2nd round PCR	58
P2	GCGCGGGCCACCTTCCTA	*Hup_OS1_7*	2nd round PCR	58
P5	TTGCAGATATTAATTCGCTGG	*Hup_OS1_27*	1st round PCR	57
P9	TAATATAAGTGCCCGCGCATAC	*Hup_OS1_7*	1st round PCR	61
*Mse*I adaptor forward	GACGATGAGTCCTGAG			
*Mse*I adaptor reverse	TACTCAGGACTCAT			
*Mse*I primer	GACGATGAGTCCTGAGTAA			

Table A.4: PCR primers for maize *Helitrons*

Name	Sequence	Element	Note	Annealing Temperature (°C)
PM1	TTTATGTATGGCTAGGATCAC	*Hip1_14*	1st round PCR	53
PM3	ATTCCCGTTGCAACGCAC	*Hip1_14*	2nd round PCR	58
PM6	AGTTATCGTAGTATACTGGT	*Hip1_5*	1st round PCR	47
PM7	TATTCCCGTTGCAACGCAC	*Hip1_5*	2nd round PCR	58
PM9	TTTGGTGGATAATTTATGTGG	*Hip2_2*	1st round PCR	51
PM10	GGTATTGTTGTGAGCCGTCGC	*Hip2_2*	2nd round PCR	58
PM11	TATGTATGGCTAGGATCAC	*Hip1_3*	1st round PCR	NA
PM12	AGTTATCATAGTATACTGG	*Hip1_3*	2nd round PCR	NA
PM13	GTATGGCTAGGATCACAAATG	*Hip1_1*	1st round PCR	51
PM14	GGTATTATATGTTCCCGTTGC	*Hip1_1*	2nd round PCR	58
PM15	TACAGAGAGCGGAGACAATC	*Hip2_1*	1st round PCR	53
PM17	GGTATTGTTGTAAGCCGTCGC	*Hip2_1*	2nd round PCR	58

Figure Legends

Figure A.1: Autoradiograph of rice transposon display. (A) Primers specific to *mPing* as positive control (Nakazaki et al. 2003). N denotes Nipponbare and G denotes Gimbozu EG4. (B) Primer P9 as 1st round PCR primer and primer P2 as 2nd round PCR primer. N denotes Nipponbare and G denotes Gimbozu EG4. (C) Primer P5 as 1st round PCR primer and primer P1 as 2nd round PCR primer. N denotes Nipponbare cells before tissue culture and NC denotes Nipponbare cells after tissue culture. C denotes C5924 cells before tissue culture and OC denotes C5924 cells after tissue culture.

Figure A.2: Autoradiograph of maize transposon display. (A) Primer PM1 as 1st round PCR primer and primer PM3 as 2nd round PCR primer. (B) Primer PM6 as first round PCR primer and primer PM7 as 2nd round PCR primer.

Figure A.1:

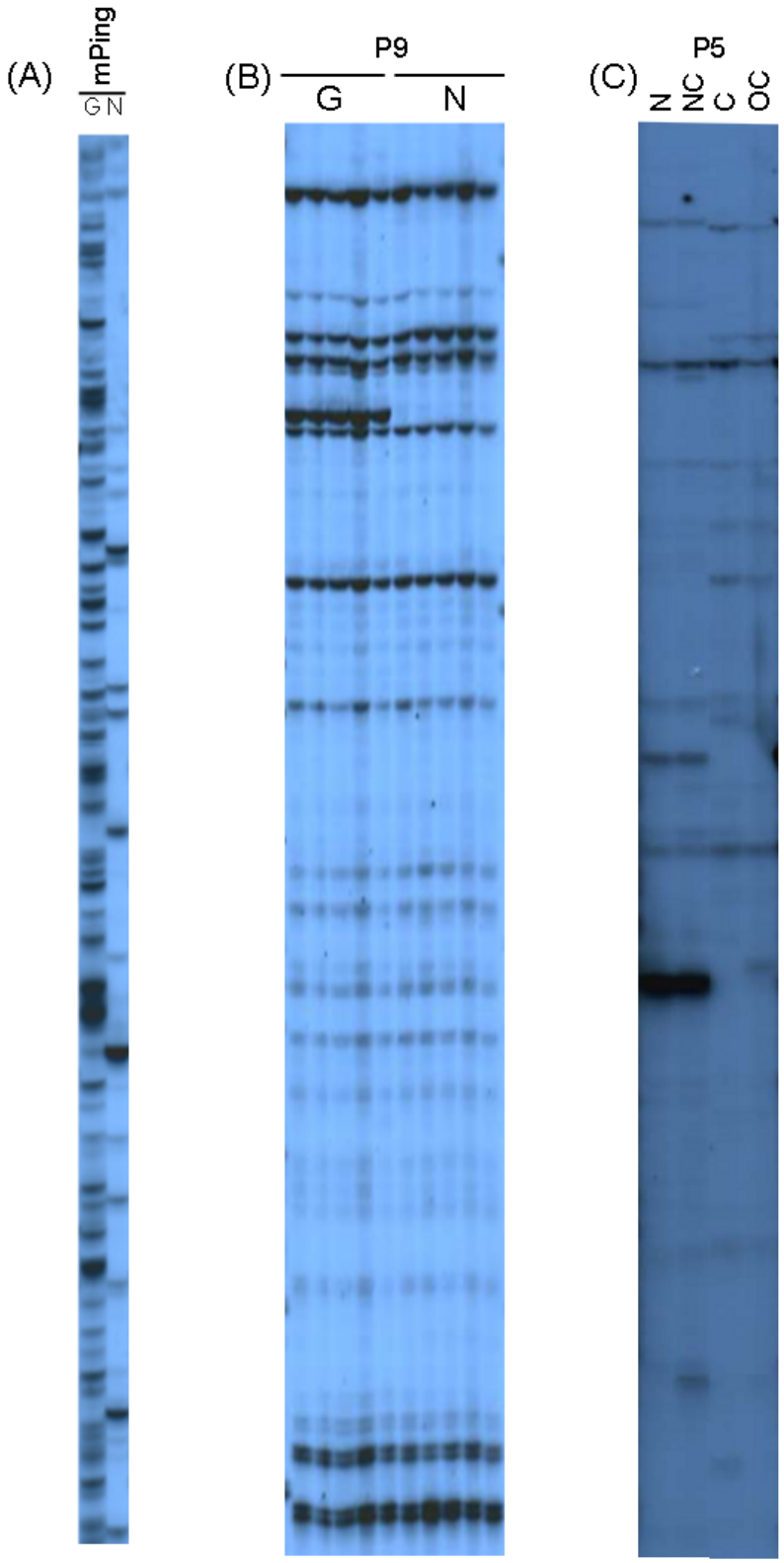


Figure A.2 (A):

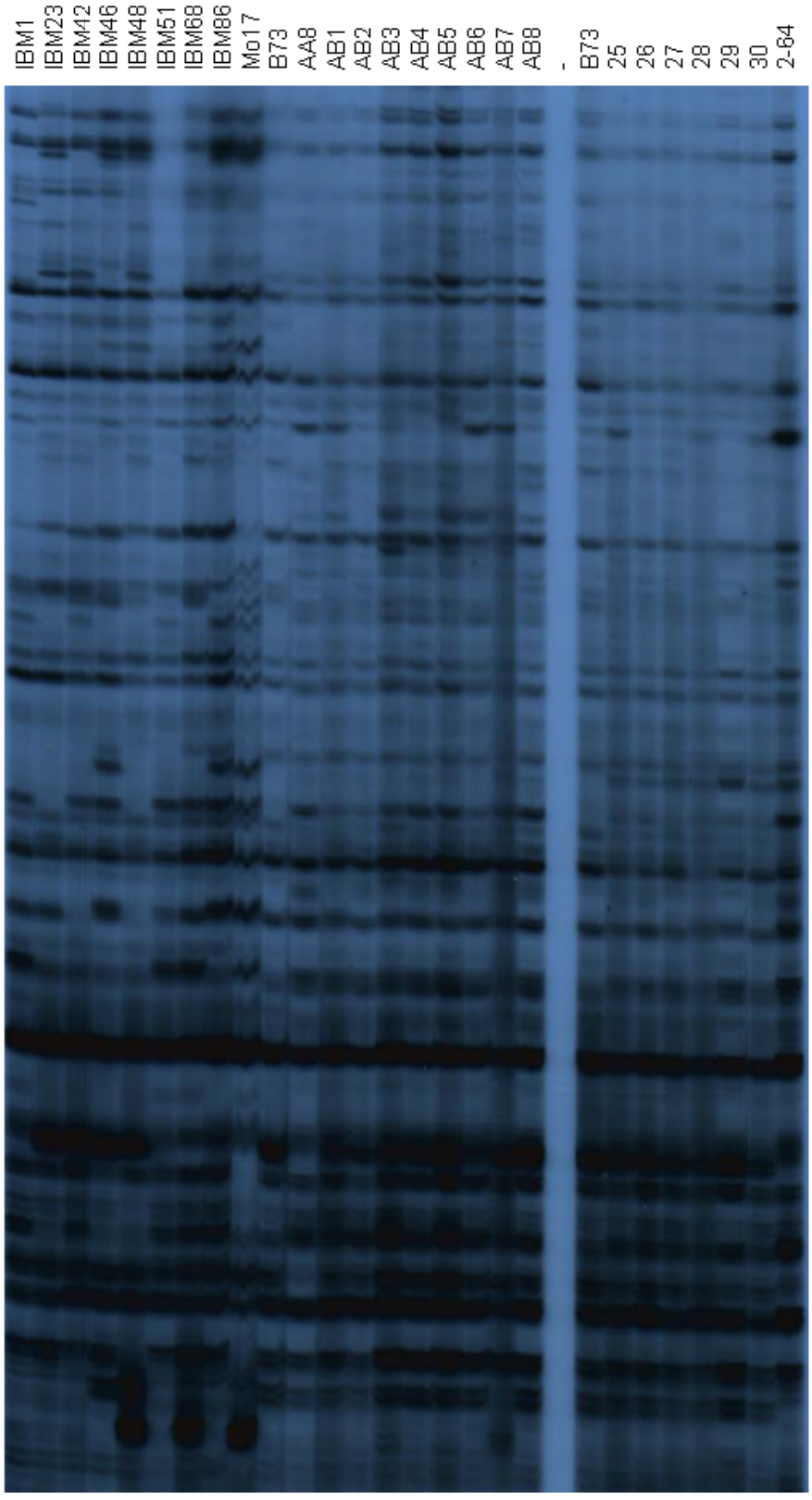
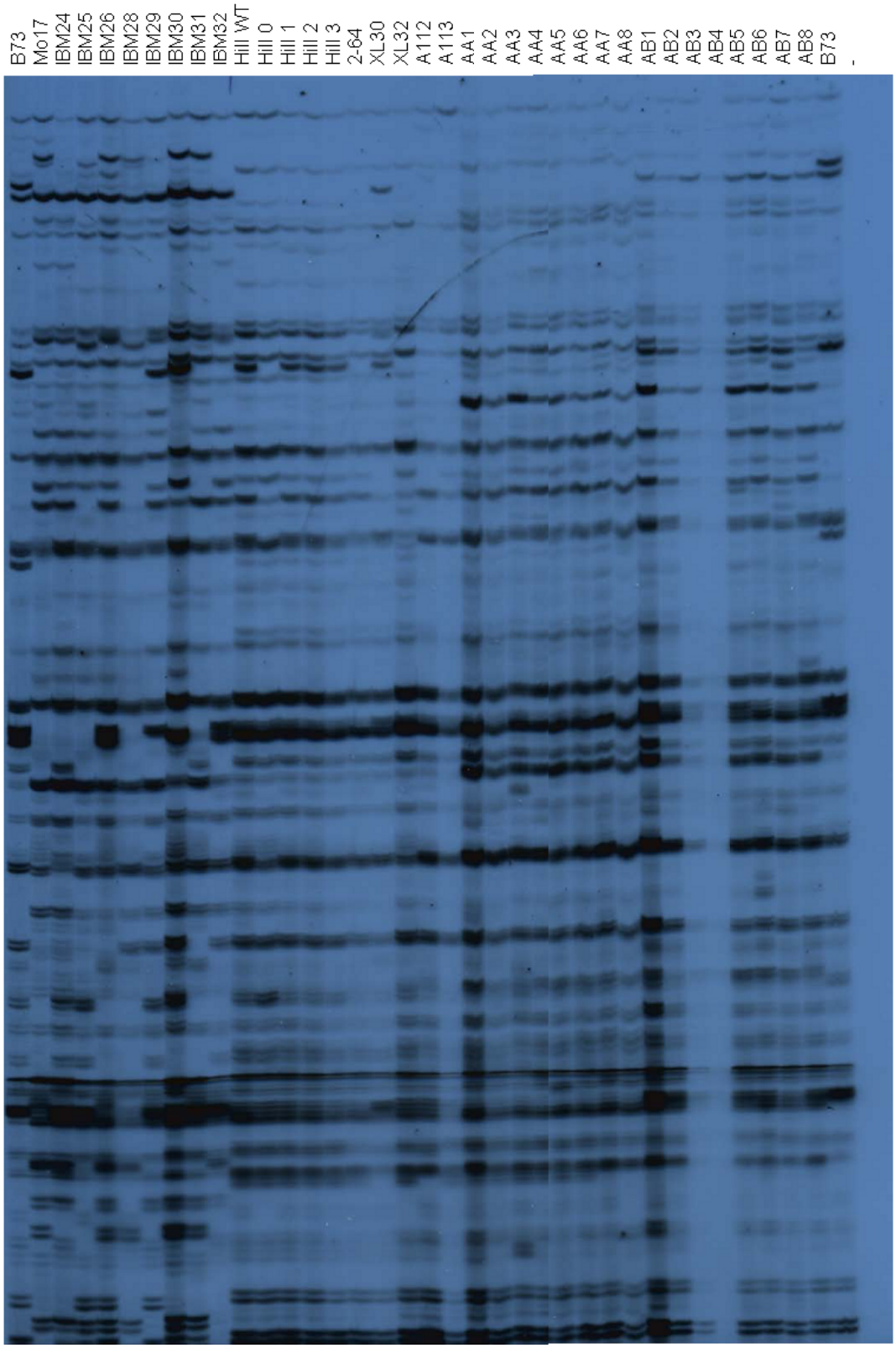

IBM1
IBM23
IBM42
IBM46
IBM48
IBM51
IBM68
IBM86
Mo17
B73
AA8
AB1
AB2
AB3
AB4
AB5
AB6
AB7
AB8
-
B73
25
26
27
28
29
30
2-64

Figure A.2 (B):


B73
Mo17
IBM24
IBM25
IBM26
IBM28
IBM29
IBM30
IBM31
IBM32
HIII WT
HIII 0
HIII 1
HIII 2
HIII 3
2-64
XL30
XL32
A112
A113
AA1
AA2
AA3
AA4
AA5
AA6
AA7
AA8
AB1
AB2
AB3
AB4
AB5
AB6
AB7
AB8
B73
-