THE CONSEQUENCES OF GENE DUPLICATION BY DNA TRANSPOSONS AND THEIR INTERACTION WITH THE HOST GENOMES AND RETROTRANSPOSONS

By

Stefan Cerbin

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Plant Breeding, Genetics, and Biotechnology-Horticulture-Doctor of Philosophy

2020

ABSTRACT

THE CONSEQUENCES OF GENE DUPLICATION BY DNA TRANSPOSONS AND THEIR INTERACTION WITH THE HOST GENOMES AND RETROTRANSPOSONS

By

Stefan Cerbin

DNA is the ultimate genetic information carrier. These sequences of nucleotides hold enormous coded data controlling all aspects of functions, including growth, development, and defense of an organism. Genes are the protein coding units that support cellular function. While gene number is similar across species, genome size varies dramatically. One source of this variation is due to transposable elements, which are DNA sequences that are capable of moving from one locus to another in the genome. These sequences are ubiquitous and provide sources of mutations for evolution. Transposable elements are classified into two classes: DNA and RNA elements (retroelements). The elements are further classified into autonomous and non-autonomous elements according to their capability to transpose. Specific elements have been shown to duplicate gene fragments and amplify in the genomes. These elements carrying genes have regulatory, evolutionary, and phenotypic effects. This dissertation illustrates examples of gene duplications by DNA transposons and their interactions with the remainder of the genome. The first entails *GingerRoot*: A novel DNA transposon encoding integrase-related transposase in plants and animals. This study reveals a unique DNA transposon located in the heterochromatic regions of the genome. The capability of duplicating gene fragments may have allowed them to be retained longer in genomic regions enriched with retrotransposons. The second comprises a study of *Nucifera nelumbo* landscape of transposable elements. In this basal dicot species, the

genic regions have been significantly expanded by the insertion of transposable elements. Interestingly, genes involved in epigenetic pathways are enriched with insertions, suggesting the co-evolution between the transposable elements and the genome surveillance machine. The third study investigates Pack-MULE *SlPM37* in *Solanum lycopersicum* and its relatives. This Pack-MULE element has achieved a higher copy number than any other Pack-MULE elements, and the possible mechanism underlining its amplification has been proposed through detailed characterization of this element and the relevant parental genes. These chapters show how genomes are comprised of varying transposons, how their context influences gene duplication, and the interactions with other genomic components including genes and other transposons. The dynamic interactions between transposable elements and their host genomes suggest the composition and abundance of transposons not only influence the genome size and genome structure, but also the path of evolution.

This dissertation is dedicated to the Cerbin family for all their inspiration.

ACKNOWLEDGEMENTS

I would like to acknowledge members of the Jiang lab that helped in-numerous times during my time, Dongmei Yin, Ann Ferguson, Dongyan Zhao, Shujun Ou, and Ning Jiang. I would also like to thank my committee members, Cornelius Barry, Eva Farre, and Min-Hao Kou. Additionally, I am thankful for the constant support of the Horticulture department staff, the greenhouse staff, and all other support staff that were tireless aides during my Ph.D. career. I would like to thank Bob VanBuren and Ching Man Wai for their help working on the *GingerRoot* project. I would especially like to thank my advisor Ning Jiang for all the time and effort she put into mentoring and teaching me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

5′ five prime

3′ three prime

A Adenine

Ac/Ds Activator/Dissociation

AGO1 Argonaute1

auto-MULE autonomous *Mutator* Like Element

ATS Aberrant Testa Shape

att attachment

BLAST Basic Local Alignment Search Tool

bp base pair

°C degree(s) Centigrade

cDNA complementary DNA

CDS Coding Sequence

Cont'd Continued

CRISPR Clustered Regularly Interspaced Short Palindromic Repeats

CTAB Cetrimonium Bromide

CYP51 Cytochrome P 450 51

C Cytosine

DCL3 Dicer Like 3

DD/E aspartic acid (D) aspartic acid (D) glutamic acid (E)

DNA Deoxyribonucleic Acid

et al. et alii/ae/a *(*and others)

etc. et cetera (unspecified additional items)

FKPM Fragments per Kilobase of exon Per Million reads mapped

G Guanine

gff general feature format

GO Gene Ontology

GR *GingerRoot*

hAT hobo Activator Tam3

HIV Human Immunodeficiency Virus

kb kilobase

LINE Long Interspersed Nuclear Element

LTR Long Terminal Repeat

Mb Megabase

MITEs Miniature Inverted repeat Transposable Elements

mRNA messenger RNA

Mu *Mutator*

MULE *Mutator*-Like Element

MYA Million Years Ago

NCBI National Center for Biotechnology Information

NN *Nelumbo nucifera*

ORF Open Reading Frame

OS/Os *Oryza sativa*

P-450 Phytochrome 450

p-value probability value

Pack-MULE MULE which carries gene or gene fragments

PIF P Instability Factor

piRNA Piwi-interacting RNA

pol polymerase

RNA Ribonucleic Acid

RNA-seq RNA sequencing

SINE Short Interspersed Nuclear Element

siRNA short interfering RNA

SRA Sequence Read Archive

sRNA small RNA

Sl *Solanum lycopersicum*

SlPM Solanum lycopersicum Pack-MULE

T Thymine

*Tam3 T*ransposable element *Antirrhinum majus* 3

TARE1 Tomato Atypical Retrotransposon Element

Tc1 Transposon *C. elegans* number 1

TE(s) Transposable Element(s)

TIR Terminal Inverted Repeat

TSD Target Site Duplication

TSS Transcription Start Site

TTS Transcription Termination Site

U5/U3 Unique5/Unique3 sequences

UTR Untranslated Region

vs versus

w/v weight per volume

CHAPTER 1

DUPLICATION OF HOST GENES BY TRANSPOSABLE ELEMENTS

The work in this chapter is part of the final publication:

Stefan Cerbin and Ning Jiang. 2018. Duplication of host genes by transposable elements. Current Opinion in Genetics and Development 49: 63-69

ABSTRACT

The availability of large amounts of genomic and transcriptome sequences have allowed systematic surveys about the host gene sequences that have been duplicated by transposable elements. It is now clear that all super-families of transposons are capable of duplicating genes or gene fragments, and such incidents have been detected in a wide spectrum of organisms. Emerging evidence suggests that a considerable portion of them function as coding or non-coding sequences, driving innovations at molecular and phenotypic levels. Interestingly, the duplication events not only have to occur in the reproductive tissues to become heritable, but the duplicated copies are also preferentially expressed in those tissues. As a result, reproductive tissues may serve as the 'incubator' for genes generated by transposable elements.

For the full text of this work, please visit https://doi.org/10.1016/j.gde.2018.03.005.

CHAPTER 2

GINGERROOT: A NOVEL DNA TRANSPOSON ENCODING INTEGRASE RELATED TRANSPOSASE IN PLANTS AND ANIMALS

The work in this chapter is part of the final publication:

Stefan Cerbin, Ching Man Wai, Robert VanBuren, and Ning Jiang. 2019. Duplication of host genes by transposable elements. Genome Biology and Evolution (11) 11: 3181-3193

ABSTRACT

Transposable elements represent the largest components of many eukaryotic genomes and different genomes harbor different combinations of elements. Here, we discovered a novel DNA transposon in the genome of the clubmoss *Selaginella lepidophylla*. Further searching for related sequences to the conserved DDE region uncovered the presence of this superfamily of elements in fish, coral, sea anemone, and other animal species. However, this element appears restricted to Bryophytes and Lycophytes in plants. This transposon, named *GingerRoot*, is associated with a 6 bp (base pair) target site duplication, and 100–150 bp terminal inverted repeats. Analysis of transposase sequences identified the DDE motif, a catalytic domain, which shows similarity to the integrase of *Gypsy*-like long terminal repeat retrotransposons, the most abundant component in plant genomes. A total of 77 intact and several hundred truncated copies of *GingerRoot* elements were identified in *S. lepidophylla*. Like *Gypsy* retrotransposons, *GingerRoots* show a lack of insertion preference near genes, which contrasts to the compact genome size of about 100 Mb. Nevertheless, a considerable portion of *GingerRoot* elements was found to carry gene fragments, suggesting the capacity of duplicating gene sequences is unlikely attributed to the proximity to genes. Elements carrying gene fragments appear to be less methylated, more diverged, and more distal to genes than those without gene fragments, indicating they are preferentially retained in gene-poor regions. This study has identified a broadly dispersed, novel DNA transposon, and the first plant DNA transposon with an integrase-related transposase, suggesting the possibility of de novo formation of *Gypsy*-like elements in plants.

For the full text of this work, please visit https://doi.org/10.1093/gbe/evz230.

CHAPTER 3

THE UNIQUE TRANSPOSON LANDSCAPE OF SACRED LOTUS (*NELUMBO NUCIFERA* GAERTN.)

Stefan Cerbin and Ning Jiang. 2020.

ABSTRACT

Sacred lotus is a basal eudicot plant that propagates through rhizomes in addition to seeds, which are well known for their longevity. Here we report the unique profile of transposable elements (TEs) in the genome. TEs account for over half of the lotus genome, and *hAT* (*Ac/Ds*) elements alone represent 9%, more than that for any reported plant. The lotus genome also harbors 14% *Copia* long terminal repeat *(*LTR) retrotransposons, and 1/3 of them are associated with non-canonical termini. Such an abundance of non-canonical LTR retrotransposons has not been reported for any other organism, providing guidance for future genome annotations. Surprisingly, *Copia* elements and long interspersed nuclear elements (LINEs) comprise the largest portion of introns in protein coding genes, with 76% of the LINEs in the genome located in introns, whereas all DNA TEs are underrepresented in introns. This contradicts previous observations that small DNA transposons are dominant in genic regions. The frequent insertion of TEs has lead to significant intron size expansion, with a total of 240 Mb for only 23,000 genes. Despite the prevalence of TEs in genic regions, some genes are associated with zero or few TEs, such as those involved in fruit ripening, photosynthesis, and stress responses. Other genes are enriched with TEs, including genes in epigenetic pathways, demonstrating the dynamic interaction between TEs and the host surveillance machine. We propose the unique growth and propagation behavior of sacred lotus may have enabled the amplification of TEs in genic regions, and there is co-evolution between the TEs and the mechanisms for genome integrity.

INTRODUCTION

Sacred lotus (*Nelumbo nucifera*) is a land plant that has adapted to an aquatic environment; it belongs to the Nelumbonaceae family and is found throughout Asia and northern Australia (Han et al. 2007; Pan et al. 2010). It provides economic value as an ornamental and food crop in Asia. In addition, parts of lotus such as the flowers, roots, and rhizomes are used for medicinal purposes (Shen-Miller 2002). Sacred lotus holds the world's record for long-term seed viability, up to 1,300 years (Shen-Miller 2002). The genome of sacred lotus variety "China Antique" was sequenced in 2013, using Illumina and 454 technologies (Ming et al. 2013), and the assembly was improved in 2018 through a linkage map and a BioNano optical map (Gui et al. 2018). The availability of the genomic sequence of sacred lotus represents an excellent resource in the evolutionary analysis of eudicots and comparative studies between dicots and monocots since sacred lotus phylogenetically lies outside the core eudicots. Phylogenetic comparisons between grape and sacred lotus suggest that it is a better model than the grape genome for inferences about the common ancestors of eudicots (Ming et al. 2013). Genomic analysis reveals that the sacred lotus genome lacks the lambda triploidization event found in all core eudicots and shows a remarkably low substitution rate and higher retention of duplicated genes compared with most other angiosperm genomes. For simplicity, sacred lotus will be referred to as "lotus" in the remainder of the manuscript.

Transposable elements (TEs) are genetic sequences first discovered over 70 years ago by Barbara McClintock (McClintock 1950). TEs move from one genomic location into another and in the process increase their copy numbers. According to the intermediate form of transposition used by the specific element, TEs are generally classified into two major groups: Class I or RNA

elements which transpose via an RNA intermediate using a copy-and-paste mechanism, and Class II or DNA elements that transpose via a DNA intermediate using a cut-and-paste mechanism (Wicker et al. 2007; Kapitonov and Jurka 2008). Based on their structural features, class I elements are further divided into two subclasses: LTR retrotransposons (LTR), most of which falls into *Gypsy* and *Copia* super-families, and non-LTR elements, which include long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) (Kumar and Bennetzen 1999). To date, eight superfamilies of DNA transposons have been identified in plants, including CACTA (*En/Spm*/DTC), *hAT* (*Ac/Ds*/DTA), *Helitron* (DHH), MULE (*Mutator*/ DTM), *PIF/Harbinger* (*Tourist*/DTH), *Tc1/Mariner* (*Stowaway*/DTT), *Sola* and *GingerRoot* (Wicker et al. 2007; Bao et al. 2009; Cerbin et al. 2019). Among these, *Sola* and *GingerRoot* have not been detected in angiosperm genomes. In addition, the coding capacity of elements for proteins involved or comprising the transposition machinery allows for further classification of elements into autonomous elements, which code for these proteins, or non-autonomous elements, which rely on the cognate autonomous elements for movement within the genome.

Due to their capacity to multiply within a host and their prevalence among plant and animal genomes, TEs have previously been implicated to contribute significantly to increases in genome size (Bennetzen and Kellogg 1997; SanMiguel et al. 1998; Piegu et al. 2006; Agren and Wright 2011). In some instances, TEs may constitute the largest part of the genome (Schnable et al. 2009; Mayer et al. 2012; Badouin et al. 2017; Bertioli et al. 2019). Dramatic differences exist in the content and diversity of different TE types between organisms. While the genomes of mammals typically contain a high proportion of non-LTR retrotransposons (Lander et al. 2001; Waterston et al. 2002), the TE composition in Arthropods is more diverse (Wu and Lu 2019). In

contrast, LTR retrotransposons largely dominate the TE landscape in plants (International Rice Genome Sequencing Project 2005; Rensing et al. 2008; Paterson et al. 2009; Schnable et al. 2009; VanBuren et al. 2018). Various computational and biological analyses of genomic information have demonstrated the critical roles of transposons in many aspects of genome evolution, gene expression, and regulation (Jordan et al. 2003; Muotri et al. 2007; Feschotte 2008; Bennetzen and Wang 2014). Moreover, TE domestication refers to the process whereby TEs lose their ability to transpose, typically through the loss of conserved terminal sequences, and evolve into new genes or adaptive sequences (Volff 2006). A variety of such events have been reported in plants (Hudson et al. 2003; Knip et al. 2012). The majority of the DNA-binding domains of plant-specific transcription factors are considered to have likely originated from endonucleases associated with transposable elements (Yamasaki et al. 2013). Taken together, these indicate that TEs are prevalent in plants and are actively interacting with other components in their host genomes.

Prior to the sequencing of the lotus genome, no information was available regarding any aspect of its repetitive sequence content. The lotus genome sequencing reported limited details on the transposons in the genome (Ming et al. 2013). Due to its position in the eudicot phylogeny, lotus may offer important biological insights in terms of its TE content, structure, distribution, and diversity. Here we report the results of a comprehensive computer-assisted identification and analysis of the repetitive content in the assembled sequence of lotus.

RESULTS

The content and diversity of TEs in the genome of sacred lotus

Identification of repeats was performed using the 798 Mb final assembly (Gui et al. 2018), with sequencing gaps excluded from the calculation of repeat fraction. The non-gap sequences (~700 Mb) account for ~75% of the estimated lotus genome (929 Mb) (Diao et al. 2006). TEs were mined using a combination of structure-based and homology-based approaches (see Materials and Methods). Approximately 55% of the genome is composed of recognizable TEs, with a total of 411,655 copies (Table 3.1).

Based on genome coverage, the majority of recognizable TE DNA in lotus is contributed by Class I/retrotransposons (36% of the genome or 65% of all TE DNA), a familiar phenomenon across the plant kingdom where the amplification of LTR elements has been suggested to contribute to genome size expansion. In lotus, the LTR retrotransposon content (29% of the genome) is comprised of a comparable number and content of *Copia* and *Gypsy* elements (14% vs. 15%, respectively; Table 3.1). Although this pattern is not unique, the *Gypsy* content considerably outnumbers the *Copia* content in the majority of sequenced plant genomes (Table S3.1). Among the 89 genomes with available *Gypsy*:*Copia* ratio, *Copia* element content outnumbers *Gypsy* elements in only 12 (13.5%) genomes (Table S1). In addition, the lotus genome contains a high coverage of non-LTR retrotransposons (6.4%), which are predominantly contributed by LINEs (Table 3.1). Only eight other plant genomes analyzed (9.0%) contain a higher fraction of non-LTR retrotransposons, ranging from 7.0% to 11.9% (Table S3.1). Two of the seven genomes are from basal angiosperm plants, *Amborella trichopoda,* and *Chara braunii*, and the remainder are from dicot plants (Table S3.1). Taken together, these results suggest a high

activity and/or retention of non-LTR and *Copia* retrotransposons in the evolution of the lotus genome in comparison to many other plants.

Class II or DNA elements comprise about 19% of the genome. This level of DNA TE content is notable, and only two other genomes, rice (20%) and red bayberry (21%), contain more DNA TEs (Jiang and Panaud 2013; Jia et al. 2019) (Table S3.2). The largest contributors to DNA element content in lotus are *hAT* elements (9% of the genome), followed by *PIF* and *Helitron* elements (occupying 3.6% and 3.5% of the genome, respectively). Over 122,000 copies of the *hAT* elements were detected, making lotus the most abundant in the genomic fraction of *hAT* elements among all plant genomes sequenced to date (Figure 3.1, Table S3.2). The lotus genome contains about twice of *hAT* elements as that in blueberry, which has the second highest *hAT* content among all sequenced species (9.0% vs. 4.4%, Table S3.2). There is a correlation between the abundance of *hAT* elements and total DNA TE content (Figure S3.1A, Pearson Correlation $r = 0.46$ $p < 0.00001$). However, if the *hAT* elements are excluded from total DNA content, the correlation between the abundance of *hAT* elements and other DNA transposons is rather weak (Figure S3.1B, Pearson Correlation $r = 0.23$ $p = 0.0288$). Although most major DNA transposon families present in angiosperms are identified, the *Tc1*/*Mariner* superfamily is absent from lotus. The absence of *Tc1*/*Mariner* elements has been reported in 11 other plant genomes, including two basal angiosperms (Table S3.2). In general, plant genomes contain limited amount of *Tc1/Mariner* elements. Among the 64 plant genomes with *Tc1*/*Mariner* elements characterized, over half (35) of them contain 0.1% or less *Tc1/Mariner* elements in the genome (Table S3.2). In addition, CACTA elements were poorly represented (0.1%) in the lotus genome.

Finally, we detected over 1,000 Pack-MULEs, which refers to *Mutator*-like elements carrying gene or gene fragments (Jiang et al. 2004), suggesting its abundance in this basal dicot.

Abundance, diversity, and domestication of *hAT* elements

DNA transposable elements belonging to the *hAT* superfamily are widespread in plant and animal genomes and have been widely used in gene tagging and functional genomics studies (Sundaresan et al. 1995; Kunze and Weil 2002). Although widespread in plants, the contribution of *hAT* elements to genomic repeat content is typically low, as this is the case for most DNA transposons. Among the 89 plant genomes with *hAT* elements detected, these elements represent ≤ 1% of the genome for 58 (65%) plants (Table S3.2). As stated above, the lotus genome is exceptional in its *hAT* content (9%) among plant genomes. The copy number of *hAT* elements (over 122,000) is the highest among all super-families of TE in the lotus genome, more than the total of all retrotransposons detected (Table 3.1).

To evaluate whether specific families of *hAT* elements have expanded in the genome, the coverage of individual families was determined. Results indicate that overall, the high abundance of *hAT* elements in lotus was not due to the massive amplification of a single or a few families, instead, it was attributed to the amplification of numerous distinct families. Despite the fact that some elements are more abundant than others, the most amplified family only contributes 0.3% of the genome, and the top 20 families contribute 3.7% of the genome. This is in contrast to the LTR retrotransposons in maize, where the top 20 families contribute up to 70% of the maize genome (Baucom et al. 2009). Among the top 20 families of *hAT* elements in lotus, only two potentially encode transposase. The remainder are all small non-autonomous elements, and the average size of these *hAT* elements is only 510 bp. This explains why *hAT* elements contribute a

much smaller total genomic sequences than retrotransposons despite its higher copy number (Table 3.1).

To test whether diversity in *hAT* transposase may reflect its successful amplification in the lotus genome, phylogenetic analysis of the most conserved domain (motif 3 which contains the E region of the catalytic DD/E motif) of the *hAT* transposase among autonomous copies was performed (Kempken and Windhofer 2001; Lazarow et al. 2012). Our analysis indicates substantial diversity among the *hAT* transposases found in the lotus genome, which contains autonomous *hAT*s from the two clades typically found in plants: *Ac*/*Tam3* and *Tag1* (Figure S3.2) (Kempken and Windhofer 2001; Robertson 2002). The majority of the *hAT* elements with a recognizable motif 3 that groups within the *Ac*/*Tam3* clade show a wide spectrum of diversity wherein various subgroups found are more closely related to *hAT* proteins from other plant species than *hAT* transposase within the genome. Overall, these results suggest that diversity within autonomous elements may have contributed to the success of the *hAT* superfamily in lotus.

As mentioned above, TE domestication is prevalent in plants. A well-known domesticated *hAT* transposase in Arabidopsis is *Daysleeper*, which is essential for growth and development (Bundock and Hooykaas 2005). Since their discovery, *Daysleeper*-like genes (*Sleeper*s) are found to be unique in angiosperms and frequently found in multiple copies (Knip et al. 2012). In the lotus genome, four *Daysleeper*-like sequences not associated with TIRs are found (here referred to as *NNSLEEPER1*, *NNSLEEPER2*, *NNSLEEPER3*, and *NNSLEEPER4*, Figure S3.2). Analysis of EST sequences suggests that all four lotus *Sleeper* genes are expressed. Further, genomic sequence comparison indicates that *NNSLEEPER1* is syntenic to *Daysleeper*

and the grape *Vinesleeper2* (Figure S3.2). Phylogenetic analysis using the conserved domain 3 of the catalytic core suggests that the *NNSLEEPER1* and *NNSLEEPER2* are paralogs that originated after the separation of lotus and grape (Figure S3.2). While *NNSLEEPER3* and *NNSLEEPER4* appear to be paralogs where the corresponding elements are absent from the grape genome, and these two sleeper genes are more closely related to the *Sleeper* gene in *Amborella trichopoda* (*AmtSleeper*) (Knip et al. 2012), also a basal angiosperm (Amborella Genome Project 2013). The conserved domain 3, which contains the catalytic site, of *NNSLEEPER3* and *NNSLEEPER4* is identical (Figure S3.2), suggesting very recent duplication and/or strong sequence conservation. The distinct origin of *NNSLEEPER1/2* and *NNSLEEPER3/4* in lotus is in contrast to that in grape and rice where all *Sleeper* genes are more phylogenetically similar and potentially derived from relatively recent sequence duplication (Figure S3.2).

LTR elements with non-canonical ends

The majority of LTR retrotransposons are classified into two major super-families: *Copia* and *Gyspy* depending on the arrangement of the genes in the *pol* region (Kumar and Bennetzen 1999). In both major superfamilies, the canonical LTR repeats found at the ends of these elements typically start with 5' -TG and ends in CA-3' (Kumar and Bennetzen 1999), which forms a short inverted repeat. In fact, many computer-assisted searches use these criteria in *de novo* searches for LTRs (McCarthy and McDonald 2003; Xu and Wang 2007; Ellinghaus et al. 2008). In Lotus, however, 111 LTR families comprising eight different non-canonical LTR ends were found. The vast majority of elements with non-canonical ends (or non-TGCA) are *Copia* elements (106 families with 14,810 copies) and collectively contributes 4.5% of the genome or 1/3 of the total coverage by *Copia* elements (Table 3.2), suggesting non-TGCA elements could

represent a considerable portion of *Copia*-like elements. Among all groups of *Copia* elements with non-canonical ends, four of them (TGCT, TGGA, TACA, and TGTA) harbor mutations in one nucleotide compared to the canonical ends, and the remainder (TGGT, TACT, TATA, and TGTT) harbor mutations in two nucleotides. Those eight ends no longer form short inverted repeat except one (TATA). Overall, the groups containing a single mutation are more abundant than those harboring two mutations (3.65% vs. 0.71%; Table 3.2). In addition, variations in constraints are observed in terms of the mutations allowed at different sites: a) no mutation was detected at the most 5' nucleotide (always "T"); b) the second nucleotide at 5' end is a purine (G or A); c) the second nucleotide at the 3' end is the most flexible, and can be C, G, or T; and d) the last nucleotide can be either "A" or "T". The most abundant non-canonical end type is found in elements where the LTR starts with the canonical 5' -TG but ends in CT-3' (referred to as TGCT LTR), where the most terminal nucleotide is not inverted. This LTR end type includes an estimated 8,869 copies, making up 2.58% of the genome, much more abundant than other types of non-TGCA elements. Consistent with our previous study (Ou and Jiang 2018), the LTRs of non-TGCA elements are shorter than the canonical TGCA elements (312 vs 431 bp, Table 2).

To determine the relationship between the non-canonical elements and canonical elements, a phylogenetic analysis was performed using the conserved integrase catalytic domain of TGCT LTR elements as well as other LTR elements in lotus and *Copia* elements in other species. It appears that the majority of the TGCT LTR families in lotus are closely related to each other and form a single clade (Figure S3.3). Although the TGCT LTR families are predominant in this clade, it also contains families with the canonical TGCA end, and two other non-canonical types (TGGT and TGGA *Copia* LTR families). Meanwhile, one TGCT LTR family (NN00285)

grouped separately in a clade that also includes other LTR families from five non-canonical LTR types (TACA, TACT, TATA, TGTA, and TGTT; Figure S3.3). Since grape is the closest eudicot sequenced genome to lotus, *Copia* elements from grape including five TGCT families identified in this study were also analyzed. Overall, the TGCT *Copia* families in grape grouped distinctly from those in lotus. In contrast, some *Copia* families in grape with canonical ends group together with their lotus counterparts (Figure S3), suggesting that the TGCT elements in the two species do not have a common origin. Taken together, most elements with non-canonical ends in lotus group with other lotus elements, and elements with different ends are intermingled in polyphyletic clades.

The prevalence of TEs in genic regions and distinct niches for different super-families of TEs

Based on the latest gene annotation (Gui et al. 2018), there are 23,810 protein coding genes in the lotus genome. After filtering out 944 potential TEs (see Materials and Methods), 22,866 genes remained. To study the distribution of TEs around genes, we examined the insertion of TEs within genic regions (from transcription start site, TSS, to transcription termination site, TTS) as well as 1 kb flanking sequences (upstream and downstream) of genes. Over 135,000 element copies were detected with 126 Mb of TE sequences in these regions, accounting for about 1/3 of all TEs in the genome. The number of TEs in the genic regions of lotus is about 4-fold of all TEs (31,189 elements) in the Arabidopsis genome according to information from TAIR10 (https://www.arabidopsis.org/) (Berardini et al. 2015). If we exclude the TEs in flanking regions, 97,599 copies with 111 Mb of TEs are within lotus genes (exons and

introns), representing 24% of the insertions and 29% of the TE length in the genome, respectively.

The insertion density (insertions per 100 kb) and genomic fraction (%) of TEs in genic regions are only slightly lower than the genomic average value (Figure 3.2A). If different genic regions are compared, it is obvious that there are few TE insertions into coding regions (Figure 3.2A, Figure 3.3), which is not surprising. The number of TEs within UTRs is also much lower than the genome-wide level, suggesting the strong selection against insertions into these sequences. The density of insertions within introns is slightly lower than the genome-wide level; however, the genome fraction of TEs in introns is comparable to the genomic average (Figure 3.2A), implying that TEs within introns are longer than the average TEs in the genome. Despite the overall lower density of TEs in genic regions than other regions, the TE density in the immediate flanking regions of genes is significantly higher than that of genome-wide level ($\chi 2$ test, $p < 1e\text{-}10$, Figure 3.2A). Nevertheless, the genomic fraction of TEs is much lower, due to the enrichment of short DNA TEs in flanking regions of genes (see below).

The number of insertions as well as the percent difference between insertion density in each region and that of genome-wide level, called enrichment index, are shown in Figure 3.3. A positive enrichment index value indicates enrichment of TE insertions, whereas a negative value indicates depletion or underrepresentation. It is obvious there is a depletion of retrotransposons in the immediate flanking regions of genes, and the bias is more obvious at the upstream regions than that at the downstream regions, with the exception of *Gypsy* elements (Figure 3.3), which is less depleted in the upstream regions than downstream regions. Among DNA transposons, MULEs are most abundant and most enriched in upstream regions, followed by *PIF* and *hAT*

elements. The relative order of preference among different TEs within 5' UTR is similar to that within 5' flanking, despite the much lower overall insertion density (Figure 3.3, Table S3.3), suggesting the specificity for the upstream region of genes extended into transcribed regions. Within the downstream regions, *PIF* elements are most abundant and most enriched, followed by MULEs and *hAT* elements. Again the distribution of different TEs within 3' UTR mimics that for 3' flanking sequences yet with a much lower overall insertion density (Table S3.3). Since the only elements enriched in flanking regions of genes are DNA transposons, which are small in size (Table 3.1), it explains why the flanking regions of genes are associated with the highest insertion density in the genome but lower TE fraction than the genomic average (Figure 3.2A). To further explore the role of element size in the distribution of TEs in flanking regions, the correlation between enrichment index and element size was tested through Pearson correlation coefficient. The result indicates a negative correlation between enrichment index and element size at 3' flanking ($r = -0.94$, $p = 0.0002$), yet the correlation at 5' flanking is much weaker ($r = -0.64$, $p = 0.063$) (Figure S3.4). This suggests selection for small TEs may have played more significant roles in the TE composition at 3' flanking than that at 5' flanking.

Among the 22,866 protein coding genes, 19,451 contain introns. The total size of introns is close to 240 Mb, representing about 30% of the genome and nearly twice of the Arabidopsis genome (Arabidopsis Genome Initiative 2000). The average length of the longest 10% of introns is 13,625 bp, which is comparable to that (13,607 bp) of ginkgo (*Ginkgo biloba*), the largest intron size reported so far in plants (Guan et al. 2016). Over half (53%) of lotus intron sequences are composed of TEs, and the largest contributions are from *Copia* LTR retrotransposons (19%) and LINEs (16%). These are followed by *hAT* elements (8%) and *Gypsy* LTR retrotransposons

(5%) (Figure 3.2B). As a result, retrotransposons occupied 40% of the intron space which is unique. The size contribution from other super-families of TEs (*Helitron*, *PIF*, MULE, and SINE) is insignificant (< 2%, Figure 3.2B). Regardless, TEs, particularly retrotransposons, are the most important factor for the expansion of intron size in lotus. Numerically, DNA elements account for 62% of the insertions, and *hAT* elements are the most abundant elements in introns, yet LINEs are most enriched in introns (Figure 3.2B, Figure 3.3). In fact, 76% of the LINEs in lotus are located in introns and this is in comparison to the overall fraction of 23% for other TE families. Such a level of concentration for a single super-family of TEs in introns is unprecedented in any other organism.

A summary of the skewed distribution of different super-families of TEs is provided in Table 3.3. It is obvious that each super-family of TEs has a unique distribution pattern. This is even true for LINEs and SINEs, which share transposition machinery (Singer 1982). Among retrotransposons, LINEs appear to have the highest specificity, which are only concentrated in introns. *Copia* elements are significantly underrepresented in upstream regions of genes and are enriched in introns. For DNA transposons, MULEs demonstrate the strongest bias, with exceptional enrichment in upstream regions of genes followed by downstream regions. Among all the super-families with considerable amplification (> 1% of the genome), MULEs is the most depleted from introns. *PIF*-like elements are also enriched in flanking regions but such preference is more evident in downstream regions. In contrast, *hAT* and *Helitron* elements have a relatively minor preference (Table 3.3), yet *hAT* elements are enriched in flanking regions of genes while *Helitron* elements are not.

Genes involved in distinct biological processes are differentially enriched with TE insertions

To investigate whether different genes have distinct tolerance/preference for TE insertions, we examined the density of TE insertions in genes involved in different biological processes according to their GO terms. The abundance of TEs associated with different genes varies dramatically (Table 3.3), with a nearly 6 fold difference in terms of average insertion number (from 2.00 to 11.08 insertions per gene). If the entire genic regions are considered, TEs are underrepresented with genes involved in processes such as fruit ripening, abscission, flower development, photosynthesis, and responses to various stimuli (endogenous, external, biotic, abiotic, chemical, and light) (Table 3.3; Table S3.4). In contrast, TEs are enriched in genes involved in DNA metabolic process, regulation of gene expression and epigenetic, protein metabolic process, cell cycle, etc.

If different genic regions are examined, there are detectable differences in terms of the TE density among different genes. Since a large portion of TE insertions reside in introns, genes that are enriched or depleted with TEs in introns are similar to the entire genic regions (Tables 3.4, 3.5, S3.4, and S3.5), albeit the variation is more dramatic (0.64 vs 9.68 TEs per gene). The number of TEs in introns per gene is positively correlated with the number of introns, which is predictable (Table 3.5 and Table S3.5). However, this relationship does not always hold. For example, genes involved in secondary metabolic process have fewer introns than that in fruit ripening (4.09 vs 5.11), yet they have three times more TE insertions (2.63 vs 0.64 per gene, Table 3.5) in introns. The number of introns only varies by one fold in different categories, yet

there is a 15 fold variation in terms of TE insertions (Table 3.5). This suggests that intron number is not the sole factor that determines the abundance of TEs in introns.

Compared with introns, the variation of TE numbers is not as dramatic in the flanking regions of genes (Table S3.6). For upstream regions, genes involved in DNA metabolic process are the second most enriched with TE insertions (Table S3.6). Surprisingly, genes involved in translation are the most enriched with TE insertions in upstream regions of genes (Table S3.6) but not in introns or downstream regions. In fact, in the downstream regions of genes, no significant enrichment or depletion of TEs is observed with genes in any category, suggesting the differentiation of TE specificity or variation of selection against insertions among genes is less significant in the downstream regions than that in upstream regions and introns.

Among the gene categories depleted of TE insertions, there are several homologous to genes that have been reportedly related to seed longevity in other plants. Most of those genes are involved in flower development or responses to various stimuli (Table S3.7), and they are associated with fewer TEs, with the most dramatic difference within introns. While an average genomic gene carries over 5 kb of TEs in their introns, these genes only have on average 800 bp of TEs in the introns, nearly 1/7th of the genomic average value. Those genes have slightly more introns than that of the genomic average (6.64 vs 6.05, Table S3.7), again suggesting that the number of introns is not the only factor responsible for the abundance of TE insertions. For example, the *Aberrant Testa Shape* (*ATS*) gene, which encodes a transcription factor that defines the polarity of Arabidopsis ovule integuments (McAbee et al. 2006), has two copies in the lotus genome. There is no TE within 1 kb flanking and introns for one of the *ATS* genes. For the

second copy, there is only one small *hAT* element (257 bp) located 742 bp upstream of TSS. Both genes have 5 introns without any TE insertions (Table S3.7).

In contrast to the above genes, genes involved in the regulation of gene expression and epigenetic pathways are enriched with TE insertions in introns (Table 3.5). Genes involved in this pathway are associated with more introns than average (9.14 vs 6.05, Table 3.5 and Table S3.5) as well as large coding regions (Zhao et al. 2017). One example of the genes in the epigenetic pathway is the *Dicer 3* gene (*DCL3*), which is known to be responsible for the generation of 24 -26 nt small RNAs and silencing of TEs (Xie et al. 2004; Slotkin and Martienssen 2007). As shown in Figure 3.4, the protein sequences of the relevant genes are well conserved in Arabidopsis, grape, and lotus, yet the gene sizes vary more than 10 fold, largely due to the expansion of introns. The introns within the Arabidopsis *DCL3* gene are all very small (< 500 bp), and the gene is 7.3 kb in length. The grape gene is 28.6 kb, largely due to the expansion of 4 introns (Figure 3.4), which are caused by insertions of TEs and the largest contribution to the size of introns in grape *DLC3* is from LINEs (6.0 kb) and a *Copia* element (5.1 kb). The lotus *DCL3* gene is over 100 kb, with numerous TEs in 11 introns, and 7 introns are longer than 5 kb. TE sequences account for 68% of the intron sequences, and the majority is from retrotransposons (60% out of 68%). No EST sequence was found to be associated with the lotus *DCL3* gene, so it is unclear whether it is expressed. The dramatic gene size variation of *DCL3* in the three species demonstrates amplification of retrotransposons could lead to a significant expansion of intron size.

DISCUSSION

Angiosperm is the most dominant plant taxon containing as many as 400,000 species and ranks second to insects in species richness (Jarvis 2007). Two major groups within angiosperms are monocots and dicots whose split was dated 150-130 million years ago (MYA) (Wikstrom et al. 2001). At present, the eudicot clade represents ~75% of the species diversity in angiosperms (Drinnan et al. 1994), and lotus occupies a key position in studies of angiosperm evolution particularly in the monocot-dicot split. The divergence from its closest sister lineage was dated at 137-125 MYA (Wikstrom et al. 2001). Therefore, it replaces the grape genome, which diverged from its sister lineage about 118-108 MYA (Wikstrom et al. 2001) as the most basal eudicot sequence released (Velasco et al. 2007; Ming et al. 2013). In addition, it lies outside the core eudicot clade and lacks a duplication event that is shared by all other sequenced eudicots (Ming et al. 2013).

In this study, computer-assisted and manual approaches were used to mine and characterize the repeat content and diversity of lotus. Over half of the genome sequence is repetitive with the bulk of this composed of TEs. This is an underestimate because the unassembled region or sequencing gaps are usually more enriched in repetitive sequences. Like other plant genomes, a large portion of the lotus genome repeat sequence is contributed by LTR retrotransposons. However, the content and diversity of some of the TE families contained in the genome provide some interesting facets.

Infrequent exchange of DNA transposons between lotus genome and other genomes reflected by TE profile

Multiple factors are involved in the success or failure of a TE family in a host genome. Novel TEs may be introduced into a host through many pathways including but not limited to hybridization and polyploidy (Kawakami et al. 2010; Parisod et al. 2010), and horizontal transfer (Bartolome et al. 2009; Schaack et al. 2010; El Baidouri et al. 2014). Unless quickly lost, novel TEs may become prolific until they are recognized by the host and subsequently silenced. TEs can be lost or become undetectable when inactive TEs accumulate mutations over time until they are no longer recognizable, or are eliminated from the genome through drastic processes such as recombination and random deletions (Tenaillon et al. 2010).

The TEs in lotus exhibits a notable gain and loss in DNA elements compared to that of other plant genomes. The lotus genome contains all major types of DNA TEs found in higher plants except the *Tc1/Mariner* family, a widespread DNA TE family found in many plant genomes (Feschotte and Wessler 2002), making it one of the 12 reported genomes devoid of these elements (Table S2). Since the half-life of TEs is only a few million years (Ma et al. 2004), it would take a much longer time for a TE superfamily to completely disappear from the genome after it loses activity. Despite the overall abundance of DNA transposons in lotus, little (0.1% of the genome) is contributed by CACTA elements. The lack of *Tc1/Mariner* elements from multiple plant genomes and the very low abundance of CACTA elements in lotus suggests that horizontal transfer events of these elements in plants are rare or unsuccessful, compared with the frequent horizontal transfer of *Tc1/Mariner* elements in animals (Robertson and Lampe 1995; Zhang, Peccoud et al. 2020). This is also in contrast to LTR retrotransposons, which may have had two million horizontal transfer events in angiosperms (El Baidouri et al. 2014). It is not

impossible that the RNA intermediate of retrotransposons may favor horizontal element transfer among different organisms.

Both classes of TEs occupy genic regions in lotus but concentrated in different domains

According to their locations in the genome, TEs can be divided into two mutually exclusive groups: one concentrated in large constitutive heterochromatic blocks found in the pericentromere, knobs, and TE islands (heterochromatic TEs); the other is frequently near or inside genes (genic TEs). The activity of those two groups of TEs is regulated by distinct silencing mechanisms (Sigman and Slotkin 2016). In general, DNA transposons, particularly miniature inverted repeat transposable elements (MITEs), target genic regions while retrotransposons are nested in heterochromatic regions (Cresse et al. 1995; SanMiguel et al. 1996; Pereira 2004; Hollister and Gaut 2009). Among plants, the lotus genome contains an exceptional TE fraction in genic regions, including 29% within genes. This is in contrast to that in Arabidopsis, where only 3% of TEs are within genes (Le et al. 2015). Theoretically, the distribution of TEs in the genome is determined by the target specificity of the elements as well as the selection pressure for retention or elimination of the insertions. Due to the presence of selective force, the distribution pattern does not always reflect the target specificity of TEs. Nevertheless, the distribution pattern of some TEs in lotus seems to be in accordance with their target specificity. For example, the exceptional enrichment of MULEs at the upstream regions (Figure 3.3) is consistent with their target specificity at the 5' end of genes (Dietrich et al. 2002; Liu et al. 2009; Jiang et al. 2011). Recently, it was proposed that such target specificity is linked to transcription initiation with RNA polymerase II (Zhang, Zhao et al. 2020).

The analysis of lotus genomes also revealed a variety of exceptions/modifications to previous observations. First, both DNA transposons and retrotransposons could contribute to a significant portion of genic regions; second, flanking regions are associated with the highest insertion density in the genome (Figure 3.2A), and most of them are from DNA transposons. This is consistent with the observation that TSS and TTS are the hot spots for DNA transposon insertions (Liu et al. 2009; Zhang, Zhao et al. 2020). Based on our analysis, it is likely both flanking regions are associated with selection for small TEs but the correlation between TE size and degree of enrichment is much more significant in downstream regions than that for the upstream regions (Figure S3.5). In addition, there is no significant differentiation of insertion density in downstream regions for genes in distinct biological processes. This may suggest that the distribution pattern of TEs in the downstream region is largely shaped by a uniform constraint for disruption of function (such as the integrity of poly-A signal), while in upstream regions there are more specific interactions between genes and TEs.

Previous studies based on model plants such as Arabidopsis and rice indicate MITEs preferentially insert into genic regions including introns (Feschotte et al. 2002). In this study, we show that intron size could be significantly expanded through the insertion of TEs. Moreover, since TEs are only recognizable for a few million years, it implies such expansion has occurred in a rather short evolutionary period. The largest contributions of intron size are from *Copia* LTR retrotransposons and LINEs. Indeed, DNA transposons are numerically more abundant than retrotransposons in introns; however, their insertion densities are actually lower in introns than in other regions in the genome (Figure 3.3), suggesting introns are unlikely the preferred targets for DNA transposons in lotus. Introns in human genomes are large and full of TEs, yet there is no

significant variation among different TEs in terms of their relative fraction in introns compared with the genome average, with the exception that LTR retrotransposons are slightly underrepresented (Sela et al. 2007). The fact that 76% of the LINEs in lotus are located in introns compared with only 23% for other TE families in the same location suggests either there are factors favoring the retention of LINEs, or these elements specifically target introns. For example, it is known that the majority of the splicing events occur co-transcriptionally (Tilgner et al. 2012), so it is not impossible for the endonuclease of LINEs to interact with the splicing machinery, therefore, direct the elements into introns.

The abundance of TEs in genic regions may be attributed to the unique propagation and growth behavior of lotus

The total genomic fraction of TEs in lotus is comparable to other plants with similar genome sizes, such as tomato, potato, sorghum, soybean, and apple (Table S3.2); nevertheless, the degree of enrichment of TEs in genic regions, particularly in introns, has not been reported for those plants. Lotus is different from plants with similar genome sizes (except potato) because of its capability to propagate through rhizomes, which allows the genome to carry a heterozygous deleterious TE insertion for an extended time instead of being selected out immediately from gametes. This provides more opportunity for a TE insertion to retain and spread in the population. Certainly, a heterozygous TE insertion will require sexual reproduction to be fixed in the genome. In general, selfing or inbreeding favors the retention of a TE insertion due to the small effective population size and limited recombination (Charlesworth and Charlesworth 1995; Wright et al. 2001).

An individual lotus flower is not self-fertile because there is a lag time between the maturation of stigmas and that of stamens in the same flower (Shen-Miller 2002). Therefore, lotus is considered to be an outcrossing plant. Nonetheless, the rhizomes of lotus plants are able to generate secondary and tertiary rhizomes. Over time, a single plant forms a network of rhizomes, their areal apexes, leaves, and flowers, and it could occupy an entire pond (Shen-Miller 2002). As a result, the chance for pollination between lotus flowers from the same plant is rather high, which favors the fixation of TE insertions even they are slightly deleterious. Although vegetative propagation applies to potato as well, its self-incompatibility may enable more effective elimination of undesirable insertions. Furthermore, it is known that TEs could be activated by γ-radiation (Nakazaki et al. 2003). Lotus fruits could be preserved in lakebeds for hundreds of years, and the seeds remain viable. During the extended preservation, the fruits have been exposed to a considerable dosage of γ-radiation cumulatively (Shen-Miller 2002), and it is conceivable that some TEs could be activated during the process and amplify in the genome.

Target specificity and TE domestication may enable the success of *hAT* elements

Although the propagation and growth behavior of lotus may have favored the retention of TEs including those in the genic regions, it is obvious that this does not represent a "magic" recipe for all TEs in the genic region, as evidenced by the extinction of *Tc1/Mariner* elements in lotus. Moreover, the correlation between the abundance of *hAT* elements and other DNA transposons is rather weak (Figure S3.1). The *hAT* superfamily includes the *Ac/Ds* elements which were the very first transposons discovered and has since been paramount in the study of TE domestication. In a study of 65 instances of traits in angiosperms generated by TEs, *hAT* elements account for 20% of these (Oliver et al. 2013), suggesting its importance in angiosperm

evolution. These include the *SLEEPER* genes from a domesticated *hAT* transposase that functions as transcriptional regulators of plant development unique to angiosperms and *Gary* genes conserved in cereal genomes (Muehlbauer et al. 2006; Knip et al. 2012). Four *SLEEPER* genes are found in the lotus genome (Figure S3.2) and all of them are expressed. Phylogenetic comparisons with other known *SLEEPER* genes show the four lotus *SLEEPER* genes belong to two distinct groups: one group contains the original *Daysleeper* gene from Arabidopsis (*NNSLEEPER-1/2*), and the other contains the *SLEEPER* gene from Amborella, a basal angiosperm species (*NNSLEEPER-3/4*, Figure S3.2). This seems to imply that there was a duplication of *SLEEPER* genes in the early stage of angiosperm, where only one copy was retained in other genomes yet both of them retained in lotus. If this is the case, it is consistent with the observation that lotus is associated with a higher retention rate of duplicated genes (Ming et al. 2013), including those derived from TEs. For this reason, lotus represents an excellent model to study the origin, function, and evolution of *hAT*-related genes.

A recent study indicates that the targeting of *hAT* elements is not as precise as MULEs (Zhang, Zhao et al. 2020), which is consistent with our current finding that, among DNA transposons *hAT* and *Helitron*s are more evenly distributed in the genic regions than other DNA transposons (Figure 3.3). The relatively low specificity may provide more potential targets for insertions or opportunities for survival. In addition, *Helitron*s are twice as big as *hAT* elements and that may partly explain why *hAT* elements are more enriched in genic regions than Helitrons (Figure 3). Unlikely MULEs, which target highly expressed genes, *hAT* elements preferentially insert into genes that are expressed at a medium level (Zhang, Zhao et al. 2020). Given theat highly expressed genes are often subject to more selective constraint (Drummond et al. 2005;

Koonin and Wolf 2010; Davidson et al. 2012), targeting moderately expressed genes may represent advantages for retention. Moreover, the intensity of epigenetic silencing of a TE family positively correlates with family copy number in plants (Hirochika et al. 2000; Cheng et al. 2006; Noreen et al. 2007), so the modest amplification of numerous individual families of *hAT* elements instead of the domination of a few families may prevent a complete silencing of transposition activity. Finally, given the presence of multiple *hAT*-related genes in the genome and the residual homolog between the *hAT*-related genes and *hAT* transposase, one may not rule out the possible conflicts between the silencing of *hAT* transposase and the *hAT*-related genes. Taken together, the unique growth behavior of lotus, the size, diversity, and target specificity of *hAT* elements as well as the presence of domesticated *hAT*-related genes may all contribute to the exceptional amplification of *hAT* elements in lotus.

Recent origin and burst of *Copia* elements with non-canonical ends in lotus

As mentioned above, most LTR elements start with 5'-TG and end in CA-3', and the conservation of this terminal dinucleotide sequence is not limited to plant retrotransposons and also exists in animal LTR retrotransposons and Class III retroviruses (Benachenhou et al. 2013). Such terminal sequence was believed to be important for element integration (Temin 1981). Two conserved motifs, called attachment (att) sites, found at the most distal ends of the element (11-12bp) function in the recognition by retroviral integrase and therefore confer specificity in integration (Masuda et al. 1998). Systematic deletions within the att sites affect the capacity of the integrase to recognize the LTRs resulting in inhibited amplification. Particularly, the mutation of the 3' end sequence (CA) significantly impaired the activity of integration of HIV cDNA (Chow et al. 1992). Moreover, a mutation in the integrase sequence can compensate for

mutations in the att site (Du et al. 1997), suggesting a co-evolutionary relationship between integrase and terminal sequences of LTR elements.

Prior to this study, we screened for non-TGCA LTR elements using an automatic tool and revealed their presence in most (42 out of 50) of the sequenced plant genomes (Ou and Jiang 2018). Nevertheless, overall those elements seemed to only account for about 1% of the LTR elements (Ou and Jiang 2018). As a result, it is unclear whether the non-canonical ends represent transient mutations or they can successfully amplify for an extended period. In this study, the improved lotus assembly combined with manual curation of repeat library lead to the finding of 8 different non-canonical ends with 14810 copies and 4.5% of the genome (15.4% of total LTR content or 32.7% of total *Copia* elements, Table 3.2), such abundance of LTR elements with non-canonical ends has not been reported for any organisms. Our analysis indicates that among the two terminal nucleotides on each side of the LTR, only the first nucleotide at the 5' end is not replaceable. All other nucleotides can be substituted without the complete abolishment of the transposition activity. The second nucleotide at the 3' end is the most flexible and can be C, G, or T. The differential constraint at the two ends suggests that they play distinct roles in integration. This is consistent with studies in retroviruses where the integrase recognizes and assembles independently at the att sites on each LTR end (Masuda et al. 1998; Bera et al. 2009).

Thus far, the high copy TGCT LTR covers over 2% of the lotus genome and is the first non-canonical LTR type reported with this level of coverage and copy number. The presence of some elements ending with TGCA among the TGCT clade (Figure S3.3) suggests that this mutation is relatively recent yet the integrase of these elements might have evolved higher specificity for TGCT than to TGCA. If this is the case, the success of TGCT elements could be

explained by the long-term co-evolution between the element and the transposition machinery in the lotus lineage. It is possible that there is an ancient lineage of *Copia* elements coding for an integrase with a higher affinity to TGCT than TGCA ends. If that is the case, one would expect the TGCT elements group phylogenetically with similar elements in other species, but this is not observed (Figure S3.2). Alternatively, the TGCT elements are derived from relatively recent mutations in lotus and somehow have achieved significant success. Our phylogenetic analysis indicates the latter is more likely the case. Unlike the *hAT* elements (Figure S3.2), where most elements group with counterparts in other species, the majority of TGCT LTR families in lotus are closely related to each other, and may represent a specific clade of LTR elements with the frequent formation of non-canonical LTR ends.

For the tomato *TARE1* elements with TATA termini, it was postulated that the change in the LTR sequence was due to a mutation in the 3' LTR end from 'G' to 'A' prior to the transposition of the element (Yin et al. 2013). It is known that the reverse transcription reaction is error-prone so it is not surprising that mutations arise prior to transposition. Moreover, sequence analysis of *Del* retrotransposons shows that the U5 att (found at the 3' end of the element) has a lower sequence identity than the U3 att (found at the 5' end) (Cruz et al. 2014). Among the 8 non-canonical ends reported here, 5 have conserved 5' dinucleotide (TG), four of which are the most abundant end types. This may suggest that the integrase is more amenable to changes at the 3' end (U5 att) compared to the 5' end (U3 att). Since the retroviral integrase recognizes and assembles independently at each att site as a heterodimer, it is likely that the integrase for these non-canonical LTR types (particularly the TGCT type) in lotus has evolved to efficiently recognize a mutated U5 att sequence. The mechanism underlining the success of non-TGCA LTR

elements in louts requires further investigation, yet it is clear that those elements are a stable subtype of the LTR elements instead of transient mutations. Since the non-canonical LTR elements could account for a considerable portion of the genome, they should not be ignored during genome annotation.

The abundance and composition of TEs may shape the direction of evolution

For a certain gene, the probability to gain a TE insertion depends on the following factors: 1) the target the specificity of TEs that are currently active in the genome. For example, if MULEs are active, a highly expressed gene would possibly have a MULE insertion at the 5' end of the gene. If *Gypsy* elements are active, few insertions would land in genic regions. 2) the impact of the insertion. If it is beneficial, it may be retained; otherwise, it is likely to be eliminated. Theoretically, small elements are less detrimental than larger elements. 3) the efficiency of the host genome to "purge" an undesirable insertion out, such as the recombination rate, mating system, etc. For an inbreeding species with a small effective population size, it is possible for a slightly deleterious insertion to spread or be fixed in the population. As a consequence, the abundance and composition of TEs in a genome will determine not only the density of TEs but also which part of the genes they inhabit.

TEs in genic regions could have a variety of genetic and epigenetic impacts on the nearby genes. Insertions in CDS is certainly most detrimental. Even if the insertion is located in a non-coding region, it is often consequential. For example, UTRs may contain elements important for transcription or translation (Juntawong et al. 2014; Srivastava et al. 2018), and a TE insertion may disrupt or dislocate such regulatory elements. TEs upstream (where promoters are located) may influence the transcription of the elements, whereas TEs in introns may interfere with

splicing, even if the insertion is not close to the donor site or acceptor site for splicing. A well-known example is the LINE element called *Karma* in oil palm. This element located in a large intron (about 35 kb in length) of a gene in the flowering pathway (Ong-Abdullah et al. 2015). When *Karma* is hypomethylated, an alternative acceptor site inside the element is effective, which causes mis-splicing of the gene and leads to infertile fruits. Moreover, longer introns are associated with a reduced level of expression (Castillo-Davis et al. 2002), so TE insertions in introns may influence both the quantity and structure of transcripts.

DNA methylation is one of the most important means to control TE activity. In Arabidopsis, there is a negative correlation between gene expression level and the density of methylated TEs (Hollister and Gaut 2009). Accordingly, there is a "trade-off" between control of TE activity and gene expression (Choi and Lee 2020). If we assume in general there is a negative impact on the function/expression of genes with more TE insertions, it appears that genes in different biological processes are influenced to a different degree. It is not surprising to observe that TEs insertion are underrepresented in genes involved in fruit ripening, photosynthesis, and response to various stimuli, in that these processes are critical for the survival and reproduction of the organism. Genes in fruit ripening are associated with much fewer insertions than genes in all other processes (Table 3.4, Table 3.5), and it is conceivable that if any insertion results in any negative impact in this process, the relevant insertion is unlikely to pass to the next generation. In contrast, genes involved in DNA metabolic process, regulation of gene expression, epigenetic, and cell cycle, are probably less critical for survival, so the insertions are more likely retained. Alternatively, some features (such as expression level) of those genes make them more attractive for TE targeting. The two possibilities are not mutually exclusive. It is ironic that genes involved

in epigenetic pathways and DNA metabolic process are the key to genome integrity, yet it seems they are poor “sentinels” for themselves. If TEs preferentially target those genes, it implies that TEs are not passively being controlled by the genome surveillance machine, but actively attack the machine to achieve their prosperity. On the other hand, if selection allows those genes to be vulnerable to TE insertion, it may suggest the activity of TEs provides benefits to the organism or the cost of controlling TE activity is over the benefit of such action, such as the negative impact on gene expression. From this point of view, the interaction between TEs and genes in the epigenetic pathway may represent a self-correction process to maintain the activity of silencing at an appropriate level.

MATERIALS AND METHODS

Construction of repeat library

The construction of the lotus repeat library was initialized using a previous version of lotus assembly (Ming et al. 2013) and supplemented with the latest assembly when it was available (Gui et al. 2018). Repetitive sequences were mined using a variety of approaches. LTR retrotransposons were collected using LTR_retriver (Ou and Jiang 2018), and all the sequences for all non-TGCA LTR elements were manually verified for its boundary, terminal sequences, and target site duplications (TSDs) before integrated into the library.

Non-autonomous DNA elements were mined using the MITE-Hunter package with parameters as recommended (Han and Wessler 2010). The sequences of exemplars of LTR elements, non-autonomous DNA elements were then used to mask the genomic sequence using RepeatMasker (http://www.repeatmasker.org/) and the repetitive sequences in the unmasked portion of the genomic DNA were further identified in a second mining step using RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html). The output of RepeatModeler contains both known and unknown repeats. The resulting sequences were first filtered to remove putative gene families using BLASTX and sequences matching non-TE genes proteins ($E < 10^{-5}$) were removed. The remaining sequences where the genome coverage is ≥ 0.05% were manually curated to determine their identity and 5' and 3' boundaries. This was done in a stepwise process. First, the relevant sequences collected through RepeatModeler were used to search and retrieve at least 10 hits (BLASTN, $E < 10^{-10}$) with the corresponding 100 bp of 5' and 3' flanking sequences. Second, recovered sequences were aligned using DIALIGN2 (Morgenstern 1999), to determine the possible boundary between elements and their flanking

sequences. In this case, a boundary was defined as the position to which sequence homology stops for over half of the aligned sequences. Finally, sequences with defined boundaries were examined for the presence of TSDs. To classify the relevant TEs, features in the terminal ends and TSD were used. Each transposon family is associated with distinct features in their terminal sequences and TSD, which can be utilized to identify the element (Wicker et al. 2007). The identification of putative autonomous elements was assisted by their homology to known transposase from Repbase (Bao et al. 2015).

Manually curated TE sequences and that collected through LTR_retriever and MITE-Hunter were merged into a single repeat library for subsequent analysis.

Estimation of copy number and genomic fraction

The lotus repeat library was used to mask the genomic sequence to determine TE coverage and copy number. If an element in the genomic sequence matched a sequence in the repeat library over the entire sequence, or if the truncation was less than 20 bp on both ends, this copy was considered to be intact. If the element contains one end (truncation less than 20 bp) it was considered as a truncated sequence or half of a copy. If no end was detected, it was considered as a fragment, and the copy number was estimated by comparing the length of the fragment to the full length of the element. For example, if a fragment of the element was 200 bp, and the intact element is 1 kb, this fragment was considered to be 0.2 copy. The genome fraction of TEs was estimated using the total sequence masked by each superfamily with overlapping regions between different entries only calculated once.

Phylogenetic analysis

To search for *hAT* elements encoding transposase, curated sequences for autonomous *hAT* families were used to search the genome. Copies that are truncated by no more than 15 bp on each end of the element or are over 2.5 Kb in length were retained and compared against a database containing known *hAT* transposase to identify elements containing the conserved motif 3 (Kempken and Windhofer 2001; Lazarow et al. 2012). For LTR elements, the conserved integrase core domain, which contains the RNase H fold catalytic motif, of representative LTR elements from each family was retrieved.

Sequences of conserved integrase core domain from LTR elements and motif 3 from *hAT* transposase were used to generate multiple alignments and resolved into lineages by generating phylogenetic trees. Multiple sequence alignment was performed by ClustalW (http://www.ebi.ac.uk/clustalw) with default parameters. Phylogenetic trees were generated using the maximum-likelihood method. Support for the internal branches of the phylogeny was assessed using 500 bootstrap replicates using MEGA (http://www.megasoftware.net). The EST sequences used in this study were from a previous study (Ming et al. 2013).

The abundance of insertion density and fraction in genic regions

The gff file for gene annotation of current assembly was downloaded from https://genomevolution.org/coge/GenomeInfo.pl?gid=35393 (accessed 11/19/2019) (Gui et al. 2018). The coordinates from the gff file were used to classify different genomic regions including 1kb upstream of genes (1 kb upstream of TSS), 5′ UTR, exons, introns 3′ UTR, and 1kb downstream (1 kb downstream of TTS). If a gene was associated with multiple models, only mRNA1 was used.

Based on the gff file, sequences were extracted and concatenated into complete CDS sequences for individual genes. The CDS sequences were masked using RepeatMasker with the lotus repeat library (see above). If 50% or more of the CDS sequence was masked, the relevant gene was considered a TE and excluded from subsequent analysis. The original gff file contained a total of 23,810 protein coding genes. Among these, 944 were considered as TEs and excluded, so the final non-TE gene dataset contains 22,866 genes.

The copy number and length of TEs in each genic region were obtained by comparison between the coordinates of the genic regions and the coordinates of the TE in the genome, based on the RepeatMasker output of the assembly. If a TE was located at the boundary of two regions, the copy number was calculated based on the length distribution on each site. For example, if a TE that is 1 kb in length is located at the boundary between an intergenic region and the adjacent upstream region, with 200 bp inside the upstream region, it was considered 0.2 copy for the upstream region and 0.8 copy for the intergenic region. The fraction of TE in each region was calculated by the total TE length divided by the total length of the relevant region. The insertion density was calculated by the total copy number in each region divided by the total length and expressed as insertions per 100 kb. The expected insertions in a region were calculated using the average density of a superfamily multiplied by the length of the region. The observed insertions were compared with expected insertions using multiple comparison testing with a Bonferroni correction at the 0.01, 0.001, and 0.0001 significance level. The percent difference from expected (enrichment index) was calculated using (observed insertions − expected insertions) ÷ (expected insertions) × 100.

The enrichment and depletion of TE insertions in genes involved in different biological processes

Using the CDS sequence of the 22,866 non-TE genes a pipeline was developed in Blast2GO to assign GO terms (Conesa et al. 2005). The CDS sequences were searched using BLAST with default setting (E value = 0.001), and the top 3 hits were retained using the UNIProt protein database. GO-Slim plant terms were assigned based on that of the top 3 hits, resulting in 18,335 genes with GO-Slim plant GO terms. The number of TE insertions (observed value) for a certain GO term was obtained by counting the TEs of all genes in this category. Thereafter, the number was compared with the expected value, which was derived from the average values of the 18,335 genes multiplied by the number of genes in this GO category. The significance of the difference between the observed value and the expected value was tested using Chi-squared multiple comparison testing at $p < 0.01$ using a Bonferroni correction.

ACKNOWLEDGEMENTS

We wish to thank Ann Ferguson for her work on a previous version of this manuscript. This study was supported by the National Science Foundation [MCB-1121650 and IOS-1740874 to N.J.]; United States Department of Agriculture National Institute of Food and Agriculture and AgBioResearch at Michigan State University [Hatch grant MICL2707 to N.J.].

APPENDIX

FIGURES

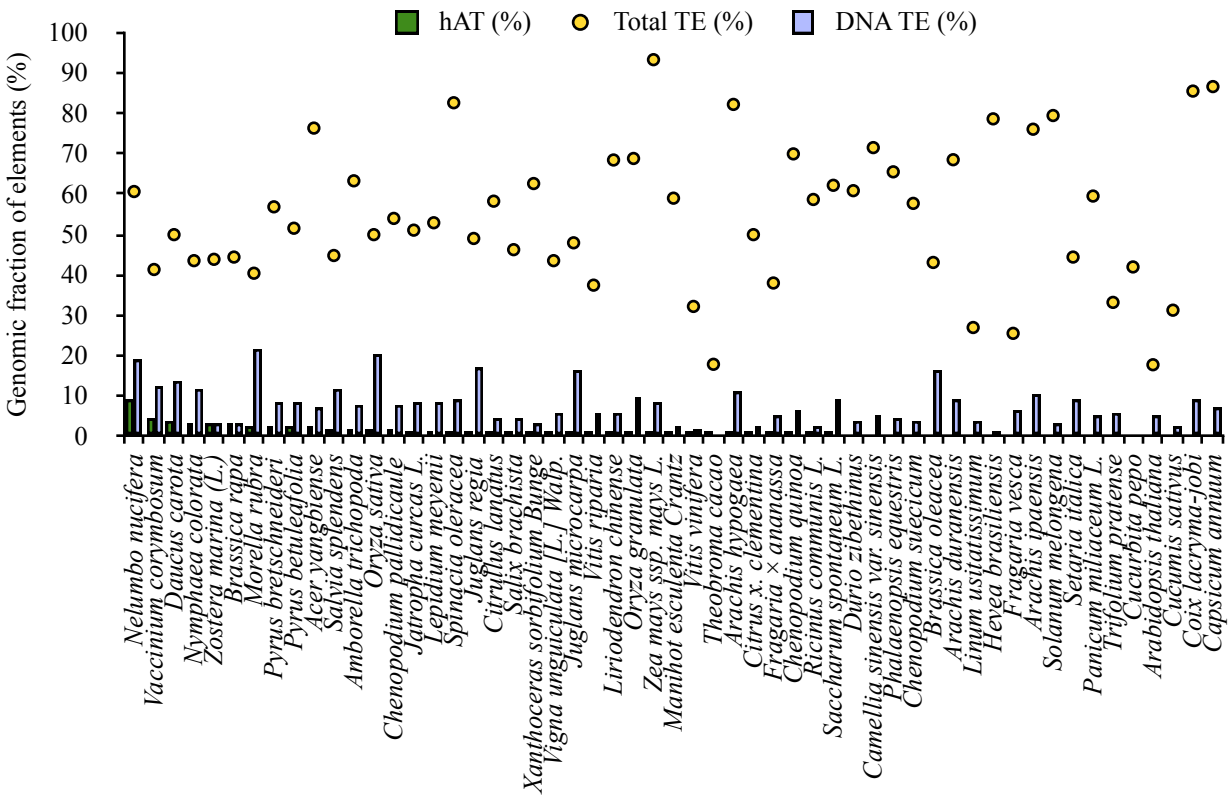


Figure 3.1. Abundance of *hAT* elements and other TEs in sequenced plant genomes. Percentage of *hAT* elements, dark green bars, percentage of DNA elements, light blue bars, and total transposon content (TE) in the genome, yellow circles. Plant genomes with greater than 0.5% hAT content shown, ordered from highest (left) to least (right) *hAT*% content.

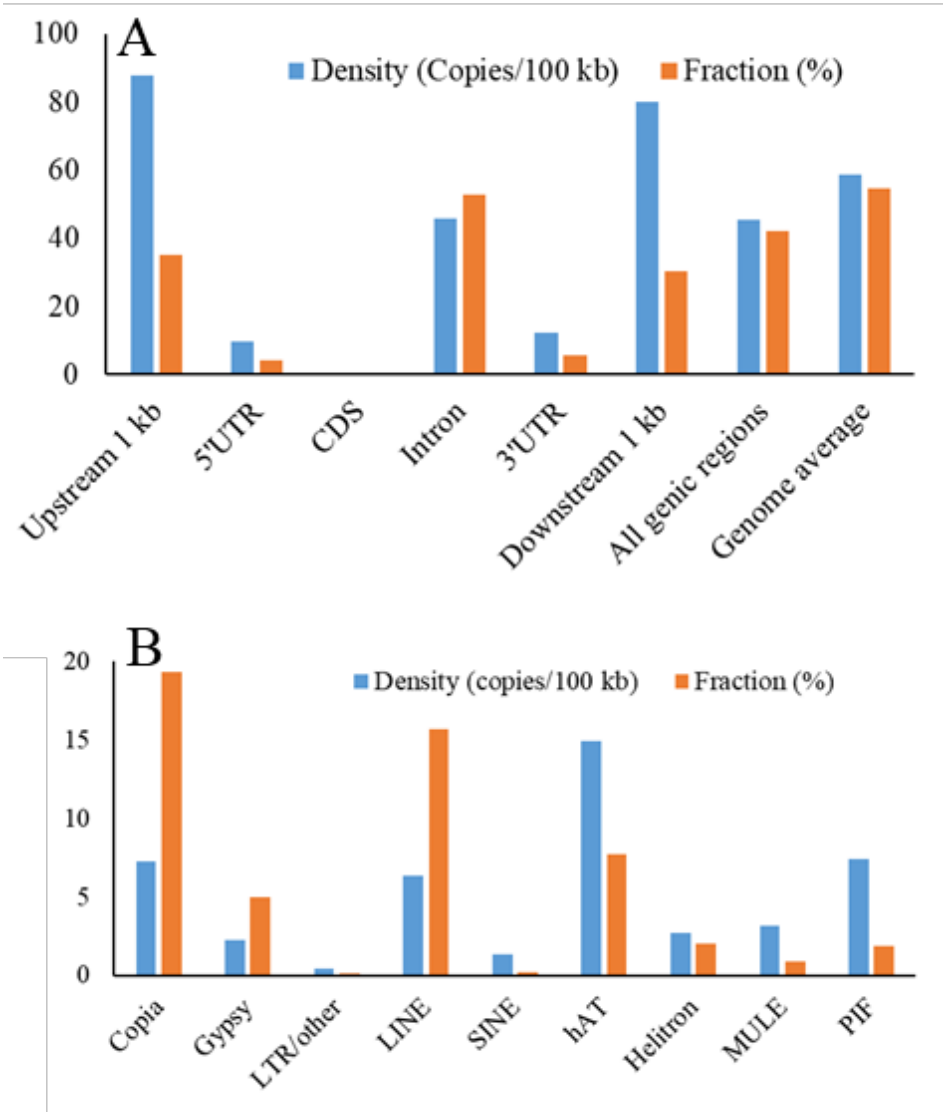
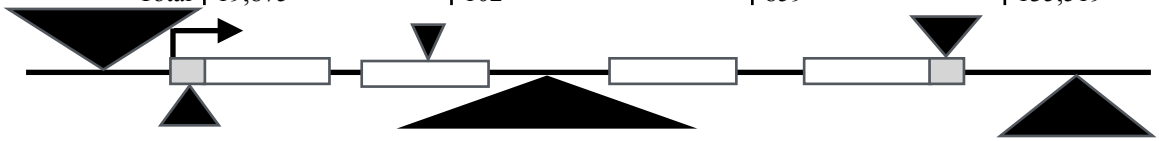


Figure 3.2. TE abundance indicated as insertion density and genome fraction in different genic regions A. Overall TE density and genome fraction in different genic regions and compared with genome-wide average. B. The insertion density and intron fraction of different superfamily of TEs in introns.

Superfamily	A. Upstream 1 kb		C. CDS		E. 3′ UTR		G. Total	
LTR/*Copia*	230***	-83	11***	-99	35***	-92	15,850***	-14
LTR/*Gypsy*	767***	-41	16***	-99	31***	-93	6,070***	-65
LTR/other	81 n/s	-23	0***	-100	12 n/s	-65	1,152***	-17
LINE	101***	-82	9***	-99	55***	-70	13,665***	84
SINE	245 n/s	0	4***	-99	4***	-95	3,517***	9
CACTA	10 n/s	-52	1***	-97	4 n/s	-41	72***	-74
hAT	5,921***	49	16***	-100	152***	-88	43,044***	-18
Helitron	730 n/s	-7	2***	-100	66***	-74	7,199***	-31
MULE	7,470***	236	27***	-99	185***	-74	18,891***	-36
PIF-Harbinger	4,320***	59	16***	-100	315***	-64	26,059***	-27
Total	19,875		102		859		135,519	



Superfamily	B. 5′ UTR		D. Intron		F. Downstream 1kb	
LTR/*Copia*	14***	-96	15,145***	18	415***	-70
LTR/*Gypsy*	36***	-90	4,692***	-61	528***	-60
LTR/other	3**	-90	893 n/s	-8	163***	55
LINE	19***	-88	13,281***	156	200***	-64
SINE	19***	-72	2,858***	26	387***	58
CACTA	1 n/s	-82	22***	-89	34 n/s	62
hAT	93***	-91	31,247**	-15	5,615***	42
Helitron	13***	-94	5,657***	-22	731 n/s	-7
MULE	285***	-53	6,666***	-68	4,258***	92
PIF-Harbinger	114***	-85	15,530***	-38	5,764***	113
Total	597		95,991		18,095	

Figure 3.3. Abundance and degree of enrichment of individual TE super-families in different genic regions

Boxes A-G are tables of TEs in different genic regions with name of super-families in the far left. For each box, the first column of numbers indicates the number of insertions in each superfamily and the total number of insertions at the bottom. These insertion numbers were compared to expected insertions (based on genome average density) and tested using Chi-squared Bonferroni correction method. P-values are denoted using *0.01, **0.001 and ***<0.0001. The second column of numbers indicates enrichment index, which is the percent difference from expected values (genome average). In the center is a genic region schematic with exons in white boxes, UTRs depicted in grey boxes, black triangles representing transposons and the arrow denoting the transcription start site.

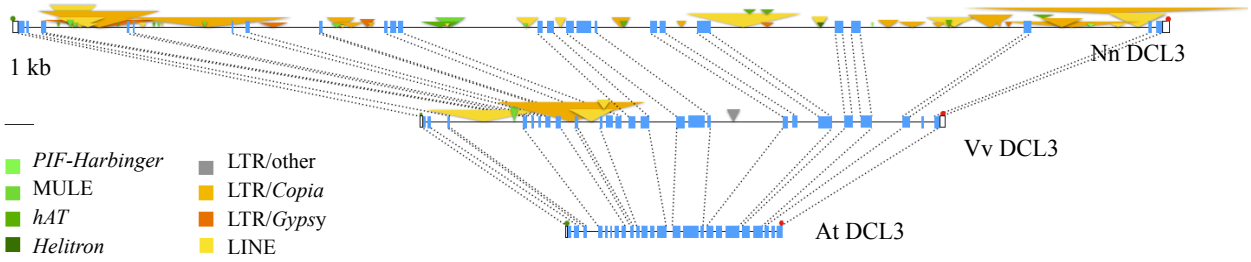


Figure 3.4. Comparison of the structure of DCL3 genes in Arabidopsis, grape and sacred lotus
Top, Nn (*Nelumbo nucifera*) DCL3 (111 kb), bottom, AT (*Arabidopsis thaliana*) DCL3 (7.3 kb, AT3G43920, TAIR10), middle, Vv (*Vitus vinifera*) DCL3 (28.6 kb, GenBank accession No. NC_012010.3, 1757859 – 1786478, GeneID:100254311). Eleven kb of non-TE sequences in the 21st intron of NnDCL3 not shown due to space limit. Blue boxes are exons, white boxes are UTRs, triangles are transposons denoted by color. Grey LTR/other, Dark orange LTR/*Copia*, orange red LTR/*Gypsy*, and LINE light orange. DNA transposon insertions, *PIF-Harbinger* green white, MULE light green, *hAT* green, and *Helitron* dark green. Triangles stacked on top of other triangles signifies a nested insertion. Green and red octagons signify transcription start and stop sites respectively. Dash lines of blue exons denote regions of homology between *Vitus vinifera* and others.

TABLES

Table 3.1. The abundance of different families of TEs in the genome of sacred lotus

Class	**Subclass**	**Superfamily**	**Length (Mb)**	**Average Element length (kb)**	**Copy Number**	**Genomic fraction (%)**	**Percent of total copy number (%)**
Class I	LTR	LTR/*Copia*	96.17	2.23	43036	13.75	10.45
		LTR/*Gypsy*	107.50	2.65	40495	15.37	9.84
		LTR/other	1.03	0.32	3262	0.15	0.79
	non-LTR	LINE	43.31	2.49	17382	6.19	4.22
		SINE	1.29	0.17	7579	0.19	1.84
		Total Class I	249.31	2.23	111754	35.64	27.15
Class II	TIR	CACTA	0.85	1.31	649	0.12	0.16
		hAT	62.95	0.51	122488	9.00	29.76
		MULE	19.41	0.28	68712	2.78	16.69
		PIF-Harbinger	24.96	0.30	83738	3.57	20.34
	non-TIR	*Helitron*	24.52	1.01	24314	3.51	5.91
		Total Class II	132.70	0.44	299901	18.97	72.85
Total TE			382.01	0.93	411655	54.61	100

Table 3.2. The abundance of *Copia* LTR retrotransposons with different termini in sacred lotus

Terminal sequence	Average LTR length (bp)	No. of families	Copy number	Genome fraction (%)
TGCA	431	146 (57.71)	28226 (65.59)	9.25 (67.27)
TGCT	313	31 (12.25)	8869 (20.61)	2.58 (18.73)
TGTA	306	17 (6.72)	1469 (3.41)	0.51 (3.69)
TACA	371	20 (7.91)	1210 (2.81)	0.38 (2.79)
TGGT	343	7 (2.77)	1001 (2.33)	0.36 (2.63)
TGGA	250	6 (2.37)	1296 (3.01)	0.33 (2.41)
TACT	316	12 (4.74)	649 (1.51)	0.23 (1.67)
TATA	221	12 (4.74)	282 (0.65)	0.09 (0.69)
TGTT	326	1 (0.40)	35 (0.82)	0.02 (0.11)
Total non-TGCA	312	106 (41.90)	14810 (34.41)	4.50 (32.73)

Number in parenthesis indicate the percent of *Copia* elements

Table 3.3. Summary of distribution preference of TEs in the genome of sacred lotus[a]

Super-family	Preference	Bias against
LTR/*Copia*	Introns (weak)	Upstream (very strong), downstream (strong)
LTR/*Gypsy*	Intergenic regions (strong)	Genic regions (strong)
LTR/other	Downstream (moderate)	N/A
LINE	Introns (strong)	Non-intron regions (strong)
SINE	Downstream (moderate), intron (weak)	N/A
hAT	Upstream and downstream (weak)	Introns (weak)
Helitron	NA	Introns (weak)
MULE	Upstream (very strong), downstream (moderate)	Introns (strong)
PIF	Downstream (strong), upstream (moderate)	Introns (weak)

[a]The copy number of CACTA elements is too low to evaluate its preference. All TEs are significantly underrepresented in regions corresponding to mature mRNAs. The strength of preference or bias is based on enrichment value (Figure 3). Preference: 10 – 50 (weak), 50 – 100 (moderate), 100 – 200 (strong), > 200 (very strong). Bias against: -10 to -25 (weak), -25 to -50 (moderate), -50 to -75 (strong), -75 to -100 (very strong).

Table 3.4. Ten biological processes in which genes are associated with the highest and lowest TE insertion densities in the entire genic region

GO Term	GO Term ID	Number of Genes	Observed Insertions	Expected Insertions	P-value*	Insertions per Gene
Depleted GO Terms						
fruit ripening	GO:0009835	47	94	292	5.39×10^{-31}	2.00
abscission	GO:0009838	74	278	459	2.65×10^{-17}	3.76
secondary metabolic process	GO:0019748	606	2735	3762	6.88×10^{-63}	4.51
response to endogenous stimulus	GO:0009719	3106	16449	19280	2.19×10^{-92}	5.30
flower development	GO:0009908	1077	5806	6685	5.73×10^{-27}	5.39
growth	GO:0040007	435	2401	2700	8.56×10^{-9}	5.52
photosynthesis	GO:0015979	253	1404	1570	2.67×10^{-5}	5.55
response to chemical	GO:0042221	5209	29729	32334	1.51×10^{-47}	5.71
response to light stimulus	GO:0009416	1427	8235	8858	3.67×10^{-11}	5.77
response to abiotic stimulus	GO:0009628	2378	13912	14761	2.81×10^{-12}	5.85
Enriched GO terms						
regulation of molecular function	GO:0065009	504	4032	3128	1.06×10^{-58}	8.00
transport	GO:0006810	3101	25235	19249	0	8.14
cellular homeostasis	GO:0019725	442	3620	2744	7.71×10^{-63}	8.19
cellular component organization	GO:0016043	4030	33049	25015	0	8.20
cell-cell signaling	GO:0007267	175	1455	1086	4.68×10^{-29}	8.31
protein metabolic process	GO:0019538	683	5891	4240	6.43×10^{-142}	8.63
cell cycle	GO:0007049	1102	9642	6840	1.61×10^{-251}	8.75
regulation of gene expression, epigenetic	GO:0040029	336	3485	2086	3.42×10^{-206}	10.37
DNA metabolic process	GO:0006259	692	7667	4295	0	11.08

*Chi-square multiple comparison tested at $p < 0.01$ with Bonferroni correction.

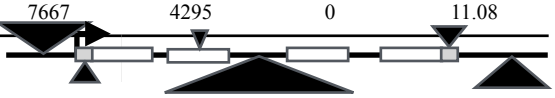


Table 3.5. Ten biological processes in which genes are associated with the highest and lowest TE insertion densities in introns

GO Term	GO Term ID	Number of Genes	Average Intron Number	Observed Insertions	Expected Insertions	P-value*	Insertions per Gene
Depleted GO Terms							
fruit ripening	GO:0009835	38	5.11	30	214	3.38×10^{-36}	0.79
abscission	GO:0009838	58	5.69	167	326	1.32×10^{-18}	2.88
secondary metabolic process	GO:0019748	499	4.09	1592	2804	5.57×10^{-116}	3.19
photosynthesis	GO:0015979	227	6.12	944	1276	1.59×10^{-20}	4.16
response to endogenous stimulus	GO:0009719	2548	6.18	11083	14319	4.53×10^{-161}	4.35
flower development	GO:0009908	925	6.27	4087	5198	1.34×10^{-53}	4.42
metabolic process	GO:0008152	748	5.65	3443	4204	8.85×10^{-32}	4.60
growth	GO:0040007	356	6.97	1662	2001	3.71×10^{-14}	4.67
response to abiotic stimulus	GO:0009628	2030	6.18	9524	11408	1.22×10^{-69}	4.69
response to chemical	GO:0042221	4307	6.24	20333	24204	1.13×10^{-136}	4.72
Enriched GO terms							
cellular protein modification process	GO:0006464	2376	7.56	15366	13353	5.36×10^{-68}	6.47
transport	GO:0006810	2855	7.59	19298	16044	1.65×10^{-145}	6.76
regulation of molecular function	GO:0065009	445	7.37	3061	2501	3.96×10^{-29}	6.88
cellular homeostasis	GO:0019725	401	7.39	2766	2254	3.61×10^{-27}	6.90
cellular component organization	GO:0016043	3652	7.88	25257	20523	1.98×10^{-239}	6.92
cell-cell signaling	GO:0007267	157	8.31	1129	882	9.95×10^{-17}	7.19
cell cycle	GO:0007049	1030	9.86	7508	5788	4.05×10^{-113}	7.29
protein metabolic process	GO:0019538	607	7.82	4550	3411	1.13×10^{-84}	7.50
regulation of gene expression, epigenetic	GO:0040029	304	9.14	2793	1708	9.16×10^{-152}	9.19
DNA metabolic process	GO:0006259	649	10.32	6283	3647	0	9.68

*Chi-square multiple comparison tested at $p < 0.01$ with Bonferroni correction.

SUPPLEMENTARY FIGURES/TABLES

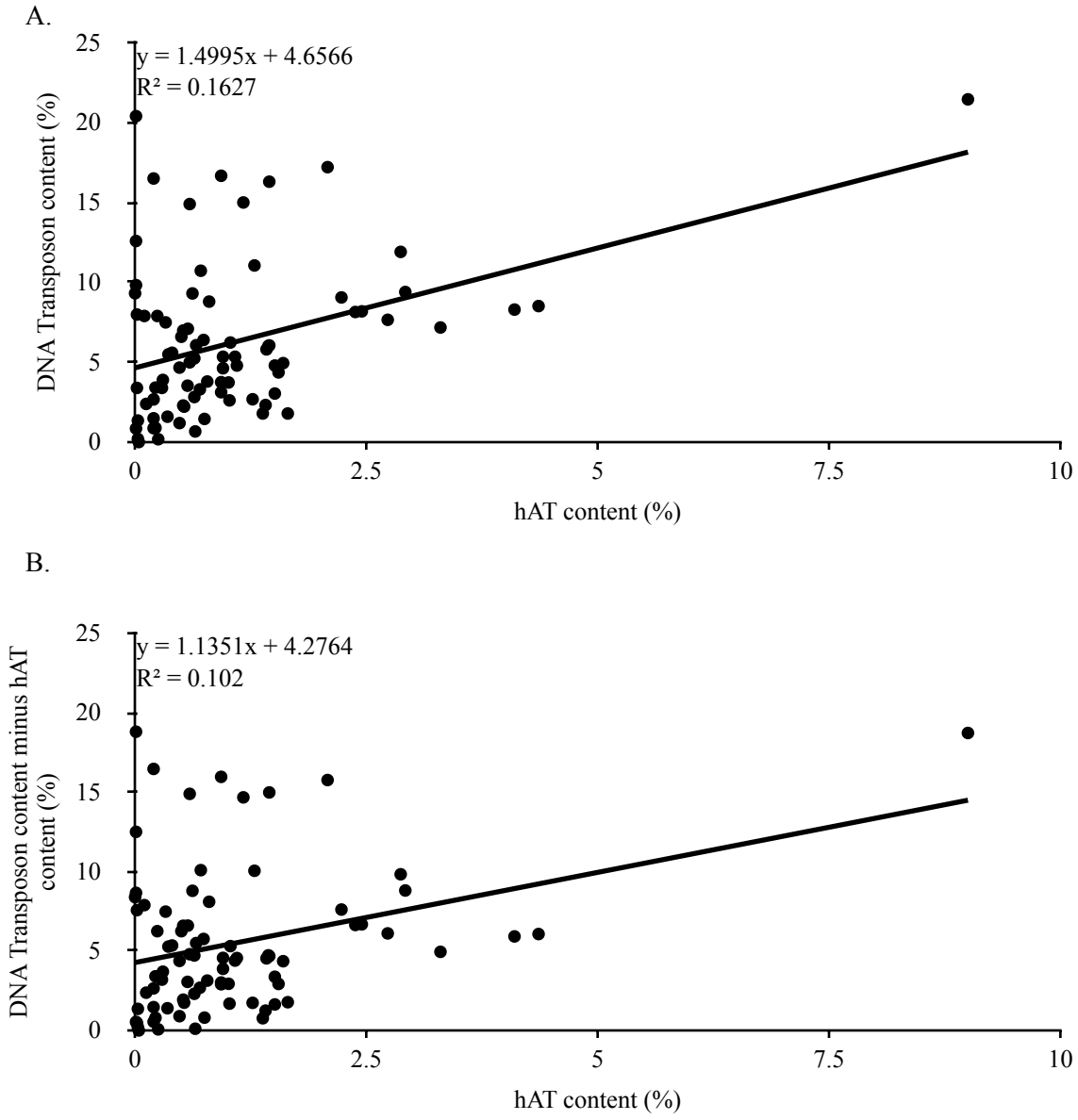


Figure S3.1. Correlation between abundance of DNA transposons and other DNA transposons. A. Genome fraction (%) of hAT elements vs. total DNA transposon content. N= 89, Pearson Correlation R = 0.4612 $p < 0.00001$. B. Genome fraction of hAT (%) vs. DNA Transposon content minus hAT content (%). N = 89, Pearson Correlation R = 0.2331 $p < 0.0288$.

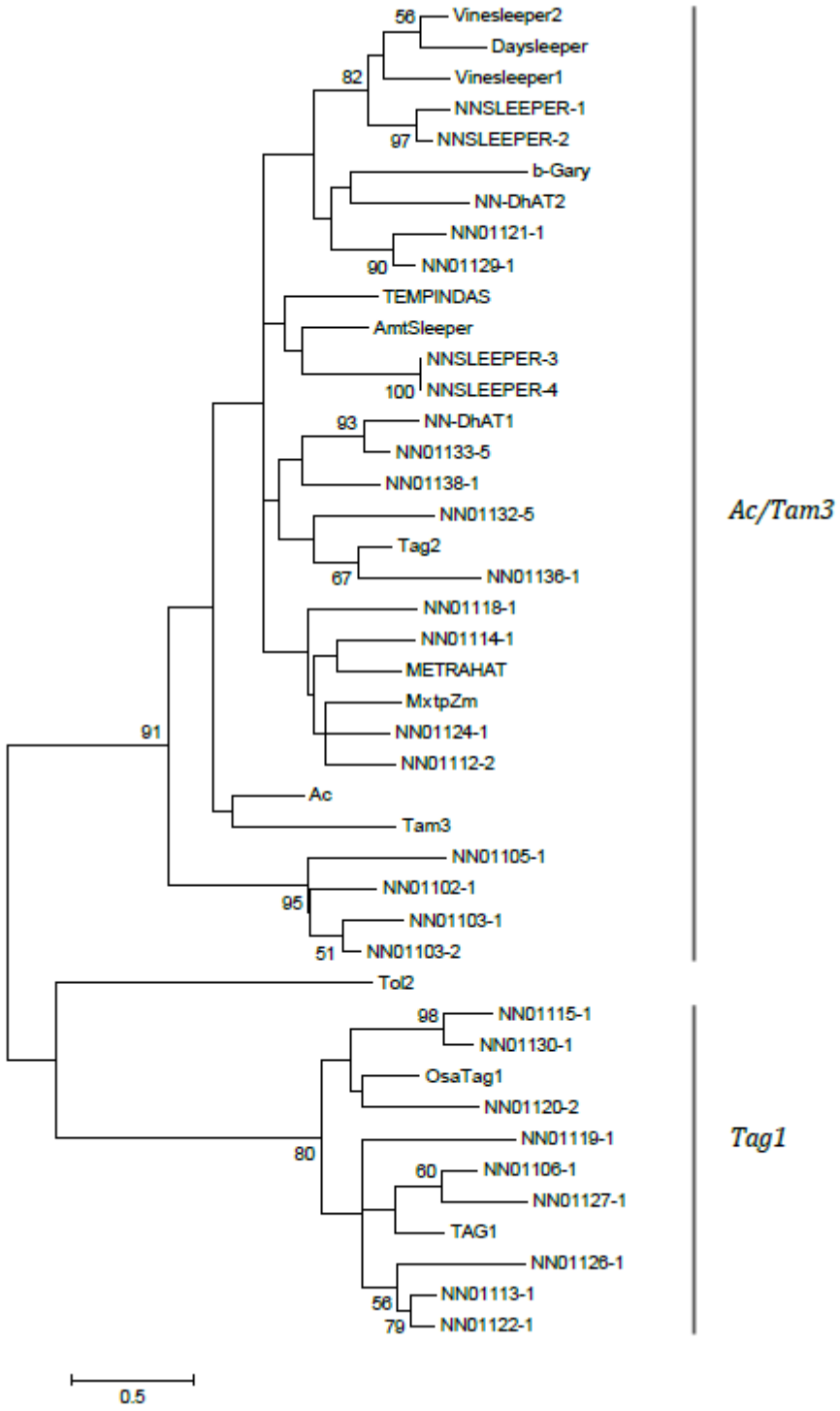


Figure S3.2. The phylogeny of motif-3 of *hAT*-like transposase in sacred lotus and other organisms. Maximum likelihood tree. Numbers next to branch indicate the % bootstrap support. NN -*Nelumbo nucifera.* Osa -*Oryza sativa.*

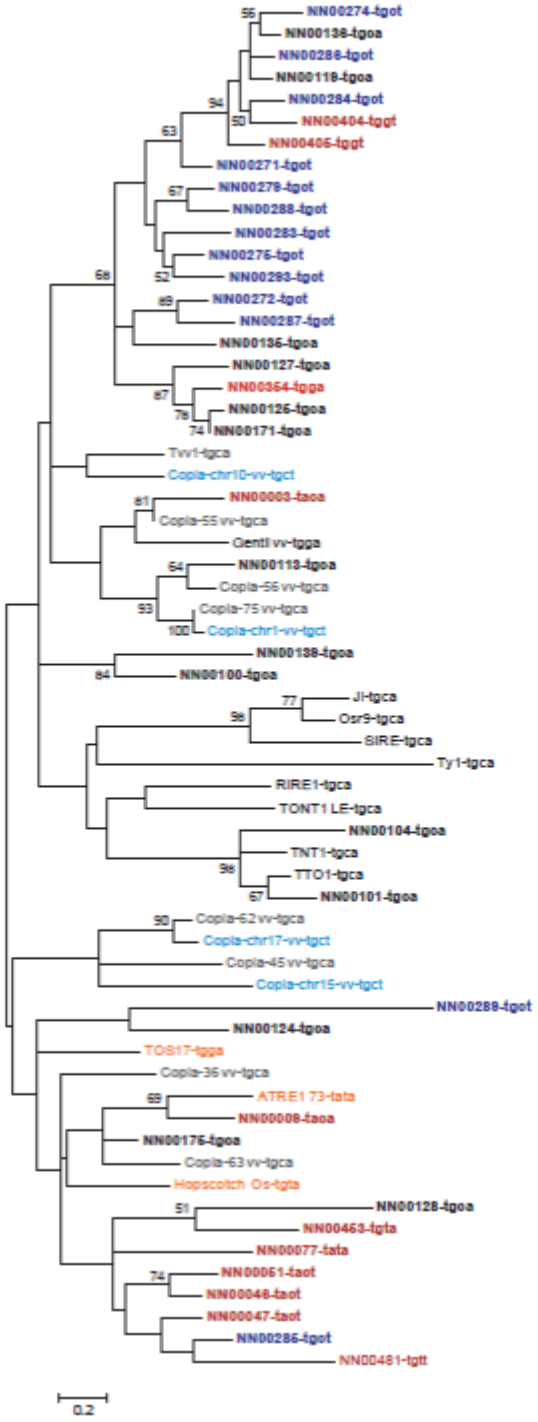


Figure S3.3. The phylogeny of *Copia* elements with different termini in sacred lotus and other organisms. Maximum likelihood tree. Numbers next to branches indication bootstrap support %. Letters following dash represent nucleotide termini.

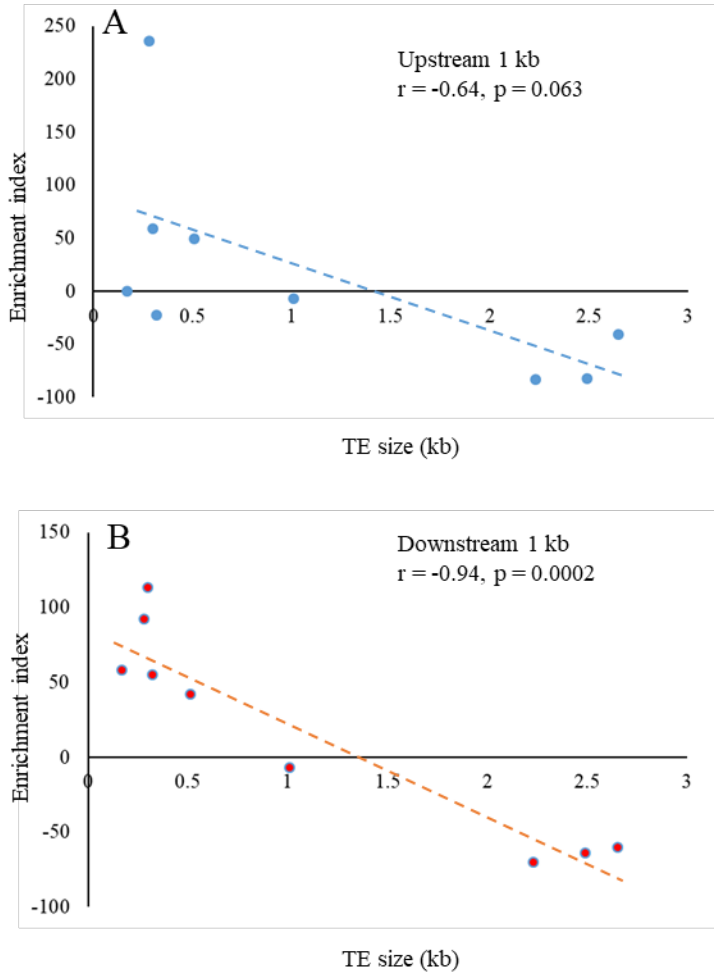


Figure S3.4. The relationship between the average size of each superfamily of TEs and the degree of enrichment in upstream (A) and downstream (B) regions of protein coding genes.

Table S3.1. Retrotransposon information of various sequenced plant genomes

Common name	*Scientific name*	*Copia* (%)	*Gypsy* (%)	*Gypsy: Copia* ratio	Non-LTR RT (%)	References[b]
DICOTS						[b]see Table S2 for references except for below
maple	*Acer yangbiense*	26.98	9.94	0.37	2.91	
carrot	*Daucus carota*	17.30	9.90	0.57	3.60	
water lily	*Nymphaea colorata*	7.95	4.91	0.62	4.84	
Savoy cabbage	*Brassica oleacea*	11.64	7.84	0.67	0.80	
red clover	*Trifolium pratense*	12.22	9.76	0.80	1.83	
flax	*Linum usitatissimum*	9.79	8.31	0.85	2.22	
maca	*Lepidium meyenii*	9.25	8.51	0.92	3.49	
yellowhorn	*Xanthoceras sorbifolium Bunge*	11.91	11.68	0.98	3.83	
cushion willow	*Salix brachista*	9.15	9.44	1.03	1.68	
spinach	*Spinacia oleracea*	25.27	27.55	1.09	2.92	
canola	*Brassica rapa*	4.85	5.34	1.10	3.28	
sacred lotus	***Nelumbo nucifera***	**13.75**	**15.37**	**1.12**	**6.38**	
strawberry	*Fragaria vesca*	5.33	6.39	1.20	0.45	
zucchini	*Cucurbita pepo*	19.80	24.20	1.22	1.71	
sweet orange	*Citrus sinensis*	7.84	9.77	1.25	0.40	
lettuce	*Lactuca sativa*	24.27	33.11	1.36	0.86	
hybrid walnut	*Juglans regia*	5.62	7.72	1.37	7.02	
barrel medic	*Medicago truncatula*	4.10	5.70	1.39	2.30	
hybrid walnut	*Juglans microcarpa*	6.05	8.90	1.47	5.83	
cowpea	*Vigna unguiculata [L.] Walp.*	16.67	25.00	1.50	0.40	
grape	*Vitis riparia*	8.33	12.66	1.52	3.61	
clementine	*Citrus x. clementina*	7.88	12.01	1.52	1.24	
thale cress	*Arabidopsis thaliana*	3.56	5.74	1.61	1.49	
sugar beet	*Beta vulgaris*	6.07	10.11	1.67	5.67	
flax	*Linum usitatissimum*	4.69	7.84	1.67	3.09	
wild pear	*Pyrus betuleafolia*	11.60	19.90	1.72	1.68	
adzuki bean	*Vigna angularis*	9.96	18.98	1.91	0.01	
red bayberry	*Morella rubra*	5.13	9.92	1.93	3.19	
scarlet sage	*Salvia splendens*	7.92	18.15	2.29	3.35	
soybean	*Glycine max*	12.47	29.52	2.37	0.25	
cotton	*Gossypium raimondii*	8.80	21.50	2.44	3.00	
goosefoot	*Chenopodium pallidicaule*	7.01	18.67	2.66	1.62	
common bean	*Phaseolus vulgaris L.*	9.37	25.12	2.68	2.70	
sunflower	*Helianthus annuus L.*	9.19	26.57	2.89	0.18	
grape	*Vitis vinifera*	6.12	17.96	2.93	0.82	
east asian tulip tree	*Liriodendron chinense*	13.08	40.45	3.09	1.70	
rubber tree	*Hevea brasiliensis*	12.73	39.79	3.13	1.50	
tomato	*Solanum lycopersicum*	6.30	19.70	3.13	0.90	
quinoa	*Chenopodium quinoa*	8.22	28.46	3.46	8.42	
poplar	*Populus trichocarpa (Torr. & Gray)*	1.79	6.96	3.89	0.54	
potato	*Solanum tuberosum*	3.80	15.20	4.00	1.00	

Table S3.1 (cont'd)						
blueberry	*Vaccinium corymbosum*	2.84	11.83	4.17	2.37	
castor bean	*Ricinus communis L.*	4.49	19.16	4.27	0.09	
strawberry	*Fragaria × ananassa*	2.69	11.94	4.44	0.85	
apple	*Malus × domestica Borkh.*	6.72	30.98	4.61	7.95	
papaya	*Carica papaya*	5.50	27.80	5.05	1.10	
n/a	*Chenopodium suecicum*	4.21	23.26	5.52	2.75	
tea	*Camellia sinensis var. sinensis*	8.24	45.85	5.56	2.31	
rubber tree	*Hevea brasiliensis*	7.65	43.36	5.67	0.37	
orchid	*Phalaenopsis equestris*	6.95	39.66	5.71	7.75	
peanut	*Arachis duranensis*	3.17	18.14	5.72	7.91	
peanut	*Arachis ipaensis*	2.98	18.56	6.23	11.86	
eggplant	*Solanum melongena*	7.52	48.43	6.44	2.96	
cassava	*Manihot esculenta Crantz*	4.87	31.91	6.55	1.19	
tea	*Camellia sinensis var. assamica*	6.02	44.27	7.35	1.60	
peanut	*Arachis hypogaea*	4.72	36.93	7.82	2.82	
ginseng	*Panax ginseng*	6.09	49.10	8.06	1.31	
durian	*Durio zibethinus*	3.20	26.20	8.19	1.32	
hot pepper	*Capsicum annuum*	6.00	51.10	8.51	0.77	
cotton	*Gossypium arboreum*	5.50	55.80	10.62	1.20	
Jaltomata	*Jaltomata sinuosa*	2.34	48.72	20.82	1.51	
MONOCOTS						
African oil palm	*Elaeis guineensis*	33.0	8.0	0.24	2.0	Al-Dous et al. 2011
banana	*Musa acuminata*	25.58	11.45	0.45	5.41	Singh et al. 2013
date palm[a]	*Phoenix dactylifera*	3.1	1.4	0.45	n/a	D'Hont et al. 2012
adlay	*Coix lacryma-jobi*	29.50	35.30	1.20	0.80	
ryegrass	*Lolium perenne*	3.87	5.98	1.55	6.12	
common eelgrass	*Zostera marina (L.)*	20.00	32.00	1.60	n/a	
sugarcane	*Saccharum spontaneum L.*	13.48	24.97	1.85	0.34	
maize	*Zea mays ssp. mays L.*	23.70	46.40	1.96	1.00	
barley	*Hordeum vulgare*	8.46	17.96	2.12	0.53	
wheat	*Triticum aestivum*	16.70	46.70	2.80	0.91	
foxtail millet	*Setaria italica*	7.18	22.14	3.08	1.98	
Brachypodium	*Brachypodium distachyon*	4.86	16.05	3.30	1.94	
sorghum	*Sorghum bicolor*	5.18	19.00	3.67	0.04	
African rice	*Oryza glaberrima*	2.70	10.41	3.86	1.53	
rice	*Oryza sativa*	3.60	15.50	4.31	1.90	
upland rice	*Oryza granulata*	5.58	37.83	6.78	0.22	
broomcorn millet	*Panicum miliaceum L.*	5.28	38.37	7.27	1.54	
BASAL ANGIOSPERM						
Amborella	*Amborella trichopoda*	13.65	25.34	1.86	7.55	
n/a	*Chara braunii*	0.00	24.00	n/a	7.70	
GYMNOSPERM						
Douglas-fir	*Pseudotsuga menziesii var. menziesii*	11.80	24.80	2.10	n/a	
Norway spruce[a]	*Picea abies*	16.00	35.00	2.19	1.00	
ginko	*Ginkgo biloba*	12.71	45.63	3.59	n/a	

Table S3.1 (cont'd)						
n/a	*Gnetum montanum*	7.45	64.67	8.68	5.92	
BRYOPHYTE						
C-ferm	*Ceratopteris richardii*	16.50	7.50	0.45	1.62	
common liverwort	*Marchantia polymorpha*	1.00	3.00	3.00	1.13	
LYCOPHYTE						
spikemoss	*Selaginella moellendorffii*	2.70	21.10	7.81	n/a	
resurrection plant	*Selaginella lepidophylla*	0.58	9.52	16.41	3.55	
ALGAE						
laver (red algae)	*Porphyra umbilicalis*	3.22	6.77	2.10	3.70	
n/a	*Volvox carteri*	0.16	0.49	3.06	2.17	

[a]Data is based on results from unassembled read;
n/a – data or common name not available
[b]see Table S2 for references

Table S3.1 references

Al-Dous EK, George B, Al-Mahmoud ME, Al-Jaber MY, Wang H, Salameh YM, Al-Azwani EK, Chaluvadi S, Pontaroli AC, Debarry J, et al. 2011. De novo genome sequencing and comparative genomics of date palm (Phoenix dactylifera). Nat. Biotechnol. 29:521–527.

D'hont A, Denoeud F, Aury JM, Baurens FC, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M, et al. 2012. The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. Nature 488:213–217.

Singh R, Ong-Abdullah M, Low ETL, Manaf MAA, Rosli R, Nookiah R, Ooi LCL, Ooi SE, Chan KL, Halim MA, et al. 2013. Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. Nature 500:335–339.

Table S3.2. Genome information and DNA transposon content of various sequenced plant genomes

Common Name	*Scientific name*	Assembled size (Mb)	Total TE (%)	DNA TE (%)	hAT (%)	Tc1/ Mariner (%)	Reference
Dicots							
sacred lotus	***Nelumbo nucifera***	**798**	**54.61**	**18.97**	**9.00**	**0.00**	**This study**
blueberry	*Vaccinium corymbosum*	1679	37.23	12.19	4.36	n/a	Colle et al. 2019
carrot	*Daucus carota*	422	45.00	13.60	4.10	1.70	Iorizzo et al. 2016
water lily	*Nymphaea colorata*	409	39.20	11.44	3.30	0.00	Zhang et al. 2020
canola	*Brassica rapa*	284	40.00	3.20	2.87	0.92	Wang et al. 2011
red bayberry	*Morella rubra*	313	36.40	21.45	2.73	0.03	Jia et al. 2019
Chinese white pear	*Pyrus bretschneideri*	533	51.20	8.52	2.45	0.00	Dong et al. 2020
wild pear	*Pyrus betuleafolia*	533	46.43	8.30	2.38	0.01	Dong et al. 2020
maple	*Acer yangbiense*	666	68.75	7.18	2.23	0.10	Yang et al. 2019
scarlet sage	*Salvia splendens*	808	40.35	11.91	2.08	1.03	Dong et al. 2018
goosefoot	*Chenopodium pallidicaule*	337	48.62	7.66	1.55	0.83	Jarvis et al. 2017
physic nut	*Jatropha curcas L.*	321	46.00	8.19	1.51	n/a	Wu et al. 2015
maca	*Lepidium meyenii*	743	47.65	8.15	1.51	0.23	Zhang et al. 2016
spinach	*Spinacia oleracea*	870	74.40	9.06	1.45	1.16	Xu et al. 2017
hybrid walnut	*Juglans regia*	1056	44.15	17.21	1.45	n/a	Zhu et al. 2019
watermelon	*Citrullus lanatus*	355	52.46	4.37	1.44	8.44×10^{-4}	Montero-Pau et al. 2018
cushion willow	*Salix brachista*	352	41.65	4.80	1.42	0.03	Chen, Hui et al. 2019
yellowhorn	*Xanthoceras sorbifolium Bunge*	440	56.39	3.05	1.41	0.02	Liang et al. 2019
cowpea	*Vigna unguiculata [L.] Walp.*	519	39.20	6.06	1.38	n/a	Lonardi et al. 2019
walnut	*Juglans microcarpa*	1056	43.18	16.30	1.32	n/a	Zhu et al. 2019
grape	*Vitis riparia*	500	33.74	6.02	1.29	0.00	Girolett et al. 2019
east asian tulip tree	*Liriodendron chinense*	1430	61.64	5.81	1.27	0.04	Chen, Hao et al. 2019
cassava	*Manihot esculenta Crantz*	746	53.13	2.33	1.08	n/a	Wang, Feng et al. 2014
grape	*Vitis vinifera*	477	29.00	1.80	1.03	0.05	Velasco et al. 2007
cacao	*Theobroma cacao*	327	16.10	n/a	1.02	0.07	Argout et al. 2010; Wei et al. 2018

Table S3.2 (cont'd)							
peanut	*Arachis hypogaea*	2540	74.03	11.06	1.01	n/a	Bertioli et al. 2019
clementine	*Citrus x. clementina*	301	45.00	2.69	0.95	0.15	Wu et al. 2014
strawberry	*Fragaria × ananassa*	805	34.22	5.35	0.95	0.01	Edger et al. 2019
quinoa	*Chenopodium quinoa*	1385	63.00	6.24	0.93	0.50	Jarvis et al. 2017
castor bean	*Ricinus communis L.*	320	52.84	2.62	0.93	n/a	Chan et al. 2010
durian	*Durio zibethinus*	715	54.80	3.74	0.80	n/a	Teh et al. 2017
tea	*Camellia sinensis var. sinensis*	3100	64.40	5.35	0.78	4.14×10^{-3}	Wei et al. 2018
peanut	*Arachis duranensis*	1250	61.70	8.80	0.70	n/a	Bertioli et al. 2016
flax	*Linum usitatissimum*	302	24.30	3.80	0.66	0.01	Wang et al. 2012
rubber tree	*Hevea brasiliensis*	1472	70.81	1.46	0.65	n/a	Liu et al. 2020
strawberry	*Fragaria vesca*	210	23.00	6.40	0.64	0.00	Shulaev et al. 2010
peanut	*Arachis duranensis*	1250	61.70	8.80	0.70	n/a	Bertioli et al. 2016
flax	*Linum usitatissimum*	302	24.30	3.80	0.66	0.01	Wang, Hobson et al. 2012
rubber tree	*Hevea brasiliensis*	1472	70.81	1.46	0.65	n/a	Liu et al. 2020
strawberry	*Fragaria vesca*	210	23.00	6.40	0.64	0.00	Shulaev et al. 2011
peanut	*Arachis ipaensis*	1560	68.50	10.73	0.64	n/a	Bertioli et al. 2016
eggplant	*Solanum melongena*	1163	71.55	3.31	0.62	0.30	Barchi et al. 2019
red clover	*Trifolium pratense*	315	29.90	6.07	0.57	0.29	Ištvánek et al. 2014
zucchini	*Cucurbita pepo*	263	37.80	0.68	0.57	n/a	Montero-Pau et al. 2018
thale cress	*Arabidopsis thaliana*	119	16.00	5.25	0.53	0.17	Hollister et al. 2011
cucumber 'Chinese long'	*Cucumis sativus*	197	28.16	2.83	0.52	n/a	Montero-Pau et al. 2018
adlay	*Coix lacryma-jobi*	1280	77.00	9.31	0.52	0.66	Kang et al. 2020
hot pepper	*Capsicum annuum*	3060	78.00	7.10	0.50	0.16	Kim et al. 2014
tea	*Camellia sinensis var. assamica*	3100	61.73	3.54	0.48	0.00	Wei et al. 2018
Jaltomata	*Jaltomata sinuosa*	1450	79.70	2.23	0.48	0.00	Wu et al. 2019
muskmelon	*Cucumis melo*	407	44.21	6.99	0.40	n/a	Montero-Pau et al. 2018
sweet orange	*Citrus sinensis*	320	20.00	2.30	0.36	0.09	Xu et al. 2013
apple	*Malus × domestica Borkh.*	604	52.00	6.60	0.35	n/a	Velasco et al. 2010

Table S3.2 (cont'd)							
cucumber	*Cucumis sativus*	244	24.00	1.20	0.30	n/a	Huang et al. 2009
ginseng	*Panax ginseng*	2980	79.52	4.67	0.29	0.21	Kim et al. 2018
lettuce	*Lactuca sativa*	2380	74.20	5.60	0.25	0.03	Reyes-Chin-Wo et al. 2017
common bean	*Phaseolus vulgaris L.*	473	45.42	5.50	0.22	n/a	Schmutz et al. 2014
cotton	*Gossypium arboreum*	1694	68.00	1.60	0.20	0.00	Li et al. 2014
potato	*Solanum tuberosum*	727	62.00	3.90	0.20	0.29	Xu et al. 2011
barrel medic	*Medicago truncatula*	262	31.00	3.40	0.20	0.04	Young et al. 2011
sunflower	*Helianthus annuus L.*	3000	74.70	0.19	0.12	0.11	Badouin et al. 2017
tomato	*Solanum lycopersicum*	760	63.00	0.90	0.10	0.79	Sato et al. 2012
soybean	*Glycine max*	973	59.00	16.50	0.04	0.07	Schmutz et al. 2010
cotton	*Gossypium raimondii*	738	61.00	1.50	0.03	0.00	Paterson et al. 2012
adzuki bean	*Vigna angularis*	538	43.10	2.68	0.03	0.09	Kang et al. 2015
poplar	*Populus trichocarpa (Torr. & Gray)*	485	44.00	2.40	0.02	n/a	Tuskan et al. 2006
African rice	*Oryza glaberrima*	316	34.00	15.00	0.33	n/a	Wang, Yu et al. 2014
Brachypodium	*Brachypodium distachyon*	272	28.00	4.80	0.24	0.07	Vogel et al. 2010
ryegrass	*Lolium perenne*	1128	29.00	3.13	0.22	0.30	Bryne et al. 2015
barley	*Hordeum vulgare*	4560	84.00	5.00	0.20	0.06	Mayer et al. 2012
wheat	*Triticum aestivum*	n/a	79.00	14.90	0.01	0.16	Wicker et al. 2018
sorghum	*Sorghum bicolor*	730	62.00	7.50	0.02	0.68	Paterson et al. 2009
Basal Angiosperm							
Amborella	*Amborella trichopoda*	706	57.00	7.90	1.65	0.00	DePamphilis et al. 2013
n/a	*Chara braunii*	1430	41.22	3.42	0.01	0.00	Nishiyama et al. 2018
Gymnosperm							
loblolly pine	*Pinus taeda*	20148	82.00	n/a	n/a	n/a	Neale et al. 2014
Norway spruce[a]	*Picea abies*	12000	70.00	1.00	n/a	n/a	Nystedt et al. 2013
n/a	*Gnetum montanum*	1240	42.16	2.64	n/a	n/a	Wan et al. 2018
Douglas-fir	*Pseudotsuga menziesii var. menziesii*	1477	69.70	7.00	n/a	n/a	Neale et al. 2017
ginko	*Ginkgo biloba*	1061	76.58	3.35	n/a	n/a	Guan et al. 2016

Table S3.2 (cont'd)							
Bryophyte							
C-fern	*Ceratopteris richardii*	4250	42.00	0.88	0.33	0.00	Marchant et al. 2019
common liverwort	*Marchantia polymorpha*	1125	42.00	0.86	0.32	0.01	Bowman et al. 2017
Lycophyte							
resurrection plant	*Selaginella lepidophylla*	109	24.61	7.99	0.42	0.07	VanBuren et al. 2018
spikemoss	*Selaginella moellendorffii*	212	37.00	1.80	0.02	n/a	Banks et al. 2011
Algae							
laver (red algae)	*Porphyra umbilicalis*	87.7	43.60	12.59	0.09	0.79	Brawley et al. 2017

n/a – data or common name not available

Table S3.2 references

Argout X, Salse J, Aury JM, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova SN, et al. 2011. The genome of *Theobroma cacao*. Nat. Genet. 43:101–108.

Badouin H, Gouzy J, Grassa CJ, Murat F, Staton SE, Cottret L, Lelandais-Brière C, Owens GL, Carrère S, Mayjonade B, et al. 2017. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature 546:148–152.

Banks JA, Nishiyama T, Hasebe M, Bowman JL, Gribskov M, DePamphilis C, Albert VA, Aono N, Aoyama T, Ambrose BA, et al. 2011. The selaginella genome identifies genetic changes associated with the evolution of vascular plants. Science (80-.). 332:960–963.

Barchi L, Pietrella M, Venturini L, Minio A, Toppino L, Acquadro A, Andolfo G, Aprea G, Avanzato C, Bassolino L, et al. 2019. A chromosome-anchored eggplant genome sequence reveals key events in Solanaceae evolution. Sci. Rep. 9.

Bertioli DJ, Jenkins J, Clevenger J, Dudchenko O, Gao D, Seijo G, Leal-Bertioli SCM, Ren L, Farmer AD, Pandey MK, et al. 2019. The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*. Nat. Genet. 51:877–884.

Bertioli DJ, Cannon SB, Froenicke L, Huang G, Farmer AD, Cannon EKS, Liu X, Gao D, Clevenger J, Dash S, et al. 2016. The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. Nat. Genet. 48:438–446.

Bowman JL, Kohchi T, Yamato KT, Jenkins J, Shu S, Ishizaki K, Yamaoka S, Nishihama R, Nakamura Y, Berger F, et al. 2017. Insights into Land Plant Evolution Garnered from the Marchantia polymorpha Genome. Cell 171:287-304.e15.

Brawley SH, Blouin NA, Ficko-Blean E, Wheeler GL, Lohr M, Goodson H V., Jenkins JW, Blaby-Haas CE, Helliwell KE, Chan CX, et al. 2017. Insights into the red algae and eukaryotic evolution from the genome of *Porphyra umbilicalis (Bangiophyceae, Rhodophyta)*. Proc. Natl. Acad. Sci. U. S. A. 114:E6361–E6370.

Byrne SL, Nagy I, Pfeifer M, Armstead I, Swain S, Studer B, Mayer K, Campbell JD, Czaban A, Hentrup S, et al. 2015. A synteny-based draft genome sequence of the forage grass *Lolium perenne*. Plant J. 84:816–826.

Cai J, Liu X, Vanneste K, Proost S, Tsai WC, Liu KW, Li-Jun C, He Y, Bian C, et al. 2015. The genome sequence of the orchid *Phalaenopsis equestris*. Nat. Genet. 47:65–72.

Chan AP, Crabtree J, Zhao Q, Lorenzi H, Orvis J, Puiu D, Melake-Berhan A, Jones KM, Redman J, Chen G, et al. 2010. Draft genome sequence of the oilseed species *Ricinus communis*. Nat. Biotechnol. 28:951–956.

Chen J Hui, Huang Y, Brachi B, Yun Q Zheng, Zhang W, Lu W, Li H Na, Li W Qing, Sun X Dong, Wang G Yan, et al. 2019. Genome-wide analysis of Cushion willow provides insights into alpine plant divergence in a biodiversity hotspot. Nat. Commun. 10.

Chen J, Hao Z, Guang X, Zhao C, Wang P, Xue L, Zhu Qihui, Yang Linfeng, Sheng Y, Zhou Y, et al. 2019. *Liriodendron* genome sheds light on angiosperm phylogeny and species–pair differentiation. Nat. Plants 5:18–25.

Colle M, Leisner CP, Wai CM, Ou S, Bird KA, Wang J, Wisecaver JH, Yocca AE, Alger EI, Tang H, et al. 2019. Haplotype-phased genome and evolution of phytonutrient pathways of tetraploid blueberry. Gigascience 8.

DePamphilis CW, Palmer JD, Rounsley S, Sankoff D, Schuster SC, Ammiraju JSS, Barbazuk WB, Chamala S, Chanderbali AS, Determann R, et al. 2013. The *Amborella* genome and the evolution of flowering plants. Science (80-.). 342.

Dohm JC, Minoche AE, Holtgräwe D, Capella-Gutiérrez S, Zakrzewski F, Tafer H, Rupp O, Sörensen TR, Stracke R, Reinhardt R, et al. 2014. The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). Nature 505:546–549.

Dong AX, Xin HB, Li ZJ, Liu H, Sun YQ, Nie S, Zhao ZN, Cui RF, Zhang RG, Yun QZ, et al. 2018. High-quality assembly of the reference genome for scarlet sage, *Salvia splendens*, an economically important ornamental plant. Gigascience 7.

Dong X, Wang Z, Tian L, Zhang Y, Qi D, Huo H, Xu J, Li Z, Liao R, Shi M, et al. 2020. De novo assembly of a wild pear (*Pyrus betuleafolia*) genome. Plant Biotechnol. J. 18:581–595.

Edger PP, Poorten TJ, VanBuren R, Hardigan MA, Colle M, McKain MR, Smith RD, Teresi SJ, Nelson ADL, Wai CM, et al. 2019. Origin and evolution of the octoploid strawberry genome. Nat. Genet. 51:541–547.

Girollet N, Rubio B, Lopez-Roques C, Valière S, Ollat N, Bert PF. 2019. De novo phased assembly of the *Vitis riparia* grape genome. Sci. Data 6.

Guan R, Zhao Y, Zhang H, Fan G, Liu X, Zhou W, Shi C, Wang Jiahao, Liu W, Liang X, et al. 2016. Draft genome of the living fossil Ginkgo biloba. Gigascience 5:49.

Hollister JD, Smith LM, Guo YL, Ott F, Weigel D, Gaut BS. 2011. Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. Proc. Natl. Acad. Sci. U. S. A. 108:2322–2327.

Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P, et al. 2009. The genome of the cucumber, *Cucumis sativus L*. Nat. Genet. 41:1275–1281.

Iorizzo M, Ellison S, Senalik D, Zeng P, Satapoomin P, Huang J, Bowman M, Iovene M, Sanseverino W, Cavagnaro P, et al. 2016. A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. Nat. Genet. 48:657–666.

Ištvánek J, Jaroš M, Krenek A, Řepková J. 2014. Genome assembly and annotation for red clover *(Trifolium pratense; Fabaceae)*. Am. J. Bot. 101:327–337.

Jarvis DE, Ho YS, Lightfoot DJ, Schmöckel SM, Li B, Borm TJA, Ohyanagi H, Mineta K, Michell CT, Saber N, et al. 2017. The genome of *Chenopodium quinoa*. Nature 542:307–312.

Jia HM, Jia HJ, Cai Q Le, Wang Y, Zhao HB, Yang WF, Wang GY, Li YH, Zhan DL, Shen YT, et al. 2019. The red bayberry genome and genetic basis of sex determination. Plant Biotechnol. J. 17:397–409.

Jiang N, Panaud O. 2013. Transposable element dynamics in rice and its wild relatives. In: Genetics and Genomics of Rice. p. 55–69.

Kang SH, Kim B, Choi BS, Lee HO, Kim NH, Lee SJ, Kim HS, Shin MJ, Kim HW, Nam K, et al. 2020. Genome Assembly and Annotation of Soft-Shelled Adlay (*Coix lacryma-jobi* Variety ma-yuen), a Cereal and Medicinal Crop in the Poaceae Family. Front. Plant Sci. 11.

Kang YJ, Satyawan D, Shim S, Lee T, Lee J, Hwang WJ, Kim SK, Lestari P, Laosatit K, Kim KH, et al. 2015. Draft genome sequence of adzuki bean, *Vigna angularis*. Sci. Rep. 5.

Kim NH, Jayakodi M, Lee SC, Choi BS, Jang W, Lee J, Kim HH, Waminal NE, Lakshmanan M, van Nguyen B, et al. 2018. Genome and evolution of the shade-requiring medicinal herb *Panax ginseng*. Plant Biotechnol. J. 16:1904–1917.

Kim S, Park M, Yeom SI, Kim YM, Lee JM, Lee HA, Seo E, Choi J, Cheong K, Kim KT, et al. 2014. Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. Nat. Genet. 46:270–278.

Li F, Fan G, Wang K, Sun F, Yuan Y, Song G, Li Q, Ma Z, Lu C, Zou C, et al. 2014. Genome sequence of the cultivated cotton *gossypium arboreum*. Nat. Genet. 46:567–572.

Liang Q, Li H, Li S, Yuan F, Sun J, Duan Q, Li Q, Zhang R, Sang YL, Wang N, et al. 2019. The genome assembly and annotation of yellowhorn (*Xanthoceras sorbifolium Bunge*). Gigascience 8.

Liu J, Shi Cong, Shi CC, Li W, Zhang QJ, Zhang Y, Li K, Lu HF, Shi Chao, Zhu ST, et al. 2020. The Chromosome-Based Rubber Tree Genome Provides New Insights into Spurge Genome Evolution and Rubber Biosynthesis. Mol. Plant 13:336–350.

Liu S, Liu Y, Yang X, Tong C, Edwards D, Parkin IAP, Zhao M, Ma J, Yu J, Huang S, et al. 2014. The *brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. Nat. Commun. 5.

Lonardi S, Muñoz-Amatriaín M, Liang Q, Shu S, Wanamaker SI, Lo S, Tanskanen J, Schulman AH, Zhu T, Luo MC, et al. 2019. The genome of cowpea (*Vigna unguiculata [L.] Walp*.). Plant J. 98:767–782.

Marchant DB, Sessa EB, Wolf PG, Heo K, Barbazuk WB, Soltis PS, Soltis DE. 2019. The C-Fern (*Ceratopteris richardii*) genome: insights into plant genome evolution with the first partial homosporous fern genome assembly. Sci. Rep. 9.

Mayer KFX, Waugh R, Langridge P, Close TJ, Wise RP, Graner A, Matsumoto T, Sato K, Schulman A, Ariyadasa R, et al. 2012. A physical, genetic and functional sequence assembly of the barley genome. Nature 491:711–716.

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly B V., Lewis KLT, et al. 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya Linnaeus*). Nature 452:991–996.

Montero-Pau J, Blanca J, Bombarely A, Ziarsolo P, Esteras C, Martí-Gómez C, Ferriol M, Gómez P, Jamilena M, Mueller L, et al. 2018. De novo assembly of the zucchini genome reveals a whole-genome duplication associated with the origin of the *Cucurbita* genus. Plant Biotechnol. J. 16:1161–1171.

Neale DB, McGuire PE, Wheeler NC, Stevens KA, Crepeau MW, Cardeno C, Zimin A V., Puiu D, Pertea GM, Sezen UU, et al. 2017. The Douglas-Fir genome sequence reveals specialization of the photosynthetic apparatus in *Pinaceae*. G3 Genes, Genomes, Genet. 7:3157–3167.

Neale DB, Wegrzyn JL, Stevens KA, Zimin A V., Puiu D, Crepeau MW, Cardeno C, Koriabine M, Holtz-Morris AE, Liechty JD, et al. 2014. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol. 15.

Nishiyama T, Sakayama H, de Vries J, Buschmann H, Saint-Marcoux D, Ullrich KK, Haas FB, Vanderstraeten L, Becker D, Lang D, et al. 2018. The *Chara* Genome: Secondary Complexity and Implications for Plant Terrestrialization. Cell 174:448-464.e24.

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, et al. 2013. The Norway spruce genome sequence and conifer genome evolution. Nature 497:579–584.

Olsen JL, Rouzé P, Verhelst B, Lin YC, Bayer T, Collen J, Dattolo E, De Paoli E, Dittami S, Maumus F, et al. 2016. The genome of the seagrass *Zostera marina* reveals angiosperm adaptation to the sea. Nature 530:331–335.

Paterson AH, Bowers JE, 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 AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D, Showmaker KC, Shu S, Udall J, et al. 2012. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. Nature 492:423–427.

Reyes-Chin-Wo S, Wang Z, Yang X, Kozik A, Arikit S, Song C, Xia L, Froenicke L, Lavelle DO, Truco MJ, et al. 2017. Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. Nat. Commun. 8.

Sato S, Tabata S, Hirakawa H, Asamizu E, Shirasawa K, Isobe S, Kaneko T, Nakamura Y, Shibata D, Aoki K, et al. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485:635–641.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al. 2010. Genome sequence of the palaeopolyploid soybean. Nature 463:178–183.

Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, Jenkins J, Shu S, Song Q, Chavarro C, et al. 2014. A reference genome for common bean and genome-wide analysis of dual domestications. Nat. Genet. 46:707–713.

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 (80-.). 326:1112–1115.

Shi C, Li W, Zhang QJ, Zhang Y, Tong Y, Li K, Liu YL, Gao LZ. 2020. The draft genome sequence of an upland wild rice species, Oryza granulata. Sci. Data 7.

Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, Jaiswal P, Mockaitis K, Liston A, Mane SP, et al. 2011. The genome of woodland strawberry (*Fragaria vesca*). Nat. Genet. 43:109–116.

Tang C, Yang M, Fang Y, Luo Y, Gao S, Xiao X, An Z, Zhou B, Zhang B, Tan X, et al. 2016. The rubber tree genome reveals new insights into rubber production and species adaptation. Nat. Plants 2.

Teh BT, Lim K, Yong CH, Ng CCY, Rao SR, Rajasegaran V, Lim WK, Ong CK, Chan K, Cheng VKY, et al. 2017. The draft genome of tropical fruit durian (*Durio zibethinus*). Nat. Genet. 49:1633–1641.

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam M, Ralph S, Rombauts S, Salamov A, et al. 2006. The genome of black cottonwood, *Populus trichocarpa (Torr. & Gray)*. Science (80-.). 313:1596–1604.

Upadhyay AK, Chacko AR, Gandhimathi A, Ghosh P, Harini K, Joseph AP, Joshi AG, Karpe SD, Kaushik S, Kuravadi N, et al. 2015. Genome sequencing of herb Tulsi *(Ocimum tenuiflorum)* unravels key genes behind its strong medicinal properties. BMC Plant Biol. 15.

Vanburen R, Wai CM, Ou S, Pardo J, Bryant D, Jiang N, Mockler TC, Edger P, Michael TP. 2018. Extreme haplotype variation in the desiccation-tolerant clubmoss *Selaginella lepidophylla*. Nat. Commun. 9:13.

Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, Bhatnagar SK, Troggio M, Pruss D, et al. 2010. The genome of the domesticated apple (*Malus × domestica Borkh.*). Nat. Genet. 42:833–839.

Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, FitzGerald LM, Vezzulli S, Reid J, et al. 2007. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS One 2.

Vogel JP, Garvin DF, Mockler TC, Schmutz J, Rokhsar D, Bevan MW, Barry K, Lucas S, Harmon-Smith M, Lail K, et al. 2010. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature 463:763–768.

Wan T, Liu ZM, Li LF, Leitch AR, Leitch IJ, Lohaus R, Liu ZJ, Xin HP, Gong YB, Liu Y, et al. 2018. A genome for gnetophytes and early evolution of seed plants. Nat. Plants 4:82–89.

Wang J, Zhang G, Liu X, Quan Z, Cheng S, Xu X, Pan S, Xie M, Zeng P, Yue Z, et al. 2012. Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. Nat. Biotechnol. 30:549–554.

Wang M, Yu Y, Haberer G, Marri PR, Fan C, Goicoechea JL, Zuccolo A, Song X, Kudrna D, Ammiraju JSS, et al. 2014. The genome sequence of African rice (*Oryza glaberrima*) and evidence for independent domestication. Nat. Genet. 46:982–988.

Wang W, Feng B, Xiao J, Xia Z, Zhou X, Li P, Zhang W, Wang Y, Møller BL, Zhang P, et al. 2014. Cassava genome from a wild ancestor to cultivated varieties. Nat. Commun. 5.

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun JH, Bancroft I, Cheng F, et al. 2011. The genome of the mesopolyploid crop species *Brassica rapa*. Nat. Genet. 43:1035–1040.

Wang Z, Hobson N, Galindo L, Zhu S, Shi D, McDill J, Yang L, Hawkins S, Neutelings G, Datla R, et al. 2012. The genome of flax (*Linum usitatissimum*) assembled de novo from short shotgun sequence reads. Plant J. 72:461–473.

Wei C, Yang H, Wang S, Zhao J, Liu C, Gao L, Xia E, Lu Y, Tai Y, She G, et al. 2018. Draft genome sequence of *Camellia sinensis var. sinensis* provides insights into the evolution of the tea genome and tea quality. Proc. Natl. Acad. Sci. U. S. A. 115:E4151–E4158.

Wicker T, Gundlach H, Spannagl M, Uauy C, Borrill P, Ramírez-González RH, De Oliveira R, Mayer KFX, Paux E, Choulet F. 2018. Impact of transposable elements on genome structure and evolution in bread wheat. Genome Biol.

Wu GA, Prochnik S, Jenkins J, Salse J, Hellsten U, Murat F, Perrier X, Ruiz M, Scalabrin S, Terol J, et al. 2014. Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. Nat. Biotechnol. 32:656–662.

Wu M, Kostyun JL, Moyle LC. 2019. Genome sequence of *Jaltomata* addresses rapid reproductive trait evolution and enhances comparative genomics in the hyper-diverse Solanaceae. Genome Biol. Evol. 11:335–349.

Wu P, Zhou C, Cheng S, Wu Z, Lu W, Han J, Chen Yanbo, Ni P, Wang Y, Xu X, et al. 2015. Integrated genome sequence and linkage map of physic nut (*Jatropha curcas L.*), a biodiesel plant. Plant J. 81:810–821.

Xu C, Jiao C, Sun H, Cai X, Wang X, Ge C, Zheng Y, Liu W, Sun X, Xu Y, et al. 2017. Draft genome of spinach and transcriptome diversity of 120 *Spinacia* accessions. Nat. Commun. 8.

Xu Q, Chen LL, Ruan X, Chen D, Zhu A, Chen C, Bertrand D, Jiao WB, Hao BH, Lyon MP, et al. 2013. The draft genome of sweet orange (*Citrus sinensis*). Nat. Genet. 45:59–66.

Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zhang G, Yang S, Li R, Wang J, et al. 2011. Genome sequence and analysis of the tuber crop potato. Nature 475:189–195.

Yang J, Wariss HM, Tao L, Zhang R, Yun Q, Hollingsworth P, Dao Z, Luo G, Guo H, Ma Y, et al. 2019. De novo genome assembly of the endangered *Acer yangbiense*, a plant species with extremely small populations endemic to Yunnan Province, China. Gigascience 8.

Young ND, Debellé F, Oldroyd GED, Geurts R, Cannon SB, Udvardi MK, Benedito VA, Mayer KFX, Gouzy J, Schoof H, et al. 2011. The *Medicago* genome provides insight into the evolution of rhizobial symbioses. Nature 480:520–524.

Zhang Jing, Tian Y, Yan L, Zhang G, Wang X, Zeng Y, Zhang Jiajin, Ma X, Tan Y, Long N, et al. 2016. Genome of Plant Maca (*Lepidium meyenii*) Illuminates Genomic Basis for High-Altitude Adaptation in the Central Andes. Mol. Plant 9:1066–1077.

Zhang L, Chen Fei, Zhang X, Li Z, Zhao Y, Lohaus R, Chang X, Dong W, Ho SYW, Liu X, et al. 2020. The water lily genome and the early evolution of flowering plants. Nature 577:79–84.

Zhang W, Zuo S, Li Z, Meng Z, Han J, Song J, Pan YB, Wang K. 2017. Isolation and characterization of centromeric repetitive DNA sequences in *Saccharum spontaneum*. Sci. Rep. 7.

Zhu T, Wang L, You FM, Rodriguez JC, Deal KR, Chen L, Li J, Chakraborty S, Balan B, Jiang CZ, et al. 2019. Sequencing a *Juglans regia* × *J. microcarpa* hybrid yields high-quality genome assemblies of parental species. Hortic. Res. 6.

Zou C, Li L, Miki D, Li D, Tang Q, Xiao L, Rajput S, Deng P, Peng L, Jia W, et al. 2019. The genome of broomcorn millet. Nat. Commun. 10.

Table S3.3. The insertion density of different TEs in different genomic regions in sacred lotus

Class	Subclass	Superfamily	Upstream 1 kb	5′ UTR	CDS	Introns	3′ UTR	Downstream 1kb	Genome wide average
			Insertions / (100 kb)						
Class I	LTR	LTR/ *Copia*	1.02	0.23	0.04	7.25	0.48	1.83	6.15
		LTR/ *Gypsy*	3.39	0.58	0.05	2.25	0.43	2.33	5.79
		LTR/other	0.36	0.05	0.00	0.43	0.17	0.72	0.47
	non-LTR	LINE	0.45	0.31	0.03	6.36	0.76	0.88	2.48
		SINE	1.08	0.31	0.01	1.37	0.06	1.71	1.08
Class II	TIR	CACTA	0.04	0.02	0.00	0.01	0.06	0.15	0.09
		hAT	26.16	1.51	0.05	14.96	2.09	24.82	17.51
		MULE	33.00	4.63	0.09	3.19	2.55	18.82	9.82
		PIF-Harbinger	19.08	1.85	0.05	7.43	4.33	25.48	11.97
	non-TIR	*Helitron*	3.22	0.21	0.01	2.71	0.91	3.23	3.48

Table S3.4. Biological processes in which genes are associated with enriched or depleted TE insertions in entire genic regions

GO Term	GO Term ID	Number of Genes	Observed Insertions	Expected Insertions	P-value*	Insertions per Gene
Depleted GO Terms						
fruit ripening	GO:0009835	47	94	292	5.39×10^{-31}	2.00
abscission	GO:0009838	74	278	459	2.65×10^{-17}	3.76
secondary metabolic process	GO:0019748	606	2735	3762	6.88×10^{-63}	4.51
response to endogenous stimulus	GO:0009719	3106	16449	19280	2.19×10^{-92}	5.30
flower development	GO:0009908	1077	5806	6685	5.73×10^{-27}	5.39
growth	GO:0040007	435	2401	2700	8.56×10^{-9}	5.52
photosynthesis	GO:0015979	253	1404	1570	2.67×10^{-5}	5.55
response to chemical	GO:0042221	5209	29729	32334	1.51×10^{-47}	5.71
response to light stimulus	GO:0009416	1427	8235	8858	3.67×10^{-11}	5.77
response to abiotic stimulus	GO:0009628	2378	13912	14761	2.81×10^{-12}	5.85
Enriched GO terms						
response to stress	GO:0006950	5529	35398	34320	5.90×10^{-9}	6.40
signal transduction	GO:0007165	2660	17489	16511	2.77×10^{-14}	6.57
nucleobase-containing compound metabolic process	GO:0006139	3700	24428	22967	5.34×10^{-22}	6.60
cell growth	GO:0016049	944	6264	5860	1.28×10^{-7}	6.64
biosynthetic process	GO:0009058	5095	34069	31626	6.05×10^{-43}	6.69
cellular process	GO:0009987	3491	23367	21670	9.17×10^{-31}	6.69
reproduction	GO:0000003	1829	12424	11353	9.10×10^{-24}	6.79
pollination	GO:0009856	483	3323	2998	2.96×10^{-9}	6.88
cell differentiation	GO:0030154	1772	12271	10999	7.69×10^{-34}	6.92
translation	GO:0006412	565	3949	3507	8.53×10^{-14}	6.99
multicellular organism development	GO:0007275	1956	14011	12141	1.43×10^{-64}	7.16
generation of precursor metabolites and energy	GO:0006091	490	3532	3042	5.95×10^{-19}	7.21
embryo development	GO:0009790	1005	7496	6238	4.33×10^{-57}	7.46
cell death	GO:0008219	661	4958	4103	1.22×10^{-40}	7.50
lipid metabolic process	GO:0006629	1262	9507	7834	9.91×10^{-80}	7.53
catabolic process	GO:0009056	2051	15680	12731	1.44×10^{-150}	7.65
cellular protein modification process	GO:0006464	2661	20399	16518	2.27×10^{-200}	7.67
regulation of molecular function	GO:0065009	504	4032	3128	1.06×10^{-58}	8.00
transport	GO:0006810	3101	25235	19249	0	8.14
cellular homeostasis	GO:0019725	442	3620	2744	7.71×10^{-63}	8.19
cellular component organization	GO:0016043	4030	33049	25015	0	8.20

Table S3.4 (cont'd)						
cell-cell signaling	GO:0007267	175	1455	1086	4.68×10^{-29}	8.31
protein metabolic process	GO:0019538	683	5891	4240	6.43×10^{-142}	8.63
cell cycle	GO:0007049	1102	9642	6840	1.61×10^{-251}	8.75
regulation of gene expression, epigenetic	GO:0040029	336	3485	2086	3.42×10^{-206}	10.37
DNA metabolic process	GO:0006259	692	7667	4295	0	11.08

*Chi-square multiple comparison tested at $p < 0.01$ with Bonferroni correction.

Table S3.5. Biological processes in which genes are associated with enriched or depleted TE insertions in introns

GO Term	GO Term ID	Number of Genes	Average Intron Number	Observed Insertions	Expected Insertions	P-value*	Insertions per Gene
Depleted GO Terms							
fruit ripening	GO:0009835	38	5.11	30	214	3.38×10^{-36}	0.79
abscission	GO:0009838	58	5.69	167	326	1.32×10^{-18}	2.88
secondary metabolic process	GO:0019748	499	4.09	1592	2804	5.57×10^{-116}	3.19
photosynthesis	GO:0015979	227	6.12	944	1276	1.59×10^{-20}	4.16
response to endogenous stimulus	GO:0009719	2548	6.18	11083	14319	4.53×10^{-161}	4.35
flower development	GO:0009908	925	6.27	4087	5198	1.34×10^{-53}	4.42
metabolic process	GO:0008152	748	5.65	3443	4204	8.85×10^{-32}	4.60
growth	GO:0040007	356	6.97	1662	2001	3.71×10^{-14}	4.67
response to abiotic stimulus	GO:0009628	2030	6.18	9524	11408	1.22×10^{-69}	4.69
response to chemical	GO:0042221	4307	6.24	20333	24204	1.13×10^{-136}	4.72
response to light stimulus	GO:0009416	1194	5.57	5649	6710	2.28×10^{-38}	4.73
cell communication	GO:0007154	363	6.43	1739	2040	2.67×10^{-11}	4.79
carbohydrate metabolic process	GO:0005975	1093	7.18	5451	6142	1.13×10^{-18}	4.99
response to external stimulus	GO:0009605	2664	6.21	13352	14971	5.74×10^{-40}	5.01
response to biotic stimulus	GO:0009607	2063	6.20	10415	11594	6.98×10^{-28}	5.05
circadian rhythm	GO:0007623	245	6.21	1239	1377	2.03×10^{-4}	5.06
post-embryonic development	GO:0009791	1665	6.22	8431	9357	1.05×10^{-21}	5.06
response to stress	GO:0006950	4702	6.51	25035	26424	1.28×10^{-17}	5.32
cellular process	GO:0009987	3087	6.26	16651	17343	1.51×10^{-7}	5.39
Enriched GO terms							
multicellular organism development	GO:0007275	1702	7.62	10320	9565	1.15×10^{-14}	6.06
catabolic process	GO:0009056	1901	7.33	11633	10683	3.93×10^{-20}	6.12
cell death	GO:0008219	597	7.14	3674	3355	3.64×10^{-8}	6.15
lipid metabolic process	GO:0006629	1120	6.90	7044	6294	3.32×10^{-21}	6.29
embryo development	GO:0009790	869	7.38	5596	4884	2.09×10^{-24}	6.44
cellular protein modification process	GO:0006464	2376	7.56	15366	13353	5.36×10^{-68}	6.47
transport	GO:0006810	2855	7.59	19298	16044	1.65×10^{-145}	6.76
regulation of molecular function	GO:0065009	445	7.37	3061	2501	3.96×10^{-29}	6.88
cellular homeostasis	GO:0019725	401	7.39	2766	2254	3.61×10^{-27}	6.90
cellular component organization	GO:0016043	3652	7.88	25257	20523	1.98×10^{-239}	6.92
cell-cell signaling	GO:0007267	157	8.31	1129	882	9.95×10^{-17}	7.19
cell cycle	GO:0007049	1030	9.86	7508	5788	4.05×10^{-113}	7.29

Table S3.5 (cont'd)							
protein metabolic process	GO:0019538	607	7.82	4550	3411	1.13×10^{-84}	7.50
regulation of gene expression, epigenetic	GO:0040029	304	9.14	2793	1708	9.16×10^{-152}	9.19
DNA metabolic process	GO:0006259	649	10.32	6283	3647	0	9.68

*Chi-square multiple comparison tested at $p < 0.01$ with Bonferroni correction.

Table S3.6. Biological processes in which genes are associated with enriched or depleted TE insertions in upstream 1 kb region

GO Term	GO Term ID	Number of Genes	Observed Insertions	Expected Insertions	P-value*	Insertions per Gene
Depleted GO Terms						
flower development	GO:0009908	1077	706	939	2.97×10^{-14}	0.66
response to endogenous stimulus	GO:0009719	3106	2237	2708	1.51×10^{-19}	0.72
cell growth	GO:0016049	944	707	823	5.33×10^{-5}	0.75
response to chemical	GO:0042221	5209	4019	4541	9.63×10^{-15}	0.77
post-embryonic development	GO:0009791	2002	1563	1745	1.29×10^{-5}	0.78
signal transduction	GO:0007165	2660	2100	2319	5.52×10^{-6}	0.79
Enriched GO Terms						
DNA metabolic process	GO:0006259	692	724	603	8.79×10^{-7}	1.05
translation	GO:0006412	565	608	493	1.96×10^{-7}	1.08

*Chi-square multiple comparison tested at $p < 0.01$ with Bonferroni correction.

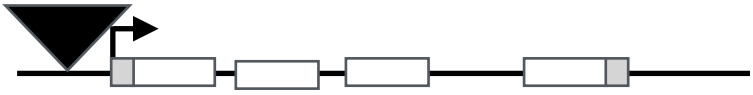


Table S3.7. The abundance of TEs in lotus genes related to seed longevity

Gene	Gene ID	Total TE length (bp) in			GO terms	Reference
		Upstream	Intron[a]	Downstream		
Single copy genes:						
ABA1	LOC104606289	777	380(15)	618	Response to light stimulus	(Clerkx et al. 2004)
ABI3	LOC104589171	0	0(5)	242	Response to endogenous stimulus	(Clerkx et al. 2004)
DOG1	LOC104611612	73	NA(0)	187	Response to chemical	(Bentsink et al. 2006)
GPX1	LOC104609149	88	1810(5)	484	Metabolic process	(Bailly et al. 1996)
Average		235	730(6.67)	383		
Duplicated genes- copy 1[b]						
ATPIMT1	LOC104591547	449	396(3)	945	Response to abiotic stimulus	(Oge et al. 2008)
ATS	LOC104595831	0	0(5)	0	Flower development	(Clerkx et al. 2004)
GPX8	GP	54	384(6)	459	Metabolic process	(Bailly et al. 1996)
GR2	LOC104602393	190	1231(11)	205	Response to chemical	(Bailly et al. 1996)
LEC1	LOC104594257	0	NA(0)	74	Response to endogenous stimulus	(Ooms et al. 1993)
LOX3	LOC104611974	0	665(8)	126	Cellular process	(Xu et al. 2015)
Average		116	535(6.60)	301		
Duplicated genes – copy 2[b]						
ATPIMT1	LOC104593032	613	3060(4)	408	Response to abiotic stimulus	
ATS	LOC104601978	257	0(5)	0	Flower development	
GPX8	LOC104608706	709	0(6)	370	Metabolic process	
GR2	LOC104602913	0	1970(10)	402	Response to chemical	
LEC1	LOC104587883	85	50(1)	831	Response to endogenous stimulus	
LOX3	LOC104596509	359	1207(9)	766	Cellular process	
Average		337	1047(5.83)	463		
Grand average		228	797(6.64)	382		
Genome average		346	5682(6.05)	299		

[a]Number in parenthesis indicate number of introns in the gene. For average value, only genes with introns are considered.
[b]Copy 1 and copy 2 are arbitrarily defined by their protein identity to the original gene from Arabidopsis or rice. The one with slightly higher identity is called copy 1 and the other is called copy 2.

Table S3.7 References

Bailly C, Benamar A, Corbineau F, Come D. 1996. Changes in malondialdehyde content and in superoxide dismutase, catalase and glutathione reductase activities in sunflower seeds as related to deterioration during accelerated aging. Physiologia Plantarum 97:104-110.

Bentsink L, Jowett J, Hanhart CJ, Koornneef M. 2006. Cloning of DOG1, a quantitative trait locus controlling seed dormancy in Arabidopsis. Proc Natl Acad Sci U S A 103:17042-17047.

Clerkx EJM, Vries HB, Ruys GJ, Groot SPC, Koornneef M. 2004. Genetic differences in seed longevity of various Arabidopsis mutants. Physiologia Plantarum 121:448-461.

Oge L, Bourdais G, Bove J, Collet B, Godin B, Granier F, Boutin JP, Job D, Jullien M, Grappin P. 2008. Protein repair L-isoaspartyl methyltransferase 1 is involved in both seed longevity and germination vigor in Arabidopsis. Plant Cell 20:3022-3037.

Ooms J, Leon-Kloosterziel KM, Bartels D, Koornneef M, Karssen CM. 1993. Acquisition of Desiccation Tolerance and Longevity in Seeds of Arabidopsis thaliana (A Comparative Study Using Abscisic Acid-Insensitive abi3 Mutants). Plant Physiol 102:1185-1191.

Xu H, Wei Y, Zhu Y, Lian L, Xie H, Cai Q, Chen Q, Lin Z, Wang Z, Xie H, et al. 2015. Antisense suppression of LOX3 gene expression in rice endosperm enhances seed longevity. Plant Biotechnol J 13:526-539.

REFERENCES

REFERENCES

Agren JA, Wright SI. 2011. Co-evolution between transposable elements and their hosts: a major factor in genome size evolution? Chromosome Res 19:777-786.

Amborella Genome Project. 2013. The Amborella genome and the evolution of flowering plants. Science 342:1241089.

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

Badouin H, Gouzy J, Grassa CJ, Murat F, Staton SE, Cottret L, Lelandais-Briere C, Owens GL, Carrere S, Mayjonade B, et al. 2017. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature 546:148-152.

Bao W, Jurka MG, Kapitonov VV, Jurka J. 2009. New superfamilies of eukaryotic DNA transposons and their internal divisions. Mol Biol Evol 26:983-993.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA 6:11.

Bartolome C, Bello X, Maside X. 2009. Widespread evidence for horizontal transfer of transposable elements across Drosophila genomes. Genome Biol 10:R22.

Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, Deragon JM, Westerman RP, Sanmiguel PJ, Bennetzen JL. 2009. Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. PLoS Genet 5:e1000732.

Benachenhou F, Sperber GO, Bongcam-Rudloff E, Andersson G, Boeke JD, Blomberg J. 2013. Conserved structure and inferred evolutionary history of long terminal repeats (LTRs). Mob DNA 4:5.

Bennetzen JL, Kellogg EA. 1997. Do Plants Have a One-Way Ticket to Genomic Obesity? Plant Cell 9:1509-1514.

Bennetzen JL, Wang H. 2014. The contributions of transposable elements to the structure, function, and evolution of plant genomes. Annu Rev Plant Biol 65:505-530.

Bera S, Pandey KK, Vora AC, Grandgenett DP. 2009. Molecular Interactions between HIV-1 integrase and the two viral DNA ends within the synaptic complex that mediates concerted integration. J Mol Biol 389:183-198.

Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E. 2015. The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. Genesis 53:474-485.

Bertioli DJ, Jenkins J, Clevenger J, Dudchenko O, Gao D, Seijo G, Leal-Bertioli SCM, Ren L, Farmer AD, Pandey MK, et al. 2019. The genome sequence of segmental allotetraploid peanut Arachis hypogaea. Nat Genet 51:877-884.

Bundock P, Hooykaas P. 2005. An Arabidopsis hAT-like transposase is essential for plant development. Nature 436:282-284.

Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. 2002. Selection for short introns in highly expressed genes. Nat Genet 31:415-418.

Cerbin S, Wai CM, VanBuren R, Jiang N. 2019. GingerRoot: A Novel DNA Transposon Encoding Integrase-Related Transposase in Plants and Animals. Genome Biol Evol 11:3181-3193.

Charlesworth D, Charlesworth B. 1995. Transposable elements in inbreeding and outbreeding populations. Genetics 140:415-417.

Cheng C, Daigen M, Hirochika H. 2006. Epigenetic regulation of the rice retrotransposon Tos17. Mol Genet Genomics 276:378-390.

Choi JY, Lee YCG. 2020. Double-edged sword: The evolutionary consequences of the epigenetic silencing of transposable elements. PLoS Genet 16:e1008872.

Chow SA, Vincent KA, Ellison V, Brown PO. 1992. Reversal of integration and DNA splicing mediated by integrase of human immunodeficiency virus. Science 255:723-726.

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674-3676.

Cresse AD, Hulbert SH, Brown WE, Lucas JR, Bennetzen JL. 1995. Mu1-related transposable elements of maize preferentially insert into low copy number DNA. Genetics 140:315-324.

Cruz GM, Metcalfe CJ, de Setta N, Cruz EA, Vieira AP, Medina R, Van Sluys MA. 2014. Virus-like attachment sites and plastic CpG islands:landmarks of diversity in plant Del retrotransposons. PLoS One 9:e97099.

Davidson RM, Gowda M, Moghe G, Lin H, Vaillancourt B, Shiu SH, Jiang N, Robin Buell C. 2012. Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. Plant J 71:492-502.

Diao Y, Chen L, Yang G, Zhou M, Song Y, Hu Z, Liu JY. 2006. Nuclear DNA C-values in 12 species in Nymphales. Caryologia 59:25-30.

Dietrich CR, Cui F, Packila ML, Li J, Ashlock DA, Nikolau BJ, Schnable PS. 2002. Maize Mu transposons are targeted to the 5' untranslated region of the gl8 gene and sequences flanking Mu target-site duplications exhibit nonrandom nucleotide composition throughout the genome. Genetics 160:697-716.

Drinnan AN, Crane PR, Hoot SB. 1994. Patterns of floral evolution in the early diversification of non-magnoliid dicotyledons (eudicots). In: Endress PK, Friis EM, editors. Early Evolution of Flowers. Vienna: Springer Vienna. p. 93-122.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A 102:14338-14343.

Du Z, Ilyinskii PO, Lally K, Desrosiers RC, Engelman A. 1997. A mutation in integrase can compensate for mutations in the simian immunodeficiency virus att site. J Virol 71:8124-8132.

El Baidouri M, Carpentier MC, Cooke R, Gao D, Lasserre E, Llauro C, Mirouze M, Picault N, Jackson SA, Panaud O. 2014. Widespread and frequent horizontal transfers of transposable elements in plants. Genome Res 24:831-838.

Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9:18.

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. Nat Rev Genet 9:397-405.

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

Feschotte C, Wessler SR. 2002. Mariner-like transposases are widespread and diverse in flowering plants. Proc Natl Acad Sci U S A 99:280-285.

Guan R, Zhao Y, Zhang H, Fan G, Liu X, Zhou W, Shi C, Wang J, Liu W, Liang X, et al. 2016. Draft genome of the living fossil Ginkgo biloba. Gigascience 5:49.

Gui S, Peng J, Wang X, Wu Z, Cao R, Salse J, Zhang H, Zhu Z, Xia Q, Quan Z, et al. 2018. Improving Nelumbo nucifera genome assemblies using high-resolution genetic maps and BioNano genome mapping reveals ancient chromosome rearrangements. Plant J 94:721-734.

Han Y, Wessler SR. 2010. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res 38:e199.

Han YC, Teng CZ, Zhong S, Zhou MQ, Hu ZL, Song YC. 2007. Genetic variation and clonal diversity in populations of Nelumbo nucifera (Nelumbonaceae) in central China detected by ISSR markers. Aquatic Botany 69-75.

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

Hollister JD, Gaut BS. 2009. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res 19:1419-1428.

Hudson ME, Lisch DR, Quail PH. 2003. The FHY3 and FAR1 genes encode transposase-related proteins involved in regulation of gene expression by the phytochrome A-signaling pathway. Plant J 34:453-471.

International Rice Genome Sequencing Project. 2005. The map-based sequence of the rice genome. Nature 436:793-800.

Jarvis CE. 2007. Order Out of Chaos: Linnaean Plant Names and Their Types. London, England: Linnean Society of London.

Jia HM, Jia HJ, Cai QL, Wang Y, Zhao HB, Yang WF, Wang GY, Li YH, Zhan DL, Shen YT, et al. 2019. The red bayberry genome and genetic basis of sex determination. Plant Biotechnol J 17:397-409.

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, Ferguson AA, Slotkin RK, Lisch D. 2011. Pack-Mutator-like transposable elements (Pack-MULEs) induce directional modification of genes through biased insertion and DNA acquisition. Proc Natl Acad Sci U S A 108:1537-1542.

Jiang N, Panaud O. 2013. Transposable element dynamics in rice and its wild relatives. In: Zhang Q, Wing RA, editors. Genetics and Genomics of Rice. New York: Springer New York. p. 55-69.

Jordan IK, Rogozin IB, Glazko GV, Koonin EV. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet 19:68-72.

Juntawong P, Girke T, Bazin J, Bailey-Serres J. 2014. Translational dynamics revealed by genome-wide profiling of ribosome footprints in Arabidopsis. Proc Natl Acad Sci U S A 111:E203-212.

Kapitonov VV, Jurka J. 2008. A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet 9:411-412; author reply 414.

Kawakami T, Strakosh SC, Zhen Y, Ungerer MC. 2010. Different scales of Ty1/copia-like retrotransposon proliferation in the genomes of three diploid hybrid sunflower species. Heredity (Edinb) 104:341-350.

Kempken F, Windhofer F. 2001. The hAT family: a versatile transposon group common to plants, fungi, animals, and man. Chromosoma 110:1-9.

Knip M, de Pater S, Hooykaas PJ. 2012. The SLEEPER genes: a transposase-derived angiosperm-specific gene family. BMC Plant Biol 12:192.

Koonin EV, Wolf YI. 2010. Constraints and plasticity in genome and molecular-phenome evolution. Nat Rev Genet 11:487-498.

Kumar A, Bennetzen JL. 1999. Plant retrotransposons. Annu Rev Genet 33:479-532.

Kunze R, Weil C. 2002. The hAT and CACTA superfamilies of plant transposons. In: Craig NL, Craigie R, Gellert M, Lambowitz AM, editors. Mobile DNA II. Washington, D. C.: ASM Press. p. 565-610.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860-921.

Lazarow K, Du ML, Weimer R, Kunze R. 2012. A hyperactive transposase of the maize transposable element activator (Ac). Genetics 191:747-756.

Le TN, Miyazaki Y, Takuno S, Saze H. 2015. Epigenetic regulation of intragenic transposable elements impacts gene transcription in Arabidopsis thaliana. Nucleic Acids Res 43:3911-3921.

Liu S, Yeh CT, Ji T, Ying K, Wu H, Tang HM, Fu Y, Nettleton D, Schnable PS. 2009. Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. PLoS Genet 5:e1000733.

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.

Masuda T, Kuroda MJ, Harada S. 1998. Specific and independent recognition of U3 and U5 att sites by human immunodeficiency virus type 1 integrase in vivo. J Virol 72:8396-8402.

Mayer KF, Waugh R, Brown JW, Schulman A, Langridge P, Platzer M, Fincher GB, Muehlbauer GJ, Sato K, Close TJ, et al. 2012. A physical, genetic and functional sequence assembly of the barley genome. Nature 491:711-716.

McAbee JM, Hill TA, Skinner DJ, Izhaki A, Hauser BA, Meister RJ, Venugopala Reddy G, Meyerowitz EM, Bowman JL, Gasser CS. 2006. ABERRANT TESTA SHAPE encodes a KANADI family member, linking polarity determination to separation and growth of Arabidopsis ovule integuments. Plant J 46:522-531.

McCarthy EM, McDonald JF. 2003. LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics 19:362-367.

McClintock B. 1950. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U S A 36:344-355.

Ming R, VanBuren R, Liu Y, Yang M, Han Y, Li LT, Zhang Q, Kim MJ, Schatz MC, Campbell M, et al. 2013. Genome of the long-living sacred lotus (Nelumbo nucifera Gaertn.). Genome Biol 14:R41.

Morgenstern B. 1999. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics 15:211-218.

Muehlbauer GJ, Bhau BS, Syed NH, Heinen S, Cho S, Marshall D, Pateyron S, Buisine N, Chalhoub B, Flavell AJ. 2006. A hAT superfamily transposase recruited by the cereal grass genome. Mol Genet Genomics 275:553-563.

Muotri AR, Marchetto MC, Coufal NG, Gage FH. 2007. The necessary junk: new functions for transposable elements. Hum Mol Genet 16 Spec No. 2:R159-167.

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.

Noreen F, Akbergenov R, Hohn T, Richert-Poggeler KR. 2007. Distinct expression of endogenous Petunia vein clearing virus and the DNA transposon dTph1 in two Petunia hybrida lines is correlated with differences in histone modification and siRNA production. Plant J 50:219-229.

Oliver KR, McComb JA, Greene WK. 2013. Transposable elements: powerful contributors to angiosperm evolution and diversity. Genome Biol Evol 5:1886-1901.

Ong-Abdullah M, Ordway JM, Jiang N, Ooi SE, Kok SY, Sarpan N, Azimi N, Hashim AT, Ishak Z, Rosli SK, et al. 2015. Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. Nature 525:533-537.

Ou S, Jiang N. 2018. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. Plant Physiol 176:1410-1422.

Pan L, Xia Q, Quan Z, Liu H, Ke W, Ding Y. 2010. Development of Novel EST–SSRs from Sacred Lotus (Nelumbo nucifera Gaertn) and Their Utilization for the Genetic Diversity Analysis of N. nucifera. Journal of Heredity 101:71-82.

Parisod C, Alix K, Just J, Petit M, Sarilar V, Mhiri C, Ainouche M, Chalhoub B, Grandbastien MA. 2010. Impact of transposable elements on the organization and function of allopolyploid genomes. New Phytol 186:37-45.

Paterson AH, Bowers JE, 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.

Pereira V. 2004. Insertion bias and purifying selection of retrotransposons in the Arabidopsis thaliana genome. Genome Biol 5:R79.

Piegu B, Guyot R, Picault N, Roulin A, Sanyal A, Kim H, Collura K, Brar DS, Jackson S, Wing RA, et al. 2006. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. Genome Res 16:1262-1269.

Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud PF, Lindquist EA, Kamisugi Y, et al. 2008. The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. Science 319:64-69.

Robertson HM. 2002. Evolution of DNA transposons. In: Craig NL, Craigie R, Gellert M, Lambowitz AM, editors. Mobile DNA II. Washington, D.C: ASM Press. p. 1093-1110.

Robertson HM, Lampe DJ. 1995. Recent horizontal transfer of a mariner transposable element among and between Diptera and Neuroptera. Mol Biol Evol 12:850-862.

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

SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, et al. 1996. Nested retrotransposons in the intergenic regions of the maize genome. Science 274:765-768.

Schaack S, Gilbert C, Feschotte C. 2010. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends Ecol Evol 25:537-546.

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.

Sela N, Mersch B, Gal-Mark N, Lev-Maor G, Hotz-Wagenblatt A, Ast G. 2007. Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome. Genome Biol 8:R127.

Shen-Miller J. 2002. Sacred lotus, the long-living fruits of China Antique. Seed Science Research 12:131-143.

Sigman MJ, Slotkin RK. 2016. The First Rule of Plant Transposable Element Silencing: Location, Location, Location. Plant Cell 28:304-313.

Singer MF. 1982. SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. Cell 28:433-434.

Slotkin RK, Martienssen R. 2007. Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet 8:272-285.

Srivastava AK, Lu Y, Zinta G, Lang Z, Zhu JK. 2018. UTR-Dependent Control of Gene Expression in Plants. Trends Plant Sci 23:248-259.

Sundaresan V, Springer P, Volpe T, Haward S, Jones JD, Dean C, Ma H, Martienssen R. 1995. Patterns of gene action in plant development revealed by enhancer trap and gene trap transposable elements. Genes Dev 9:1797-1810.

Temin HM. 1981. Structure, variation and synthesis of retrovirus long terminal repeat. Cell 27:1-3.

Tenaillon MI, Hollister JD, Gaut BS. 2010. A triptych of the evolution of plant transposable elements. Trends Plant Sci 15:471-478.

Tilgner H, Knowles DG, Johnson R, Davis CA, Chakrabortty S, Djebali S, Curado J, Snyder M, Gingeras TR, Guigo R. 2012. Deep sequencing of subcellular RNA fractions shows

splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. Genome Res 22:1616-1625.

VanBuren R, Wai CM, Ou S, Pardo J, Bryant D, Jiang N, Mockler TC, Edger P, Michael TP. 2018. Extreme haplotype variation in the desiccation-tolerant clubmoss Selaginella lepidophylla. Nat Commun 9:13.

Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, Fitzgerald LM, Vezzulli S, Reid J, et al. 2007. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS One 2:e1326.

Volff JN. 2006. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays 28:913-922.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520-562.

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

Wikstrom N, Savolainen V, Chase MW. 2001. Evolution of the angiosperms: calibrating the family tree. Proc Biol Sci 268:2211-2220.

Wright SI, Le QH, Schoen DJ, Bureau TE. 2001. Population dynamics of an Ac-like transposable element in self- and cross-pollinating arabidopsis. Genetics 158:1279-1288.

Wu C, Lu J. 2019. Diversification of Transposable Elements in Arthropods and Its Impact on Genome Evolution. Genes (Basel) 10.

Xie Z, Johansen LK, Gustafson AM, Kasschau KD, Lellis AD, Zilberman D, Jacobsen SE, Carrington JC. 2004. Genetic and functional diversification of small RNA pathways in plants. PLoS Biol 2:E104.

Xu Z, Wang H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35:W265-268.

Yamasaki K, Kigawa T, Seki M, Shinozaki K, Yokoyama S. 2013. DNA-binding domains of plant-specific transcription factors: structure, function, and evolution. Trends Plant Sci 18:267-276.

Yin H, Liu J, Xu Y, Liu X, Zhang S, Ma J, Du J. 2013. TARE1, a mutated Copia-like LTR retrotransposon followed by recent massive amplification in tomato. PLoS One 8:e68587.

Zhang HH, Peccoud J, Xu MR, Zhang XG, Gilbert C. 2020. Horizontal transfer and evolution of transposable elements in vertebrates. Nat Commun 11:1362.

Zhang X, Zhao M, McCarty DR, Lisch D. 2020. Transposable elements employ distinct integration strategies with respect to transcriptional landscapes in eukaryotic genomes. Nucleic Acids Res 48:6685-6698.

Zhao D, Hamilton JP, Hardigan M, Yin D, He T, Vaillancourt B, Reynoso M, Pauluzzi G, Funkhouser S, Cui Y, et al. 2017. Analysis of Ribosome-Associated mRNAs in Rice Reveals the Importance of Transcript Size and GC Content in Translation. G3 (Bethesda) 7:203-219.

CHAPTER 4

AMPLIFICATION OF A PACK-MULE TRANSPOSON THROUGH ACQUISITION OF A GENE FRAGMENT FROM AN *ARGONAUTE1* (*AGO1*) GENE IN THE TOMATO CLADE

ABSTRACT

Pack-MULEs refer to *Mutator*-like transposable elements carrying genes or gene fragments. Despite their prevalence in plants, Pack-MULEs are typically associated with low copy numbers, ranging from one to a few. Here we report a Pack-MULE, called *SlPM37* that achieved a relatively high copy number, 47/255 (18.4%) of the total Pack-MULEs in tomato. The *SlPM37* elements include two subtypes: one contains a gene fragment from a *Cytochrome P-450 51* (*CYP51*) gene, and the other contains the *CYP51* fragment as well as an additional fragment from an *Argonaute1* (*AGO1*) gene, presumably through step-wise acquisition. Querying other closely related genomes show that the copy number of *SlPM37* varies from 31 – 59 copies, predominantly comprised of *SlPM37* Pack-MULEs with both *AGO1* and *CYP51* gene fragments. However, in the distant relative potato, there are only two copies of *SlPM37* with solely the *CYP51* fragment, suggesting the amplification of *SlPM37* was enabled through the acquisition of the *AGO1* fragment. Based on the location of these gene fragments within genes, we used Nanostring expression analysis to probe leaf expression of *AGO1* and *CYP51* across the tomato clade. In this manner, we show that the expression of *AGO1* and *CYP51* is not correlated with the copy number of *SlPM37*. These results suggest that the capture of the *AGO1* sequence allowed the Pack-MULE to increase its copy number. This study illuminates that transposons could benefit from hijacking gene fragments which favor the proliferation of the element.

INTRODUCTION

Tomato is a diploid plant that is globally cultivated for its fleshy fruit. The center of origin for tomato is in South America and is classified into a 17 species clade (Darwin et al.

2003; Peralta et al. 2008; Haak et al. 2014). In the 8 million years since tomato's most recent common ancestor with potato, the clade has diversified across habitats, morphology, and biotic and abiotic adaptations (Nesbit and Tanksley 2002; Darwin et al. 2003; Spooner et al. 2005; Peralta et al. 2008; Sato et al. 2012; Särkinen et al. 2013). Along with the tomato and potato genomes, seven other species have genomic resources available for analysis (https://solgenomics.net/).

Transposable elements (TEs) are DNA sequences that are capable of moving around in the genome. Barbara McClintock discovered TEs in maize and labeled the "controlling elements" for their effects on recombination (McClintock 1950). Transposons are divided into two classes, Class I retrotransposons are mobilized by an RNA intermediate during transposition and are referred to as copy and paste elements (Lisch 2013). Class II or DNA transposons are mobilized by a DNA intermediate where the element is cut out of the original locus and inserts into a second locus, hence called cut and paste transposition mechanism.

Within both classes, individual elements are further classified by their structure, protein-coding capacity, and terminal sequences. If an element encodes all the factors to transpose they are classified as autonomous. If an element is lacking transposition proteins or these proteins are mutated, resulting in non-functional proteins, the element is labeled non-autonomous and requires transposition machinery from an autonomous element to transpose (Craig 2002). Further structural differences demonstrate the divergence of elements. One such feature is the order of the Open Reading Frame (ORF) in *Copia* and *Gypsy* elements in plants (Wicker and Keller 2007). Several structural elements in the ORFs, such as transposase amino acid sequence and similarity can be utilized as classification tools. Other transposons, such as DNA elements,

contain an open reading frame (ORF) encoding transposase, which recognizes TE sequences and catalyzes DNA cleavage for transposition (Kidwell and Lisch 2000). Also, element termini called Terminal Inverted Repeats (TIRs) or direct repeats are important for identifying and classifying elements (Craig 2002). TIR sequences are important for mobilizations in DNA elements since the transposase recognizes sequences in the TIR to cleave and excise the element to translocate (Kidwell and Lisch 2000). Overall differing structures and mechanisms result in a panoply of element diversity across genomes.

Along with polyploidy, transposons are important in determining genome size (Eddy 2012). These genome size variations are due to transposon amplification, which is primarily driven by retrotransposons (Bennetzen et al. 2005; Zou et al. 2009). Understanding transposon composition, and genomic interactions are important for understanding genetic variation and shared evolutionary histories. For example in tomato, a *Rider* retrotransposon is responsible for the elongated versus round fruit shape differences (Xiao et al. 2008). Another example is the insertion of *Gret1* retrotransposon modified MYB transcription factors leading to berry skin color in grape. (Koybashi et al. 2004; Kobayashi et al. 2005). These studies show that transposons are agents of change facilitating genome recombination, methylation, and rearrangements (Lisch and Slotkin 2011).

In addition to the activities mentioned above, TEs are capable of acquiring/duplicating entire or part of the protein coding genes. In human, retrotransposon L1 has been shown to transduce nearby cellular genes or other genomic sequences, contributing to the duplication of 1% of the human genome (Moran et al. 1999; Pickeral et al. 2000). Since retrotransposons transpose through an RNA mechanism, the resulting gene sequences are not associated with

introns. If functional genes are derived from sequences duplicated by retrotransposons, those are called “retrogenes”, with the hallmark of lack of introns. In plants, several families of TEs have been reported to duplicate and mobilize genes or gene fragments (Kawasaki et al. 2004; Morgante et al. 2005; Wang et al. 2006). Among those, one group of elements is called Pack-MULEs, referring to *Mutator* Like Elements that harbor genes or gene fragments (Jiang et al. 2004). Pack-MULEs can contain fragments from multiple genes, including intron sequences. The gene fragments found in Pack-MULEs, referred to as acquired regions, and the genes where the acquired regions derived from are called parental genes. Additionally, Pack-MULEs have been shown to acquire fragments from highly and widely expressed parental genes (Ferguson et al. 2013).

The first example of Pack-MULEs was reported in maize more than 30 years ago (Talbert and Chandler 1988). For an extended period of time, the formation of Pack-MULEs was considered a rare curiosity (Lisch 2005). Thanks to the availability of entire plant genomes, it was shown that Pack-MULEs were prevalent and abundant in plants (Jiang et al. 2004; Holligan et al. 2006; Jiang et al. 2011), thereby representing a new mechanism of gene duplication. Despite the prevalence of Pack-MULEs, the acquired gene fragments inside Pack-MULEs remain low copy numbers. For example, 58% of the acquired gene fragments in rice are only present in one individual member of Pack-MULEs (Hanada et al. 2009), suggesting constant shuffling of sequences inside Pack-MULEs or selection against the amplification of individual Pack-MULEs to high copy numbers.

In this study, we characterized a Pack-MULE in the genome of tomato, where about 50 copies of this element were identified. This copy number is higher than that of any Pack-MULEs

reported previously. Through analyses of its structure, distribution among different species, and expression of the parental genes, we reconstructed the history of acquisition and amplification of this unusual Pack-MULE. This information may facilitate the understanding of the factors that limit or favor the amplification of Pack-MULEs as well as the consequence of amplification of Pack-MULEs.

RESULTS

SlPM37 structure

In tomato 255 copies of Pack-MULEs of which one named *S. lycopersicum* Pack-MULE 37 (*SlPM37*) has 47 intact copies or 18.4% of all the Pack-MULEs identified in the genome, (Table 4.1). These elements range in size from 669-1333 base pairs (bp). Overall the elements are AT-rich, with a GC content of 34.4%. *SlPM37* contains gene fragments from two parental genes, *Argonaute 1B (AGO1) (Solyc03g098280)* and *Cytochrome P 450 (CYP51) (Solyc01g008110)* (Figure 4.1, Table 4.1). The *CYP51* fragments are from exons 1 and 2, and the *AGO1* fragment is from exons 6, 7, 8, and 9. Inside the Pack-MULEs both acquired regions lack intron sequence (Figure 4.1). The acquired *AGO1* fragment covers bp 1511-2035 (525 bp in length) in AGO1 transcript; the elements show various similarities from 84.60% identity in *SlPM37-030* to 98.21% in *SlPM37-008* to the parental gene *AGO1* respectively. The acquired *CYP51* fragments match bp 98-294 and 313-348 in the *CYP51* cDNA sequence, showing a range of similarities from 81.45% in *SlPM37-012* to 91.70% in *SlPM37-008*. Overall the acquired *AGO1* fragments are slightly more similar to the corresponding regions in the parental gene than the acquired *CYP51* fragment (average identity 91.19% vs 87.22%, t-test $p = 2.00 \times 10^{-12}$, Table

4.1). This suggests the *CYP51* fragment was acquired before the *AGO1* fragment or it evolves under less selective constraint. All of the *CYP51* fragments except those in *SlPM37-012, SlPM37-023*, and *SlPM37-035* are separated by the *AGO1* fragment (Figure 4.1, Table 4.1). Therefore, in tomato and tomato relatives the Pack-MULE has two different compositions; *SlPM37-013* has both *AGO1* and *CYP51* fragments. In contrast, *SlPM37-035* only has the *CYP51* fragment and lacks the *AGO1* fragment, and in total there are three elements in tomato with only *CYP51* fragments (Table 4.1). Also in *SlPM37-012, SlPM37-023,* and *SlPM37-035,* there are small deletions in the acquired CYP51 region, corresponding to 294-328 bp in the cDNA of the parental gene.

In several elements, we identified a region at the 5′ end of the second or downstream *CYP51* fragment and the 3′ end of the *AGO1* fragment that overlaps. The *CYP51/AGO1* region of microhomology between the two genes occurs in elements *SlPM37-003, SlPM37-009, SlPM37-009, SlPM37-010, SlPM37-011, SlPM37-021, SlPM37-033*, and *SlPM37-047*. In all these regions the overlap is six nucleotide TTGTCT. However, in all cases, the elements have a mutation differing from the parental *CYP51* from C to T in the second nucleotide.

Compared to elements with only *CYP51* fragment, about 100 bp *CYP51* sequence was lost in the regions adjacent to the *AGO1* fragment within the elements with both gene fragments. As a result, within elements with two gene fragments, the total length of the *CYP51* fragment is about 200 bp in length (or less), much shorter than the *AGO1* fragment. In addition, we observed in several elements a piece of unaligned DNA between the 3′ end of the acquired *AGO1* fragment and the 5′ end of the second *CYP51* fragment (grey box in Figure 4.1). In *SlPM37-032,* which has the longest acquired *AGO1* fragment, the unaligned DNA sequence between the *AGO1* and

CYP51 fragments is 9 bp in length and occurs from bp 650-658. In other elements, this filler DNA ranges from 28 bp in *SlPM37-011* to only 3 bp in *SlPM37-013.* The unaligned sequence was unrelated to either *AGO1* or *CYP51* and the sequence was of undetermined origin in the tomato genome. Two elements related to *SlPM37-035* were identified in potato. The potato elements did not have regions matching the *AGO1* gene in potato or tomato supporting the presence of copies similar to *SlPM37-035* but not to *SlPM37-013* in potato. Compared to the parental gene, there is no deletion in the acquired *CYP51* fragment inside the potato Pack-MULE element, suggesting the deletion in the *CYP51* region within the tomato elements occurred after the potato-tomato divergence. The parental *CYP51* gene in potato showed 96% similarity to the tomato gene. The acquired regions in the two potato elements show 94.32% and 94.21% similarity to the potato CYP51 gene (LOC102601151). These two *CYP51* parental genes show higher similarity (96%) to each other than the *SlPM37 CYP51* fragments show to the parental gene (Table 4.1).

Moderate variation of *SlPM37* copy numbers among tomato and its close relatives

Using available genome sequences across the tomato clade the copy number of *SlPM37* was estimated. Overall there is less than a 2-fold difference in copies per genome. From 31 copies, in *S. pennellii,* to 59 copies, in *S. chilense*, Figure 4.2. The high copy number in the tomato clade contrasts with only two copies in potato, showing that *SlPM37* has increased in copy number after the divergence of tomato and potato. Further delving into the structure of these elements shows they contain *AGO1/CYP51* fragment elements and elements with only the *CYP51* fragment. However, *CYP51* exclusive elements exhibit a much lower copy number than

the *AGO1/CYP51* elements in any species, ranging from 0 to 6 copies or 0 to 25% of the total intact *SlPM37* elements (Figure 4.2).

A second approach using DNA blotting was used to confirm and extend the copy number data. DNA blotting allows for querying tomato species that do not have sequenced genomes available. We then queried the copy number of *SlPM37* across the tomato clade and in other solanaceous relatives such as *Solanum dulcamara* (Särkinen et al. 2013). The results using a *CYP51* or *AGO1* probe, in Figure 4.3, show that all species in the tomato clade have copies of *SlPM37,* corroborating the bioinformatics approach. In species evolutionary more distant to tomato than potato, *S. dulcamara, S. melongena, C. annuum,* and *P. axillaris* do not show a Pack-MULE copy on the blot, Figure 4.3. This result was confirmed by a bioinformatic search, where no significant matches to *SlPM37* elements were found in available genomes of *S. melongena, C. annuum,* and *P. axillaris*. This corresponds to a recent divergence time as potato and tomato diverged from the tomato 8 MYA, while *S. dulcamara*, eggplant, pepper, and petunia diverged more distantly, 13.7, 19.1, and 31.2 MYA respectively (Bohs 2005; Wang et al. 2008; Särkinen et al. 2013). When looking at the individual species, the banding pattern between *S. lycopersicum* and *S. pimpinellifolium* is similar while banding patterns vary among the other species. These data combined with the bioinformatic approach suggest *SlPM37* elements are present in high copy numbers in the tomato clade, comprised of both *AGO1/CYP51* fragments in the elements as opposed to elements contains exclusively the *CYP51* fragment elements (Figure 4.3). Also, this approach shows *SlPM37* has amplified and increased in copy number over time since the divergence of tomato and potato.

Conservation and polymorphism of *SlPM37* insertions in tomato and its relatives

Using the genomic locations of the identified *SlPM37* copies in the tomato clade we queried the flanking regions to compare insertion locations in other genomes. The data shown are comparing flanking sequences from *S. lycopersicum SlPM37* elements to flanking regions in other genomes. Using the elements in tomato, the flanking sequences for 45 elements were unambiguously identified in the close relative *S. pimpinellifolium,* and all of them were associated with the corresponding elements in tomato (Table 4.2). In *S. galapagense* 42 copies shared flanking regions while flanking sequences of 5 copies were undetermined. In the more distant *S. arcanum*, flanking regions of 19 copies were identified but the elements were absent, 13 copies showed shared flanking regions and 15 were undetermined. In further distantly related *S. chilense,* 19 flanking regions were found without the elements, 9 elements were present, and flanking sequences of 19 elements were undetermined. For *S. habrochaites* and *S. pennellii,* 15 and 22 element copies were absent and 4 and 6 were shared with *S. lycopersicum* respectively, while 28 and 19 copies in these two genomes were undetermined when compared to *S. lycopsericum SlPM37* elements. These data show how insertions are shared among the closely related *S. lycopersicum*, *S. pimpinellifolium*, and *S. galapagense*. Among more distantly related species only a subset of element insertions is conserved showing these elements are amplified over evolutionary time.

Phylogeny of *SlPM37* in tomato

After aligning *SlPM37* sequences, phylogenetic relationships of the elements were built using a maximum likelihood approach. The resulting Maximum Likelihood tree shows the 47 elements are grouped into 3 clades (Figure 4.4). Delving into elements with shared acquired fragments, the three elements that only have the *CYP51* fragment (*SlPM37-0012*, *SlPM37-023*,

and *SlPM37-035*), *SlPM37-012,* and *SlPM37-035* group together. Also many are unresolved due to their shared similarity. The elements overall have varying levels of homology. Several elements share 98% similarity for example *SlPM37-047* and *SlPM37-003*. *SlPM37-030* is the shortest element at 669 bp and *SlPM37-046* is one of the longer elements at 1,330 bp. These elements are diverse and show shared similarities due to evolutionary history, yet it is clear that only elements with both *AGO1* and *CYP51* fragments have amplified recently (Figure 4.4). The phylogram shows several clades of *SlPM37* elements grouping into unresolved groups. Additionally, the acquired and unacquired *AGO1* and *CYP51* fragments are seen as long branches clustered outside of the short branches of the *SlPM37* elements. This phylogram shows that some individual elements are closely related and that the tomato and potato parental genes are more distantly related to *SlPM37* elements than to each other. This would be expected due to the functional constraints on the genes and drift in the *SlPM37* elements.

sRNA mapping to *SlPM37* elements and parental genes

To investigate whether *SlPM37* elements were associated with any sRNAs, sRNA datasets were used to query the presence of reads mapping to *SlPM37*. As shown in Table 4.3, the abundance of sRNAs matching elements and parental genes varies among experiments. All sRNAs mapping to *SlPM37* elements were combined due to the dearth of sRNA reads mapping to individual elements. The summary results for *S. lycopersicum* show 197, 108, and 5,142 reads mapping to the unacquired regions of *CYP51*, *AGO1*, and *SlPM37* elements respectively and these correspond to 0.43, 0.21, and 0.36 FPKM respectively (Table 4.3). Please note the reads for *SlPM37*were divided by its copy number when calculating FPKM so collectively there were many more reads mapping to a certain region in *SlPM37* than that in parental genes despite the

roughly comparable FPKM values. For the acquired regions the summary number of reads mapping were 37, 77, 751, and 411 for *CYP51, AGO1, SlPM37 CYP51*, and *SlPM37 AGO1* fragments respectively, corresponding to 0.40, 0.51, 0.34, and 0.19 FPKM respectively (Table 4.3). These results show reads mapping to both the acquired and unacquired regions across *S. lycopersicum.* In the *S. lycopersicum x S. pennellii* introgression experiments the summary results for reads show 91, 21, and 57 reads mapping to the unacquired regions of *CYP51, AGO1,* and *SlPM37* elements respectively and these correspond to 0.24 0.05, and 0.00 FPKM respectively (Table 4.3). For the acquired regions the summary number of reads mapping were 12, 7, 0, and 92 for *CYP51, AGO1, SlPM37 CYP51*, and *SlPM37 AGO1* fragments respectively, corresponding to 0.16, 0.06, 0.00, and 0.05 FPKM respectively (Table 4.3). While the datasets include introgression lines the specific locations of the *SlPM37* elements in the introgressed regions are unknown. In the *S. habrochaites* dataset the sRNA reads show 4, 6, and 114 reads mapping to the unacquired regions of *CYP51, AGO1*, and *SlPM37* elements respectively and these correspond to 0.14, 0.19, and 0.13 FPKM respectively (Table 4.3). The results for the acquired regions reads mapping were 0, 3, 7, and 5 for *CYP51, AGO1, SlPM37 CYP51*, and *SlPM37 AGO1* fragments respectively, corresponding to 0.00, 0.32, 0.00, and 0.55 FPKM respectively (Table 4.3). Overall the Pack-MULEs have a high number of reads of sRNA present, showing targeting of these elements in the sRNA pathway. These results show sRNAs are present and targeting the unacquired and acquired regions at similar levels in the Pack-MULEs. For parental genes, the acquired region in *AGO1* demonstrated higher sRNA read density than that of the non-acquired region in tomato leaves (0.51 vs 0.21, Table 4.3), while that value for CYP51 was comparable (0.40 vs 0.43).

Expression of parental genes in the tomato clade

Gene expression may vary due to copy number differences or via transcriptional repression mechanisms. In order to elucidate the expression across the tomato clade, a Nanostring probe set to both acquired and unacquired *AGO1* and *CYP51* regions were designed (Supplementary Table 4.2). Across species, a wide variation in expression of both *AGO1* and *CYP51* is observed (Figure 4.5). However, in all the species, the abundance of *CYP51* transcripts is approximately twice as high as that for *AGO1*, which is consistent with previous data showing higher expression of the *CYP51* parental gene with 143.22 RPKM in leaves, 155.08 RKPM in fully opened flowers, while *AGO1* expression showed 47.83 RPKM in leaves, 75.69 RPKM in fully opened flowers) (Sato et al. 2013). The *AGO1* parental gene shows higher but nonsignificant read counts in the acquired region than the average in the unacquired region (328 vs. 353; t-test $p = 0.485$). Correlating the expression to the copy number of *SlPM37* did not show a significant correlation either with the *AGO1* acquired or unacquired regions (Pearson correlation coefficient $R= 0.429$, $p = 0.404$, and $R = 0.343$, $p = 0.506$, respectively). *The CYP51* unacquired regions did not show any significant differences to the acquired region; t-test, $p = 0.103$. Further, testing the correlation of the read counts to *SlPM37* copy number in the genomes showed no significant relationships with the *CYP51* unacquired regions and the acquired region; Pearson correlation coefficient $R = 0.267$, $p = 0.609$, $R = 0.271$, $p = 0.604$, and $R = 0.311$, $p = 0.550$, respectively. The above results indicate parental gene expression does not significantly vary with *SlPM37* copy number among different species.

DISCUSSION

Structural variation of *SlPM37* due to acquisition of gene fragments and new insights into acquisition mechanisms

Despite the prevalence of Pack-MULEs in plants, the mechanism of sequence acquisition by Pack-MULEs is still unclear. According to one model, novel sequences can be introduced into the element through the nicks formed when the Pack-MULE sequences are present as stem loops due to the presence of long TIRs (Bennetzen and Springer 1994). Alternatively, genomic sequences can be acquired when a template is switched during gap-repair processes (Engels et al. 1990; Yamashita 1999), and the micro-homology between the sequence inside the element and the template sequence would favor the template switch. In this study, two types of *SlPM37* elements were found in the tomato genome. One is only associated with the *CYP51* fragment and the other contains both *AGO1* and *CYP51* fragments. Several lines of evidence suggest that the *AGO1* fragment was acquired after the acquisition of the *CYP51* fragment in a step-wise fashion: 1) the *AGO1* fragment is located inside the *CYP51* fragment; 2) elements with *AGO1* and *CYP51* fragments are present only in tomato clade, while the elements with only *CYP51* fragments are found in both tomato and potato; 3) in general the *AGO1* fragments are more similar to the parental gene than the *CYP51* fragment (Table 4.1). The micro-homology of 6 bp (TTGTCT) between the *AGO1* and *CYP51* fragment seems to be consistent with the template switch mechanism proposed by previous models (Engles et al. 1990; Springer and Bennetzen 1994; Yamashita et al. 1999; Bennetzen 2005). In addition, Pack-MULEs tend to acquire fragments from genes that are widely expressed (Ferguson et al. 2013). Both the *CYP51* and the *AGO1*

parental genes are broadly expressed across tissues in tomato (Sato et al. 2013), suggesting these acquisition events share observed Pack-MULE characteristics.

MULEs are DNA elements, which use a DNA intermediate for transposition. Our previous study indicates that in some cases, both exon and intron sequences are acquired/ duplicated inside Pack-MULEs, suggesting the acquisition occurred at the DNA level, not cDNA level (Jiang et al. 2004). In this study, the acquired *AGO1* fragment encompasses four exons so it is expected to detect sequences of three introns inside the elements if the acquisition is at the DNA level. Nevertheless, intron sequences are completely absent from the acquired regions. Certainly, we could not rule out the possibility that the initial acquisition was at the DNA level and later intron sequences inside Pack-MULEs were lost. However, if the acquisition of *AGO1* fragment occurred after the divergence of tomato and potato, it is unlikely all the intron sequences were completely lost in a few million years. Moreover, there is no evidence any of the Pack-MULEs have coding capacity so it is unclear why exon sequence inside the Pack-MULE is selectively retained over introns. On the other hand, if the acquired sequence is introduced into the element through a template switch, it is possible for a cDNA sequence to be used as a template. If that is the case, some of the "retrogenes" in the genome might have been created by DNA transposons, not retrotransposons.

Temporal mode of amplification of *SlPM37*

As mentioned above, it is likely that the acquisition of the *CYP51* fragment occurred prior to that of the *AGO1* fragment. Based on the distribution of the two types of *SlPM37* elements (with *CYP51* fragment only and with both gene fragments), the most parsimonious speculation is that *CYP51* fragment was acquired before the divergence of tomato and potato,

about 8 MYA (Wang et al. 2008; Särkinen et al. 2013), followed by the acquisition of *AGO1* fragment, likely after the divergence of tomato and potato. Since all the species in the tomato clade share a subset of the *SlPM37* elements, it suggests the acquisition of *AGO1* fragment, and the initial amplification of *SlPM37* occurred before the speciation of the tomato clade. In addition to the conserved insertion of *SlPM37* elements, there are polymorphic insertions inside the tomato clade, indicating the amplification continued after the speciation in this clade. Banding pattern difference was observed between *S. lycopersicum* and *S. peruvianum* (Figure 4.3), which have diverged around 2.0-2.9 MYA (Särkinen et al. 2013). In contrast, no polymorphic insertion was found between *S. lycopersicum* and *S. galapagens* (divergence between 0.19 – 0.25 MYA), suggesting amplification of *SLPM37* terminated between 0.25 to 2 MYA. In other words, *SlPM37* had been active till a time point between 0.25 to 2 MYA in the linage that led to cultivated tomato. After a TE family loses activity, their copy number tends to decline due to illegitimate recombination, random deletion, and other mutation processes that render them unrecognizable. So it is conceivable prior to the loss of transposition activity, the copy number of *SlPM37*was higher than that was detected from current genomes and some family members were lost over time.

***SlPM37* elements have been subject to epigenetic regulation through small RNA mediated mechanisms**

Due to their intrinsic feature as mutagens, extensive amplification of TEs is detrimental to the host organisms. To safeguard the genomes from the deleterious effects generated by TEs, a variety of epigenetic mechanisms have been developed to silence TEs, including small-RNA–mediated DNA methylation (Slotkin and Martienssen 2007). The sRNA search in this study

uncovered multiple reads mapping in *SlPM37*, both *AGO1* and *CYP51* regions, and the unacquired regions, demonstrating that the tomato genome and other related genomes are producing sRNA matching *SlPM37* sequences. These reads varied in length from 18-51 bp with an average of 30 bp, suggesting that multiple sRNA pathways are generated from these loci (Ito 2012; Kumar and Sathishkumar 2017). The evidence suggests that the genome is generating unique sRNAs to the Pack-MULE acquired regions. sRNA reads map across experiments demonstrating specific fragments target not only the TIR regions but additionally the acquired regions. This indicates *SlPM37*elements have been subject to sRNA mediated silencing mechanisms despite the fact that they lost transposition activity 0.25 to 2 MYA.

In addition to sRNAs matching the *SlPM37* elements, sRNA reads mapping to the parental genes were also detected, albeit with a much lower abundance (Table 4.3). It remains a question whether such a level of sRNAs is sufficient to cause any meaningful impact on gene expression. Moreover, it is unclear whether sRNAs for the acquired regions could have "cross talk" between the Pack-MULEs and the parental genes. For the *AGO1* gene, it appears that the acquired region was associated with more abundant sRNAs (FPKM 0.51 vs 0.21) while no difference was observed for *CYP51*. This seems to imply the *AGO1* acquired region has been influenced by the amplification of *SlPM37* whereas the effect on *CYP51* was not obvious. Such differentiation is understandable in that in most *SlPM37* elements, the acquired *AGO1* fragment is much longer than the *CYP51* fragment. In addition, the acquired *AGO1* fragment inside the *SlPM37* is more similar to the parental gene than the *CYP51* counterpart is (Table 4.1). Accordingly, longer acquisition combined with higher similarity allows "cross talk" to occur for *AGO1* but not *CYP51*. Nevertheless, no correlation between expression level and copy number

of *SlPM37* cross the tomato clade was observed under our experimental conditions. This suggests, the cross talk, if exists, is not robust enough to mask other regulatory mechanisms. In a previous study, it was shown that the abundance of sRNAs, either the total sRNAs or the sRNAs shared between the Pack-MULE and the parental gene, declined with the age of the Pack-MULE (Hanada et al. 2009). Although there is no unambiguous evidence for the modulation of parental gene expression by *SlPM37* elements at the current time, we cannot rule out the possibility for its occurrence in the past – when *SlPM37* reached a maximum copy number after amplification and were more similar to the parental genes.

Possible evolutionary trajectory of *SlPM37*

As mentioned above, two distinct genes were acquired by *SlPM37*. *AGO1* is critical in the sRNA pathway (Carbonell 2017). *AGO1* loads a guide RNA and binds complementary transcripts, cleaving the transcripts, and thus reducing the expression of the target genes (Bohmert et al. 1998; Hutvagner and Simard 2008; Carbonell 2017). Specifically, *AGO1* functions in the sRNA pathway to bind and then cleave RNA transcripts (Bohmert et al. 1998). The *CYP51* gene family is involved in the sterol 14α-demethylase pathway. These genes are widely conserved across species due to their essential metabolic function (Kushiro et al. 2001; Lepesheva and Waterman 2007; Nelson et al. 2008). Based on current evidence, the *CYP51* exclusive *SlPM37* elements are always associated with low copy numbers, either with or without the presence of elements with both gene fragments. Therefore, the amplification of *SlPM37* should be attributed to the acquisition of the *AGO1* fragment. The occurrence of high copies of Pack-MULE in multiple genomes suggests that the element has some characteristic that either

prevents genomic control or promotes transposition/retention resulting in an increase in copy number.

It was demonstrated that *AGO1* mutants lead to increased resistance to *Botrytis cinerea* in arabidopsis (Weiberg et al. 2013). This is because knockout of *AGO1* prevents the pathogen from hijacking the host RNA interference machinery to selectively silencing host immunity genes. If we consider in the early stage of speciation of tomato clade, the *SlPM37* elements were rapidly increasing their copy number, whereas the elements were highly similar to the *AGO1* gene. The amplification of *SlPM37* triggered the host silencing mechanism, resulting in large amounts of sRNAs matching both the elements and the parental genes. The host-pathogen interaction parallels *SlPM37* interactions with the genome, whereby the *AGO1* fragments inside the Pack-MULEs create a knockdown expression effect in the plant enhancing the disease resistance of the plant (Figure 4.6). A potential explanation is that *SlPM37* enhanced disease resistance, which would increase the fitness of the plant leading to the proliferation of plants with high copy numbers of *SlPM37*. As the elements mutated the resistance benefit is decreased and the pathogen resistance is attenuated. The cycle continued as *SlPM37* increased in copy number which led to enhanced silencing and eventually *SlPM37* lost transposition activity. A mechanism of this sort could explain the continuous increases in copy number across the tomato clade. If that is the case the *SlPM37* elements have been interacting with the genome in a novel and fascinating manner.

MATERIALS AND METHODS

Plant Materials

Tomato and wild relative species seeds were obtained from the CM Rick tomato genetics resource center (https://tgrc.ucdavis.edu/index.aspx). Seeds were germinated on Whatman #1 filter paper using distilled water in petri dishes and transferred to potting soil: perlite mix 3:1. Plants were grown in the MSU research greenhouses, East Lansing Michigan. Plant accessions used were *S. lycopersicum* Moneymaker, *S. pimpinellifolium* LA1589, *S. galapagense* LA0528, *S. cheesmaniae* LA1037, *S. chmielewskii* LA1318, *S. arcanum* LA1706, *S. neorickii* LA1326, *S. huaylasense* LA1982, *S. peruvianum* LA0098, *S. corneliomulleri* LA1293, *S. chilense* LA1932, *S. habrochaites* LA177, *S. pennellii* LA1926, *S. juglandifolium* LA2788, *S. lycopersocides* LA2408, *S. sitens* LA4113. Other species were obtained from the USDA germplasm repository *S. dulcamara* PI 643457, *S. tuberosum* RH89-039, *S. melongena* PI 643457, *Capsicum annum* PI 592831, and *Petunia axillaris* PI 667515.

Genome Sequences and *SlPM37* Search

*Solanum lycopersicum*v4.0 (Sato et al. 2012) (https://solgenomics.net/organism/Solanum_lycopersicum/genome), *S. pennellii* (Bolger et al. 2014), and *S. tuberosum*v3.0 (Xu et al. 2011) genomes were downloaded from NCBI. *S. chilense* (Stam et al. 2019), *S. pimpinellifolium* (Razali et al. 2018)*, S. galapagense,* and *S. melongena* (Barchi et al. 2019) were downloaded from the Sol Genomics FTP portal (Fernandez-Pozo et al. 2015). *S. arcanum* (Aflitos et al. 2014), and *S. habrochaite*s (Aflitos et al. 2014) were downloaded from European biotechnology institute (Cook *et al.* 2020).

The *SlPM37* elements in those genomes were identified as described previously with slight modifications (Ferguson and Jiang 2012). Briefly, a tomato MULE TIR library from the previous study (Ferguson and Jiang 2012) was used to search for intact MULEs with TSDs with the following criteria: (1) distance between the external TIRs is not larger than 20 kb and there is no sequencing gap between the TIRs (all final *SLPM37* elements are < 1.4 kb, Table 1), (2) TIRs must be at least 50 bp long, (3) truncations at the external ends of TIRs must be no more than 15 bp, and (4) presence of a 9–11 bp TSD with no more than 2 mismatches. The retrieved sequences of intact MULEs were searched against the sequences of *CYP51* and *AGO1B* gene in tomato to extract the *SlPM37* elements. *CYP51* gene identified was NP_001234537.2 (Solyc01g008110) (https://solgenomics.net/). For table 4.1 the cDNA was used for *CYP51*. In potato the *CYP51* gene searched was XM_006348474.2 (https://www.ncbi.nlm.nih.gov/). *AGO1* gene identified was Solyc03g098280, (https://solgenomics.net/). Elements that did not pass the criteria above (for example, one TIR truncated more than 15 bp or there was a 3-bp mismatch in the TSD) but fully alignable (indel < 100 bp) with at least one intact element, containing at least one gene fragment, were considered as truncated elements.

Presence and absence of *SlPM37* elements in the genome of wild relatives

Based on the coordinates of intact *SlPM37* elements in tomato, two sets of sequences were built. One set contains junction sequences of the insert site (100 bp of upstream flanking sequence plus 100 bp of 5' end of TIR sequence, 100 bp 3' end of TIR sequence plus 100 bp of downstream flanking sequence). Although TIR sequences are highly repetitive, each insertion site represents a unique combination of flanking and element sequences, so the entire junction sequence remains one copy in the genome unless there is segmental duplication encompassing

the insertion site, which has not been detected for any of the intact *SlPM37*elements. The second set of sequences resemble the "empty site" (the sequences status prior to the insertion), i.e. 100 bp upstream flanking sequence plus 100 bp downstream flanking sequence with one copy of TSD excluded. The two sets of sequences were used to search other genomes. If any of the entire junction sequences matched a locus with identity over the cutoff (see Table 4.2) and only half of the "empty site" was returned, this element was considered to be present in the relevant genome. If none of the junction sequences were retrieved and the entire "empty site" matched one single locus and the identity was over the cutoff, the relevant element was considered to be absent from the genome. If the entire "empty site" matched multiple loci, it implied the flanking sequences were repetitive. In this case, if the identity between the best match and the query was over the cutoff, and the identity between the second match and the query was at least 5% lower, the element was considered to be absent from the genome. All other cases were considered "undetermined".

sRNA search

NCBI sequence read archive (SRA) was used to search for sRNA matches to each of the 47 *SlPM37* elements against an sRNA dataset from tomato and other genomes using default settings (Altschul *et al.* 1990). Additionally, parental genes *AGO1* and *CYP51* were searched with BLAST using default parameters. Resulting matches were categorized into acquired, unacquired, and acquired overlapping regions for both *AGO1* and *CYP* based on their coordinates inside *SlPM37* and genes. Overlapping acquired regions were counted into both genes. FPKM was calculated using the number of fragments / (length of transcript/1000) / (total reads/10^6).

DNA Blotting

DNA was extracted from fully expanded leaves using a modified CTAB protocol (Porebski *et al.* 1997). The extracted DNA was digested at 37°C for 6 hours using the HindIII restriction enzyme to conduct Southern DNA blot analysis (New England Biolabs, Massachusetts, USA). The digested DNA was resolved on a 1% (w/v) agarose gel, followed by transfer of the DNA to a nylon membrane (GE Healthcare, Pittsburgh, PA) using capillary flow. The resulting blot was probed for the following sequences; *AGO1* acquired sequence amplified from genomic DNA with primers forward- CGATAATGGCCGAGAAAAAGATTG, reverse- GCCTAGATATTGCATCAACTAG. *AGO1* unacquired region with primers forward- GTCCTACAATCATTTTTGGTGC, reverse- CCTGTATGCCAGCATGACTAC. *CYP51* acquired region forward- TTCAGCTATGGAGTTAGGTGAC, reverse -CAGAAACTTCTGGACCAATGAAG. *CYP51* unacquired region forward- TAGCCAACAAGAGGTTTATCAG, reverse- AGAGCAGACACATCGTCAAAG. The resulting image was exposed for 4 hours, manually developed, and digitally scanned.

Phylogenetics

To construct a phylogeny for 47 *SlPM37* elements the entire element sequences were extracted from the tomato genome. No nested insertion of other TEs was detected inside any of the intact elements. These sequences were aligned using MUSCLE (Edgar 2004) with default settings in MEGAX (Tamura et al. 2011). The aligned sequences were then used to construct a Maximum Likelihood phylogram using MEGAX with default parameters and 1,000 bootstraps (Tamura et al. 2011). The resulting trees are shown with bootstrap percentages greater than 50%. The midpoint was used to root the tree.

Gene Expression

Nanostring Technologies was used to develop a probeset to measure transcripts from varying genomes. Target genes were identified and specific probes developed for acquired and unacquired regions of the parental genes, *AGO1* and *CYP51* using Nanostring quality control software. From these gene targets Nanostring Technologies, Washington, USA, manufactured a probe set for each gene using the *S. lycopersicum*.v2.40 genome, (Supplementary Table 4.1). Leaves from two separate plants were used as biological replicates. RNA samples were extracted from fully expanded leaves using a Trizol method with reagents from Invitrogen Thermo Fisher, California USA (Rio et al. 2010). RNA was quantified and quality controlled using a Qubit DNA fluorometer from Invitrogen Thermo Fisher, California USA. Nanostring expression experiment was conducted at the Michigan State University Research Technology Support Facility, East Lansing, Michigan USA. Read counts were analyzed using nSolver software from Nanostring Technologies, Washington USA (Kulkarni 2011).

ACKNOWLEDGEMENTS

This study was supported by National Science Foundation [MCB-1121650, IOS-1126998, IOS-1740874 to N.J.]; United States Department of Agriculture National Institute of Food and Agriculture and AgBioResearch at Michigan State University [Hatch grant MICL02707 to N.J.].

We would like to thank the CM Rick Tomato Genetics Resource Center at the University of California Davis for providing seeds for the tomato accessions.

APPENDIX

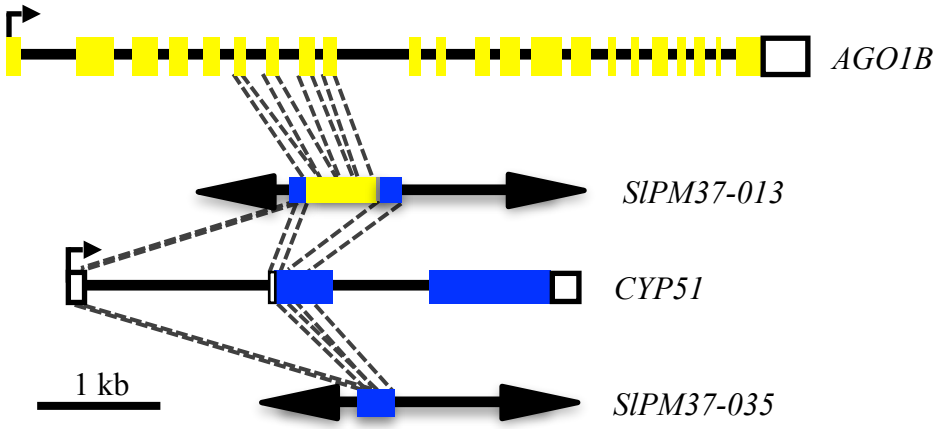
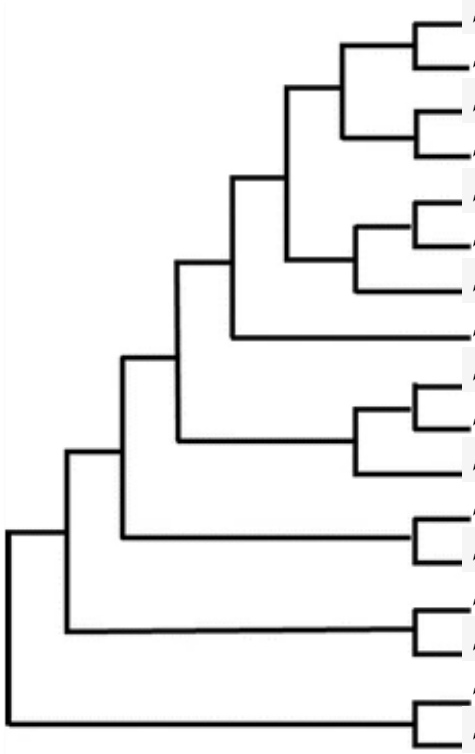


Figure 4.1. Schematic of *SlPM37* in tomato and acquired genes. Top, *AGO1* gene in tomato, yellow boxes are exons, black arrow is transcription start site (TSS), and white box is the UTR. Middle, *SlPM37*, black arrows are TIRs, Blue boxes are regions acquired from CYP51, yellow box is region acquired from *AGO1*, and the grey box is an unknown sequence. Bottom, *CYP51*, Blue boxes are exons, the arrow is TSS, and white box is the UTR. Dashed lines between parental genes and *SlPM37* are the acquired regions, not exon are present but introns were not found. *SlPM37-013* has 44 similar copies in the tomato genome. *SlPM37-035* has 3 similar copies in the tomato genome and 2 element copies in the potato genome.



Species	Intact elements with TSD		
	***CYP51* only**	**All**	**Total**
S. lycopersicum	3	47	55
S. pimpinellifolium	3	46	54
S. galapagense	0	5	49
S. cheesmaniae	-	-	-
S. chmielewskii	-	-	-
S. arcanum	4	24	39
S. neorickii	-	-	-
S. huaylasense	-	-	-
S. peruvianum	-	-	-
S. corneliomulleri	-	-	-
S. chilense	6	18	59
S. habrochaites	1	44	55
S. pennellii	4	19	31
S. ochranthum	-	-	-
S. juglandifolium	-	-	-
S. lycopersicoides	-	-	-
S. sitiens	-	-	-

Figure 4.2. Phylogeny of the tomato clade and estimated copy number of *SlPM37* in tomato and its close relatives. Left, phylogeny of the tomato clade, right estimated copy number of *SlPM37* in available genomes. "*CYP51* only" indicates number of intact elements only contain *CYP51* fragments. "All" indicates all intact *SlPM37* elements. "Total" includes both intact and truncated elements.

Left phylogram adapted from "Interspecific reproductive barriers in the tomato clade: opportunities to decipher mechanisms of reproductive isolation" Bedinger et al. 2011, *Sexual Plant Reproduction*, 24, 171-180.

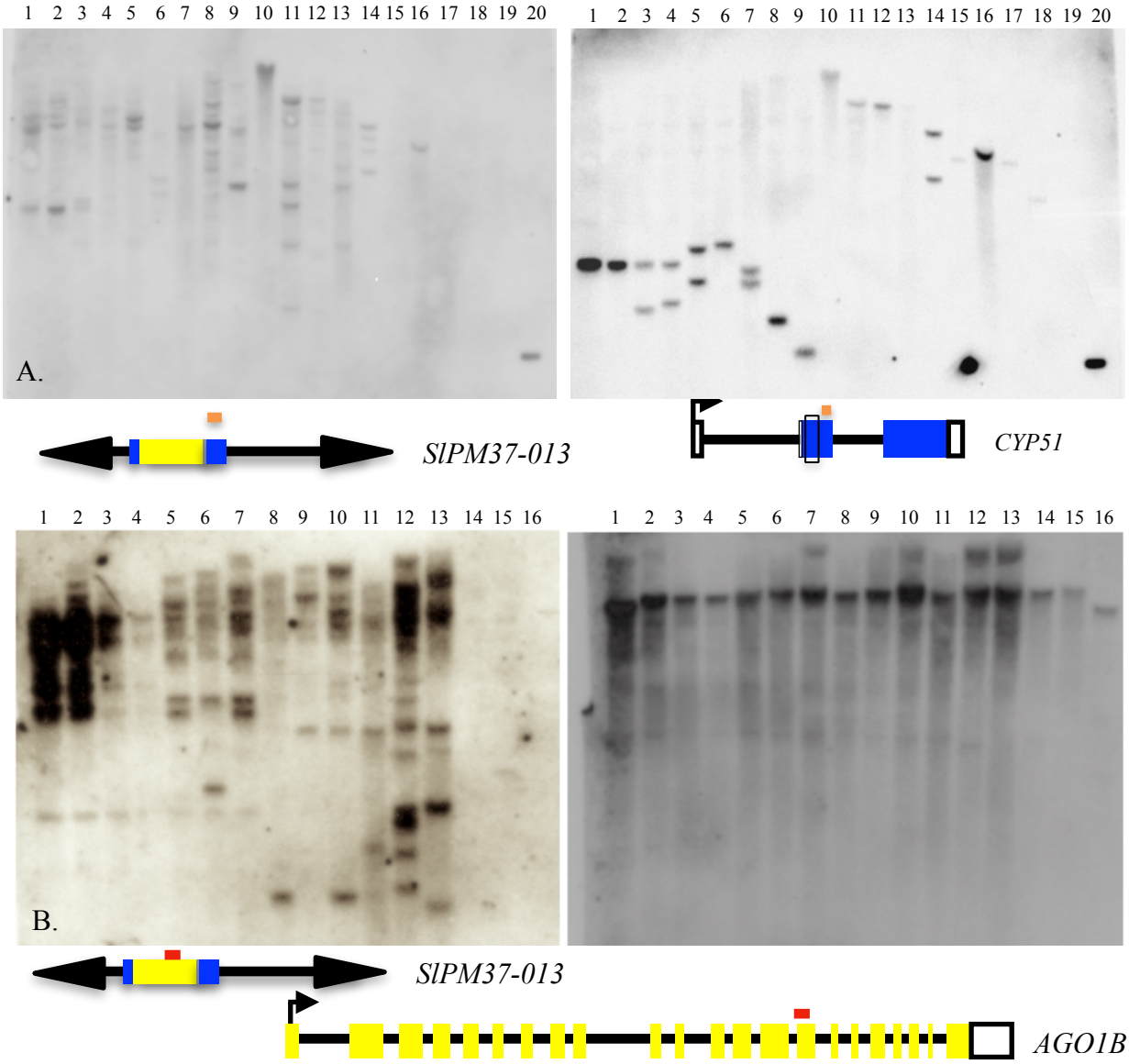


Figure 4.3. Identification of *SlPM37* elements in tomato and wild relatives using DNA blotting. A. Left, the image of DNA blot using the acquired region of *CYP51* as a probe; Right, the same blot using the unacquired region of *CYP51* as a probe. The species queried are 1 *S. lycopersicum,* 2 *S. pimpinellifolium,* 3 *S. arcanum,* 4 *S. huylasense,* 5 *S. peruvianum,* 6 *S. corneliomulleri,* 7 *S. chilense,* 8 *S. habrochaites*, 9 *S. pennellii,* 10 *S. juglandifolium,* 11 *S. ochranthum,* 12 *S. lycopersocides,* 13 *S. sitiens,* 14 *S. tuberosum,* 15 *S. dulcamara,*16 *S. melongena,* 17 *C. annum,*18 *P. axillaris,* 19 H_20, 20 1kb+ ladder. B. Left, the image of DNA blot using the acquired region of *AGO1* as a probe; Right, the same blot using the unacquired region of *AGO1* as a probe. Bottom, left, schematic of *SlPM37* with probed region in orange above the second *CYP51* region. Bottom, right, schematic of *AGO1* and *CYP51* gene with the *CYP51* probed region in orange *AGO1* in red, and acquired region in black box. 1 *S. lycopersicum,* 2 *S. pimpinellifolium,* 3 *S. arcanum,* 4 *S. huylasense,* 5 *S. peruvianum,* 6 *S. corneliomulleri,* 7 *S. chilense,*

Figure 4.3 (cont'd) 8 *S. habrochaites*, 9 *S. pennellii,* 10 *S. juglandifolium,* 11 *S. ochranthum,* 12 *S. lycopersocides,* 13 *S. sitiens,* 14 *S. tuberosum,* 15 *S. dulcamara,*16 *S. melongena*.

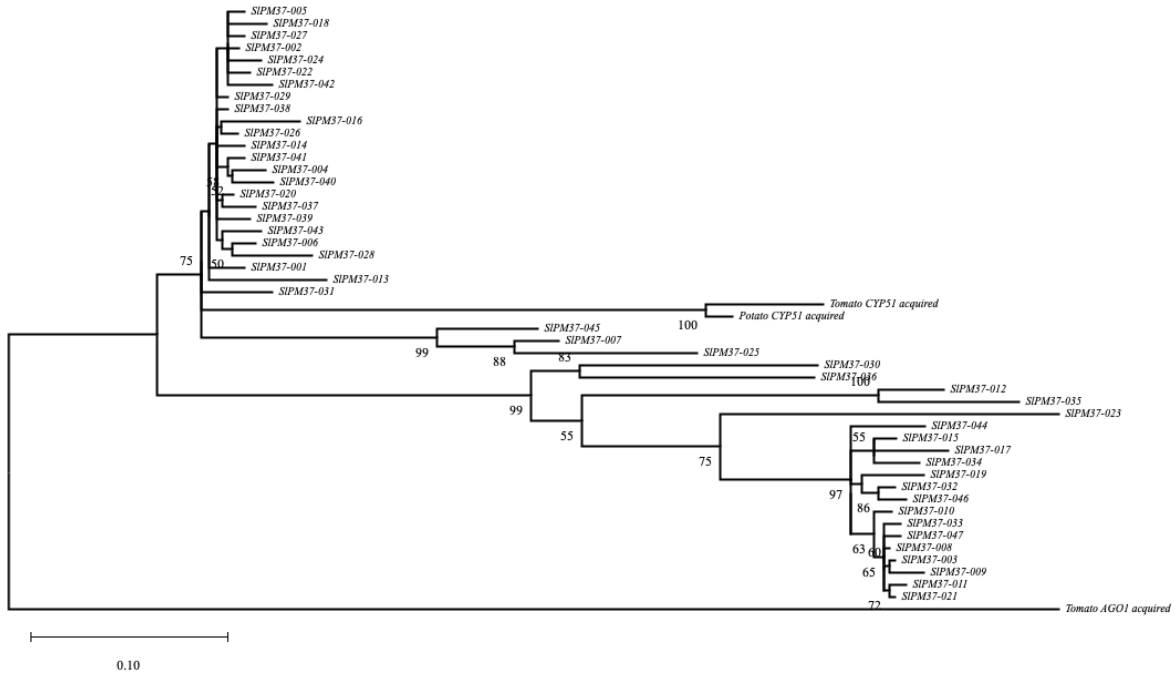


Figure 4.4. Maximum Likelihood phylogeny of 47 *SlPM37* intact elements in tomato and *AGO1/CYP51* acquired and unacquired fragments from tomato and potato. Maximum Likelihood method and General Time Reversible model +Gamma +Invariable, log-likelihood (-8005.50) shown. The percentage of trees in which the associated taxa clustered together greater than 50% is shown next to the branches using 1,000 bootstraps.

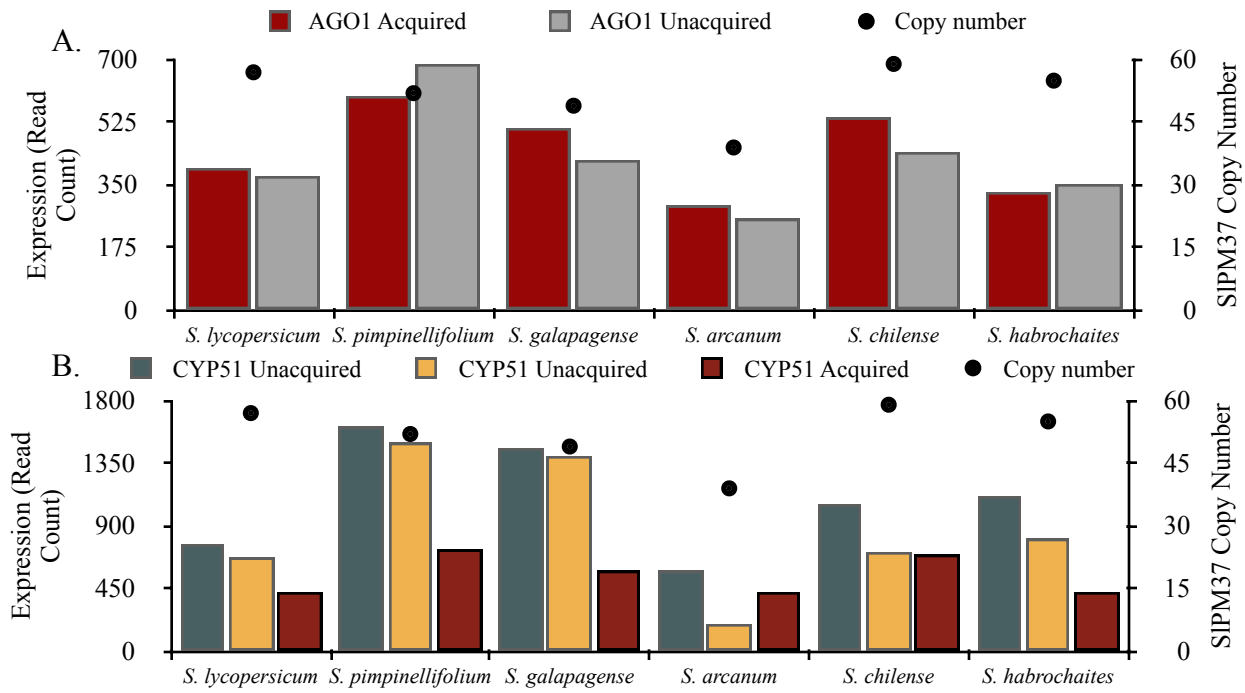


Figure 4.5. The transcript abundance of *CYP51* and *AGO1* (parental genes) across the tomato clade. A. The read abundance of *AGO1* acquired region, red bars, and *AGO1* unacquired regions, grey bars. B. The read abundance of *CYP51* unacquired region 1, blue bars, *CYP51* unacquired region 2, orange bars, and *CYP51* acquired region, red bars. For both plots, x-axis specifies species in the tomato clade, left y-axis, colored columns, represents the read count of the probes, right y-axis, black circles, represents the copy number of *SlPM37* in the genome.

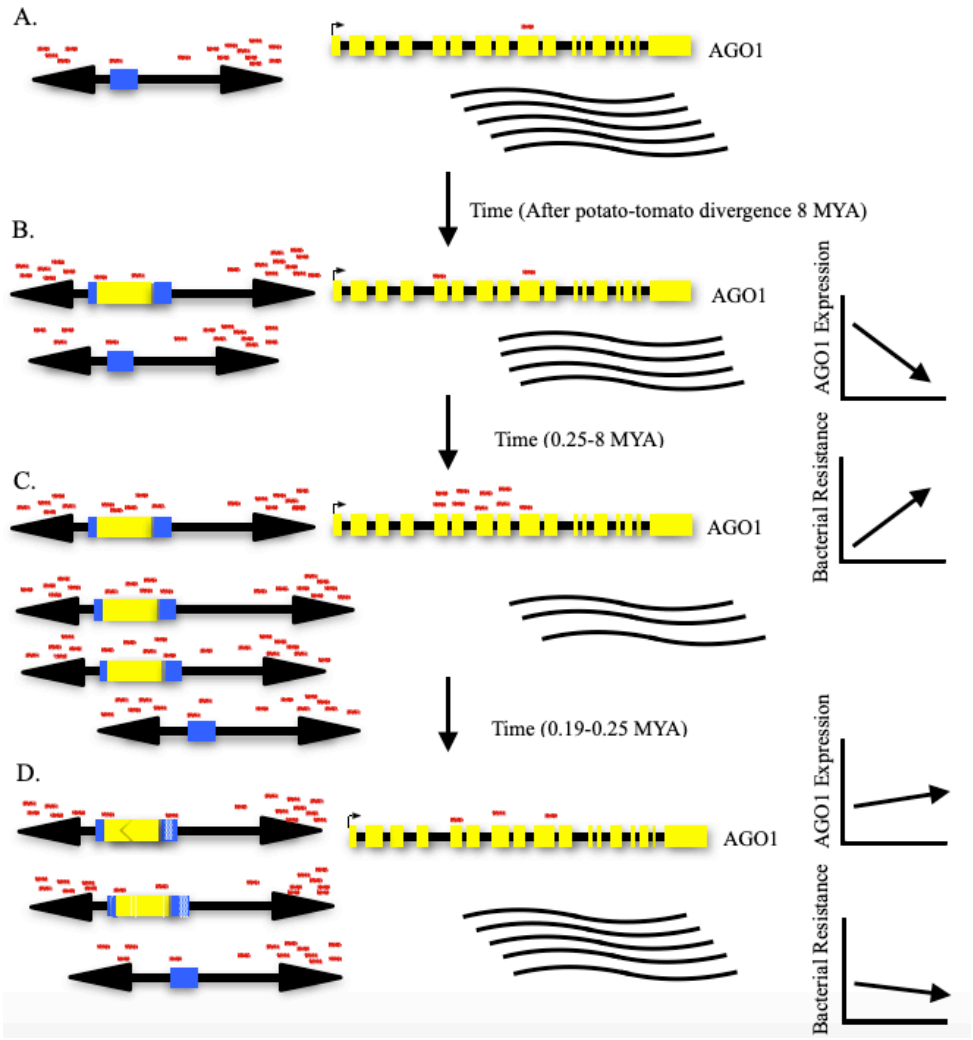


Figure 4.6. Proposed model of *SlPM37* evolution and genome interaction. Panel A. 8 MYA, before the tomato potato divergence. *SlPM37* in potato-tomato ancestor present in low copy numbers and generating sRNAs to TIR and CYP51 regions. Panel B. After potato-tomato divergence, 0.25-8MYA. *SlPM37* acquired *AGO1* fragment. sRNAs are generated for acquired regions targeting the *AGO1* parental gene, which reduces expression. Lower *AGO1* expression enhances disease resistance to bacterial pathogens. Plant fitness increases and *SlPM37* increases in copy number. Panel C. 0.19-0.25 MYA. *SlPM37* mobilization dissipates. Copy number is at a maximum. *AGO1* expression is low, bacterial, occurring before the *S. lycopersicum* - *S. galapagense* divergence. Panel D. Present time. *SlPM37* drifts and accumulates mutations, leading to a reduction in sRNAs targeting *AGO1*. *SlPM37* elements are mutated and decrease in copy number increasing AGO1 expression and decreases bacterial pathogen resistance. Left, *SlPM37* element, black arrows are TIRs, Blue boxes are regions acquired from *CYP51*, yellow box is region acquired from AGO1, and grey box is an unknown sequence, red lines are sRNAs. Middle, *AGO1* gene in tomato, yellow boxes are exons, black arrow is transcription start site

Figure 4.6 (cont'd) (TSS), wavy lines are transcripts. Right, graphical trend over time; top expression of *AGO1* parental gene, x-axis is time and y-axis is expression, bottom resistance of plants to bacterial pathogens, x-axis is time and y-axis is resistance.

TABLES

Table 4.1. *SlPM37* elements in the *S. lycopersicum* genome

Pack-MULE	Coordinates	Element length (bp)	*AGO1* parental fragment region (nt)	% Identity to *AGO1*	*CYP51* parental fragment region 1 (nt)	*CYP51* parental fragment region 2 (nt)	% Identity to *CYP51*
SlPM37-001	ch01_15558628_15559670	1,043	1511-1678	89.88	100-207	313-416	87.96
SlPM37-002	ch01_37871624_37872627	1,004	1530-1657	91.41	100-219	330-416	88.15
SlPM37-003	ch01_59663426_59664433	1,008	1530-1659	90.77	100-219	330-416	87.68
SlPM37-004	ch01_68124762_68125816	1,055	1511-1678	91.67	100-214	313-416	87.05
SlPM37-005	ch01_77868162_77869164	1,003	1530-1668	89.93	100-219	329-416	85.98
SlPM37-006	ch02_34002447_34003456	1,010	1511-1593	95.18	100-219	315-416	89.09
SlPM37-007	ch02_38288185_38289160	976	1751-1871	89.34	98-284		85.71
SlPM37-008	ch02_46699154_46700098	945	1530-1585	98.21	100-219	330-416	91.70
SlPM37-009	ch03_24901966_24902968	1,003	1530-1659	92.31	100-219	330-416	85.30
SlPM37-010	ch03_28206977_28207984	1,008	1530-1668	89.21	100-219	330-416	86.60
SlPM37-011	ch03_29197809_29198810	1,002	1530-1659	92.31	100-219	330-416	86.60
SlPM37-012	ch03_30395077_30396238	1,162	n/a		100-294	328-444	81.45
SlPM37-013	ch03_45753282_45754614	1,333	1511-1908	90.70	100-231	315-416	86.97
SlPM37-014	ch03_48936176_48937186	1,011	1511-1583	95.89	100-219	314-416	88.16
SlPM37-015	ch04_11690518_11691570	1,053	1511-1678	91.07	100-218	313-416	87.28
SlPM37-016	ch04_12027845_12029145	1,301	1511-1908	88.69	100-231	313-385	86.12
SlPM37-017	ch04_43537256_43538309	1,054	1511-1678	86.31	102-219	313-416	87.84
SlPM37-018	ch05_56976753_56977752	1,000	1530-1668	89.93	100-219	330-416	87.56
SlPM37-019	ch06_921288_922225	938	1511-1678	92.26	140-219	313-416	91.30
SlPM37-020	ch06_30895281_30896336	1,056	1511-1678	91.67	100-219	313-416	87.78
SlPM37-021	ch06_32305437_32306441	1,005	1530-1659	92.31	100-219	330-416	87.26
SlPM37-022	ch06_41414970_41415968	999	1530-1668	89.93	100-214	330-416	87.02
SlPM37-023	ch07_20785892_20786853	962	n/a		100-294	328-458	82.98
SlPM37-024	ch07_43927601_43928597	997	1530-1657	92.19	100-219	330-416	87.20
SlPM37-025	ch07_56760107_56761073	967	1757-2035	87.19	103-254		84.41

Table 4.1 (cont'd)

SlPM37-026	ch08_26562756_26563807	1,052	1513-1678	92.17	100-218	313-416	88.44
SlPM37-027	ch08_46434641_46435651	1,011	1530-1668	89.93	100-219	330-416	86.79
SlPM37-028	ch08_47946384_47947435	1,052	1511-1676	91.57	100-214	313-416	87.94
SlPM37-029	ch08_53851285_53852296	1,012	1511-1583	95.89	100-219	314-416	88.39
SlPM37-030	ch08_59794413_59795081	669	1958-2035	84.60	194-254		90.16
SlPM37-031	ch09_57393923_57394928	1,006	1511-1625	96.52	100-219	315-416	90.90
SlPM37-032	ch09_58475895_58477221	1,327	1511-1902	91.07	100-231	313-416	87.87
SlPM37-033	ch10_12882621_12883629	1,009	1530-1659	91.54	100-219	330-416	87.26
SlPM37-034	ch10_21761307_21762367	1,061	1511-1678	88.10	100-219	313-416	87.28
SlPM37-035	ch10_41769299_41770460	1,162	n/a		100-294	328-443	86.81
SlPM37-036	ch10_55157080_55158142	1,063	1645-2035	91.20	99-248		84.87
SlPM37-037	ch10_56429531_56430587	1,057	1511-1678	89.29	100-219	313-416	88.55
SlPM37-038	ch10_57687556_57688609	1,054	1511-1678	91.67	100-219	313-416	88.93
SlPM37-039	ch11_12891362_12892378	1,017	1508-1583	94.74	100-219	316-416	87.17
SlPM37-040	ch11_18525863_18526881	1,019	1511-1678	92.26	100-224	313-412	86.09
SlPM37-041	ch11_18535153_18536216	1,064	1511-1678	91.07	100-219	313-416	88.11
SlPM37-042	ch11_33270892_33271898	1,007	1530-1668	89.93	100-219	330-416	85.71
SlPM37-043	ch11_47472501_47473547	1,047	1511-1682	91.28	100-219	314-416	87.72
SlPM37-044	ch11_50292469_50293523	1,055	1511-1678	89.29	100-219	313-416	86.90
SlPM37-045	ch11_51749518_51750577	1,060	1651-2023	90.58	98-248		87.58
SlPM37-046	ch12_37376899_37378228	1,330	1511-1908	90.70	100-231	313-416	83.68
SlPM37-047	ch12_55080570_55081559	990	1530-1659	90.77	100-219	330-416	87.26
	Total length of Elements (bp)	49,019	Average % Identity to *AGO1* Parental	91.19		Average % Identity to *CYP51* Parental	87.22

Table 4.2. Conservation of intact *SlPM37* element insertions in tomato and its relatives

Species	Intact elements with flanking sequences in other species			Flanking sequences not determined	Identity cutoff (%)
	Absent	Present	% shared		
S. pimpinellifolium	0	45	100	2	97
S. galapagense	0	42	100	5	96
S. arcanum	19	13	40.6	15	94
S. chilense	19	9	32	19	93
S. habrochaites	15	4	21	28	93
S. pennellii	22	6	21	19	92

Table 4.3. sRNA reads mapping to *SlPM37*, *CYP51*, and *AGO1* in public databases

Organism		Number of reads mapping to unacquired region			FPKM unacquired			Number of reads mapping to acquired region				FPKM acquired				Database	Reference	Number of reads in database
	Tissue	Parental CYP	Parental AGO	SlPM	Parental CYP	Parental AGO	SlPM	Parental CYP	Parental AGO	SlPM CYP	SlPM AGO	Parental CYP	Parental AGO	SlPM CYP	SlPM AGO			
S. lycopersicum	Leaf	85	42	2801	0.75	0.33	0.79	16	30	501.50	377.50	0.69	0.80	0.86	0.69	SRX375787 SRX375786 SRX375785 SRX375784	Kravchik et al. 2014	72,723,946
S. lycopersicum	Leaf	17	6	0	0.77	0.24	0.00	0	2	0	18	0.00	0.27	0.00	0.17	SRX026449 SRX026450	Shivaprasad et al. 2012	14,166,899
S. lycopersicum	Leaf	2	1	2327	0.02	0.01	0.77	0	0	1	1	0.00	0.00	0.00	0.00	SRX099432 SRX099433	Zhong et al. 2013	61,817,189
S. lycopersicum	Leaf	80	54	14	0.98	0.59	0.01	20	44	248.50	14.50	1.21	1.64	0.59	0.04	SRX871217	Omidvar et al. 2015	52,194,436
S. lycopersicum	Leaf	13	5	0	0.09	0.03	0.00	1	1	0	0	0.03	0.02	0.00	0.00	SRX3197422 SRX3197423 SRX3197424 SRX3197425	Prigigallo et al. 2019	91,542,750
Summary	Leaf	197	108	5142	0.43	0.21	0.36	37	77	751	411	0.40	0.51	0.32	0.19			292,445,220
S. lycopersicum x *S. pennellii*	Leaf	40	9	13	0.28	0.06	0.00	6	3	0	53	0.21	0.06	0.00	0.08	SRX026453 SRX026454 SRX026455 SRX026456 SRX026457	Shivaprasad et al. 2012	92,276,277
S. lycopersicum x *S. pennellii*	Leaf	51	12	27	0.25	0.05	0.00	6	4	0	39	0.15	0.06	0.00	0.04	SRX026458 SRX026466	Shivaprasad et al. 2012	14,577,194
S. lycopersicum x *S. pennellii*	Leaf	0	0	17	0.00	0.00	0.02	0	0	0	0	0.00	0.00	0.00	0.00	SRX328502	n/a	19,940,788
Summary	Leaf	91	21	57	0.46	0.09	0.01	12	7	0	92	0.30	0.11	0.00	0.10			126,794,259
S. lycopersicum x *S. habrochaites*	Leaf	4	6	114	0.14	0.19	0.13	0	3	0	75	0.00	0.32	0.00	0.55	SRX3071584	n/a	18,280,589

SUPPLEMENTARY TABLE

Supplementary Table S4.1. nanoString Probes

Probe Name	Pack-MULE region	Accession #	Target Sequence
AGO1 a	Unacquired Upstream	NM_001279332.2	TGGAGGGCAGGGTCGTGGCCAGCGTCCACCCCAGCAGCATCAACATGAAGGTGGCTACCAAGGTGGAGGACAGGGTCGCGGCATGCGTCCTCCTCCCCAG
AGO1 b	Acquired	NM_001279332.2	TTTCATGATGCCATGACAAAGCTGCAGCCGCTATCGAAGGAGCTTGATCTCCTGGTCGCTATCTTGCCAGACAATAATGGCTCTCTTTATGGTGATCTGA
CYP51 a	Unacquired Downstream	NM_001247608.2	TCTGCTCTCTTCCATGACCTGGACAATGGGATGCTCCCTATCAGTGTTATCTTTCCCTACCTTCCAATTCCAGCCCATCGCCGACGTGACAATGCCAGGA
CYP51 b	Unacquired Downstream	NM_001247608.2	GTCACTTGCTGAGAAATTTCGAGTTTGAACTGATCTCGCCTTTCCCTGAAATTGACTGGAACGCTATGGTTGTCGGTGTCAAAGGGGAAGTAATGGTGAA
CYP51 c	Acquired	NM_001247608.2	ATGTGGGGTTGCTTTTAGTTGCAACCCTTTTGGTAGCAAAGCTCATATCTGCACTAATTATGCCCAGATCTAAGAAACGTTTGCCTCCAGTGGTTAAGGC

REFERENCES

REFERENCES

Aflitos S, Schijlen E, De Jong H, De Ridder D, Smit S, Finkers R, Wang J, Zhang G, Li N, Mao L, et al. 2014. Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing. Plant J. 80:136–148.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic Local Alignment Search Tool. J. Mol. Biol. 215:403–410.

Barchi L, Pietrella M, Venturini L, Minio A, Toppino L, Acquadro A, Andolfo G, Aprea G, Avanzato C, Bassolino L, et al. 2019. A chromosome-anchored eggplant genome sequence reveals key events in Solanaceae evolution. Sci. Rep. 9.

Bedinger PA, Chetelat RT, McClure B, Moyle LC, Rose JKC, Stack SM, van der Knaap E, Baek YS, Lopez-Casado G, Covey PA, et al. 2011. Interspecific reproductive barriers in the tomato clade: Opportunities to decipher mechanisms of reproductive isolation. Sex. Plant Reprod. 24:171–187.

Bennetzen JL, Springer PS. 1994. The generation of Mutator transposable element subfamilies in maize. Theor. Appl. Genet. Int. J. Plant Breed. Res. 87:657–667.

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. In: Annals of Botany.Vol. 95 Oxford University Press pp. 127–132.

Bohmert K, Camus I, Bellinni, Bouchez D, Cabouche M, and Benning C. 1998. AGO1 defines a novel locus of Arabidopsis controlling leaf development. EMBO J. 17:170–180.

Bohs L. 2005. Major clades in Solanum based on ndhF sequence data. Monogr. Syst. Bot. from Missouri Bot. Gard. 104:27–50.

Bolger A, Scossa F, Bolger ME, Lanz C, Maumus F, Tohge T, Quesneville H, Alseekh S, Sørensen I, Lichtenstein G, et al. 2014. The genome of the stress-tolerant wild tomato species Solanum pennellii. Nat. Genet. 46:1034–1038.

Carbonell A. 2017. Plant ARGONAUTES: Features, functions, and unknowns. In: Methods in Molecular Biology. Vol. 1640.

Cook CE, Stroe O, Cochrane G, Birney E, Apweiler R. 2020. The European Bioinformatics Institute in 2020: Building a global infrastructure of interconnected data resources for the life sciences. Nucleic Acids Res. 48:D17–D23.

Craig NL. 2002. Mobile DNA: an Introduction. In: Mobile DNA II. American Society of Microbiology pp. 3–11.

Darwin SC, Knapp S, Peralta IE. 2003. Taxonomy of tomatoes in the galápagos islands: Native and introduced species of solarium section *lycopersicon* (*solanaceae*). Syst. Biodivers. 1:29–53.

Eddy SR. 2012. The C-value paradox, junk DNA and ENCODE. Curr. Biol. 22:R898–R899.

Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Engels WR, Johnson-Schlitz DM, Eggleston WB, Sved J. 1990. High-frequency P element loss in Drosophila is homolog dependent. Cell 62:515–525.

Ferguson AA, Jiang N. 2012. Mutator -like elements with multiple long terminal inverted repeats in plants. Comp. Funct. Genomics 2012.

Ferguson AA, Zhao D, Jiang N. 2013. Selective Acquisition and Retention of Genomic Sequences by Pack-Mutator-Like Elements Based on Guanine-Cytosine Content and the Breadth of Expression. Plant Physiol. 163:1419–1432.

Fernandez-Pozo N, Menda N, Edwards JD, et al. 2015. The Sol Genomics Network (SGN)-from genotype to phenotype to breeding. Nucleic Acids Res. 43:D1036–D1041.

Haak DC, Kostyun JL, Moyle LC. 2014. Merging ecology and genomics to dissect diversity in wild tomatoes and their relatives. Ecol. Genomics. 781:273–298.

Hanada K, Vallejo V, Nobuta K, Slotkin RK, Lisch D, Meyers BC, Shiu S-HH, Jiang N. 2009. The Functional Role of Pack-MULEs in Rice Inferred from Purifying Selection and Expression Profile. Plant Cell 21:25–38.

Holligan D, Zhang X, Jiang N, Pritham EJ, Wessler SR. 2006. The transposable element landscape of the model legume Lotus japonicus. Genetics 174:2215–2228.

Hutvagner G, Simard MJ. 2008. Argonaute proteins: Key players in RNA silencing. Nat. Rev. Mol. Cell Biol. 9:22–32.

Ito H. 2012. Small RNAs and transposon silencing in plants. Dev. Growth Differ. 54:100–107.

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, Ferguson AA, Slotkin RK, Lisch D. 2011. Pack-Mutator-like transposable elements (Pack-MULEs) induce directional modification of genes through biased insertion and DNA acquisition. Proc. Natl. Acad. Sci. U. S. A. 108:1537–1542.

Kawasaki S, Nitasaka E. 2004. Characterization of Tpn1 family in the Japanese morning glory: En/Spm-related transposable elements capturing host genes. Plant Cell Physiol. 45:933–944.

Kidwell MG, Lisch DR. 2000. Transposable elements and host genome evolution. Trends Ecol. Evol. 15:95–99.

Kobayashi S, Goto-Yamamoto N, Hirochika H. 2004. Retrotransposon-Induced Mutations in Grape Skin Color. Science (80-.). 304:982.

Kobayashi S, Goto-Yamamoto N, Hirochika H. 2005. Association of VvmybA1 gene expression with anthocyanin production in grape (Vitis vinifera) skin-color mutants. J. Japanese Soc. Hortic. Sci. 74:196–203.

Kravchik M, Damodharan S, Stav R, Arazi T. 2014. Generation and characterization of a tomato DCL3-silencing mutant. Plant Sci. 221–222:81–89.

Kulkarni MM. 2011. Digital multiplexed gene expression analysis using the nanostring ncounter system. Curr. Protoc. Mol. Biol. 25B.10.1-25B.10.17.

Kumar RS, Sathishkumar SR. 2017. Small RNAs: Master Regulators of Epigenetic Silencing in Plants. In: Rajewsky N., Jurga S., Barciszewski J. (eds) Plant Epigenetics. RNA Technologies. Springer, Cham. p. 89–106.

Kushiro M, Nakano T, Sato K, Yamagishi K, Asami T, Nakano A, Takatsuto S, Fujioka S, Ebizuka Y, Yoshida S. 2001. Obtusifoliol 14α-Demethylase (CYP51) Antisense Arabidopsis Shows Slow Growth and Long Life. Biochem. Biophys. Res. Commun. 285:98–104.

Lepesheva GI, Waterman MR. 2007. Sterol 14α-demethylase cytochrome P450 (CYP51), a P450 in all biological kingdoms. Biochim. Biophys. Acta - Gen. Subj. 1770:467–477.

Lisch D. 2005. Pack-MULEs: Theft on a massive scale. BioEssays 27:353–355.

Lisch D. 2013. How important are transposons for plant evolution? Nat. Rev. Genet. 14:49–61.

Lisch D, Slotkin RK. 2011. Strategies for Silencing and Escape: The Ancient Struggle Between Transposable Elements and Their Hosts. Int. Rev. Cell Mol. Biol. 292:119–152.

McClintock B. 1950. The origin and behavior of mutable loci in maize. Proc. Natl. Acad. Sci. U. S. A. 36:344–355.

Moran J V., DeBerardinis RJ, Kazazian HH. 1999. Exon shuffling by L1 retrotransposition. Science (80). 283:1530–1534.

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.

Nelson DR, Ming R, Alam M, Schuler MA. 2008. Comparison of Cytochrome P450 Genes from Six Plant Genomes. Trop. Plant Biol. 1:216–235.

Nesbitt CT, Tanksley SD. 2002. Comparative sequencing in the genus lycopersicon: Implications for the evolution of fruit size in the domestication of cultivated tomatoes. Genetics 162:365–379.

Omidvar V, Mohorianu I, Dalmay T, Fellner M. 2015. Identification of miRNAs with potential roles in regulation of anther development and male-sterility in 7B-1 male-sterile tomato mutant. BMC Genomics 16.

Peralta IE, Spooner DM, Knapp S. 2008. The taxonomy of tomatoes: a revision of wild tomatoes (*Solanum* section *Lycopersicon*) and their outgroup relatives in sections *Juglandifolium* and *Lycopersicoides*. Syst. Bot. Monogr. 84:1–186.

Pickeral OK, Makałowski W, Boguski MS, Boeke JD. 2000. Frequent human genomic DNA transduction driven by line-1 retrotransposition. Genome Res. 10:411–415.

Porebski S, Bailey LG, Baum BR. 1997. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. Plant Mol. Biol. Report. 15:8–15.

Prigigallo MI, Križnik M, De Paola D, Catalano D, Gruden K, Finetti-Sialer MM, Cillo F. 2019. Potato virus y infection alters small RNA metabolism and immune response in tomato. Viruses 11.

Razali R, Bougouffa S, Morton MJL, Lightfoot DJ, Alam I, Essack M, Arold ST, Kamau AA, Schmöckel SM, Pailles Y, et al. 2018. The genome sequence of the wild tomato *solanum pimpinellifolium* provides insights into salinity tolerance. Front. Plant Sci. 9.

Rio DC, Ares M, Hannon GJ, Nilsen TW. 2010. Purification of RNA using TRIzol (TRI Reagent). Cold Spring Harb. Protoc. 5.

Särkinen T, Bohs L, Olmstead RG, Knapp S. 2013. A phylogenetic framework for evolutionary study of the nightshades (*Solanaceae*): A dated 1000-tip tree. BMC Evol. Biol. 13.

Sato S, Tabata S, Hirakawa H, Asamizu E, Shirasawa K, Isobe S, Kaneko T, Nakamura Y, Shibata D, Aoki K, et al. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485:635–641.

Shivaprasad P V., Dunn RM, Santos BACM, Bassett A, Baulcombe DC. 2012. Extraordinary transgressive phenotypes of hybrid tomato are influenced by epigenetics and small silencing RNAs. EMBO J. 31:257–266.

Slotkin RK, Martienssen R. 2007. Transposable elements and the epigenetic regulation of the genome. Nat. Rev. Genet. 8:272–285.

Spooner DM, Peralta IE, Knapp S. 2005. Comparison of AFLPs with other markers for phylogenetic inference in wild tomatoes [*Solanum L.* section *Lycopersicon* (Mill.) Wettst.]. Taxon. 54:43–61.

Stam R, Nosenko T, Hörger AC, Stephan W, Seidel M, Kuhn JMM, Haberer G, Tellier A. 2019. The de novo reference genome and transcriptome assemblies of the wild tomato species *solanum chilense* highlights birth and death of NLR genes between tomato species. G3 Genes, Genomes, Genet. 9:3933–3941.

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

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28:2731–2739.

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

Wang Y, Diehl A, Wu F, Vrebalov J, Giovannoni J, Siepel A, Tanksley SD. 2008. Sequencing and comparative analysis of a conserved syntenic segment in the solanaceae. Genetics 180:391–408.

Weiberg A, Wang M, Lin F-M, Zhao H, Zhang Z, Kaloshian I, Huang H-D, Jin H. 2013. Fungal Small RNAs Suppress Plant Immunity by Hijacking Host RNA Interference Pathways. Science (80-.). 342:118–123.

Wicker T, Keller B. 2007. Genome-wide comparative analysis of *copia* retrotransposons in *Triticeae*, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual *copia* families. Genome Res. 17:1072–1081.

Xiao H, Jiang N, Schaffner E, Stockinger EJ, Van Der Knaap E. 2008. A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. Science (80). 319:1527–1530.

Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zhang G, Yang S, Li R, Wang J, et al. 2011. Genome sequence and analysis of the tuber crop potato. Nature. 475:189–195.

Yamashita S, Takano-Shimizu T, Kitamura K, Mikami T, Kishima Y. 1999. Resistance to gap repair of the transposon Tam3 in Antirrhinum majus: A role of the end regions. Genetics 153:1899–1908.

Zhong S, Fei Z, Chen YR, Zheng Y, Huang M, Vrebalov J, McQuinn R, Gapper N, Liu B, Xiang J, et al. 2013. Single-base resolution methylomes of tomato fruit development reveal epigenome modifications associated with ripening. Nat. Biotechnol. 31:154–159.

Zou J, Gong H, Yang T-J, Meng J. 2009. Retrotransposons - a major driving force in plant genome evolution and a useful tool for genome analysis. J. Crop Sci. Biotechnol. 12:1–8.

CHAPTER 5

CONCLUSION:
TRANSPOSABLE ELEMENTS IMPACT GENOMES VIA GENE DUPLICATIONS AND INTERACTIONS WITH OTHER COMPONENTS IN THE GENOMES

INTRODUCTION

Transposons as mobile DNA sequences have changed our understanding of genomes. The interaction between the host genome and transposons has influenced the evolution of genes and genomes. This interaction has a sophisticated back and forth seen over evolutionary time, which leads genomes to display unique transposon landscapes, gene combinations, and gene regulation. The abundance of LTR/*Gypsy* and LTR/*Copia* elements in sequenced Angiosperms shows the dominance of the copy and paste elements as successful elements to reach high copy numbers. However, the dominance of superfamilies can lead to overlooking the other fascinating transposon stories hiding in genomes. Among which is the interesting original discovery of transposons in maize, a genome dominated by *Gypsy/Copia* elements, by Barbara McClintock (McClintock 1950). McClintock identified *Ac/Ds* DNA elements, as their recombination and "controlling elements" characteristics interact with the genome.

One major challenge in studying genomes and their repetitive content is difficult to assemble. Once these are resolved transposon composition and evolutionary history presents unresolved questions. For example, some superfamilies, such as *Tc1/Mariner* are absent in some angiosperm genomes. This is not well understood, as it could be due to element loss, lack of horizontal gene transfer, or other evolutionary forces that are not well understood. These genomic variances in transposon biology present opportunities to study and resolve differences and mechanisms of evolution of transposons.

Another topic of research in transposons is the mechanisms, function, and evolutionary impacts of gene and gene fragments in transposons. The presence of these fragments is of active interest as these gene and gene fragments may be sources of novel genes, new regulatory

sequences. The importance of studying the mechanisms of gene capture may yield insights into new molecular tools, novel breeding mechanisms, and powerful gene editors.

DISCUSSION

In the chapters above, I have outlined several different interactions among transposons and genomes. Transposons have duplicated gene and gene fragments and these elements require specific necessities to persist in the germ line. Specifically, TEs require mobilization and proliferation in germ cell lineages to be passed from one generation to another in sexual reproduction. Otherwise, elements can persist in asexually propagated plant tissues such as the lotus rhizomes. In the second chapter, I describe a new superfamily, *GingerRoot,* which shows the unique biology of a DNA element. *GingerRoot* elements are found in plants and animals and have a similar genomic niche as *Gypsy-like* LTR elements. The interaction of *GingerRoot* with the genome is seen through gene fragment acquisition, DNA methylation patterns, transposon location, and expression evidence. This interaction between transposons and the genome shows how an element can occupy genomic sequence in a small genome to be successful. In the third chapter, studying the lotus genome, I described the transposons in lotus and how transposons have proliferated in introns. The dogma in plant transposon biology states RNA elements are high copy number elements found in the heterochromatic regions of the genome (Bennetzen 2000). In lotus, there are many RNA elements in genic regions showcasing counter dogma genomic localization. In chapter four, the gene fragment carrying Pack-MULEs, *SlPM37*, show specific gene capture events in Pack-MULEs alters element copy number and interacts with parental genes. These studies show how gene fragments in transposons affect the interaction with

the genome. These specific interactions demonstrate that unique genomic contexts illustrate the dynamics of transposons and that rules or patterns observed from some genomes may not apply to other genomes. With this knowledge, the unlocking of more genomic secrets awaits further exploration as more sequences emerge.

Advances in sequencing are essential for transposon biology

Every project I have worked on relies on accurate and detailed sequencing efforts. The ability to sequence and assemble complex plant genomes with accuracy has improved as read lengths and assemblers have advanced and costs have decreased (Amarasinghe et al. 2020). In all my projects, the data stemmed directly from newly sequenced genomes. After sequencing, the assembly and annotation are vitally important. This is especially true for repeats as resolving highly similar repeats can change evolutionary time frames when estimating time divergences and copy numbers. While working on the sacred lotus assembly, annotation and sequencing limitations resulted in challenges due to the genomic composition. The project was aided due to improvements in the assembly. For example, the *Dicer*-Like 3 gene in lotus covered over 100 kb and the ORF predicted by the annotation was quite different from my comparative gene sequences using grape and Arabidopsis. While this may be due to the presence of TEs in introns, this difficulty impacts confidence in the annotation. Additionally, it also restricts the analysis of new or novel genes if they are not well supported by transcriptomes or other evidence. Expectedly, this will be less of an issue as more genomes, transcriptomes, and predictive software are available for comparative analysis. Annotation methods can be improved to remedy differences in genomes.

Transposons acquire gene duplications as an evolutionary mechanism to persist in the genome

The capability of transposons to duplicate genes facilitates novel evolutionary pathways in the genome. This duplication also allows for evolutionary adaption and retention over time. The ability of genes to be partially duplicated or combined with other genes allows novel genes to evolve. The evolutionary influence of transposons can vary from stress induced elements (Feng et al. 2013; Dubin et al. 2018), novel fragment creation (Bennetzen et al. 2000; Jiang et al. 2004), and regulatory effects (Lisch 2009; Martienssen and Chandler 2013). In the case of *GingerRoot,* several elements carry fragments from multiple genes. Elements with gene fragments were more likely to be more distal from genes, and older than other *GingerRoots* without gene fragments. This suggests non-random retention of *GingerRoot* elements and could be utilized for regulatory reasons. Gene duplication influences elements and genomes, allowing the genome to create novel fragments and potential for evolutionary source material. Another previously discovered process is in rice as Pack-MULEs are an integral part of upstream flanking regions of genes (Ferguson et al. 2013). These duplication events are a mechanism for transposon persistence.

Transposons and genome interactions show mechanisms for genome diversity

The interaction of transposons and genomes are fundamental to understanding evolution at a genome level. In plants, transposons contribute to genome size along with whole genome duplication (Figure 5.1). Primarily, the LTR *Gypsy/Copia* elements are the largest fraction of many large plant genomes. In these genomes, *Gypsy/Copia* elements have found a favorable environment to expand and persist. The interaction between RNA element and genes combines

the copy and paste biology along with the genome's ability to sustain a high load of elements without disrupting gene functions. In contrast, smaller genomes either are less flexible with transposon mutagenesis and/or have robust and rapid transposon removal mechanisms. Specifically, the *S. lepidophylla* genome demonstrates mechanisms closer to the latter, whereby TE removal constrains the genome size to 110 Mb, reducing the abundance of *Gypsy/Copia* elements. The genomic interaction between TEs and the genome further allows the class II cut and paste DNA element *GingerRoot* to persist in the genome by potentially reducing LTR proliferation and inter superfamily competition. This high removal mechanism allows *GingerRoot* to persist and not be obliterated by nested insertions from *Gypsy/Copia* elements. This interaction between elements and the genome defines the variation, abundance, and gene context in the genome.

In the tomato clade, the interaction of *SlPM37* and the genome has evolved over time due to sequence acquisition. After the *SlPM37* acquired the *AGO1* fragment, the copy number increased, which was potentially due to *AGO1* parental gene expression differences. This is a direct interaction between the Pack-MULE composition and a hypothesized regulation of the Pack-MULE using sRNA or other epigenetic pathways. There are copies of *SlPM37* with only the *CYP51* fragment but these are found in low copy number, which is in contrast to the 50+ copy number of the *AGO1* fragment variant. The resulting Pack-MULE composition with the *AGO1* fragment allows the proliferation of the element. This observation could be tested empirically with novel Pack-MULEs, that would generate sRNA to the parental gene.

FUTURE WORK

The work presented above offers opportunities and further hypotheses to test as new findings lead to more questions. Three topics to potentially explore in the future would be (a) the transposition of *GingerRoot* in other genomes, (b) the intron composition of other genomes with large intron sizes, and (c) changing copy number of *SlPM37* in tomato to interrogate expression profiles of parental genes. These follow up experiments could have impacts on transposon activity, genome evolution, and gene regulation.

GingerRoot **transposition in other genomes**

GingerRoot was shown to have reads that map to their transposase in publicly available transcriptome datasets. These data show that the element may be active in the species in the SRA datasets. The ability of *GingerRoot* to transpose is worthy of study for two reasons. One, transposition is rare to observe and can be used to study transposition mechanisms and to find recent elements. Two, a potentially influential project may be to see if the *GingerRoot* elements in these genomes are harboring gene fragments. If they are, then *GingerRoot* could be studied to see whether any of the acquired fragments are highly similar to the parental genes and to potentially observe acquisition events, which hereto have not been observed in Pack-MULEs and therefore, would greatly improve our understanding of gene fragment acquisition of DNA elements.

Intron composition

Lotus has 50% transposon content, additionally, introns are enriched with insertions and overall introns are majority composed of transposons, which is notable and uncommon. This observation is contrary to transposon dogma of RNA elements in gene poor regions and DNA

elements in genic regions in plants. Genomes can tolerate RNA elements not only in intergenic regions but as large parts of introns. Large introns have also been seen in ginkgo (Guan et al. 2016), and elucidating other genomes with large introns may yield insights into differences in LINE elements targeting or other biological insights that have resulted in high TE frequency in introns. Another project stemming from the lotus genome is the abundance of introns that are comprised of retroelements. Other genomes such as the ginkgo, *Ginkgo biloba,* (Guan et al. 2016), white spruce, *Picea glauca,* (Warren et al. 2015), and loblolly pine, *Pinus taeda,* (Voronova et al. 2020), report long introns comprised of transposable elements. Other newly sequenced genomes could be mined for average intron length from a list of sequenced genomes. These genomes could be mined for the TE content in introns. Additionally, these observations warrant testing whether genes with introns high or low in TEs in lotus are expressed differentially in lotus. Further, the composition and evolutionary history of these elements in lotus could be traced to identify founder elements. The presence and prevalence of similar lotus elements in other genomes could test whether these elements are targeting introns. Within the transposon, there could be a domain targeting introns, which could be explored through transposition studies with deletion mutants in the targeting region as seen in prior work in yeast (Zou et al. 1996; Asif-Laidin et al. 2020). Further, the lotus genome could have sequence bias in introns that these LINEs and other elements are targeting. These hypotheses warrant further exploration in order to understand the biochemical and evolutionary interactions of TEs in introns.

SlPM37 copy number manipulation

In the tomato genome, we observed a high copy number Pack-MULE that was amplified across the tomato clade. The observation of the *AGO1* acquisition appears to be the important fragment that led to *SlPM37* proliferation. As this high copy number has not been observed in other genomes with Pack-MULEs, it suggests that the *AGO1* fragment is crucial for amplification.

The genome transposon interaction can be further studied by testing the hypothesis that Pack-MULE copy number influences gene regulation by the generation of CRISPR knockout, varying the copies present. This could be done by targeting regions of the *SlPM37* in tomato to either or both of the *AGO1* or *CYP51* regions, additionally targeting multiple copies in the genome. This would produce plants with variable numbers of copies ideally ranging from a few copies to the wild type copy number. These plants could then be sequenced for sRNA reads and map them back to the Pack-MULE. Additionally, point mutations could be introduced to specifically map reads to the Pack-MULE and investigate whether there is evidence for the influence of the regulation of these elements.

Further, the acquisition of sequences by Pack-MULE is poorly understood. An observance of a MULE element acquiring a gene fragment in real time has not been reported. While two studies have shown transposition of MULEs in yeast and a mosquito *Mutator* element, no acquisition of gene sequences has been observed (Zhao et al. 2015; Liu et al. 2017). *SlPM37* shows pieces of potential evidence of how the acquisition may function molecularly. First, the *CYP51* and *AGO1* have a short alignment overlap in the 5′ *CYP51*/*AGO1* boundary. Microhomology between the *CYP51* and *AGO1* maybe because the acquisition is formed using a

microhomology related repair pathway (Jia et al. 2013; García-Medel et al. 2019). Potentially, analyzing current auto-MULEs for microhomology targets and looking at sequence conservation may yield insights into viable candidates or designer constructs for acquisition experiments. The potential amplification of Pack-MULEs could elucidate the potential evolution of regulation and gene formation pathways.

APPENDIX

FIGURES

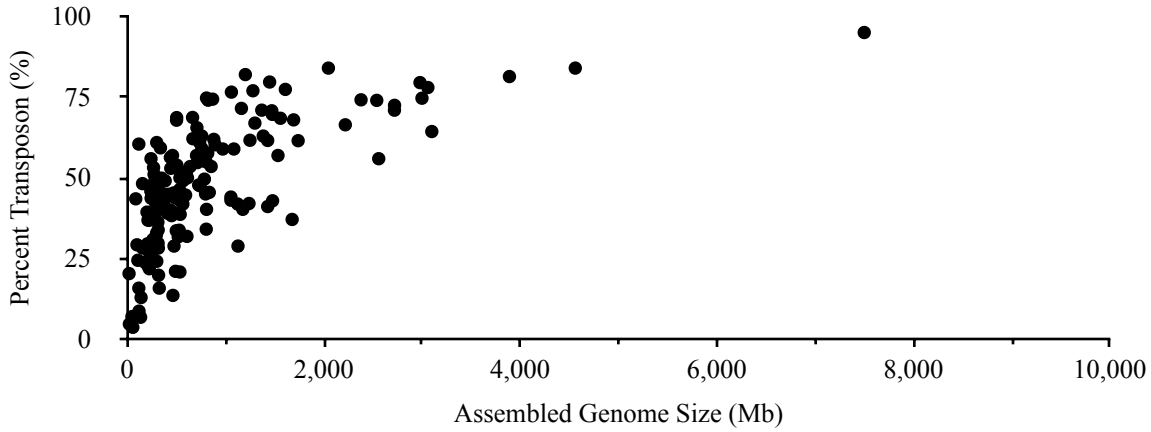


Figure 5.1. Genome size vs. percent transposon in selected plant genomes.

REFERENCES

REFERENCES

Amarasinghe SL et al. 2020. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 21:1–16. doi: 10.1186/s13059-020-1935-5.

Asif−Laidin A et al. 2020. A small targeting domain in Ty1 integrase is sufficient to direct retrotransposon integration upstream of tRNA genes. EMBO J. 39. doi: 10.15252/embj.2019104337.

Bennetzen JL. 2000. Transposable element contributions to plant gene and genome evolution. Plant Mol. Biol. 42:251–269. doi: 10.1023/A:1006344508454.

Dubin MJ, Mittelsten Scheid O, Becker C. 2018. Transposons: a blessing curse. Curr. Opin. Plant Biol. 42:23–29. doi: 10.1016/j.pbi.2018.01.003.

Feng G, Leem YE, Levin HL. 2013. Transposon integration enhances expression of stress response genes. Nucleic Acids Res. 41:775–789. doi: 10.1093/nar/gks1185.

Ferguson AA, Zhao D, Jiang N. 2013. Selective Acquisition and Retention of Genomic Sequences by Pack-Mutator-Like Elements Based on Guanine-Cytosine Content and the Breadth of Expression. Plant Physiol. 163:1419–1432. doi: 10.1104/pp.113.223271.

García-Medel PL, Baruch-Torres N, Peralta-Castro A, Trasviña-Arenas CH, Torres-Larios A, Brieba LG. 2019. Plant organellar DNA polymerases repair double-stranded breaks by microhomology-mediated end-joining. Nucleic Acids Res. 47:3028–3044. doi: 10.1093/nar/gkz039.

Guan R, Zhao Y, Zhang H, Fan G, Liu X, Zhou W, Shi C, Wang Jiahao, Liu W, Liang X, et al. 2016. Draft genome of the living fossil Ginkgo biloba. Gigascience 5:49. doi: 10.1186/s13742-016-0154-1. doi: 10.1186/s13742-016-0154-1.

Jia Q, Dulk-Ras A Den, Shen H, Hooykaas PJJ, de Pater S. 2013. Poly(ADP-ribose)polymerases are involved in microhomology mediated back-up non-homologous end joining in Arabidopsis thaliana. Plant Mol. Biol. 82:339–351. doi: 10.1007/s11103-013-0065-9

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. 2004. Pack-MULE transposable elements mediate gene evolution in plants. Nature. 431:569–573. doi: 10.1038/nature02953.

Lisch D. 2009. Epigenetic regulation of transposable elements in plants. Annu. Rev. Plant Biol. 60:43–66. doi: 10.1146/annurev.arplant.59.032607.092744.

Liu K, Wessler SR. 2017. Functional characterization of the active Mutator-like transposable element, *Muta1* from the mosquito *Aedes aegypti*. Mob. DNA 8. doi:10.1186/s13100-016-0084-6

Martienssen RA, Chandler VL. 2013. Molecular Mechanisms of Transposon Epigenetic Regulation. In: Plant Transposons and Genome Dynamics in Evolution. pp. 71–92. doi: 10.1002/9781118500156.ch5.

McClintock B. 1950. The origin and behavior of mutable loci in maize. Proc. Natl. Acad. Sci. U. S. A. 36:344–355. doi: 10.1073/pnas.36.6.344.

Stival Sena J et al. 2014. Evolution of gene structure in the conifer *Picea glauca*: A comparative analysis of the impact of intron size. BMC Plant Biol. 14:1–16. doi: 10.1186/1471-2229-14-95.

Voronova A, Rendón-Anaya M, Ingvarsson P, Kalendar R, Ruņģis D. 2020. Comparative study of pine reference genomes reveals transposable element interconnected gene networks. doi: 10.21203/rs.3.rs-34803/v1.

Warren RL et al. 2015. Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. Plant J. 83:189–212. doi: 10.1111/tpj.12886.

Zhao D, Ferguson A, Jiang N. 2015. Transposition of a rice mutator-like element in the yeast saccharomyces cerevisiae. Plant Cell. 27:132–148. doi: 10.1105/tpc.114.128488.

Zou S, Ke N, Kim JM, Voytas DF. 1996. The Saccharomyces retrotransposon Ty5 integrates preferentially into regions of silent chromatin at the telomeres and mating loci. Genes Dev. 10:634–645. doi: 10.1101/gad.10.5.634.