TRANSPOSABLE ELEMENTS DRIVE LINEAGE-SPECIFIC PATTERNS OF GENOME EVOLUTION IN THE ASTERACEAE

by

SPENCER EVAN STATON

(Under the Direction of John M. Burke)

ABSTRACT

Transposable elements (TEs) make up the most abundant and dynamic component of plant genomes and they represent important sources of evolutionary novelty. Despite these common genomic features among plants, fundamental questions about TE evolution are still poorly understood, in part because so few plant species have been characterized in a phylogenetic framework with respect to TEs. An important objective therefore is to understand the specific contributions of TEs in a large number of plant species in order to construct a more complete picture of evolution in the plant kingdom. In this work, I present novel computational methods and software for analyzing TEs in unexplored genomes, and I demonstrate the utility of these developments by explaining patterns of TE evolution in the plant family Asteraceae. The Asteraceae is the largest family of flowering plants, and has very recent evolutionary origin. These features, along with a global distribution of species adapted to many different environments, make this family an excellent system to investigate evolutionary processes. By using a novel repeat finding method described herein, I show that Asteraceae genomes differ in the abundance and diversity of TEs from the most closely related family Calyceraceae, and each subfamily of the Asteraceae exhibits unique patterns of TE evolution. From the base of the family

to the most derived lineages of the Asteraceae, there is a linear increase in the amount of one type of TE, *Gypsy*, and there is a linear decrease in the amount of *Copia* TEs. This pattern is driven, in part, by a marked increase in the genomic dominance of certain *Gypsy* TE families at the base of tribe Heliantheae, and these events have lead to a decrease in the diversity of TEs in this tribe. Contrary to the near universal species-area relationship in ecological studies, I show that larger genomes may be a product of unequal contributions of TE families and do not necessarily support a greater diversity of TEs. Taken together, these findings highlight the importance of broad taxonomic sampling in a phylogenetic framework for understanding the mechanisms contributing to the evolution of TEs across the plant kingdom.

INDEX WORDS: Asteraceae, transposable elements, speciation, genome evolution, phylogenomics, bioinformatics, biodiversity

TRANSPOSABLE ELEMENTS DRIVE LINEAGE-SPECIFIC PATTERNS OF GENOME EVOLUTION IN THE ASTERACEAE

by

SPENCER EVAN STATON

Bachelor of Science, Reinhardt University, 2005

Master of Botany, Miami University, 2008

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

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2014

TRANSPOSABLE ELEMENTS DRIVE LINEAGE-SPECIFIC PATTERNS OF GENOME EVOLUTION IN THE ASTERACEAE

by

SPENCER EVAN STATON

Major Professor: John M. Burke

Committee: Jeffrey L. Bennetzen
Kelly Dyer
Jim Leebens-Mack
Xiaoyu Zhang

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2014

DEDICATION

To my family, whose support and encouragement helped make this work possible, and to my loving wife, for always believing in me.

ACKNOWLEDGEMENTS

I want to thank my advisor, John Burke, for allowing me the room to pursue my own research questions, and my lab mates for feedback and discussion on manuscripts along the way. The majority of my work at UGA was supported by the NSF-funded Compositae Genome Project and I am thankful for the opportunities and experience from working with such a talented group of researchers. I am indebted to the many people that contribute their time to developing or maintaining free open source software, without which this work would not be possible.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Barbara McClintock revealed one of greatest discoveries in science when she described how variation in maize kernel pigmentation was inherited and that the causative factors were mobile genetic elements called transposable elements (TEs; she called them controlling elements). It is due to the visible mutations of TEs that led to their discovery and characterization in maize in the 1940's (McClintock 1950). The decades of work following their discovery revealed the mutagenic nature of TEs in maize and other model systems such as *Drosophila* (Charlesworth and Langley 1989). In fact, much of the theoretical basis of TE evolution is based on the relationship between a few deleterious TEs in *Drosophila melanogaster* (Charlesworth and Langley 1989) and it is largely an antagonistic relationship that initially defined the interaction between TEs and their hosts. However, the analysis of large comparative data sets with modern computational techniques have revealed that there are many levels of interactions that take place between TEs and their hosts. It is now appreciated that the typical eukaryotic genome is rich with TE diversity that may have unique functional and demographic properties.

Being ancient components of eukaryotic genomes, TEs have developed into a diverse array of structures with many different mechanisms for replication and survival (reviewed in Wicker *et al*., 2007). Briefly, I will highlight some of the main distinctions. Though split into just two classes based on their structure and function, the behavior and composition of TEs within the each class is exceptionally diverse. Class I TEs, the retrotransposons, use a RNA intermediate for transposition, and are divided into two groups depending on the presence of

long terminal repeats (LTR) that flank the internal coding domains (*i.e*., the so-called LTR and non-LTR elements; Kumar and Bennetzen 1999). Replication and movement of Class II TEs, the DNA transposons, involves a catalytic domain that initiates excision and transposition of the element by the activity of a transposase (Wicker *et al*., 2007). This "cut-and-paste" replication is in contrast to the "copy-and-paste" mechanism of retrotransposons. While the mechanisms of replication are not shared between Class I and II TEs, there is a common feature of all TEs in that they have the ability to replicate faster than the host organism. For some classes of TEs (i.e., the retrotransposons) this means they can become large proportions of the nuclear genome if unregulated. Interestingly, plant genomes exhibit an almost universal pattern of ongoing LTR retrotransposon activity, with the exception of the *Amborella* and Norway spruce genomes (Amborella Genome Project 2013; Nystedt *et al*., 2013), and this has lead to TE-induced increases in genome size in many species (Piegu *et al*., 2006; Ungerer *et al*., 2006; Hawkins *et al*., 2008).

Transposable elements are confined to the nucleus of a host species and may only enter a new host through sexual transmission or rare horizontal transfer events (Schaak *et al*., 2010). Though confined to the same nucleus, individual TE families may be active over different evolutionary time scales, and occupy unique niches within the genome reflecting specific needs for their maintenance and survival. These two basic life history traits are thought to play a role in the differential survival of TEs lineages.

The diversity of TEs within a genome may be explained using the same principles that govern biodiversity in any ecosystem (Venner *et al*., 2009; Brookfield 2005; Baucom *et al*., 2009a). Thus, the genome has been referred to as the "genome ecosystem" to describe the biodiversity and structural complexity within the nuclear genome of an organism (Kidwell and

Lisch 1997; Le Rouzic *et al*., 2007a; Venner *et al*., 2009; Brookfield 2005). This analogy has allowed for a refinement of TE classification and opened the door for describing population dynamics within the genome on spatial and temporal scales with the aid of existing ecological and evolutionary models (Le Rouzic and Capy 2005; Venner *et al*., 2009; Serra *et al*., 2013). This is somewhat of a paradigm shift, as the primary theoretical literature that still guides TE description is based on the fact that TEs are "selfish DNA" and do not directly contribute to host function (Doolittle and Sapienza 1980; Orgel and Crick 1980). Given the deleterious nature of TEs, it is not surprising that numerous molecular mechanisms have evolved for suppressing their activity (reviewed in Lisch 2009). In addition to host-encoded mechanisms for TE control, certain classes of TEs also contain structural features that contribute to their removal from the genome (Devos *et al*., 2002). Thus, it is somewhat puzzling that TEs have been so successful in colonizing nearly all forms of life. While all available evidence suggests that TE insertions are generally deleterious there are certainly a multitude of interactions that exist between TEs and their hosts aside from an antagonistic one. Transposable element-host interactions are probably best described as a comprising a continuum from purely antagonistic to adaptive by the functions they may provide as co-opted TEs (Kidwell and Lisch 2000; Feschotte 2008). Thus, applying mathematical and ecological modeling with thorough phylogenetic analysis may help to uncover TE contributions to evolutionary innovations and offer some new perspective on biodiversity.

The role of external influences on transposable element evolution

The demography of species is shaped by a combination of biotic and abiotic factors. The outcome of variation in these factors over time is that species abundance and geographic distribution may change. It is clear that factors influencing species demography also influence

the patterns of TE evolution, though it is not clear in most cases which interactions are more important in shaping patterns of TE diversity (i.e., host-mediated events or environmental-mediated), since bursts of TE activity have been attributed to environmental stress alone (Wessler 1996; Kalendar *et al*., 2000). Thus, a key component to understanding how TEs are able to successfully increase in number within a population and persist between generations is to combine population genetic data to investigate the demographic history of a species and genomic data to explore patterns of change due to selection or drift. Analysis of genomic patterns of TE demography combined with TE frequency polymorphism data from natural populations of *Drosophila*, human, and *Arabidopsis lyrata* have revealed important aspects of TE distribution and persistence (Petrov *et al*., 2003; Han *et al*., 2005; Lockton *et al*., 2008). First, the migration patterns and demographic history of species are strong determinants of TE dynamics within genomes and between species populations (Le Rouzic *et al*., 2007b; Lockton *et al.*, 2008). Theoretical work also suggests that migration rates and population size are important factors in TE evolution due to their effect on how selection operates in the genome (Deceliere *et al*., 2005; Le Rouzic *et al*., 2007b). Specifically, periods of reduced population size are thought to characterize the history of most species and these periods of fluctuation in effective population size may trigger TE amplification due to a reduction in the efficacy of selection (Le Rouzic *et al*., 2007b; Lynch 2007). Whereas traditional theoretical work on the maintenance of TEs was based on the idea of an equilibrium being reached with the host through transposition-selection balance, these findings provide another possibility for TEs to obtain evolutionary stability through the stochastic process of genetic drift (Le Rouzic *et al*., 2007b; Lockton *et al*., 2008). The relative roles of selection and drift in TE evolution may vary

due to the unique history of a species and the relative importance of each process may be influenced by which form of reproduction characterizes the host species.

For some taxa, particularly plants, reproduction may take many forms and even the composition of mating types in a single population may vary from one year to the next. This variation may be an important component of TE persistence because bursts of TE activity may be induced by hybridization in plants and animals (Ungerer *et al*., 2006; O'Neill *et al*., 1998; Labrador *et al*., 1999), and crosses between different strains or populations (Kidwell and Lisch 2001; Rangwala *et al*., 2007). The mechanisms behind these events have been investigated in many taxa and appear to involve the loss of TE regulation through epigenetic modifications (O'Neill *et al*., 1998; Yoder *et al*., 1997), or a TE may go through a period of intragenomic selection for activity once in a new genomic environment (Gregory 2005; Venner *et al*., 2009). Thus, a key component to TE survival over evolutionary time scales may be to find a naïve environment (i.e., genomes in a new population) through dispersal or horizontal transfer, or evolve a method for avoiding host silencing mechanisms (discussed below). The distribution of TEs within the genome of outcrossing species such as *Drosophila melanogaster* and partial selfers such as *Caenorhabditits elegans* or *Arabidopsis thaliana* will differ due, in part, to different rates of recombination. Specifically, there is a negative correlation between TE accumulation and recombination rate in *D. melanogaster* (Petrov *et al*., 2003) whereas this pattern is not present in *C. elegans* or *A. thaliana* (Duret *et al*., 2000; Wright *et al*., 2003). Presumably, this pattern reflects the fact that ectopic recombination is less frequent in inbred species and occurs more frequently in heterozygotes (Hollister and Gaut 2009).

One generality of many species is that environmental stress and population bottlenecks (or population reductions in less extreme cases) often follow the occupation of a new niche (or

a period of environmental transition). This is consistent with classical theory on ecological speciation as well as the role of environmental disturbance in hybrid speciation (Rieseberg 1997). These events, created by dispersal or hybridization and then occupation of a new niche, may be following by periods of dramatic increases in TE numbers (Ungerer *et al*., 2006; Ungerer *et al*., 2009). Massive TE amplification can lead to genome restructuring, alter gene expression, and generate novel phenotypes (O'Neill *et al*., 1998; Feschotte 2008; Rebollo *et al*., 2010). Thus, TEs may be an important component of the speciation process by contributing to population divergence and reproductive isolation (Robello *et al*., 2010). While many TE insertions, in fact most, will be deleterious to the host, this process allows TEs to increase in number rapidly in (small) populations. While the mechanism will vary depending on the taxa, host silencing and periods of inactivity will typically follow periods of TE amplification. Though transposable elements are pervasive, their evolutionary survival depends solely on transmission, which will be determined by the mating system and population structure in host populations.

The role of transposable elements as drivers of host evolution

In addition to being mutagenic in nature, TEs also have a creative role in generating adaptive variation. It is apparent that TEs were involved in the evolution of the vertebrate immune system, the evolution of sex, and the evolution of gene regulation, for example (Gregory 2005; Slotkin *et al*., 2012). Transposable elements also create new genes by shuffling exons, causing alternative splicing, and picking up gene fragments (Bennetzen 2005; Feschotte 2008). In addition, TEs may be important components of chromosomes through their ability to promote heterochromatin formation in a directed fashion (Gao *et al*., 2008), and thus, have a direct influence on chromosome behavior (reviewed in Gregory 2005). In addition to

molecular variation, TEs are major sources of structural variation by creating chromosomal rearrangements in plants and animals (Lim 1998; Zhang *et al*., 2011). Rearrangements are thought to protect environmentally adapted gene complexes and promote speciation by sequence and expression divergence and reproductive isolation due to structural differences (Rieseberg 2001; Rieseberg and Willis 2007). The foregoing examples suggest a potential role for TEs in the process of speciation (Robello *et al*., 2010); this has lead to the idea that TEs may promote divergence by displacing populations from an adaptive peak on the fitness landscape (Zeh *et al*., 2009).

Interactions within the genome

The mutagenic role of TEs has been explored throughout prokaryotic and eukaryotic taxa and it is generally accepted that they create mutations that would not have otherwise arisen. The nature of mutations caused by TEs can typically be ascribed to a specific class or subclass of elements and it is apparent that eukaryotic genome structure may arise, in part, due to interactions between the different subclasses of TEs. Because there is some cost on the host associated with the act of transposition, TEs have developed adaptations for avoiding negative purifying selection by targeted insertion or limiting their own activity (Chaboissier *et al*., 1998; Peterson-Burch *et al*., 2004; Gao *et al*., 2008; Gregory 2005). For example, some non-autonomous TEs use the replication machinery of intact elements and avoid the cost associated with transposition (e.g., Jiang *et al.*, 2003). It is not clear whether this type of competition between elements is evolutionarily stable though theoretical work suggests that competition between elements can help to explain their stochastic loss from populations (Le Rouzic and Capy 2005; Le Rouzic *et al*., 2007a). The host may even exploit these ancient conflicts by using products synthesized from one TE type to silence another (Cam *et al*., 2008). It is also

possible that there is some level of cooperation between element families (Le Rouzic *et al.*, 2005), but more work needs to be done in this area.

A product of the interactions between TE families and the host is that many families have evolved specific insertion preferences (i.e., niche partitioning) possibly as method of competition avoidance as much as a way to reduce impact on host function. A transposition event followed by targeted insertion could have a significant impact on the diversity and genome landscape of the targeted area over evolutionary timescales. Similar types of habitat modification can be seen in natural ecosystems. For example, ecosystem engineers such as the North American beaver and Spartina grass have the ability to create and maintain habitat, thereby allowing for population expansion (Gurney and Lawton 1996). Thus, it seems at least tenable that targeted insertion could minimize the fitness effects for the host, allow for TE population growth and provide one model for genome size growth by the non-random activity of TEs. In addition, it seems that non-random insertion may be a common feature of TEs (Peterson-Burch *et al*., 2004; Gao *et al*., 2008; Baucom *et al*., 2009a; Naito *et al*., 2009). These patterns are likely driven by a combination of interactions with host-encoded factors, as well as competition between TE element families for space and other resources.

Implications for particular classes of transposable elements

The distribution and abundance of TEs in the genome is determined by a number of host-encoded factors, and is dependent on the type of TE since replication mode and size (i.e., length) do vary between classes of TEs (Dolgin and Charlesworth 2008; Morgan 2001; Lockton *et al*., 2008; Wicker *et al*., 2007). For example, retrotransposons, which are generally much longer that DNA transposons, typically avoid gene-rich regions, while DNA transposons routinely insert into or near genes (Duret *et al*., 2003; Naito *et a*l., 2009). Because the timing of replication, the

enzymes involved, and localization preferences among TE classes vary, it will be important to identify environmental and host-encoded factors that specifically influence the demography and evolutionary persistence of each TE class. One important factor for long-term survival is the production of proteins that have the ability to recognize the source element. The non-LTR LINE elements have this ability while DNA transposons do not, and this may help to explain why Class II elements are prone to extinction (partly by being unable to avoid non-autonomous parasites) while LINE elements have persisted for hundreds of millions of years in some animal lineages. There are many functional and structural properties that differ between TE types and there may be general mechanisms of activation as McClintock envisioned (McClintock 1984). However, a consideration of the genomic context of each type is likely to reveal a more unique picture of the characteristics of TEs.

LITERATURE CITED

Amborella genome project. 2013. The Amborella genome and the evolution of flowering plants. *Science* 342, doi: 10.1126/science.1241089.

Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, Deragon JM, Westerman RP, SanMiguel PJ, Bennetzen JL. 2009a. Exceptional diversity, non-random distribution, and rapid evolution or retroelements in the B73 maize genome. *PLoS Genetics*, 5(11), 1-13.

Bennetzen JL. 2005. Transposable elements, gene creation and genome rearrangement in flowering plants. *Curr. Op. Gen. Dev.* 15(6), 621-62.

Brookfield JYF. 2005. The ecology of the genome – mobile DNA elements and their hosts. *Nature Rev.* 6, 128-136.

Cam HP, Noma K, Ebina H, Levin HL, and Grewal SIS. 2008. Host genome surveillance for retrotransposons by transposon-derived proteins. *Nature* 451, 431-436.

Chaboissier MC, Bucheton A, and Finnegan DJ 1998. *Proc. Natl. Acad. Sci. U.S.A.* 95(20), 11781-11785.

Charlesworth B, Langley CH. 1989. The population genetics of Drosophila transposable elements. *Ann. Rev. Gen.* 23, 251-287.

Deceliere G, Charles S, Biemont C. 2005. The dynamics of transposable elements in structured populations. *Genetics* 169, 467-474.

Devos KM, Brown JKM, and Bennetzen JL. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis. Genome Res.* 12, 1075-1079.

Dolgin E, and Charlesworth B. 2008. The effects of recombination rate on the distribution and abundance of transposable elements. *Genetics* 178, 2169-2177.

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

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

Gao X, Hou Y, Ebina H, Levin HL, and Voytas DF. 2008. Chromodomains direct integration of retrotransposons to heterochromatin. *Genome Res.* 18, 359-369.

Gregory TR. 2005. Evolution of the genome. Elsevier, Inc.

Gurney WSC, and Lawton JH. 1996. The population dynamics of ecosystem engineers. *Oikos* 76, 273-283.

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nature Rev. Gen.* 9, 397-405.

Han KD, Xing JC, Wang H, Hedges DJ, Garber RK, Cordaux R, Batzer MA. 2005. Under the genomic radar: The stealth model of Alu amplification. *Genome Res.* 15, 655-664.

Hawkins JS, Grover CE, Wendel JF. 2008. Repeated big bangs and the expanding universe: Directionality in plant genome size evolution. *Plant Science* 174, 557-562.

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.

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

Kalendar R, Tanskanen J, Immonen S, Nevo E, Shulman AH. 2000. Genome evolution in wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. *Proc. Natl. Acad. Sci. U.S.A.* 97, 6603-6607.

Kidwell, M.G. and Lisch, D. 1997. Transposable elements as sources of variation in animals and plants. *Proc. Natl. Acad. Sci. U.S.A.* 94, 7704-7711.

Kidwell MG, Lisch DR. 2000. Transposable elements and host genome evolution. *Trends Ecol. Evol.* 15(3), 95-99.

Kim JK. 1998. Intrachromosomal rearrangements mediated by *hobo* transposons in *Drosophila melanogaster. Proc. Natl. Acad. Sci. U.S.A.* 85, 9153-9157.

Kumar, A, Bennetzen, J.L. 1999. Plant retrotransposons. *Annu. Rev. Genet.* 33, 479-532.

Lisch D. 2009. Epigenetic regulation of transposable elements in plants. *Ann. Rev. Plant Biol.* 60, 43-66.

Labrador M, Farre M, Utzet F, Fontadevila A. 1999. Interspecific hybridization increases transposition rates of Osvaldo. *Mol. Biol. Evol.* 16, 931-937.

Le Rouzic A, and Capy P. 2005. Population models of competition between transposable element subfamilies. *Genetics* 174, 785-793.

Le Rouzic A, Dupas S, Capy P. 2007a. Genome ecosystem and transposable element species. *Gene* 390, 214-220.

Le Rouzic A, Boutin TS, Capy P. 2007b. Long-term evolution of transposable elements. *Proc. Natl. Acad. Sci. U.S.A* 104(49), 19375-19380.

Lockton S, Rossi-Ibarra J, Gaut BS. 2008. Demography and weak selection drive patterns of transposable element diversity in natural populations of Arabidopsis lyrata. *Proc. Natl. Acad. Sci. U.S.A.* 105(37), 13965-13970.

Lynch M. 2007. The origins of genome architecture. Sinauer Associates, Inc.

McClintock B. 1950. The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci. U.S.A.* 36(6), 344-355.

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

Morgan MT. 2001. Transposable element number in mixed mating populations. *Gen. Res.* 77, 261-275.

Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, Richardson AO, Okumoto Y, Tanisaka T, Wessler SR. 2009. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature* 461, 1130-1134.

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, et al. 2013. The Norway spruce genome sequence and conifer genome evolution. *Nature* doi:10.1038/nature12211.

O'Neill RJ, O Neill MJ, Graves JAM. 1998. Undermethylation associated with retroelement activation and chromosome remodeling in an interspecific mammalian hybrid. *Nature* 393, 68–72.

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

Peterson-Burch BD, Nettleton D, Voytas DF. 2004. Genomic neighborhoods for Arbidopsis retrotransposons: a role for targeted integration in the distribution of the Metaviridae. *Genome Biol.* 5, R78.

Petrov DA, Aminetzach YT, Davis JC, Bensasson D, Hirsh AE. 2003. Size matters: Non-LTR retrotransposable elements and ectopic recombination in Drosophila. *Mol. Biol. Evol.* 20, 880-892.

Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, Collura K, et al. 2006. Doubling genome size without polyploidization: dynamics of retrotransposon-mediated genome expansions in Oryza australensis, a wild relative of rice. *Genome Res.* 16, 1262-1269.

Rieseberg LH. 1997. Hybrid origins of plant species. *Ann. Rev. Ecol. Sys.* 28, 259-289.

Rieseberg LH. 2001. Chromosomal rearrangements and speciation. *Trends Ecol. Evol.* 16(7), 351-358.

Rieseberg LH and Willis JH. 2007. Plant Speciation. *Science* 317, 910-914.

Robello R, Horard B, Hubert B, and Vieira C. 2010. Jumping genes and epigenetics: Towards new species. *Gene* 454, 1-7.

Schaak S, Gilbert C, and Feschotte C. 2010. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends Ecol. Evol.* 25(9), 537-546.

Serra F, Becher V, and Dopazo H. 2013. Neutral theory predicts the relative abundance and diversity of genetic elements in a broad array of eukaryotic genomes. *PLOS One*, 8:6.

Slotkin RK, Nuthikattu S, Jiang N. 2012. The impact of transposable elements on gene and genome evolution. *Plant genome diversity* Vol. 1. Springer-Verlag Wien.

Ungerer MC, Strakosh SC, and Zhen Y. 2006. Genome expansion in three hybrid sunflower species is associated with retrotransposon proliferation. *Curr. Biol.* 16: R872-R873.

Ungerer MC, Strakosh SC, and Stimpson KM. 2009. Proliferation of Ty3/gypsy-like retrotransposons in hybrid sunflower taxa inferred from phylogenic data. *BMC Biol.* 7, 40.

Venner S, Feschotte C, Biemont C. 2009. Dynamcis of transposable elements: towards a community ecology of the genome. *Trends Gen.* 739, 1-7.

Wessler SR. 1996. Plant retrotransposons: Turned on by stress. *Curr. Biol.* 6, 959–961.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, et al., 2007. A unified classification system for eukaryotic transposable elements. *Nature Rev Gen.* 12, 973-982.

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

Yoder JA, Walsh CP, Bestor TH. 1997. Cytosine methylation and the ecology of intragenomic parasites. *Trends Gen.* 13, 335-340.

Zeh DW, Zeh JA, Ishida Y. 2009. Transposable elements and an epigenetic basis for punctuated equilibria. *Bioessays* 31, 715-726.

Zhang J, Yu C, Krishnaswamy L, Peterson T. 2011. Transposable elements as catalysts for chromosome rearrangements. *Methods Mol. Biol.* 701, 315-326.

CHAPTER II

THE SUNFLOWER (*HELIANTHUS ANNUUS* L.) GENOME REFLECTS A RECENT HISTORY OF BIASED ACCUMULATION OF TRANSPOSABLE ELEMENTS[1]

[1]Staton S.E., Bakken B.H., Blackman B.K., Chapman M.A., Kane N.C., Tang S., Ungerer M.C., Knapp S.J., Rieseberg L.H., Burke J.M. 2012. *The Plant Journal*, 72(1), 142-153. Reprinted here with permission of the publisher.

SUMMARY

Aside from polyploidy, transposable elements are the major drivers of genome size increases in plants. Thus, understanding the diversity and evolutionary dynamics of transposable elements in sunflower (*Helianthus annuus* L.), especially given its large genome size (ca. 3.5 Gb) and the well-documented cases of amplification of certain transposons within the genus, is of considerable importance for understanding the evolutionary history of this emerging model species. By analyzing approximately 25% of the sunflower genome from random sequence reads and assembled BAC clones, we show that it is composed of over 81% transposable elements, 77% of which are LTR retrotransposons. Moreover, the LTR retrotransposon fraction in BAC clones harboring genes is disproportionately composed of chromodomain-containing *Gypsy* LTR retrotransposons ('chromoviruses'), and the majority of the intact chromoviruses contain tandem chromodomain duplications. We show that there is a bias in the efficacy of homologous recombination in removing LTR retrotransposon DNA, thereby providing insight into the mechanisms associated with TE composition in the sunflower genome. We also show that the vast majority of observed LTR retrotransposon insertions have likely occurred since the origin of this species, providing further evidence that biased LTR retrotransposon activity has played a major role in shaping the chromatin and DNA landscape of the sunflower genome. While our findings on LTR retrotransposon age and structure could be influenced by the selection of BAC clones analyzed, a global analysis of random sequence reads indicates that the evolutionary patterns described herein apply to the sunflower genome as a whole.

INTRODUCTION

Transposable elements (TEs) are mobile DNA sequences that are present in the nuclear genomes of virtually all eukaryotes. A common feature of TEs is the potential to replicate faster than the host, thereby allowing them to increase in abundance, sometimes drastically (e.g., Naito *et al.,* 2009; Belyayev *et al.,* 2010), from one generation to the next. Variation in TE amplification rates can thus generate enormous variation in TE content within and between the genomes of even closely related species (e.g., Piegu *et al*., 2006; Ungerer *et al.,* 2006; Wicker *et al*., 2009). Differences in TE abundance amongst genomes may be explained by differences in host-encoded mechanisms that limit transposition, modes of TE replication, or specific properties that limit TE removal from the genome (Lippman *et al.,* 2004; Du *et al.,* 2010).

Class I TEs (i.e., retrotransposons) replicate through an RNA intermediate that is reverse transcribed into a DNA copy that can insert elsewhere in the genome (Kumar and Bennetzen 1999). These elements can be classified into five taxonomic orders (Wicker *et al.,* 2007). The most abundant and diverse order in plants, the long terminal repeat retrotransposons (LTR-RTs), is primarily composed of two superfamilies, *Ty1*/*copia* and *Ty3*/*gypsy* (referred to hereafter as *Copia* and *Gypsy*, respectively; Wicker *et al.,* 2007) which can be distinguished based on the order of their coding domains as well as the similarity of their reverse transcriptase sequences (Xiong and Eickbush 1990; Kumar and Bennetzen 1999). Certain *Gypsy* clades exhibit an extra coding domain known as the 'chromodomain', which is thought to confer insertion site specificity (Gao *et al.,* 2008). Though *Copia* and *Gypsy* elements are present in all plant genomes (Suoniemi *et al.,* 1998; Voytas *et al.,* 1992), their relative proportions vary between species (Hua-Van *et al.,* 2011). This variation may result from different insertion site preferences (Peterson-Burch *et al.,* 2004; Gao *et al.,* 2008), but could also be driven by variation in the efficacy of illegitimate

recombination and/or unequal homologous recombination in removing LTR-RTs from the genome (Devos *et al.,* 2002; Ma *et al.,* 2004). In contrast, Class II TEs (i.e., DNA transposons) use a DNA-based enzymatic method for excision and transposition of the parent copy without creating a new copy (Wicker *et al.,* 2007). Consequently, Class II TEs are generally less abundant than retrotransposons.

Despite their differences in genomic abundance, both retrotransposons and DNA transposons are potent sources of genetic variation (e.g., McClintock 1984; Hilbrict *et al*., 2008; Zeh *et al*., 2009). Transposable elements also have a large impact on, and appear to be integral components of, the chromatin landscape of the host genome (Biemont 2009). In *Arabidopsis thaliana*, for example, epigenetic regulation of TEs and tandem repeats contributes to genome organization and the regulation of neighboring genes (e.g., Lippman *et al.*, 2004; Hollister and Gaut 2009), and TEs also contribute to expression divergence between *Arabidopsis* species (Pereira *et al.,* 2009; Warenfors *et al.,* 2010; Hollister *et al.,* 2011). Given the potential influence of TEs on the structure and function of plant genomes, we investigated their contribution to sunflower (*Helianthus annuus* L.) genome evolution.

Sunflower is a diploid (2n = 34) species with an estimated genome size of ~3.5 Gb (Baack *et al*., 2005). Because the total number of retrotransposon copies in the genome of a plant species typically correlates with genome size (Bennetzen 2000; Bennetzen 2007; Devos 2010), we expected the sunflower genome to contain billions of bases pairs of retrotransposon DNA. Indeed, previous studies have suggested that the sunflower genome may be 62 – 78% repetitive (Cavallini *et al.,* 2010; Kane *et al.,* 2011), and a few studies have also investigated the genomic organization of retrotransposons in sunflower. Retrotransposons are known to be transcriptionally active in this species (Vukich *et al.,* 2009; Cavallini *et al.,* 2010; Kawakami *et al.,* 2011) and fluorescent *in situ*

hybridization studies have indicated that the *Gypsy* and *Copia* superfamilies are enriched in the heterochromatic regions of the pericentromeres and telomeres, respectively (Santini *et al.,* 2002; Natali *et al.,* 2006; Staton *et al.,* 2009). This genomic organization of *Gypsy* elements has been conserved in hybrid sunflower species derived from the common sunflower, despite massive amplifications of these elements in the hybrid species' genomes (Ungerer *et al.,* 2006; Staton *et al.,* 2009).

Many basic questions about the contributions of transposons to sunflower genome evolution remain unanswered, however, because previous studies have relied on *in situ* hybridization techniques that only offered chromosome level resolution (Natali *et al.,* 2006; Staton *et al.,* 2009; Cavallini *et al.,* 2010). For example, what has been the evolutionary time scale over which these sequences have been active? Were these, and the majority of other LTR-RT sequences, present in the common ancestor of sunflower and related species, or do they arise following the origin of the sunflower lineage (0.74 – 1.67 million years ago [MYA]; Heesacker *et al*., 2009)? Also, given that the sunflower genome is ca. 1 Gb larger than the maize genome, what type of TE diversity resides in the sunflower genome? And what is the relative importance of selective removal vs. selective amplification of TEs in shaping sunflower genome composition?

Here, we address these questions through a global survey of sequence composition and a fine-scale analysis of genomic structure. Specifically, we interrogated a large set of whole genome shotgun (WGS) sequence reads representing approximately 25% of the sunflower genome as well as the sequences of 21 unique bacterial artificial chromosome (BAC) clones. The random sequence reads allowed us to generate an unbiased and accurate estimate of sunflower genome composition, while the BAC sequences allowed for a detailed analysis of full-length TEs. We show that the sunflower genome is highly biased towards one superfamily of LTR-RTs,

discuss the diversity of LTR-RT families identified in this study, and investigate the evolutionary time scales over which all types of LTR-RTs in this species appear to have been active. The sunflower-specific repeats identified in this study will aid in efforts to assemble the sunflower genome, which is currently being sequenced (Kane *et al.,* 2011), and will greatly improve future repeat-masking and gene annotation efforts in the Asteraceae.

RESULTS

Sunflower genome composition

We investigated repeat content and abundance in a collection of whole genome shotgun (WGS) reads corresponding to 0.23X coverage of the sunflower genome. Through our analyses, we estimated that the sunflower genome is at least 81.1 ± 1.1% (mean ± SD) TEs and ribosomal repeats with 77.7 ± 1.8% being comprised of LTR-RTs, 57.9 ± 1.4% of which belong to the *Gypsy* superfamily (see Experimental Procedures; Figure 2.1a). Conversely, Subclass I (comprised of all terminal inverted repeat transposons) of Class II TEs and *Helitrons* (which are the only Class II - Subclass II TEs found in plants) accounted for just 1.3 ± 0.4% and 0.7 ± 1.6% of the genome, respectively (Figure 2.1a). Non-LTR retrotransposons appeared to occupy even less genomic space than Class II TEs, accounting for only 0.6 ± 0.4% of the sunflower genome, and were almost entirely composed of LINE-like lineages. Our graph-based analyses found that ~15% of the genome was single-copy, as represented by singletons, and an additional 4% of the genome was described as multi-copy genic sequences or low-copy transposable element families. The most abundant Class II-Subclass I TEs were the hAT and *Mutator* superfamilies, which comprised 0.38 ± 0.04% and 0.11 ± 0.06% of the genome, respectively (Figure 2.1b).

In addition to analyzing the WGS data for repeat composition and abundance we also analyzed the repeat composition of 21 BAC clones (ca. 2.5 Mb), twenty of which were selected for sequencing because they carry genes of interest (see Experimental Procedures). To characterize the diversity and demography of LTR retrotransposons in these BACs, we utilized both model-based and structure-based methods. All BAC clones were composed of, on average, 40.3% intact LTR-RTs with *Gypsy* families alone accounting for over 30% of the BAC clone sequences (Table 2.1; see Table A2.1). The lower frequency of TEs in the BAC data was likely due to the fact that the majority of these clones were selected for sequencing because they contained genes of interest, as noted above. We identified 16 families of LTR-RTs based on coding domain and terminal repeat similarity from intact and fragmented elements. The largest family, RLG-iketas, accounted for 19% of the LTR-RTs contained in the BAC clones analyzed. Consistent with the much lower frequency of the Class II transposable elements observed in the WGS dataset, the BAC sequences contained only a single *Mutator* element, four putative *Helitrons* families of two to four copies per family, and four putative MITE families of five to eight copies per family. In total, *Helitrons* and MITEs accounted for just 0.09% and 0.12% of the total BAC sequences, respectively. To further investigate the genomic abundance of specific LTR-RT families identified in the BAC clones, we compared an index of *k*-mers from the WGS reads to the BAC clones (see Experimental Procedures). In agreement with our estimates of family-level abundance based on the BAC clones, the WGS data has a high frequency of sequences matching the coding domains of *Gypsy* elements relative to *Copia* elements (Figure 2.2; see Figure A2.1).

Demography of LTR retrotransposons in the sunflower genome

To better understand the dynamics of LTR-RTs during sunflower genome evolution, we analyzed the structure and age of all elements from the BAC clones analyzed, including those not belonging to any of the 16 families described here. The *Copia* superfamily had a higher percentage of solo LTRs compared to *Gypsy* elements (Table 2.1; Table A2.2). While this result could potentially be an artifact of the non-random sample of BAC clones analyzed, Cavallini *et al.,*(2010) also reported a similar finding using a hybridization-based approach. In addition, an analysis of solo LTRs on a genome-wide scale revealed that *Copia* solo LTRs and truncated elements appear to be more abundant than those from *Gypsy* elements, compared to intact elements (Table 2.1). The average length of the solo LTRs was just 200 bp, while the average length of all LTRs was 1346 bp (Table A2.2). All truncated LTR-RTs and solo LTRs appeared to have arisen within the past 1.4 MY (0 – 1.4 MY for solo LTRs and 0.28 – 1.18 MY for truncated copies; as determined by the method of Vitte *et al.,* 2007). In addition, an analysis of the age distribution of all LTR-RTs found that the majority of copies identified in this study arose within the past 1 MY (Figure 2.3). Although many LTR-RT families were quite young (mean = 0.70 MY), the mean age of individual families was greater than 2 MY in some cases (e.g., RLG-kefe; Table A2.2, Figure 2.3).

The chromodomain-containing *Gypsy* families accounted for over 55% of all *Gypsy* elements, and these particular *Gypsy* families were characterized by an absence of solo LTRs in our data set. Moreover, all but one family (RLG-ryse; Table A2.2) contained all of the coding domains necessary for activity. Although the BAC clones analyzed represent a non-random sample of the genome, this finding is unlikely to be artifactual, as a comparison to the WGS reads revealed a high frequency of sequences matching to the chromoviruses, including the

chromovirus coding domains, identified in this study (Figure 2.2; Figure A2.1). We infer that these retrotransposons are likely to be autonomous based on the presence of multiple intact domains and translated ORFs longer than 500 amino acids in 81.8% of the elements (22.7% contained translated ORFs longer than 1000 amino acids; see also Bachlava *et al*., 2011) as well as evidence of transcriptional activity. Indeed, all chromoviruses also had at least 8 and as many as 26 unique matches to sunflower ESTs for a total of 574 unique ESTs matching the chromovirus sequences identified in this study (e.g., Figure 2.2; Figure A2.1) indicating that these sequences are expressed. This is in contrast to the *Copia* domain organization where only the reverse transcriptase and integrase were detectable. This latter finding may be related to the fact that the average age of *Copia* retrotransposons identified in this study was approximately twice the average age of the *Gypsy* superfamily described here (963,000 yrs vs. 552,000 yrs).

Phylogenetic diversity and structure of chromoviruses in sunflower

Because over half of the ~3.5 Gb sunflower genome is likely composed of LTR retrotransposons belonging to a phylogenetic clade referred to as the chromoviruses, we asked whether there were yet unknown novel clades of chromoviruses in sunflower. We also pursued this question because previous studies of chromovirus diversity have focused on a biased sample of plant genomes limited mainly to cereal crops and a few model dicot species (Gorinsek *et al*., 2004; Novikova *et al*., 2008). The phylogenetic placement of sunflower chromovirus sequences indicates that all sequences fall into known clades with nearly all sequences belonging to the Tekay clade, while a single sequence falls in the Reina clade (see Appendices II).

The two recognized groups of chromodomains—Group I and Group II—are defined by the presence of three aromatic residues (Gao *et al.*, 2008). All plant chromodomains appear to lack the first of these residues, and some plant species also lack the third aromatic residue

(Gorinsek *et al*., 2004; Gao *et al*., 2008; Novikova *et al*., 2008). As in other plant chromoviruses, sunflower chromodomains lack the first aromatic residue (position 6; Figure 2.4) but contain the second aromatic residue, which is characteristic of Group II chromodomains. One chromodomain (RLG-wimu-2; Figure 2.4) does contain a tryptophan at the third site, though this is not uncharacteristic of Group II chromodomains (Gao *et al*., 2008). By aligning the chromodomains from sunflower with predicted chromodomain secondary structures, we inferred the structure of these domains (Ball *et al.,* 1997; Figure 2.4). This alignment of chromodomains revealed the presence of duplications of entire chromodomains within individual retrotransposons in the sunflower genome. Nearly 85% (28/33) of the chromoviruses contained a single duplication of the chromodomain varying in length from 49 to 56 amino acids. Additionally, two chromoviruses from different BAC clones contained three perfect tandem duplications of a 53 amino acid chromodomain; the amino acid sequence of the chromodomain for these two retrotransposons varied by a single residue at position 51. In contrast, only 9% (3/33) of the chromoviruses contained just one chromodomain (52 - 53 amino acids). This pattern is also evident when looking at the whole genome level. For example, of the 4318 unique WGS reads with homology to a chromodomain, 74.4% were derived from a duplicated chromodomain (23.43% [1012/4318] with homology to a tandem chromodomain, 50.97% [2201/4318] with homology to more than two tandem chromodomains), as compared to 25.6% (1105/4318) being derived from a solo chromodomain. A phylogenetic analysis of duplications for all chromoviruses in sunflower revealed no evidence for multiple origins of tandem chromodomains (data not shown).

DISCUSSION

It is evident that the sunflower genome contains many thousands of retrotransposon copies (this study; Santini *et al.,* 2002; Natali *et al.,* 2006; Ungerer *et al.,* 2006), and numerous retrotransposon families are transcriptionally active in both cultivated (Vukich *et al.,* 2009) and wild populations (Kawakami *et al.,* 2011). However, there is a paucity of information regarding TE diversity and the mechanisms influencing the abundance of individual TE families in the sunflower genome. Thus, it seems clear that a comprehensive analysis of the diversity and dynamics of TEs would yield valuable insights into the role of TEs in the evolution of this important species.

Sunflower genome composition: pattern and process

Sunflower is distantly related to any plant species for which there is a curated set of genomic repeats (e.g., the estimated divergence time from *A. thaliana* is ~120 MY – i.e. the divergence time between Asterids and Rosids; Cenci *et al*., 2010). Therefore, to create a library of repeats for sunflower, we relied on a *de novo* repeat finding method rather than strictly homology-based methods (Novak *et al.,* 2010). To assess the composition of the sunflower genome we analyzed over 811 Mb of WGS reads (~0.23X; see Experimental Procedures) using the method of Novak *et al.,*(2010). LTR-RTs were the most abundant form of DNA in the sunflower genome, with the *Gypsy* superfamily alone accounting for ~58% of the genome (see also Cavallini *et al.,* 2010). Interestingly, analysis of intact LTR-RTs in BAC clone sequences revealed that the largest density of all LTR-RT insertions has occurred within the last 1 MY. That is, they arose since, or concomitantly with, the origin of sunflower as a species (Figure 2.3; Heesacker *et al*., 2009). Although this dating procedure is an approximation and may not reflect the true time since insertion, the finding of recent insertions is concordant with a previous study demonstrating that

LTR-RTs are transcriptionally active in multiple wild populations of *H. annuus* and other annual sunflower species (Kawakami *et al.,* 2011). Though many insertions likely predate the origin of the *H. annuus* lineage (Figure 2.3), all insertions are within the age estimates for the origin of the genus *Helianthus* (i.e. the extant lineages arose 1.7 – 8.2 MYA; Schilling 1997). Thus, the diversity and dynamics of LTR-RTs presented here likely reflect properties unique to the sunflower lineage, a finding consistent those of Buti *et al.* (2011) where LTR-RT age was analyzed in three gene-harboring BAC clones. Biases towards recent (i.e. < 5 MY) LTR-RT insertions have also been noted in other plant genomes (Ma and Bennetzen 2004; Vitte *et al.,* 2007; Wang and Liu 2008; Du *et al.,* 2010), and this pattern likely reflects an ongoing struggle (i.e. 'genomic turnover') between the addition and removal of repetitive elements (Ma and Bennetzen 2004).

We investigated how this process may have shaped the sunflower genome by analyzing the structure of LTR-RTs in order to assess the relative efficacy of unequal homologous recombination and illegitimate recombination in counteracting expansion of the sunflower genome. Formation of solo LTRs and truncated elements results from unequal homologous recombination between LTRs of a single LTR-RT or between elements at different genomic locations, respectively (Devos *et al.,* 2002; Bennetzen *et al.,* 2005), and this process appears to have been an effective DNA removal mechanism in the rice and barley genomes (Vitte *et al.,* 2003; Shirasu *et al.,* 2000). However, the process of illegitimate recombination, which involves microhomology and occurs independently of the normal recombinational machinery, may have a greater impact on counteracting genome expansion through the formation of truncated elements (Chantret *et al.,* 2005), as appears to be the case in *Arabidopsis thaliana* (Devos *et al.,* 2002) and *Medicago truncatula* (Wang and Liu 2008).

In sunflower, solo LTRs and truncated LTR-RTs appeared to be in lower abundance than full-length elements (0.14:1.0:0.6 ratio of solo LTR:intact LRT-RT:truncated LTR-RT for all sunflower LTR-RTs; Table A2.2), as has been observed in maize (0.2:1.0 ratio of solo LTR:intact LTR-RT; SanMiguel *et al.,* 1996; Devos *et al.,* 2002). Solo LTRs were also biased towards the *Copia* superfamily and the majority of *Copia* solo LTRs analyzed (10/15) showed no divergence, suggesting a recent origin in our data set. In addition, a ratio of greater than 2:1 for LTR:reverse transcriptase sequences on a whole genome scale could indicate that 1) *Copia* solo LTRs are more abundant that intact elements, 2) there is paucity of coding domains for *Copia* elements in the genome, or 3) both of these factors are contributing to the observed patterns, and the latter possibility is supported by our results from the analysis of 21 BAC clones (Table 2.1; Table A2.2). These differences in solo LTR formation between superfamilies may be driven by insertion preferences and LTR length—e.g., elements containing longer LTRs may be biased towards solo LTR formation (Vitte *et al.,* 2003; Du *et al.,* 2010)—though *Copia* LTRs are half the length of *Gypsy* LTRs on average. In addition, the solo LTR fragments detected in this study averaged only 200 bp in length, which may reflect selection against the removal of larger stretches of DNA in genic regions (Tian *et al.,* 2009). Despite finding a paucity of solo LTRs, however, we did find a large number of deletions (278 total, ranging from 10-17 bp each) flanked by short (4-9 bp) direct repeats (Figure A2.2; Table A2.3).

Though results from analyses of genomic structure can vary depending on the genomic regions being analyzed (e.g., Ma and Bennetzen 2004, 2006), the foregoing findings highlight important processes that may be contributing to sunflower genome evolution. First, the observed bias in sunflower genome composition appears to have been driven, at least in part, by the selective removal of *Copia* LTR-RTs, as opposed to solely resulting from amplification of *Gypsy*

elements (Table A2.2). This result is supported by hybridization-based studies using *Gypsy* and *Copia* LTR sequences in sunflower (Cavallini *et al.*, 2010), and may have a significant impact on TE composition because solo LTR formation may remove more LTR-RT DNA than illegitimate recombination alone over short evolutionary time scales (Devos *et al.,* 2002). However, the frequency of putative illegitimate recombination events we analyzed for the *Gypsy* and *Copia* superfamilies was proportional to their abundance (Table A2.2; Table A2.3). Second, our observation that solo LTRs were rare in regions harboring genes, where they might be expected to be more abundant (Tian *et al.,* 2009; Du *et al.,* 2010), suggests that illegitimate recombination may play an important role in regulating the DNA content in the sunflower genome. The high percentage of small deletions associated with sunflower LTR-RTs was also strongly suggestive of illegitimate recombination (Figure A2.2; Table A2.3). Even so, the relative importance of unequal homologous recombination and illegitimate recombination likely varies over evolutionary time (Tian *et al.,* 2009), and further investigation of the nature of recombination in sunflower will be required to determine the absolute genomic impact of these processes.

We also found a disproportional abundance of LINE-like lineages of Non-LTR retrotransposons, as compared with the abundance of SINE-like lineages in our WGS data. In contrast, despite a slight bias towards the hAT superfamily, all types of Class II (Subclass I) TEs appear in nearly equal abundance (Figure 2.1b). This variation in proportionality may indicate differences in insertion preferences and host control between Class I and Class II TEs in sunflower.

Chromovirus structures and their potential impact on the sunflower genome

Chromoviruses appear to be the most abundant lineage of *Gypsy* LTR-RTs among flowering plants (Gorinsek *et al.,* 2004; Kordis 2005); this pattern was concordant with our

observations in sunflower where over 55% of intact *Gypsy* elements identified in the BAC sequences contained a chromodomain. Based on work in *Schizosaccharomyces pombe,* it has been shown that chromodomains mediate the integration of chromovirus sequences by interacting with di-methyl and tri-methylated lysine-9 residues on histone H3, an epigenetic mark of heterochromatin (Gao *et al.,* 2008). Notably, the most highly conserved residues of chromodomains in sunflower chromoviruses, four of which are invariant, reside within the regions predicted to mediate interactions with methylated lysine residues on histone H3 (Figure 2.4; Jacobs and Khorasanizadeh 2002; Nielsen *et al.*, 2002).

Interestingly, nearly 85% of the chromovirus sequences identified in the BAC sequences contain at least one tandem duplication of the chromodomain, and nearly 75% of the chromodomain-derived sequences identified in the WGS reads appear to have been derived from tandem arrays of chromodomains. Given that tandem chromodomains recognize methylated lysine-4 on histone H3 in *Drosophila* and humans, which is a mark of transcriptionally active euchromatin (Flanagan *et al.,* 2005; Flanagan *et al.,* 2007), and that the abundance of elements with duplicated chromodomains is marginally higher in gene-containing BACs vs. the genome as a whole, it is tempting to infer that a similar function could be employed by certain sunflower chromovirus sequences. Analyses of randomly selected BAC clones could provide insight into the genome-wide co-occurrence of chromoviruses and genes. This finding also raises the possibility that chromatin remodeling factors associated with sunflower chromoviruses could potentially lend to their stability in the genome (Lippman *et al.,* 2004), and help to explain the biased composition of TEs in the sunflower genome.

Whether these findings represent yet unknown active targeting mechanisms for chromoviruses or are the result of aberrant integration due to mutations (i.e. duplication of the chromodomain), it is evident that these sequences have played an active and presumably ongoing role in shaping the sunflower genome.

EXPERIMENTAL PROCEDURES

WGS and BAC clone sequencing

In order to obtain an unbiased estimate sunflower genome composition, 2,325,196 random genomic sequences (i.e. WGS sequences; mean length 403 bp, GC 39.05%; ~811 Mb total) were generated via Roche 454 GS FLX sequencing of a highly inbred line derived from sunflower cultivar HA412-HO (PI 642777) using the XLR (Titanium) chemistry. With the exception of sequences showing similarity to rDNA genes and organellar genomes (see below), all of these sequences were used in the analysis of genome composition.

Twenty-one bacterial artificial chromosome (BAC) clones from sunflower cultivar HA383 (PI 578872) were selected for sequencing based on the presence of genes of evolutionary and/or agronomic importance (Table A2.1). BAC clones were prepared using standard protocols (Bachlava *et al*., 2011; Blackman *et al.,* 2011). Sixteen of these BAC clones were sequenced using a Sanger shotgun approach at either Washington University or the Joint Genome Institute with automatic and manual finishing. Assembly and editing were carried out with Phrap and Consed, respectively (Ewing *et al*., 1998; Ewing and Green 1998; Gordon *et al*., 1998). Four additional clones were sequenced in the Georgia Genomics Facility using a Roche 454 GS FLX sequencer with XLR (Titanium) sequencing chemistry. Final assemblies were generated with MIRA (v3.0.3; Chevreux 1999; see Appendices II for details). The final BAC clone was selected

by probing the same sunflower BAC library (filter Ha_HBa_A) with a *Gypsy* integrase sequence fragment and selecting a clone address exhibiting a strong hybridization signal. Sequencing, assembly, and editing of this BAC clone were performed at the Clemson University Genomics Institute (CUGI). The WGS and BAC clone sequences described above are available for download at sunflower.uga.edu/data.

Repeat identification from WGS and BAC clone sequences

All sequences containing chloroplast, mitochondrial, or ribosomal fragments were removed using BLAST similarity searches and custom Perl scripts (Altschul *et al.,* 1990); low complexity sequences were removed with the DUST algorithm (Hancock and Armstrong 1994). First, to identify putative repeat families, a graph-based clustering method was applied to the cleaned, reduced set of genomic sequences (2,088,836 in total; Novak *et al*., 2010). Despite having removed ribosomal and low complexity sequences, clustering was not feasible on the full data set due to computational requirements, so the data were split into four subsets containing ~500k sequences each. Briefly, clustering was performed by first using an all-by-all search with mgblast with the following parameters: -F "m D" -D 4 -p 85 -W18 -UT -X40 -KT -JF -v90000000 -b90000000 -C80 -H 320 -a 8 (Pertea *et al*., 2003; Novak *et al*., 2010). Next, a custom script was used to select read pairs that had at least 90% identity and covered at least 15% of the length of the matching sequences. The bitscore for read pairs that passed these thresholds was used for clustering with the methods and software described by Novak *et al.,*(2010). Lastly, all clusters containing at least 500 reads were assembled using the Roche gsAssembler software (version 2.5.3; 454 Life Sciences, Branford, CT), and contigs were searched for coding domains with HMMscan (version 2.3.2; Eddy 1998) using the translated nucleotide sequences as a query against the Pfam database (release 24.0; Finn *et al*., 2010). We also performed nucleotide

searches (BLASTN searches with an e-value of 1e-5) with the contigs using a custom repeat database–comprised of Repbase release 15.06 (Jurka *et al*., 2005), mips-REdat version 4.3 (Spannagl *et al*., 2007), and the JCVI maize characterized repeats V4.0 (http://maize.jcvi.org/repeat_db.shtml)–as the target. The size and composition of clusters for each of the four subsets showed very little variation with respect to abundance; thus, we have reported the abundance of each transposable element type as an average of the subsets, as well as the standard deviation for each estimate.

The program LTR_Finder (Xu and Wang 2007) was used with default settings, and executed with the batch_ltrfinder.pl script from the DAWGPAWS package (Estill and Bennetzen 2009), in order to discover intact LTR retrotransposons from the BAC clones. In addition, the program LTR*harvest* (version 1.3.4; Ellinghaus *et al*., 2008) was used to discover LTR-RTs using the default settings except for the following parameter changes: -mintsd 4 -mindistltr 4000 -maxlenltr 4000. Given that Ellinghaus *et al.,*(2008) demonstrated a higher rate of true positive recovery with LTR*harvest* when combined with a clustering step as compared to other LTR-RT prediction methods and that LTR_Finder recovered a low percentage of elements with TSDs, the output of LTR*harvest* was used to search for binding sites and coding domains. To identify coding regions within the predicted retrotransposons, the program LTR*digest* (Steinbiss *et al*., 2009) was run on the LTR-RTs predicted by LTR*harvest*. Complete, or intact, LTR-RTs were defined as having at minimum two flanking TSDs, two nearly intact LTRs, a primer binding site, and a poly purine tract (see Ma *et al.,* 2004). Solo LTRs and truncated LTR-RTs were identified by searching the BAC clone sequences with the full-length LTR-RTs (see Appendicies II). Putative sites of illegitimate recombination were identified by first, aligning all full-length members of an LTR-RT family (see below) and then comparing (with the BLAST program

"bl2seq") the 20 bp of sequence upstream and downstream of gap sites for direct repeats. To eliminate artifacts, we only analyzed gap sites >10 bp that were flanked by direct repeats >4 bp which had no more than 2 non-matching bases intervening the matching repeats and a gap (see also Devos *et al.,* 2002; Ma *et al.,* 2004). Deletions shared by more than one element were assumed to represent an ancestral event and were counted once (Ma *et al.,* 2004).

LTR-RT superfamilies (e.g., *Gypsy* and *Copia*) were constructed using evidence from matches to HMMs for the RVT domain and matches to the custom repeat database described above. LTR-RT families were identified by clustering separately the primer binding site, 5' LTR sequence, and internal coding domains (i.e. gag, reverse transcriptase, integrase, RNase H, and chromodomain) with Vmatch (http://vmatch.de) following the methods described in Steinbiss *et al.,*(2009). All LTR-RT families were named according to Wicker *et al.,*(2007). Each LTR-RT copy that could not be unambiguously assigned to a family but could be assigned to a superfamily (see Wicker *et al*., 2007) was classified as RLG-X or RLC-X for *Gypsy* unclassified or *Copia* unclassified, respectively. The procedure for dating each LTR-RT family was adapted from (Vitte *et al.,* 2007; Baucom *et al., 2*009; see also SanMiguel *et al.,* 1998). Briefly, the K80 model (Kimura 1980) within the BaseML module of PAML v4.2a (Yang 2007) was used to obtain a likelihood divergence estimate for each LTR-RT based on the similarity of the two LTRs. This divergence value (which we will refer to as *d*) was used to determine age with the formula $T = d/2r$, where $r = 1.0 \times 10^{-8}$ as determined for host encoded genes (Strasburg and Rieseberg 2008), and the multiplier of two accounts for the elevated rates of evolution of TEs as compared to genes (Baucom *et al.,* 2009). Putative Class II transposons and *Helitrons* were identified using MITEHunter as well as through similarity searches using HMMER and InterProScan (Eddy 1998; Zdobnov and Apweiler 2001), and Helsearch (Yang and Bennetzen 2009), respectively.

To compare the frequency of intact repeats identified from BAC clones to their frequency in the whole genome, we generated 20-mers for each BAC clone and compared those sequences to an index of 20-mers from all of the WGS reads using Tallymer (Kurtz *et al.,* 2008). Plotting the relationship between the length of *k*-mers and the uniqueness ratio for each value of *k* from 1 – 100 revealed a natural inflection at *k=20*, similar to the maize genome (Kurtz *et al.,* 2008), representing a value that would maximize the information and resolution in the *k*-mers being compared (Kurtz *et al.,* 2008). Custom Perl scripts were then used to format matches between the WGS index and BAC clone 20-mers for viewing in GBrowse v2.40 (Figure 2; Stein *et al.,* 2002). The genome-wide frequency of solo LTRs was estimated with similarity searches using BLAST where the WGS read set was the subject and the LTR and reverse transcriptase sequences (from intact LTR-RTs identified in the BAC clones) were used as the query (see Appendicies II). This same procedure was used for determining the relative frequency of chromodomain duplications in the genome wherein the sequences of single and tandemly duplicated chromodomains (identified in the BAC clone sequences) where used to interrogate the WGS reads. A unique match in the WGS reads was scored as single if it had only a single matching region up to the length of a chromodomain, tandem matches were scored by the presence of two (or more) regions where one match begins at the end site of the previous match. All scripts described herein are available upon request.

ACKNOWLEDGEMENTS

We kindly thank Dr. Dusan Kordis for sharing plant chromovirus sequences as well as Navdeep Gill and members of the Burke Lab for comments on an earlier version of the manuscript. This work was supported by grants from the National Science Foundation (DBI-0820451 to JMB, SJK, and LHR and DEB-0742993 to MCU) as well as the USDA National Institute of Food and Agriculture (2008-35300-19263 to JMB).

LITERATURE CITED

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215, 403-410.

Baack EJ, Whitney KD, and Rieseberg LH. 2005. Hybridization and genome size evolution: timing and magnitude of nuclear DNA content increases in *Helianthus* homoploid hybrid species. *New Phytol.* 167, 623-630.

Bachlava E, Radwan OE, Abratti G, Tang S, Gao W, Heesacker AF, et al. 2011. Downy mildew (Pl8 and Pl14) and rust (RAdv) resistance genes reside in close proximity to tandemly duplicated clusters of non-TIR-like NBS-LRR-encoding genes on sunflower chromosomes 1 and 13. *Theor. Appl. Genet.* 122, 1211-1221.

Ball LJ, Murzina NV, Broadhurst RW, Raine AR, Archer SJ, Stott FJ, Murzin AG, Singh PB, Domaille PJ and Laue ED. 1997. Structure of the chromatin binding (chromo) domain from mouse modifier protein 1. *EMBO J.* 16, 2473-2481.

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

Belyayev A, Kalendar R, Brodsky L, Nevo E, Schulman AH, and Olga Raskina. 2010. Transposable elements in a marginal plant population: temporal fluctuations provide new insights into genome evolution of wild diploid wheat. *Mobile DNA,* 6, 1.

Bennetzen JL. 2000. Transposable element contributions to plant gene and genome evolution. *Plant Mol. Biol.* 42, 251-269.

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

Bennetzen JL, Ma J, and Devos KM. 2005. Mechanisms of recent genome size variation in flowering plants. *Plant Mol. Biol.* 95, 127-132.

Biemont C. 2009. Are transposable elements simply silenced or are they under house arrest? *Trends Genet.* 25, 333-334.

Blackman BK, Rasmussen DA, Strasburg JL, Raduski AR, Burke JM, Knapp SJ, Michaels SD, and Rieseberg LH. 2011. Contributions of flowering time genes to sunflower domestication and improvement. *Genetics* 187, 271-287.

Bohne A, Brunet F, Galiana-Arnoux D, Schultheis C, and Volff J-N. 2008. Transposable elements as drivers of genomic and biological diversity in vertebrates. *Chrom. Res.* 16, 203-215.

Buti M, Giordani T, Cattonaro F, Cossu RM, Pistelli L, Vukich M, Morgante M, Cavallini A, and Natali L. 2010. Temporal dynamics in the evolution of the sunflower genome as revealed

by sequencing and annotation of three large genomic regions. *Theor. Appl. Genet.* 5, 779-791.

Cavallini A, Natali L, Zuccolo A, Giordani T, Jurman I, Ferrillo V., et al. 2010 Analysis of transposons and repeat composition of the sunflower (*Helianthus annuus* L.) genome. *Theor. Appl. Genet.* 120, 491-508.

Cenci A, Combes MC, and Lashermes P. 2010. Comparative sequence analysis reveals that Coffea (Asterids) and Vitis (Rosids) derive from the same paleo-hexaploid ancestral genome. *Mol. Genet. Genomics* 283, 493-501.

Chantret N, Salse J, Sabot F, Rahman S, Bellec A, Laubin B, Dubois I, et al. 2005. Molecular basis of evolutionary events that shaped the *Hardness* locus in diploid and polyploidy wheat speces (Triticum and Aegilops). *The Plant Cell,* 17, 1033-1045.

Chevreux B, Wetter T, and Suhai S. 1999. Genome sequence assembly using trace signals and additional sequence information. *Comp Sci. Biol: Proc. German Conf. Bioinf. (GCB)* 99, 45-56.

Chueh AC, Northrop EL, Brettingham-Moore KH, Choo KHA, and Wong LH. 2009. LINE retrotransposon RNA is an essential structural and functional epigenetic component of a core neocentromeric chromatin. *PLOS Genetics,* 5.

Devos, K.M. 2010. Grass genome organization and evolution. *Curr. Opin. Plant Biol.* 13, 139-145.

Devos KM, Brown JKM, and Bennetzen JL. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis. Genome Res.* 12, 1075-1079.

Donoghue MTA, Keshavaiah C, Swamidatta SH, Spillane C. 2011. Evolutionary origins of Brassicaceae specific genes in Arabidopsis thaliana. *BMC Evol. Biol.* 11, 47.

Du J, Tian Z, Hans CS, Laten HM, Cannon SB, Jackson SA, Shoemaker RC, and Ma J. 2010. Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison. *Plant J.* 63, 584-598.

Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics* 14, 755-763.

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

Estill JC, and Bennetzen JL. 2009. The DAWGPAWS pipeline for the annotation of genes and transposable elements in plant genomes. *Plant Methods*, 5, 8.

Ewing B, and Green P. 1998 Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.* 8, 186–194.

Ewing B, Hillier L, Wendl MC, and Green P. 1998. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Res*. 8, 175–185.

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

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, et al. 2010. The Pfam protein families database. *Nucleic Acids Res.* 38, D211-D222.

Flanagan JF, Mi L, Chruszcz M, Cymborowski M, Clines KL, Kim Y, Minor W, Rastinejad F, and Khorasanizadeh S. 2005. Double chromodomains cooperate to recognize the methylated histone H3 tail. *Nature,* 438, 1181-1185.

Flanagan JF, Blus BJ, Kim D, Clines KL, Rastinejad F, and Khorasanizadeh S. 2007. Molecular implications of evolutionary differences in CHD double chromodomains. *J. Mol. Biol.* 369, 334-342.

Gao X, Hou Y, Ebina H, Levin HL, and Voytas DF. 2008. Chromodomains direct integration of retrotransposons to heterochromatin. *Genome Res.* 18, 359-369.

Gorinsek B, Gubensek F, and Kordis D. 2004. Evolutionary genomics of chromoviruses in eukaryotes. *Mol. Biol. Evol.* 21, 781-798.

Gordon D, Abajian C, and Green P. 1998. Consed: a graphical tool for sequence finishing. *Genome Res.* 8, 195-202.

Hancock JM, and Armstrong JS. 1994. SIMPLE34: an improved and enhanced implementation for VAX and Sun computers of the SIMPLE algorithm for analysis of clustered repetitive motifs in nucleotide sequences. *Comput. Appl. Biosci.* 10, 67-70.

Heesacker AF, Bachlava E, Brunick RL, Burke JM, Rieseberg LH, and Knapp SJ. 2009. Karyotypic evolution of the common and silverleaf sunflower genomes. *Plant Gen.* 2, 233-246.

Hernandez-Hernandez A, Rincon-Arano H, Recillas-Targa F, Ortiz R, Valdes-Quezada, C, Echeverria OM, et al. 2008. Differential distribution and association of repeat DNA sequences in the lateral element of the synaptonemal complex in rat spermatocytes. *Chromosoma* 117, 77-87.

Hernandez-Pinzon I, de Jesus E, Santiago N, and Casacuberta JM. 2009. The frequent transcriptional readthrough of the Tobacco Tnt1 retrotransposon and its possible implications for the control of resistance genes. *J. Mol. Evol.* 68, 269-278.

Hilbrict T, Varotto S, Sgaramella V, Bartels D, Salamini F, and Furini, A. 2008. Retrotransposons and siRNA have a role in the evolution of desiccation tolerance leading to resurrection of the plant *Craterostigma plantagineum*. *New Phytol.* 179, 877-887.

Hollister JD, and 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.

Hollister JD, Smith LM, Guo Y, Ott F, Weigel D, and 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.

Hua-Van A, Le Rouzic A, Boutin TS, Filee J, and Capy P. 2011. The struggle for life of the genome's selfish architects. *Biol. Direct* 6, 19.

Jacobs SA, and Khorasanizadeh S. 2002. Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail. *Science,* 295, 2080-2083.

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

Kane MC, Gill N, King ME, Bowers JE, Berges H, Gouzy J, Bachlava E, et al. 2011. Progress towards a reference genome for sunflower. *Botany* 89, 429-437.

Kawakami T, Dhakal P, Katterhenry AN, Heatherington CA, Ungerer MC. 2011. Transposable element proliferation and genome expansion are rare in contemporary sunflower hybrid populations despite widespread transcriptional activity of LTR retrotransposons. *Genome Biol. Evol.* 3, 156-167.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111-120.

Kobayashi S, Goto-Yamamoto N, and Hirochika H. 2004. Retrotransposon-induced mutations in grape skin color. *Science* 304, 982.

Kordis D. 2005. A genomic perspective on the chromodomain-containing retrotransposons: Chromoviruses. *Gene* 347, 161-173.

Kumar A, and Bennetzen JL. 1999. Plant Retrotransposons. *Annu. Rev. Genet.* 33, 479-532.

Kurtz S, Narechania A, Stein JC, and Ware D. 2008. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics,* 9, 517.

Lippman Z, Gendrel A, Black M, Vaughn MW, Dedhia N, McCombie WR, et al. 2004 Role of

transposable elements in heterochromatin and epigenetic control. *Science* 430, 471-476.

Ma J, and Bennetzen JL. 2004. Recent growth and divergence of rice nuclear genomes. *Proc. Natl Acad. Sci. U.S.A.* 101, 12404-12410.

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

McClintock B. 1984. The significance of responses of the genome to challenge. *Science* 226, 792-801.

Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, Richardson AO, Okumoto Y, Tanisaka T, and Wessler SR. 2009. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature* 461, 1130-1134.

Natali L, Santini S, Giordani T, Minelli S, Maestrini P, Cionini PG, and Cavallini A. 2006. Distribution of Ty3-*gypsy*- and Ty1-*copia*-like DNA sequences in the genus *Helianthus* and other Asteraceae. *Genome* 49, 64-72.

Nielsen PR, Nietlispach D, Mott HR, Callaghan J, Bannister A, Kouzarides T, Murzin, AG, Murzina NV, and Laue E.D. 2002. Structure of the HP1 chromodomain bound to histone H3 methylated at lysine 9. *Nature* 416, 103-107.

Novak P, Neumann P, and Macas J. 2010. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinf.* 11, 378.

Novikova O, Mayorov V, Smyshlyaev G, Fursov M, Adkison L, Pisarenko O, and Blinov A. 2008. Novel clades of chromodomain-containing Gypsy LTR retrotransposons from mosses (Bryophyta). *Plant J.* 56, 562-574.

O'Neill RJW, O'Neill MJ, and Graves JAM. 1998. Undermethylation associated with retroelement activation and chromosome remodeling in an interspecific mammalian hybrid. *Nature* 393, 68-72.

Pearlman RE, Tsao N, and Moens PB. 1992. Synaptonemal complexes from DNase-treated rat pachytene chromosomes contain (GT)n and LINE/SINE sequences. *Genetics* 130, 865-72.

Pereira V, Enard D, and Eyre-Walker A. 2009. The effect of transposable element insertions on gene expression evolution in rodents. *PLoS ONE,* 4, e4321.

Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, et al. 2003. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* 19, 651-652.

Peterson-Burch BD, Nettleton D, and Voytas DF. 2004. Genomic neighborhoods for *Arabidopsis* retrotransposons: a role for targeted integration in the distribution of the Metaviridae. *Genome Biol.* 5, R78.

Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, Collura H, et al. 2006. Doubling genome size without polyploidization: dynamics of retrotransposon-mediated genome expansions in *Oryza australensis*, a wild relative of rice. *Genome Res.* 16, 1262-1269.

Rieseberg LH. 2006. Hybrid speciation in wild sunflowers. *Ann. Missouri Bot. Gard.* 93, 34-48.

SanMiguel P, Tikhonov A, Jin Y, Motchoulskaia N, Zakharov D, Melake-Berhan A, et al.1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science* 274, 765-768.

Santini S, Cavallini A, Natali L, Minelli S, Maggini F, and Cioini PG. 2002. *Ty1*- and *Ty3/gypsy*-like retrotransposon sequences in *Helianthus* species. *Chromosoma* 111, 192-200.

Schilling EE. 1997. Phylogenetic analysis of *Helianthus* (Asteraceae) based on chloroplast restriction-site data. *Theor. Appl. Genet.* 94, 925-933.

Shirasu K, Schulman AH, Lahaye T, and Schulze-Lefert P. 2000. A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Res.* 10, 908–915.

Spannagl M, Noubibou O, Haase D, Yang L, Gundlach H, Hindemitt T, Klee K, Haberer G, Schoof H, and Mayer KFX. 2007. MIPSPlantsDB—plant database resource for integrative and comparative plant genome research. *Nucleic Acids Res.* 35, D834-D840.

Suoniemi A, Tanskanen J, and Schulman AH. 1998. Gypsy-like retrotransposons are widespread in the plant kingdom. *The Plant Journal,* 13, 699-705.

Staton SE, Ungerer MC, Moore RC. 2009. The genomic organization of Ty3/gypsy-like retrotransposons in Helianthus (Asteraceae) homoploid hybrid species. *Amer. J. Bot.* 96, 1646-1655.

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, et al. 2002. The generic genome browser: a building block for a model organism system database. *Genome Res.* 10, 1599-1610.

Steinbiss S, Willhoeft U, Gremme G, and Kurtz S. 2009. Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Res.* 37, 7002-7013.

Strasburg J, and Rieseberg LH. 2008. Molecular demographic history of the annual sunflowers Helianthus annuus and H. petiolaris - large effective population sizes and rates of long-term gene flow. *Evolution* 62, 1936-1950.

Tian Z, Rizzon C, Du J, Zhu L, Bennetzen JL, Jackson SA, Gaut BS, and Ma J. 2009. Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? *Genome Res.* 19, 2221-2230.

Ungerer MC, Strakosh SC, and Zhen Y**.** 2006. Genome expansion in three hybrid sunflower species is associated with retrotransposon proliferation. *Current Biology,* 16, R872-R873.

Ungerer MC, Strakosh SC, and Stimpson KM. 2009. Proliferation of Ty3/gypsy-like retrotransposons in hybrid sunflower taxa inferred from phylogenic data. *BMC Biol.* 7, 40.

van der Heijden WG, and Bortvin A. 2009. Transient relaxation of transposon silencing at the onset of mammalian meiosis. *Epigenetics* 4, 76-79.

Vitte C, and Panaud O. 2003. Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice Oryza sativa L. *Mol. Biol. Evol.* 20, 528–540.

Vitte C, Panaud O, and Quesneville H. 2007. LTR retrotransposons in rice (*Oryza sativa*, L.): recent burst amplifications followed by rapid DNA loss. *BMC Genomics* 8, 218.

Voytas DF, Cummings MP, Konieczny A, Ausubel FM, and Rodermel SR. 1992. Copia-like retrotransoposons are ubiquitous among plants. *Proc. Natl Acad. Sci. U.S.A.* 89, 7124-7128.

Vukich M, Giordani T, Natali L, and Cavallini A. 2009. *Copia* and *Gypsy* retrotransposons activity in sunflower (*Helianthus annuus* L.). *BMC Plant Biol.* 9, 150.

Wang H, and Liu J. 2008. LTR retrotransposon landscape in *Medicago truncatula*: more rapid removal than in rice, *BMC Genomics,* 9, 382.

Warenfors M, Pereira V, and Eyre-Walker A. 2010. Transposable elements: Insertion pattern and impact on gene expression evolution in Hominids. *Mol. Biol. Evol.* 27, 1955-1962.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, et al. 2007. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Gen.* 12, 973-982.

Wicker T, Taudien S, Houben A, Keller B, Graner A, Platzer M, and Stein N. 2009. A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. *Plant J.* 59, 712-722.

Winckler T, Szafranski K, and Glöckner G. 2005. Transfer RNA gene-targeted integration: an adaptation of retrotransposable elements to survive in the compact Dictyostelium discoideum genome. *Cytogenet. Genome Res.* 110, 288-298.

Xiao H, Jiang N, Schaffner E, Stockinger EJ, and van der Knapp E. 2008. A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science,* 319, 1527-1530.

Xiong Y, and Eickbush TH. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.* 9, 3353–3362.

Xu Z, and Wang H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 35, W265-W268.

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

Yang L, and Bennetzen JL. 2009. Distribution, diversity, evolution and survival of Helitrons in the maize genome. *Proc. Natl Acad. Sci. U.S.A.* 106, 19922-19927.

Zdobnov EM, and Apweiler R. 2001. InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17, 847-848.

Zeh DW, Zeh JA, and Ishida Y. 2009. Transposable elements and an epigenetic basis for punctuated equilibria. *Bioessays* 31, 715-726.

TABLES FOR CHAPTER II

Table 2.1. Statistics for LTR retrotransposon superfamilies derived from BAC clone sequences (top) and WGS reads (bottom).

Superfamily	Count	Overall length[1]	LTR length[1]	Percent of BACs[2]	Solo:FL:TR[3]
Copia	*28*	*9061*	*775*	*9.86 ± 10.6*	*0.53:1:0.03*
Gypsy	*79*	*9918*	*1551*	*30.47 ± 26.7*	*0.15:1:0.07*
Total	*107*	*9693*	*1346*	*40.33 ± 24.0*	*0.14:1:0.06*

Superfamily	Percent of WGS reads[2]	LTR:RVT[4]
Copia	*19.83 ± 2.8*	*2.27:1*
Gypsy	*57.93 ± 1.4*	*1.53:1*
Total	*77.75 ± 1.84*	*1.9:1*

1 – Lengths are presented as the average (in bp). 2 – Percent composition of BAC clones and WGS reads along with the standard deviation for each superfamily. 3 – Ratio of solo LTRs (Solo) to full-length (FL) to truncated (TR) LTR retrotransposon copies. 4 – The ratio of BLAST hits for LTR sequences (LTR) to reverse transcriptase (RVT) sequences from the WGS reads (see Experimental Procedure

FIGURES FOR CHAPTER II

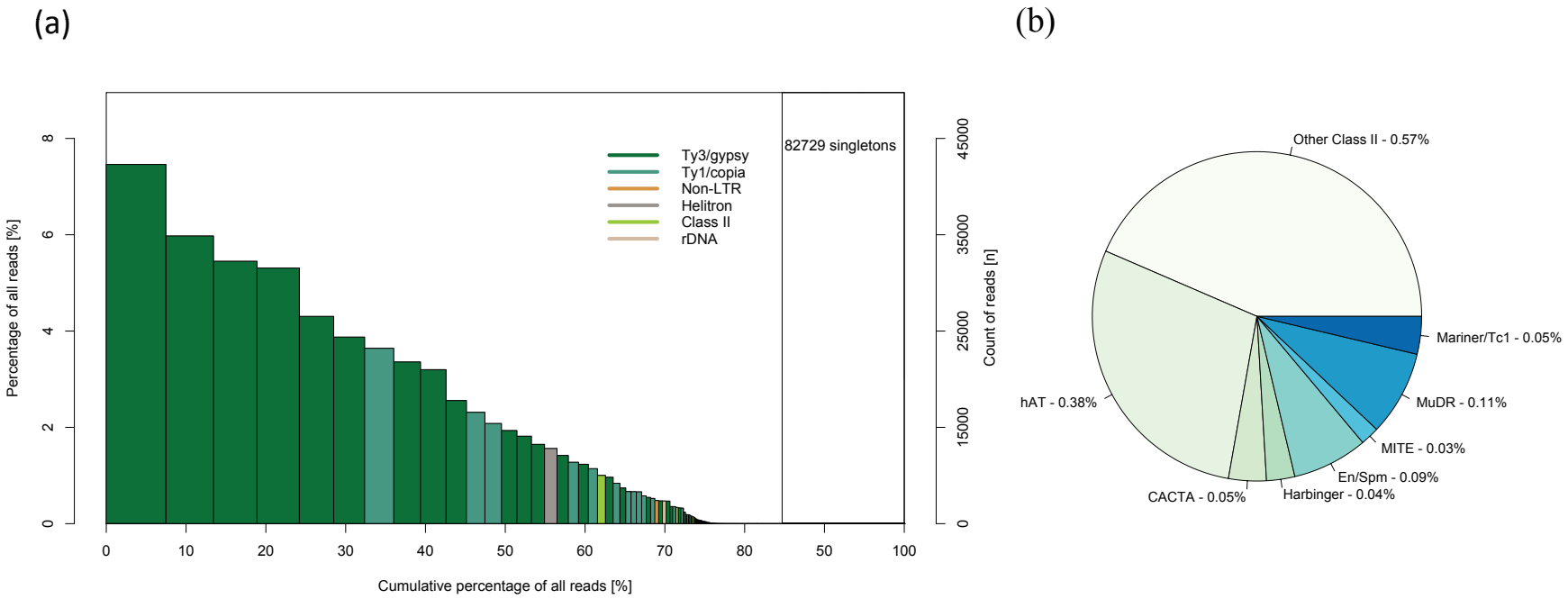


Figure 2.1. Repeat abundance based on 540,574 reads (a subset of all the WGS reads; see Experimental Procedures). a) Each bar in the histogram shows the individual size (height) of each cluster and the size relative to the total (width). The composition of each cluster is indicated by color, and single-copy, unclustered sequences are reflected to the right of the vertical bar. b) The genomic composition of Subclass I of Class II TEs. Genome-wide abundance of each superfamily based on the same subset of WGS reads, as in (a), is shown since their low abundance made them difficult to visualize in (a).

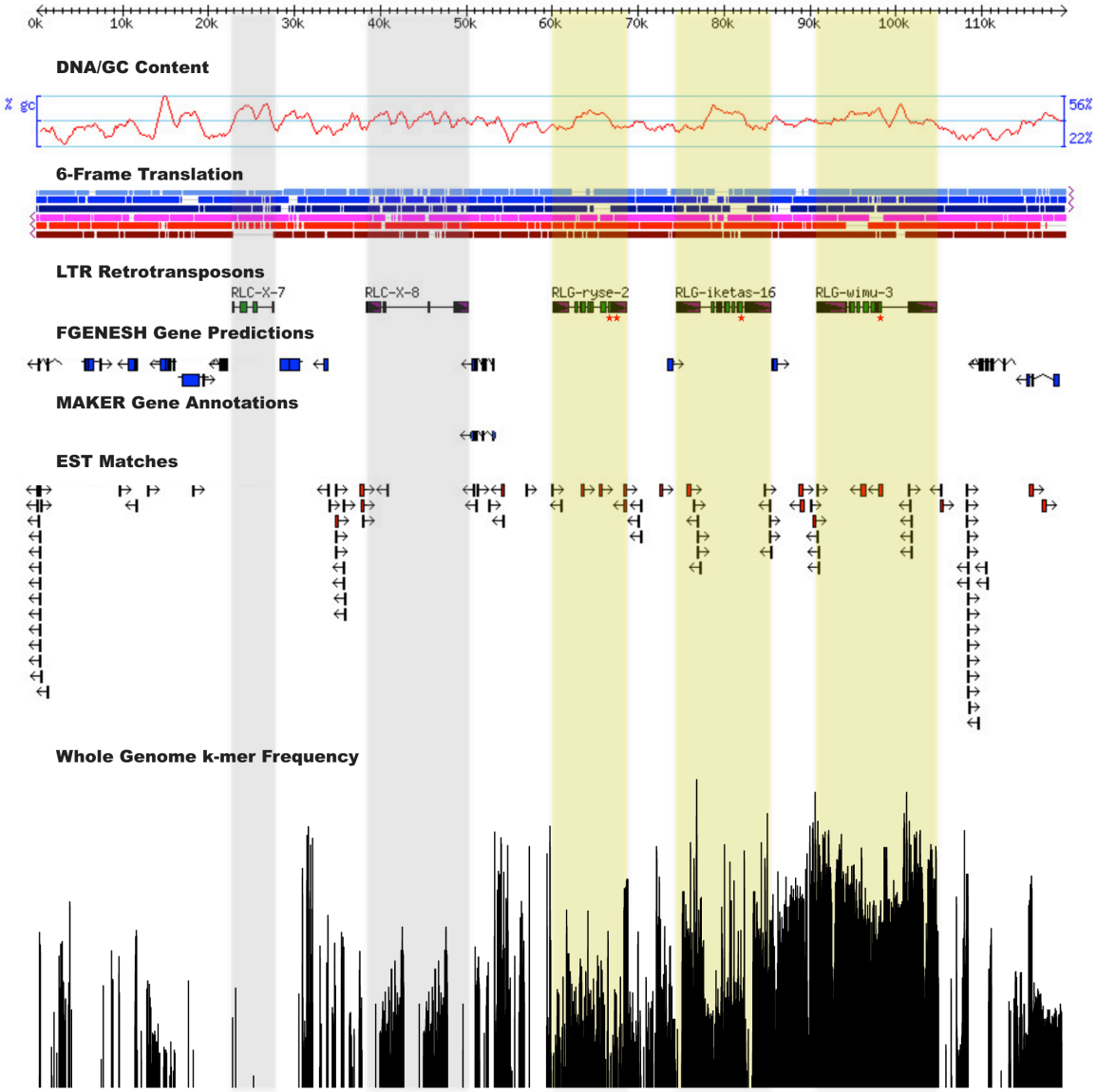

0k
10k
20k
30k
40k
50k
60k
70k
80k
90k
100k
110k
DNA/GC Content
% gc
56%
22%
6-Frame Translation
LTR Retrotransposons
RLC-X-7
RLC-X-8
RLG-ryse-2
RLG-iketas-16
RLG-wimu-3
FGENESH Gene Predictions
MAKER Gene Annotations
EST Matches
Whole Genome k-mer Frequency

Figure 2.2. Fine-scale structure of BAC clone 254L24 (see Table A2.1). The track displaying LTR retrotransposons demonstrates the characteristic lack of coding domains (green) for *Copia* elements (columns shaded in grey) as compared to the prevalence of coding domains found in *Gypsy* (columns shaded in yellow) chromovirus sequences (location of chromodomains indicated with a star). The name above each element denotes its family designation (see Table A2.2). The bias in EST matches to *Gypsy* elements and the biased genomic abundance of these sequences are shown in the tracks below the predicted genes. Gene predictions were made using FGENESH (http://www.softberry.com/) and MAKER (http://gmod.org/wiki/MAKER). The relative genome-wide frequency (plotted on a log scale) of genomic elements in this region is shown in the bottom track.

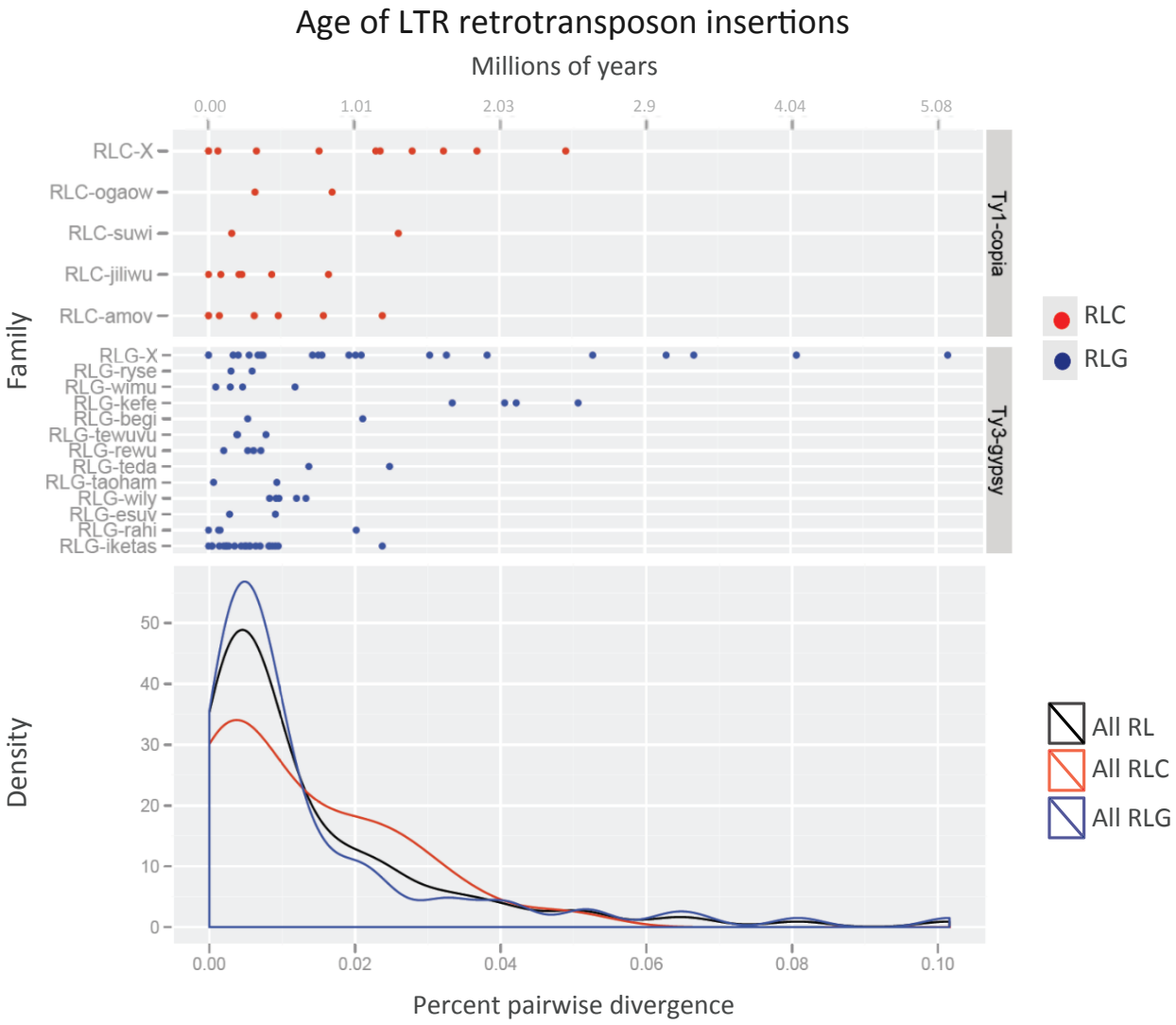


Figure 2.3. LTR retrotransposon insertion age distribution. The top panel shows the divergence between the LTRs of each individual retrotransposon insertion by family, while the bottom panel shows the same for each superfamily. The values along the lower x-axis represent the level of nucleotide divergence between the LTRs of each LTR-RT while the values along the top x-axis represent the corresponding age of each element.

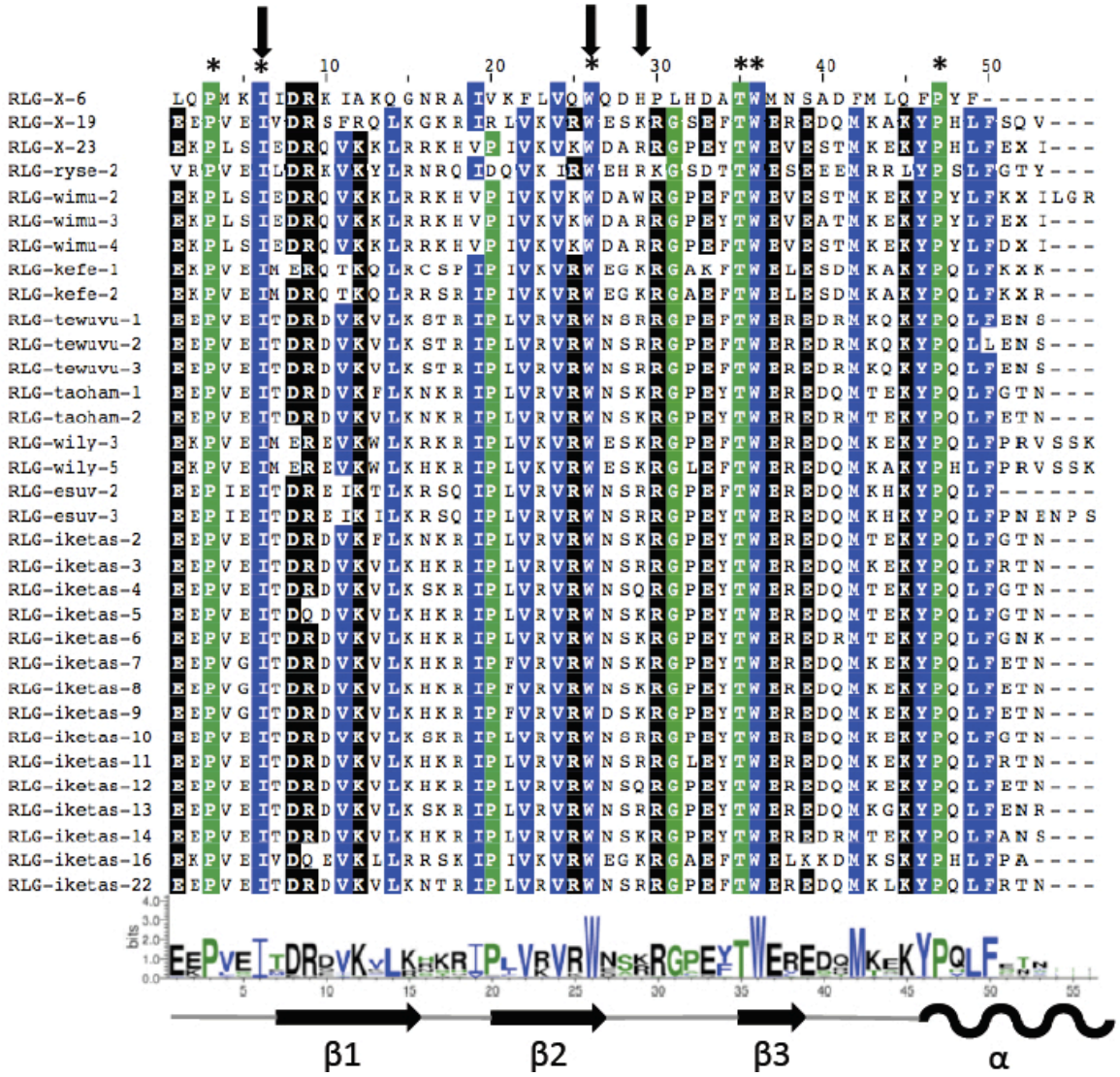


Figure 2.4. Alignment of sunflower chromodomain sequences. Only a single domain for each chromovirus was used in the alignment. Residues with conservation levels above 80% are highlighted, and the composition of each position is indicated in the sequence logo below the alignment. Aromatic residues characteristic of chromodomains are indicated with arrows (top), invariant sites are indicated with asterisks (top), and the predicted secondary structure is shown below the alignment.

CHAPTER III

NON-NEUTRAL PATTERNS OF TRANSPOSABLE ELEMENT EVOLUTION DRIVE GENOME DIVERGENCE AND TURNOVER IN THE ASTERACEAE[1]

[1]Staton S.E. and Burke J.M. To be submitted to *Proceedings of the National Academy of Sciences of the United States of America*.

SUMMARY

The transposable element (TE) content of species in the plant kingdom varies from near zero in the genome of *Utricularia gibba* (Ibarra-Laclette *et al*., 2013), to more than 80% in many species (SanMiguel *et al*., 1996; Staton *et al*., 2012; Qin *et al.*, 2014). An important question, therefore, is to understand whether these changes in genome composition represent common mechanisms or stochastic variation. The major obstacles to investigating mechanisms of TE evolution have been a lack of comparative genomic data sets and statistical methods for measuring changes in TE composition between species. Though changes in DNA sequencing technologies have made available many comparative data sets, adequate computational methods for leveraging large-scale genomics data for identifying TE families have hitherto been unavailable. In this study, we present a novel computational toolkit, Transposome, for identifying TE families from unassembled DNA sequence reads, and employ this methodology for the identification of patterns of TE evolution in 14 species in the plant family Asteraceae and 1 outgroup species in the family Calyceraceae. Our findings indicate that TE families in the Asteraceae exhibit distinct patterns of non-neutral evolution, and that there has been a directional increase in copy number of *Gypsy* retrotransposons since the origin of the Asteraceae. This biased pattern of genome evolution has had a significant impact on the diversity and abundance distribution of TEs in a lineage-specific manner.

INTRODUCTION

An enigmatic feature of eukaryotic genomes is that nearly all species contain transposable elements (TEs), yet there is a remarkable amount of variation in the composition of TEs between species (Bennetzen 2000; Bennetzen *et al*., 2005). This property of eukaryotic genomes has parallels with ecological communities (Brookfield 2005; Venner *et al*., 2009), which vary in the abundance and diversity of species. While it has been shown that niche differences are a factor in shaping species diversity (Pielou 1975; Tokeshi 1990), it is believed by some that neutral processes may explain the assembly of communities over evolutionary time scales (Hubbell 2001). Given the ubiquitous nature of TEs and their contributions to eukaryotic genome evolution (Gregory 2005; Slotkin *et al*., 2012), an important question is whether or not similar mechanisms operate to shape the genome landscape.

One theory for the variation in TEs between species is that random processes govern TE evolution and chance alone determines the properties of each TE lineage (Lynch 2007). However, there is strong evidence that TEs integrate in non-random locations in the genome and may show signs of positive selection (Gao *et al.*, 2008; Baucom *et al*., 2009a-b; Nellaker *et al*., 2012). It is important to understand the mechanistic basis behind these patterns because TE activity may have a profound impact on the evolution of their host lineages. For example, species radiations in teleost fishes and vertebrates appear to be associated with genome repatterning and TE amplification events (Volff *et al*., 2001; Ray *et al*., 2006; Bohne *et al*., 2008). In the *Taterillus* genus of gerbils there have been six species that have differentiated in the past 0.4 million years making this example likely the most recent radiation of mammalian species (Dobigny *et*

al., 2004). Species in *Taterillus* are divergent by 39 large chromosomal changes, and LINE-1 elements are distributed non-randomly between the chromosomal breakpoints with the most recently divergent species showing the greatest amount of LINE-1 accumulation (Dobigny *et al*., 2004). Also, waves of TE amplification are associated with the radiation and subsequent speciation of four genera of salmonid fishes (de Boer *et al*., 2007). Similarly, massive retrotransposon amplification appears to coincide with speciation events in hybrid sunflower species (Ungerer *et al*., 2006), and non-random patterns of retrotransposon accumulation in the hybrid species' genomes indicate a potential mechanism for chromosomal divergence between species (Staton *et al*., 2009). Taken together, these results suggest that studying the properties of TE evolution may indicate the timing and nature of important evolutionary transitions. Thus, we are keenly interested in understanding the nature of TEs in the plant family Asteraceae, which represents unparalleled species diversity in the plant kingdom.

The Asteraceae is the largest family of vascular plants, comprised of more than 23,600 species, or 8% of all plant species (Stevens 2001). The consensus view is that the Asteraceae originated in South America 40-50 million years ago, which is somewhat surprising given the large number of species in this family (Kim *et al*., 2005). From South America, the Asteraceae spread to Central America and Africa, and the family currently has a worldwide distribution, being found on every continent except Antarctica (Panero and Funk 2008). There are 12 recognized subfamilies in the Asteraceae, though four of those subfamilies, the Mutisioideae, Carduoideae, Cichorioideae, and Asteroideae, contain 99% of the species (Panero and Funk 2008). Within the Asteraceae, there is exceptional diversity in the ecological distribution of species. For example, there are

narrow endemics, and also species such as sunflower and dandelion that are found widely distributed on multiple continents. Though most species in the Asteraceae are herbaceous, there are also many shrub and tree species (Panero and Funk 2008). However, this plant family is perhaps best known for the numerous ornamental and agronomically important species such as sunflower, safflower, lettuce, and globe artichoke. Given the recent evolutionary origin of such a large plant family, combined with the global distribution of species, the Asteraceae represent an excellent system to study plant adaptation and speciation. However, very little is known about genome evolution and the diversity of TEs in the Asteraceae, aside from studies in sunflower and cytogenetic studies involving one type of TE (e.g., Santini *et al*., 2002; Natali *et al*., 2006; Cavallini *et al*., 2010; Staton *et al*., 2012). Thus, any description of genome evolution in this plant family will be incomplete without an accurate picture of the TE diversity in Asteraceae genomes.

Our primary interests in this study are to understand what the major features of genomes are throughout the Asteraceae, and to explore the mechanistic basis of TE evolution in different lineages of this family. For example, it is known that there is a major bias in genome composition towards *Gypsy* DNA in the common sunflower *Helianthus annuus* (Staton *et al*., 2012), but do other Asteraceae genomes exhibit similar patterns? More importantly, what are the mechanisms contributing to TE community structure in plants? We address this question by analyzing the relative abundance of TEs in 14 species in the Asteraceae from five separate subfamilies and one outgroup species using whole-genome shotgun sequencing data. We use phylogenetic and linear models to investigate whether there have been patterns of TE evolution specific to certain lineages in the Asteraceae. How might changes in TE abundance influence the diversity of TEs in

a genome? To address this question, we use ecological measures of community diversity, along with simulation-based approaches, to better understand the genomic impact of TE amplification events. Taken together, these approaches represent a novel approach to studying TE properties by employing both descriptive statistical approaches along with phylogenetic and ecological models to investigate the mechanisms of genome community assembly.

RESULTS

Transposable element composition in the Asteraceae

With a novel repeat finding method, we determined that Asteraceae genomes are on average 69.9 ± 5.3% TEs (Mean ± SD), with 53.19 ± 19.1% of genomes being comprised of LTR retrotransposons (LTR-RTs; Figure 3.1). As expected for plant species, Class II TEs and Non-LTR-RTs were lower in abundance relative to LTR-RTs, comprising just 0.60 ± 0.7% and 0.82 ± 1.1% of each genome, respectively. The outgroup species *Nasanthus patagonicus* exhibited comparable patterns of LTR-RT abundance (47.3 ± 3.3%) and total repeat abundance (62.0 ± 0.1%) as the Asteraceae, but contains a much higher abundance of Class II TEs (2.9 ± 0.1%) and Non-LTR-RTs (2.0 ± 0.2%). Interestingly, despite the apparent differences in Non-LTR-RT abundance between the Asteraceae and *Nasanthus patagonicus*, there is a common pattern shared by all but one species, which is that LINE-like sequences are more prevalent (by a factor of at least 2:1) than other Non-LTR-RT types. The one species that does not fit this pattern is *Fulcaldea stuessyi*, belonging to the Barnadesioideae (the most basal subfamily of the Asteraceae), which appears to possess a greater abundance of SINE-like sequences than other Non-

LTR-RT types. In addition to this exception at the base of the Asteraceae, the *Nasanthus patagonicus* genome appears to contain a far greater abundance of endogenous retroviruses (ERVs; 1.2 ± 0.4%) than Asteraceae genomes on average (0.06 ± 0.09%), though it is likely these sequences represent novel LTR-RTs since plant ERV sequences are more closely related to LTR-RTs than to the *Retroviridae* (Peterson-Burch et al., 2000). Contrasting the widespread nature of the aforementioned TE types, Penelope transposons are characterized by a sparse distribution throughout eukaryotes (Arkhipova 2006). Consistent with this finding, Penelope transposons were found in all but two species in the Asteraceae (*Fulcaldea stuessyi* and *Phoebanthus tenuifolius*), and ERV-like sequences were absent from four species (*Fulcaldea stuessyi, Conoclinium coelestinum, Phoebanthus tenuifolius,* and *Helianthus argophyllus*).

In agreement with previous studies (Cavallini *et al*., 2010; Staton *et al*., 2012), we found a large bias in TE content in the genome of *Helianthus annuus*, which is composed primarily of *Gypsy* elements (60.0 ± 3.3%). This bias appears to be shared by all members of the subfamily Asteroideae, including all species of the genus *Helianthus* analyzed here (62.4 ± 2.7%), and the most basal member of the tribe Heliantheae, *Phoebanthus tenuifolius* (67.5 ± 5.6%; Figure 3.1). We found a significant linear increase in the genomic proportion of *Gypsy* LTR-RTs from the base of the Asteraceae to the most derived subfamily, the Asteroideae using a generalized least squares test (P = <2.2e-16; r^2 = 0.9964; Figure 3.2). *Copia* TEs exhibit an inverse pattern (P = 2.831e-12; r^2 = 0.9151; Figure 3.2), to that of *Gypsy* with species at the base of Asteraceae containing proportionally more *Copia* DNA than those species in the Asteroioideae. These phylogenetic patterns remained significant when considering only one *Helianthus* species

(*Helianthus annuus*) in the analysis. In order to further test the significance of the patterns, we compared the proportion of TEs at the superfamily and family level along the phylogenetic tree to what would be expected under a Brownian motion model, and we assessed significance of these results using phylogenetic independent contrasts (PICs). We detected significant phylogenetic signal (i.e., $P < 0.05$), *K,* for ten superfamilies of TEs (Figure 3.3). Notably, *Copia* TEs as a whole show significantly more phylogenetic signal (i.e., $K \geq 1$), while *Gypsy* shows significantly less phylogenetic signal (i.e., $K =< 1$). At the individual TE family level, we found more LTR-RT families exhibiting significant phylogenetic signal (7 *Copia* families, 10 *Gypsy* families, 1 *ERV1* family) than either Non-LTR-RTs (3 *L1*-like families, 3 *CR1* families, 1, NeSL family) or Class II TEs (1 *hAT* family, 2 *Mariner/Tc1* family, 1 *Helitron* family), though the average phylogenetic signal for Class II TE families was much higher ($K = 3.26 \pm 0$) than either LTR-RTs ($K = 1.78 \pm 1.13$) or Non-LTR-RTs ($K = 3.19 \pm 0.16$) (Figure A3.1; Table A3.1).

Properties of individual TE family evolution

We investigated the mechanisms of genome community assembly over large time scales by analyzing the rank abundance/dominance (RAD) for all TE families in each species in this study. We considered five ecological models and present the best statistical model (as determined by a Bayesian Information Criterion). The predominant pattern across the Asteraceae is a lognormal-like distribution of TE family abundances, though it is evident that many species exhibit strikingly different distributions (Figure 3.4). For example, we found that *Fulcaldea stuessyi*, a member of the subfamily Barnadesioideae, has a very equal distribution of TE families in terms of abundance ($0.33 \pm 0.52\%$), while

members of the subfamily Asteroideae have a very uneven distribution, being composed of only highly abundant families and many rare families (0.92 ± 2.44%). Specifically, five species in the Heliantheae show TE family distributions best fit by a straight line (i.e., the Niche preemption model), where there is a clear pattern of dominance by a small number of TE families, while the remaining genome space is occupied by many rare TE families with low abundance (Figure 3.4). The dominance of TE families in the Heliantheae is evident when considering that the top 10 TE families in this group account for nearly 2X the genomic proportion (51.46 ± 3.14%) as the top 10 TE families in the rest of the Asteraceae (26.77 ± 9.10%).

While the RAD models described above demonstrate global patterns of abundance and dominance of TE families, these plots are unlabeled and do not allow investigation of specific changes in rank abundance. To infer what specific TE families have contributed to the rank abundance patterns observed in this study, and in the marked change in rank abundance and dominance within the Heliantheae in particular, we analyzed the rank of TE families sorted by abundance in the Asteraceae as a whole (Figure 3.5) as compared to the abundance of TE families within the Heliantheae. With the exception of the top two TE families being the same in abundance across all species (though there are two distinct groups of species sharing the same order of those families), there appears to be no phylogenetic patterns of rank abundance that are shared across the Heliantheae (Figure 3.6). At the superfamily level however, it is clear that at least the top four highest-ranking TE families in the each species in the Heliantheae are *Gypsy*.

Impact of TE family abundance on TE diversity

To investigate the potential impact of changes in TE abundance on patterns of genome community diversity, we measured the correlation of changes in TE family abundance and TE richness with genome size. As expected for plant species (Bennetzen 2000; Bennetzen 2007; Devos 2010), the abundance of retrotransposon DNA is strongly correlated with genome size (P = 6.06e-4; Figure A3.2). These patterns are also significant when considering the non-independence of the species with a phylogenetic generalized least squares test (*Copia,* P = 0.0009; *Gypsy,* P = <0.0001; Figure A3.3). However, while we did find a positive correlation with genome size and TE family size, we did not find such a correlation with genome size and TE richness (Figure 3.7). To investigate the impact of genome dominance by some TE families on genome community structure, we calculated Shannon's diversity and evenness of TE families for each species in this study (Figure A3.4). This latter analysis is far more informative of genome community patterns of evolution than looking at TE richness alone. For example, it is clear that the major shift in genome composition at the base of Heliantheae was facilitated by an unequal contribution of TE families, thereby leading to a reduction in Shannon's diversity and evenness. This result is further supported by a marked increase in the average TE family size in the Heliantheae, which is accompanied by a decrease in TE richness (Figure 3.8).

DISCUSSION

It is universally true that TEs may vary in abundance and type between eukaryotes. For example, the percent composition of TEs in the human genome is >50% (Lander et

al., 2001), though TEs are completely absent from the genomes of some unicellular eukaryotes (DeBarry and Kissinger 2011). In addition, the TE composition is 4% in the genome of *Saccharomyces cerevisiae* (Kim *et al*., 1998), which contains only LTR retrotransposons, but may be >80% in some plant genomes (e.g., SanMiguel *et al*., 1996; Schnable *et al*., 2009; Staton *et al*., 2012; Qin *et al*., 2014), which are composed of hundreds of families of both Class I and Class II TEs (Baucom *et al*., 2009a). There is also a disparity with respect to the copy number of TEs, and the number of active TE copies. For example, mammalian genomes contain high copy number TE families though only a few recently active TE families have been discovered (Furano *et al*., 2004). Conversely, there are numerous active TE families in the genomes of *Drosophila* and pufferfish, but these families only contain a few copies each (Eickbush and Furano 2002; Hua-Van *et al*., 2005). Given this variation in TE susceptibility among eukaryotes, it is important to understand the time scales and phylogenetic patterns over which different classes of TEs are active because they are important sources of variation. For example, TEs may influence macroevolutionary processes by rapidly restructuring genomes and driving gene expression divergence between species (Xie *et al*., 2010; Warenfors *et al*., 2010; Hollister *et al*., 2011). Given the contributions of TEs in these systems, our focus in this study is to gain an understanding of the nature of TEs in the plant family Asteraceae in a phylogenetic context.

Transposable elements and genome content in the Asteraceae

Species in the Asteraceae vary tremendously in the genomic composition of TEs, especially with respect to LTR-RTs. (Figure 3.1). It is not surprising that the greatest magnitude of change in genome content involves LTR-RTs given that these sequences

account for the largest portion of each genome. However, it is interesting that we see such strong linear patterns of change in genome content at the LTR-RT superfamily level from the base of the Asteraceae to the crown lineages (Figure 3.2). In the broad sense, these patterns fit Hubbell's idea of zero-sum change, which predicts that an increase in abundance in one member of a community will result in a proportional decrease in the abundance of another (Hubbell 2001). Though TE activity may lead to expansion of the nuclear genome (SanMiguel *et al*., 1996; Piegu *et al*., 2006; Ungerer *et al*., 2006), the inverse patterns of change in *Gypsy* and *Copia* abundance in the Asteraceae reflect that there are a finite number of insertion sites in the genome, and increases in copy number of one or more TE families may result in the loss or inactivation of other TE copies.

We detected significant phylogenetic signal for both Class I and Class II TEs at both the superfamily and family level (Figure 3.3; Figure A3.1; Table A3.1). These results indicate that the genomes of closely related species are more similar in the composition and abundance of certain TE types than expected by chance. When considering the variation in genome content between the basal and most derived lineages of the Asteraceae (Figure 3.1), this result is expected. However, it is clear that very different mechanisms contribute to these phylogenetic patterns. For example, the phylogenetic signal seen in Penelope retrotransposons and ERVs is likely a product of the sparse distribution of those sequences. The genomic composition of ERVs in *Nasanthus patagonicus* appears high relative to the Asteraceae, though it is not uncommon for plant species. For example, the genomic percentage of ERVs is 2.4% in the *Amborella* genome (Amborella genome project 2013); twice that of *Nasanthas patagonicus.* Alternatively, *Gypsy* elements are found in all species in the Asteraceae, but there is a clear increase in

the abundance of several *Gypsy* families at the base of Heliantheae creating a phylogenetic pattern shared by all members of this tribe. The inverse pattern can be seen for the *Copia* superfamily, which also shows significant phylogenetic signal (Figure 3.3), where a linear decrease in these sequences from the Barnadesioideae to the Asteroideae likely contributes to phylogenetic patterns across the family. The foregoing results indicate that no single mechanism can explain these patterns of genome evolution in the Asteraceae. Specifically, species in the basal subfamilies of the Asteraceae are strikingly different in the composition of TEs compared with the crown subfamilies, with those species in the basal subfamilies containing a greater abundance of Non-LTR retrotransposons and DNA transposons. Could the greater TE diversity at the base of the Asteraceae and in the outgroup species *Nasanthus patagonicus* be a result of the age of those lineages, or could there be other mechanisms influencing the abundance and diversity of the genome community?

Transposable element families and genome community assembly

Although ecosystems vary in the abundance and diversity of species, a common pattern is that communities exhibit a very similar distribution in the rank abundance of species (Hubbell 2001). Specifically, most communities exhibit a log-normal distribution of species abundance with few species having large abundance, many rare species with very low abundance, and numerous species having abundances in between these extremes (Hubbell 2001). Interestingly, one study has shown that eukaryotic genomes appear to exhibit similar log-normal distributions of genetic elements, suggesting that neutral processes may best explain community assembly over evolutionary timescales, regardless of the system (Serra *et al*., 2013). Do Asteraceae genomes also exhibit a log-normal

distribution of TE family abundances, and are these patterns shared across the family? All species in this study, including the outgroup species *Nasanthus patagonicus*, from the basal lineages of the Asteraceae to the base of the tribe Heliantheae exhibit similar log-normal-like distributions of TE family abundance (Figure 3.4). However, there is a very marked break at the base of Heliantheae with all species in this tribe exhibiting numerous highly abundant TE families and many rare families. This type of distribution has been used to describe communities with poor habitat (Keeley and Fotheringham 2003) or early succession of species (Whittaker 1972) following disturbance (Nummelin 1998). What biological change facilitated the major genomic transitions in the Heliantheae? It is tempting to speculate that a whole genome duplication event at the base of the Heliantheae (Barker *et al*., 2008), or directed integration of *Gypsy* elements may have contributed to these patterns (Peterson-Burch *et al*., 2004; Gao *et al*., 2008; Staton *et al*., 2012), but clearly more work will be required to gain a deeper understanding of the underlying processes.

Mechanisms of change in the genome-wide level of transposable elements

Major transitions in genome content are evident in each subfamily of the Asteraceae (Figure 3.1). What mechanism best explains the patterns of TE abundance in the Asteraceae? The coexistence of species may be facilitated by niche differentiation (Hutchinson 1959), and this type of model best explains the TE abundance data we see for species in the tribe Heliantheae. However, the TE abundance and diversity for this group of species indicates a very biased composition of *Gypsy* TEs (Figure 3.6). What has been the impact of linear increases in *Gypsy* TEs on the global diversity of TEs? The biased accumulation of *Gypsy* TEs in the Heliantheae has had at least two major

influences on the genome community of TEs. First, the correlation we see with TE family size and genome size (Figure 3.7) indicates an unequal contribution of TE families to the genome community. Second, it is clear that the linear pattern of increase in *Gypsy* is driven by only a few TE families (Figure 3.3), which has lead to an increase in the average size of a TE family and a decrease in overall TE richness (Figure 3.8). Interestingly, we don't see different TE families dominating each *Helianthus* genome as with species of *Gossypium* (Hawkins *et al*., 2006), and the top two TE families are shared by all *Helianthus* species. This may indicate that a single event at the base of Heliantheae could have lead to genomic change, and the patterns we see in each *Helianthus* species are shared by phylogenetic history rather than being independent events leading to similar patterns in each species. Alternatively, *Gypsy* elements may have evolved features allowing them to outcompete other TEs or avoid host-silencing mechanisms. Future investigations into these questions will surely lead to a greater understanding of the processes contributing to the success of the Asteraceae, and to processes contributing to the evolution of TE diversity in the plant kingdom.

EXPERIMENTAL PROCEDURES

Taxon sampling and WGS sequencing

In order to investigate patterns of genome evolution across the Asteraceae we generated Illumina paired-end sequence data (100 bp in length; 400 bp insert size) from fifteen taxa. The estimated genome coverage for each species ranged from 0.42 – 3.52 (Table A3.2). The species were sampled from every major subfamily of the Asteraceae, including an outgroup species, *Nasanthus patagonicus* (Table 3.1). In addition, five of the

taxa were selected from the genus *Helianthus* in order to investigate patterns of genome evolution among closely related species, and to increase our understanding of the evolutionary history of the most well-studied species in the family, *Helianthus annuus*, for which there have been numerous studies about TE properties. Taxon sampling and library preparation methods are described in Mandel et al. (2014).

Repeat identification from WGS sequences

Prior to analysis, all WGS reads were treated with PRINSEQ (version 0.19.4; Schmieder and Edwards, 2011) with the parameters '-min_len 40 –noniupac –min_qual_mean 15 –lc_method entropy –lc_threshold 60 –trim_ns_right 10 –ns_max_p 20' to remove low quality and short sequences. After quality filtering, we screened all chloroplast and mitochondria derived sequences from the WGS reads using the complete chloroplast genome sequence for *Helianthus annuus* cultivar line HA383 (Genbank accession number DQ383815) and a database of 10 complete plant mitochondria genome sequences obtained from Genbank, respectively. One million paired-end reads were sampled randomly from each set of screened reads and interleaved with Pairfq (version 0.09; https://github.com/sestaton/Pairfq) prior to analysis. Repeat identification was carried out by performing an all by all BLAST according to Staton et al. (2012) with the one million randomly sampled paired-end reads, followed by clustering using the Louvain method (Blondel *et al*., 2008). Annotation of clusters was performed using blastn (Camacho *et al*., 2009) against RepBase (version 18.01; Jurka *et al*., 2005) and a set of full-length LTR-RTs described by Staton et al. (2012). Our repeat identification methods are implemented using the Transposome software (version 0.03; Staton and Burke 2014) we developed for this study. We performed three replicates of the above

sampling and annotation procedure with Transposome for each species to minimize the statistical error in our estimates of genome composition.

In order to investigate the effect of varying levels of genome coverage, we simulated 10 different levels of genome coverage from the *Helianthus annuus* WGS reads ranging from 0.056 – 5.1%, with 3 replicates at each level (total of 30 read sets). The coefficient of variation in the inferred genomic composition of each TE family was measured at each level of genome coverage after analysis with Transposome to infer the appropriate level of sampling; this allowed us to maximize the level of TE diversity being captured.

Genome size estimation and prediction of changes in genome composition

In order to determine the genomic contribution of each TE family to the species in this study, and estimate the magnitude of change across the Asteraceae, we calculated genome size according to Hu et al. (2011), with modifications. Using WU-BLAST with parameters "M=1 N=-3 -Q -R 1" we mapped a reference transcriptome of 11 species from the Compositae Genome Project database (http://compgenomics.ucdavis.edu/) to 5 million WGS reads for each species, and calculated the coverage of each transcript using the formula:

$$Cov_i = \frac{N}{L}$$

where *N* is the total length of reads mapped and *L* is the transcript length. The genome size (*Cval*) for each species was then determined by the formula:

$$Cval = P \times \left(\frac{n \times l}{mean(Cov_i)}\right)$$

where *P* is the ploidy level, *n* is the total number of reads, and *l* is the read length. In the

above formula, only alignments over 60 base pairs in length and over 70 percent identity were considered. These values were chosen from a permutation test using all possible alignments from lengths 50-100 and percent identity thresholds from 50-100, comparing observed to expected values. The mean coverage (Cov_i) was trimmed to remove the top 10 % of transcripts by coverage. The estimated genome size for each species, along with the published prediction (if available), is a shown in (Table A2; Figure A5).

The genomic contribution of each TE superfamily was calculated from the annotation summary file generated by Transposome (Figure 3.1), and was used to determine the magnitude of change in TE composition in each species. Generalized least squares tests were performed to estimate directional change in TE content in the Asteraceae (Figure 3.2). We calculated Shannon's evenness and diversity statistics using the R package Vegan (Oksanen *et al*., 2013) to investigate the influence of genome size change on TE diversity statistics.

Phylogenetic patterns of TE family evolution

In addition to analyzing statistical patterns of repeat abundance, we also explored a mechanistic basis for TE evolution in the Asteraceae through the use of community ecology models. First, rank abundance analysis was performed using the R package Vegan (Oksanen *et al*., 2013) to test our null hypothesis of neutral evolution of TE families in the Asteraceae (Serra *et al*., 2013). Second, a phylogenetic generalized least squares (pgls) test was conducted using caper (Orme *et al*., 2012) to test for the association of changes in TE composition with particular phylogenetic divisions within the Asteraceae and genome size. The phylogenetic tree used in the pgls analyses was generated from an alignment of 763 nuclear loci sequenced by a novel targeted

enrichment method (Mandel *et al*., 2014). The model we tested was:

Log(Genome size) ~ Log(S)*

where *S** is the superfamily percent genomic abundance.

To further investigate the mechanisms and timing of shifts in genome content, we calculated phylogenetic signal for each TE family by using a descriptive statistic called *K* along with phylogenetic independent contrasts (PICs; Blomberg *et al*., 2003; Felsenstein 1985). These calculations were performed using the R package picante (Kembel *et al*., 2010), and all statistical analyses and plotting were performed in R (R Core Team, 2013).

ACKNOWLEDGEMENTS

We thank Jennifer Mandel (Memphis University) for assistance with DNA isolation and sequencing; Vicki Funk (Smithsonian Institute) for sharing plant specimens; Francois Serra (Centre Nacional d'Anàlisi Genòmica) for valuable assistance with Ecolopy for performing neutral simulations.

LITERATURE CITED

Akhipova I. 2006. Distribution and phylogeny of Penelope-like in Eukaryotes. *Systematic Biology* 55(6), 875-878.

Amborella genome project. 2013. The *Amborella* genome and the evolution of flowering plants. *Science* 342, doi: 10.1126/science.1241089.

Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore RW, Knapp SJ, and Rieseberg LH. 2008. Multiple paleopolypoidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Mol. Bio. Evol.* 25(11), 2445-2455.

Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, Deragon JM, Westerman RP, SanMiguel PJ, Bennetzen JL. 2009a. Exceptional diversity, non-random distribution, and rapid evolution or retroelements in the B73 maize genome. *PloS Genetics* 5(11), 1-13.

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

Bennetzen J.L. 2000. Transposable element contributions to plant gene and genome evolution. *Plant Mol. Biol.* 42, 251-269.

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

Bennetzen JL, Ma J, and Devos KM. 2005. Mechanisms of recent genome size variation in flowering plants. *Plant Mol. Biol.* 95, 127-132.

DeBarry J, and Kissinger JC. 2011. Jumbled genomes: missing Apicomplexan synteny. *Mol. Biol. Evol.* 28(10), 2855-2871.

Blomberg SP, Garland Jr. T, and Ives AR. 2003. Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. *Evolution* 57(4), 717-745.

Blondel VD, Guillaume J, Lambiotte R, Lefebvre E. 2008. Fast unfolding of communities in large networks. *J. Stat. Mech.* P10008.

Dobigny G, Ozouf-Costaz C, Waters PD, Bonillo C, Coutanceau JP, and Volobouev V. 2004. LINE-1 amplification accompanies explosive genome repatterning in rodents. *Chromosome Res.* 12(8), 787-793.

Bohne A, Brunet F, Galiana-Arnoux D, Schultheis C, and Volff J. 2008. Transposable elements as drivers of genomic and biological diversity in vertebrates. *Chromosome Res.* 16, 203-215.

Brookfield JYF. 2005. The ecology of the genome – mobile DNA elements and their hosts. *Nature Rev.* 6, 128-136.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, and Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics,* 10, 421.

Cavallini A, Natali L, Zuccolo A, Giordani T, Jurman I, Ferrillo V. et al. 2010. Analysis of transposons and repeat composition of the sunflower (Helianthus annuus L.) genome. *Theor. Appl. Genet.* 120, 491–508.

Devos KM. 2010. Grass genome organization and evolution. *Curr. Opin. Plant Biol.* 13(2), 139-145.

de Boer JG, Yazawa R, Davidson WS, and Koop BF. 2007. Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. *BMC Genomics* 8, 422.

Eickbush TH, and Furano AV. 2002. Fruit flies and humans respond differently to retrotransposons. *Curr. Opin. Gen. Dev.* 12, 669-674.

Felsenstein J. 1985. Phylogenies and the comparative method. *American Naturalist.* 125:1-15.

Furano AV, Duvernell DD, and Boissinot S. 2004. L1 (LINE-1) diversity differs dramatically between mammals and fish. *Trends Gen.* 20, 9-14.

Gao X, Hou Y, Ebina H, Levin HL, and Voytas DF. 2008. Chromodomains direct integration of retrotransposons to heterochromatin. *Genome Res.* 18, 359-369.

Gregory TR. 2005. Evolution of the genome. Elsevier, Inc.

Hawkins JS, Kim H, Nason JD, Wing RA, and Wendel JF. 2006. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. *Genome Res.* 16(10), 1252-1261.

Hollister JD, Smith LM, Guo Y, Ott F, Weigel D, and 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.

Hu H, Bandyopadhyay PK, Olivera BM, and Yandell M. 2011. Characterization of the *Conus bullatus* genome and its venom-duct transcriptome. *BMC Genomics,* 12, 60.

Hua-Van A, Le Rouzic A, Maisonhaute C, and Capy P. 2005. Abundance, distribution and dynamics of retrotransposable elements: similarities and differences. *Cytogen. Genome Res.* 110, 426-440.

Hubbell SP. 2001. The Unified Neutral Theory of Biodiversity and Biogeography. Princeton University Press.

Hutchinson GE. 1959. Homage to Santa Rosalia, or why are there so many kinds of animals? *American Naturalist* 93, 145-159.

Ibarra-Laclette E, Lyons E, Hernández-Guzmán G, Pérez-Torres CA, Carretero-Paulet L, Chang T, Lan T, Welch AJ, Juárez MJA, Simpson J, et al. 2014. Architecture and evolution of a miniature plant genome. *Nature* 498, 94-98.

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

Keeley JE, Fotheringham CJ. 2003. Species–area relationships in Mediterranean climate plant communities. *Journal Biogeogr.* 30, 1629–1657.

Kim JM, Vanguri S, Boeke JD, Gabriel A, and Voytas DF. 1998. Transposable elements and genome organization: A comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. *Genome Res.* 8, 464-478.

Kim KJ, Choi KS, and Jansen RK. 2005. Two chloroplast DNA inversions originated simultaneously during the early evolution of the sunflower family (Asteraceae). *Mol. Biol. Evol.* 22, 1783–1792.

Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, and Webb SP. 2010. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* 26, 1463-1464.

Lander E, 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* 15(409), 860-921.

Lynch M. 2007. The origins of genome architecture. Sinauer Associates, Inc.

Mandel J, Dikow RB, Funk VA, Masalia R, Staton SE, Kozik A, Michelmore RW, Rieseberg LH, and Burke JM. 2014. A target enrichment method for gathering phylogenetic information from hundreds of loci: An example from the Compositae. *App. Plant Sci.* 2, 1300085. doi:10.3732/apps.1300085.

Natali L, Santini S, Giordani T, Minelli S, Maestrini P, Cionini PG, and Cavallini A. 2006. Distribution of Ty3-*Gypsy*- and Ty1-*Copia*-like DNA sequences in the genus *Helianthus* and other Asteraceae**.** *Genome* 49, 64-72.

Nellåker C, Keane TM, Yalcin B, Wong K, Agam A, Belgard TG, Flint J, Adams DJ, Frankel WN, and Ponting CP. 2012. The genomic landscape shaped by selection on transposable elements across 18 mouse strains. *Genome Biol.* 13, R45.

Nummelin M, 1998. Log-normal distribution of species abundances is not a universal indicator of rain forest disturbance. *Journal Appl. Ecol.* 35, 454–457.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, and Wagner H. 2013. vegan: Community Ecology Package. R package version 2.0-7. http://CRAN.R-project.org/package=vegan.

Orme D, Freckleton R, Thomas G, Petzoldt T, Fritz S, Isaac L, and Pearse W. 2012. caper: Comparative Analyses of Phylogenetics and Evolution in R. R package version 0.5. http://CRAN.R-project.org/package=caper.

Panero JL, and Funk VA. 2008. The value of sampling anomalous taxa in phylogenetic studies: major clades of the Asteraceae revealed. *Mol. Phylo. Evol.* 47, 757-782.

Peterson-Burch BD, Wright DA, Laten HM, and Voytas DF. 2000. Retroviruses in plants? *Trends in Gen.* 16, 151-152.

Peterson-Burch BD, Nettleton D, Voytas DF. 2004. Genomic neighborhoods for Arbidopsis retrotransposons: a role for targeted integration in the distribution of the Metaviridae. *Genome Biol.* 5, R78.

Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, Collura K, et al. 2006. Doubling genome size without polyploidization: dynamics of retrotransposon-mediated genome expansions in Oryza australensis, a wild relative of rice. *Genome Res.* 16: 1262-1269.

Pielou, E.C. 1975. Ecological Diversity. Wiley-Interscience, New York.

Qin C, Yu C, Shen Y, Fang X, Chen L, Min J, Cheng J, Zhao S, Xu M, Luo Y, et al. 2014. Whole-genome sequencing of cultivated and wild peppers provides insights into Capsicum domestication and specialization. *Proc. Natl. Acad. Sci. U.S.A.* doi: 10.1073/pnas.1400975111.

R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.

Ray DA, Xing J, Salem AH, and Batzer MA. 2006. SINEs of a nearly perfect character. *Syst. Biol.* 55, 928Y935.

SanMiguel P, Tikhonov A, Jin Y, Motchoulskaia N, Zakharov D, Melake-Berhan A, et al. 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science* 274, 765-768.

Santini S, Cavallini A, Natali L, Minelli S, Maggini F, Cionini PG. 2002. Ty1/*Copia*- and Ty3/*Gypsy*-like DNA sequences in *Helianthus* species. *Chromosoma* 111, 192-200.

Schmieder R and Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27:863-864.

Schnable P, Ware D, Fulton RS, Stein JC, Wei F, Pastemak S, Liang C, et al., 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* 326, 1112-1115.

Serra F, Becher V, and Dopazo H. 2013. Neutral theory predicts the relative abundance and diversity of genetic elements in a broad array of eukaryotic genomes. *PLOS One* 8, 6.

Slotkin RK, Nuthikattu S, Jiang N. 2012. The impact of transposable elements on gene and genome evolution. *Plant genome diversity* Vol. 1. Springer-Verlag Wien.

Staton SE, Ungerer MC, Moore RC. 2009. The genomic organization of Ty3/gypsy-like retrotransposons in Helianthus (Asteraceae) homoploid hybrid species. *Amer. J. Bot.* 96(9), 1646-1655.

Staton SE, Hartman Bakken B, Blackman B, Chapman M, Kane N, Tang S, Ungerer M, Knapp S, Rieseberg L, Burke J. 2012. The sunflower (Helianthus annuus L.) genome reflects a recent history of biased accumulation of transposable elements. *The Plant Jour.* 72, 142-153.

Staton SE and Burke JM. 2014. Transposome: Annotation of transposable element families from unassembled sequence reads. http://sestaton.github.io/Transposome.

Stevens PF, 2001 onwards. Angiosperm Phylogeny Website. Version 8, June 2007. Available online at <http://www.mobot.org/MOBOT/research/APweb/>.

Tokeshi M. 1990. Niche apportionment or random assortment – species abundance patters explained. *Jour. Animal Ecol.* 59(3), 1129-1146.

Ungerer MC, Strakosh SC, and Zhen Y. 2006. Genome expansion in three hybrid sunflower species is associated with retrotransposon proliferation. *Curr. Biol.* 16, R872-R873.

Venner S, Feschotte C, Biemont C. 2009. Dynamcis of transposable elements: towards a community ecology of the genome. *Trends Gen.* 739, 1-7.

Volff JN, Korting C, Meyer A, Schartl M. 2001. Evolution and discontinuous distribution of Rex3 retrotransposons in fish. *Mol. Biol. Evol*. 18, 427–431.

Warenfors M, Pereira V, and Eyre-Walker A. 2010. Transposable elements: Insertion pattern and impact on gene expression evolution in Hominids. Molecular Biology and Evolution 27, 1955-1962.

Whittaker RH. 1972. Evolution and measurement of species diversity. *Taxon* 21, 213–251.

Xie D, Chen C, Ptaszek LM, Xiao S, Cao X, Fang F, Ng HH, Lewin HA, Cowan C, and Zhong S. 2010. Rewirable gene regulatory networks in the preimplantation embryonic development of three mammalian species. Genome Research 20, 804-815.

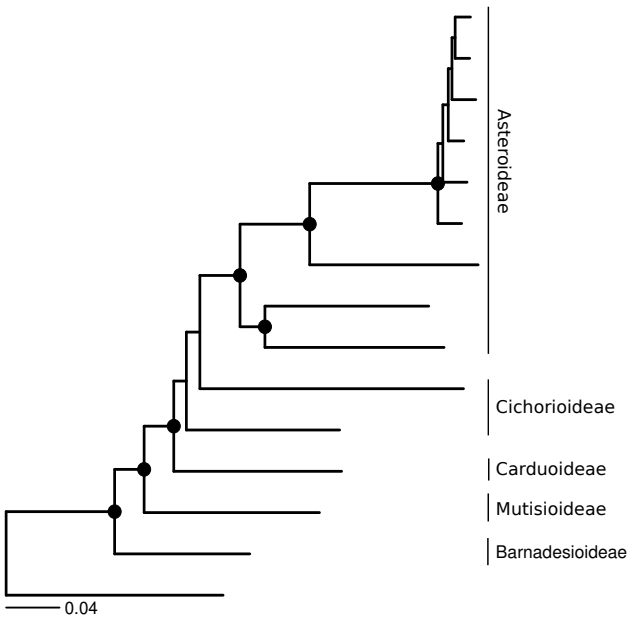
TABLES FOR CHAPTER III

Table 3.1. The percent genomic abundance of repeat types in the Asteraceae. Shown for each species are the mean abundance of each repeat type along with the standard deviation taken from three replicate analyses (see Experimental Procedures).

Subfamily	Tribe	Genus	Species	LTR-RTs	Non-LTR_RTs	Class II	ERV	Total repeat %
Calyceraceae	Calyceraceae	*Nastanthus*	*patagonicus*	47.33 ± 3.32	2.03 ± 0.20	2.93 ± 0.20	1.22 ± 0.44	62.02 ± 0.11
Barnadesioideae	Barnadesieae	*Fulcaldea*	*stuessyi*	33.88 ± 0.54	0.31 ± 0.11	0.69 ± 0.05	--	59.55 ± 0.08
Mutisioideae	Mutisieae	*Gerbera*	*hybrida*	44.60 ± 1.45	0.53 ± 0.02	0.70 ± 0.02	0.02 ± 0.01	71.54 ± 0.08
Carduoideae	Cardueae	*Carthamus*	*tinctorius*	25.24 ± 0.99	2.46 ± 0.52	2.09 ± 0.14	0.02 ± 0.01	64.01 ± 0.10
Cichorioideae	Cichorieae	*Taraxacum*	*kok-saghyz*	50.66 ± 2.14	3.21 ± 0.29	1.28 ± 0.03	0.06 ± 0.02	68.65 ± 0.03
Cichorioideae	Vernonieae	*Centrapalus*	*pauciflorus*	26.31 ± 0.78	0.50 ± 0.03	1.62 ± 0.07	0.21 ± 0.08	72.10 ± 0.05
Cichorioideae	Senecioneae	*Senecio*	*vulgaris*	27.90 ± 0.72	2.49 ± 0.53	0.51 ± 0.04	0.24 ± 0.04	66.72 ± 0.04
Cichorioideae	Gnaphalieae	*Pseudognaphalium*	*obtusifolium*	52.24 ± 1.36	0.81 ± 0.04	0.45 ± 0.02	0.13 ± 0.05	70.43 ± 0.04
Asteroideae	Eupatorieae	*Conoclinium*	*coelestinum*	44.36 ± 1.63	1.15 ± 0.14	1.09 ± 0.03	--	60.84 ± 0.07
Asteroideae	Heliantheae	*Phoebanthus*	*tenuifolius*	76.47 ± 4.28	2.95e-5 ± 9.81e-4	3.93e-3 ± 1.13e-3	--	71.17 ± 0.03
Asteroideae	Heliantheae	*Helianthus*	*porteri*	72.90 ± 3.62	0.02 ± 2.72e-3	0.01 ± 5.23e-4	7.70e-4 ± 4.44e-4	75.12 ± 0.64
Asteroideae	Heliantheae	*Helianthus*	*verticillatus*	72.25 ± 3.44	9.94e-3 ± 4.17e-4	0.01 ± 8.88e-4	2.44e-3 ± 3.62e-4	73.33 ± 1.20
Asteroideae	Heliantheae	*Helianthus*	*niveus ssp. tephrodes*	72.70 ± 2.63	0.01 ± 1.82e-3	6.03e-3 ± 3.36e-4	3.75e-4 ± 2.17e-4	73.51 ± 1.21
Asteroideae	Heliantheae	*Helianthus*	*argophyllus*	73.38 ± 2.93	7.03e-3 ± 6.23e-4	7.73e-3 ± 6.83e-4	--	74.14 ± 0.74
Asteroideae	Heliantheae	*Helianthus*	*annuus*	71.81 ± 2.80	6.60e-3 ± 1.43e-3	5.61e-3 ± 1.70e-4	3.29e-4 ± 1.90e-4	72.03 ± 0.01
				52.80 ± 19.15	0.90 ± 1.10	0.76 ± 0.89	0.13 ± 0.31	69.41 ± 5.54

FIGURES FOR CHAPTER III

A)



B)

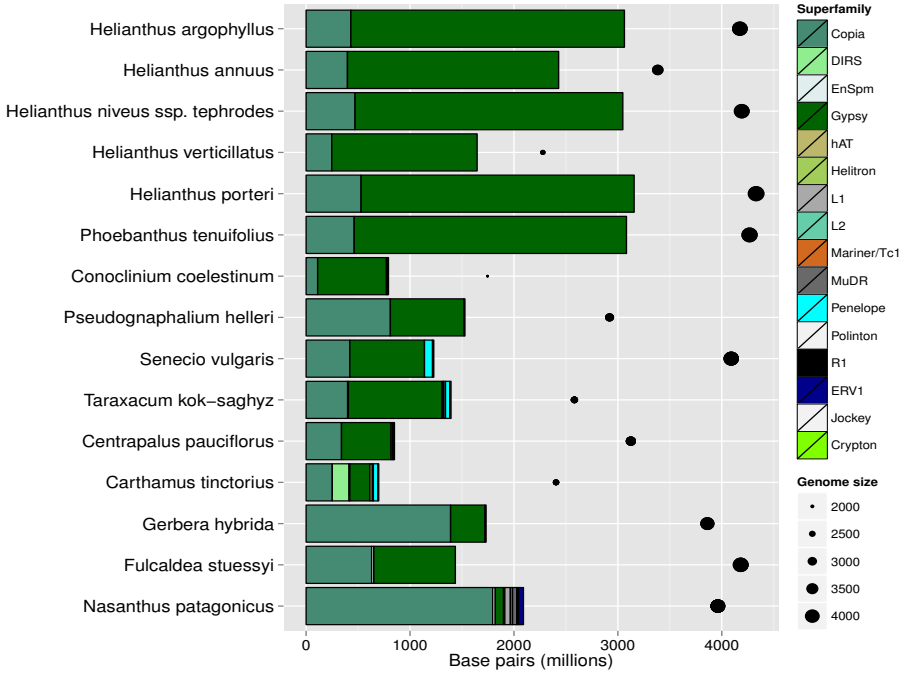
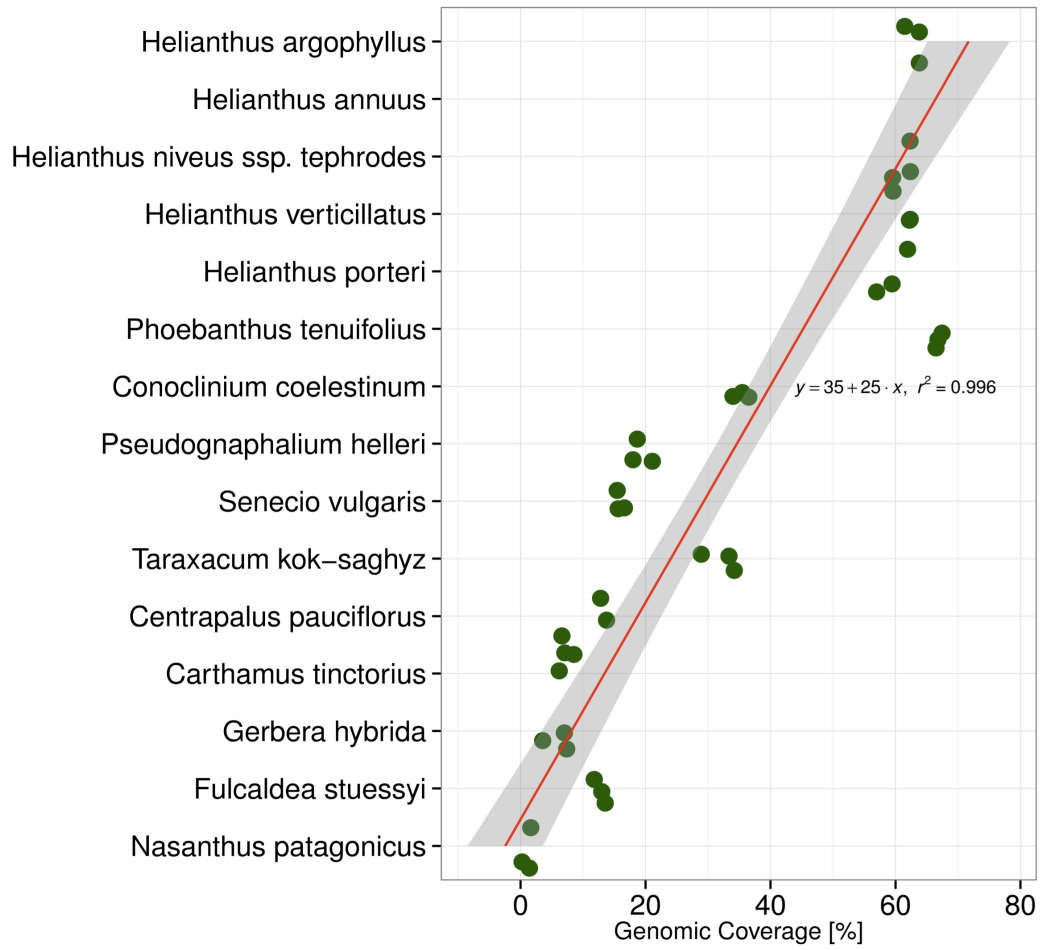


Figure 3.1. Genomic contribution of TE superfamilies in the Asteraceae. A) Phylogenetic tree of 14 Asteraceae species and one outgroup species derived from 763 nuclear loci (see Experimental Procedures). Filled circles indicate nodes with >75% bootstrap support; to the right of the tree are the subfamilies to which each species belongs. B) Barplot of the genomic composition of TE superfamilies. The x-axis indicates abundance in base pairs for each species, shown along the y-axis. Filled circles indicate the genome size for each species. Superfamilies by order and class: Copia, Gypsy, ERV, and DIRS are LTR-RTs; Helitron is in subclass II of Class II; EnSpm, MuDR, hAT, Mariner/Tc1, and Polinton are TIR Class II TEs; Crypton are unique Class II elements in the order Crypton; L1, L2, and Jockey are LINE Non-LTR-RTs; Penelope TEs belong in the unique Penelope order of retrotransposons; R1 are a group of Non-LTR-RTs that insert into rDNA genes.

A)


Helianthus argophyllus
Helianthus annuus
Helianthus niveus ssp. tephrodes
Helianthus verticillatus
Helianthus porteri
Phoebanthus tenuifolius
Conoclinium coelestinum
Pseudognaphalium helleri
Senecio vulgaris
Taraxacum kok–saghyz
Centrapalus pauciflorus
Carthamus tinctorius
Gerbera hybrida
Fulcaldea stuessyi
Nasanthus patagonicus
$y = 35 + 25 \cdot x,\ r^2 = 0.996$
0
20
40
60
80
Genomic Coverage [%]

B)

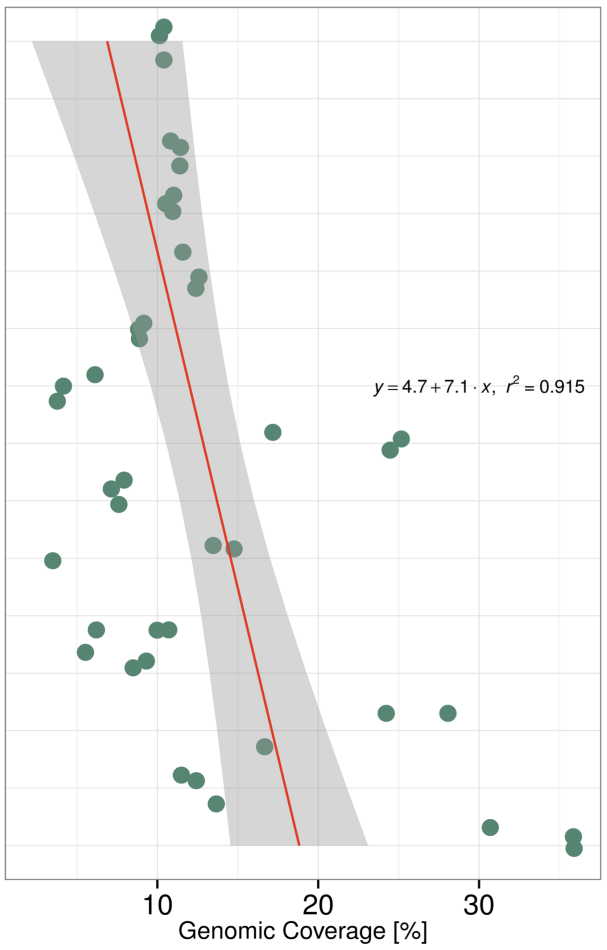

$y = 4.7 + 7.1 \cdot x,\ r^2 = 0.915$
10
20
30
Genomic Coverage [%]

Figure 3.2. Linear change in genomic composition of LTR-RTs. Shown in phylogenetic order starting with the outgroup (bottom of the y-axis) to the most derived lineages of the Asteraceae in this study (top of the y-axis) are the change in genomic proportion (shown along the x-axis) of A) *Gypsy* and B) *Copia* TEs.

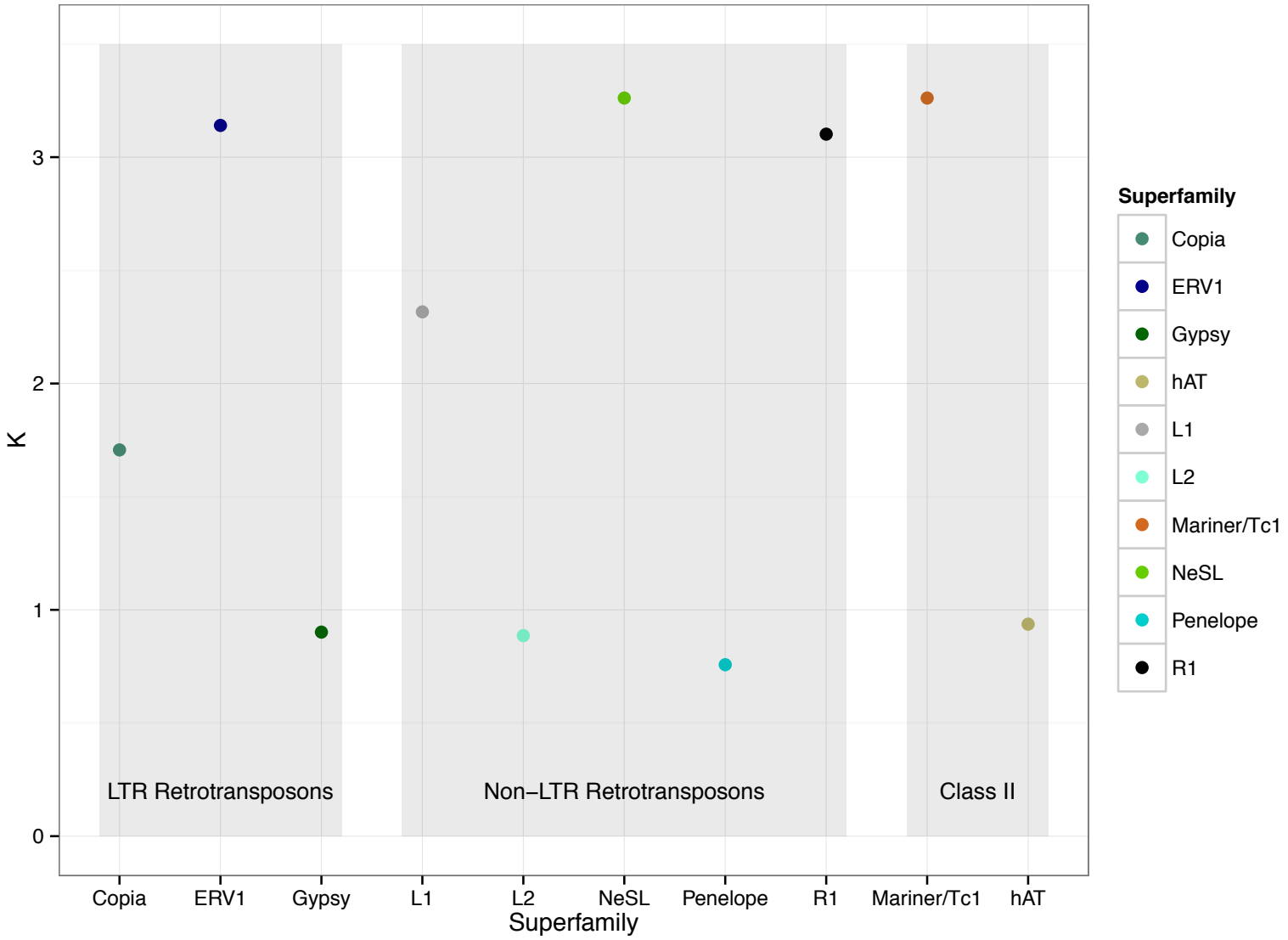

K
0
1
2
3
LTR Retrotransposons
Non−LTR Retrotransposons
Class II
Copia
ERV1
Gypsy
L1
L2
NeSL
Penelope
R1
Mariner/Tc1
hAT
Superfamily
Superfamily
Copia
ERV1
Gypsy
hAT
L1
L2
Mariner/Tc1
NeSL
Penelope
R1

Figure 3.3. TE superfamilies showing significant phylogenetic signal (K). Each TE superfamily (shown along the x-axis) is grouped by order (gray boxes). The y-axis shows the absolute level of signal displayed for each superfamily.

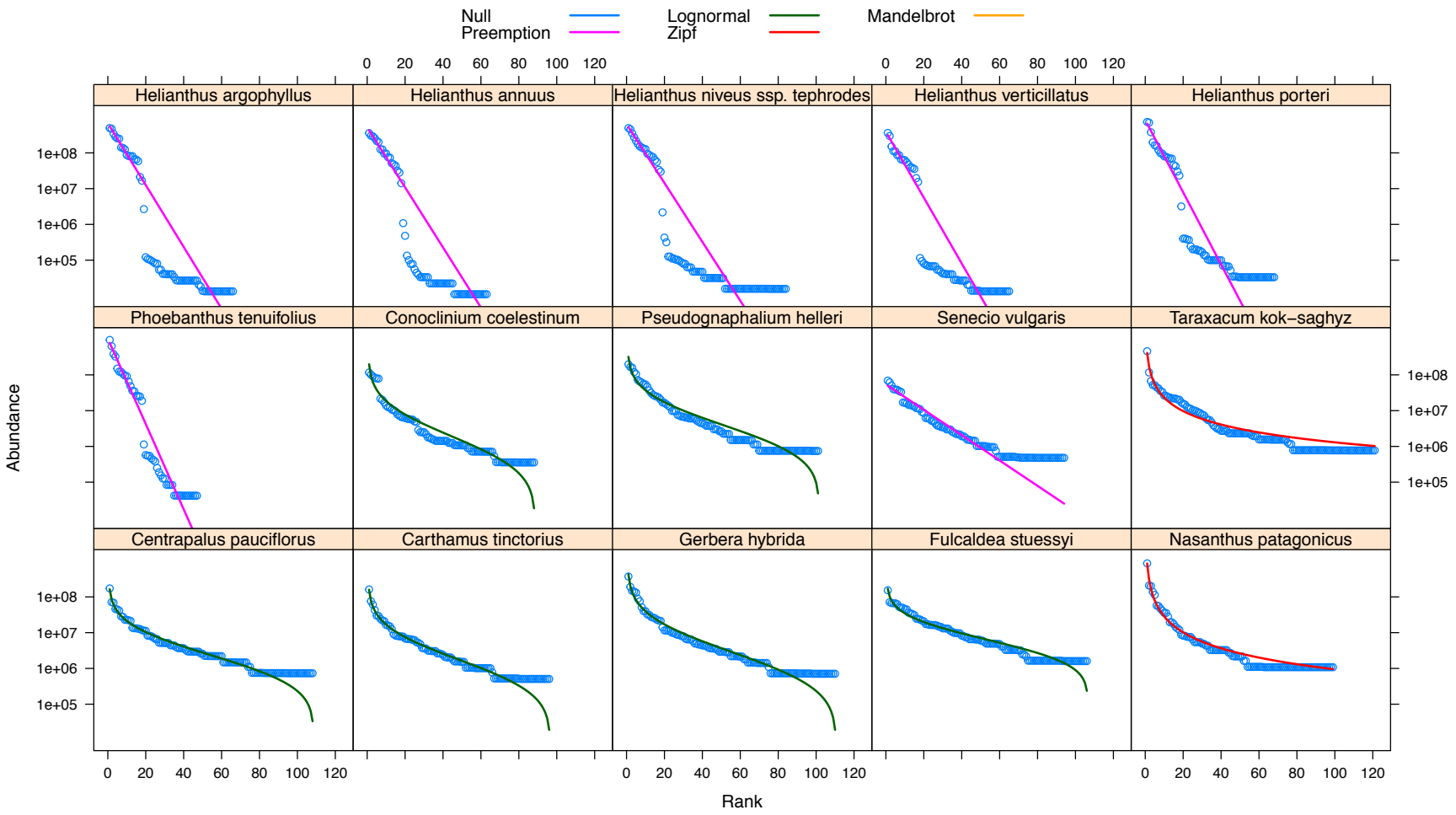

Null
Preemption
Lognormal
Zipf
Mandelbrot
Helianthus argophyllus
Helianthus annuus
Helianthus niveus ssp. tephrodes
Helianthus verticillatus
Helianthus porteri
Phoebanthus tenuifolius
Conoclinium coelestinum
Pseudognaphalium helleri
Senecio vulgaris
Taraxacum kok−saghyz
Centrapalus pauciflorus
Carthamus tinctorius
Gerbera hybrida
Fulcaldea stuessyi
Nasanthus patagonicus
Abundance
Rank
1e+08
1e+07
1e+06
1e+05
0
20
40
60
80
100
120

Figure 3.4. RAD plot of TE family abundance. Species are presented in phylogenetic order starting with the outgroup in the bottom right panel and the moving left to the most derived lineages of the Asteraceae in this study being displayed at the top left. The x-axis depicts the rank order of TEs by abundance, with rank 1 being given to the most abundant family, rank 2 given the second most abundant family, and so on. The y-axis depicts the log abundance of each TE family. Above the plots are the 5 ecological models used to test the fit of observed abundance. The colored line in each panel represents the best-fit model to each distribution as determined by BIC (see Experimental Procedures).

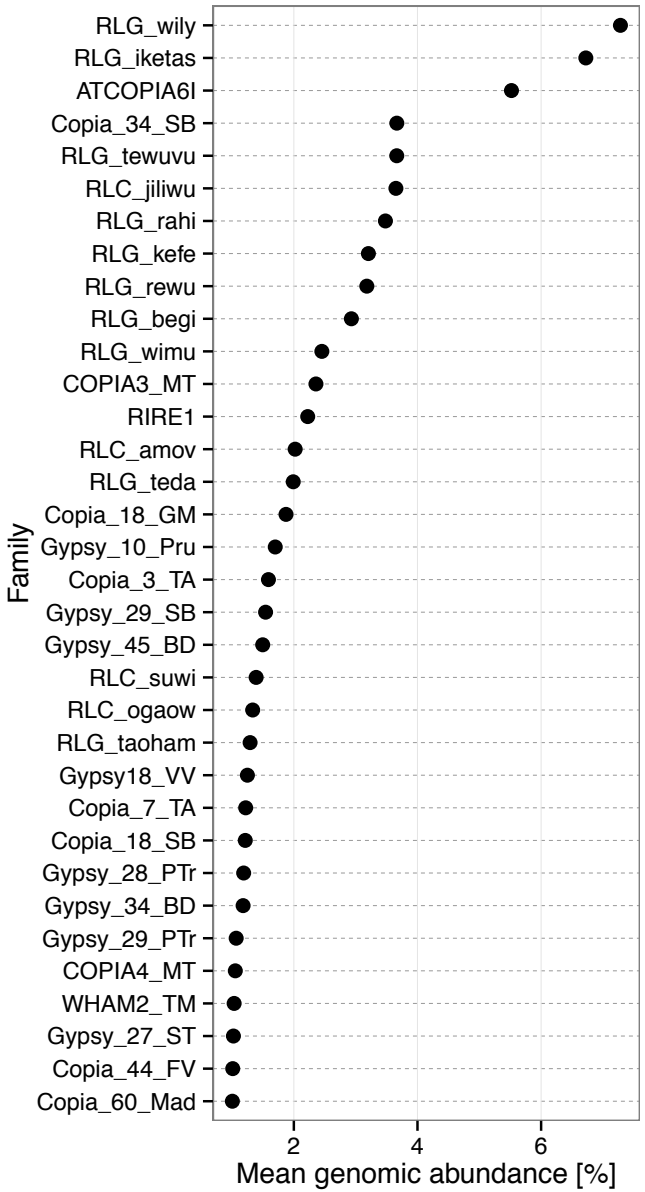


Figure 3.5. Rank abundance of TE families in the Asteraceae. The y-axis depicts the most abundant TE families in the Asteraceae, listed in decreasing rank abundance from the top the y-axis. The x-axis shows the average percent genomic abundance of each TE family in the Asteraceae.

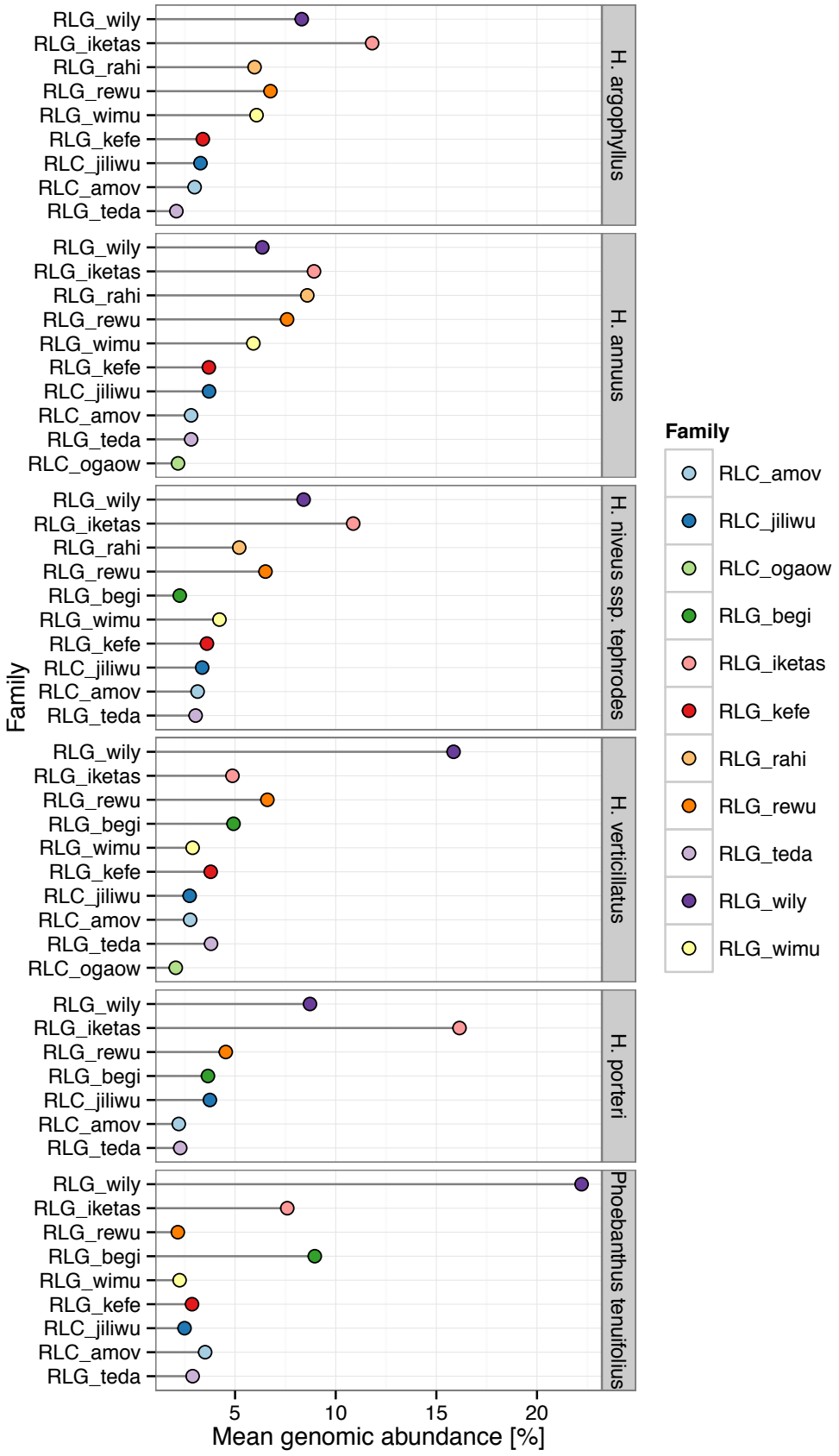
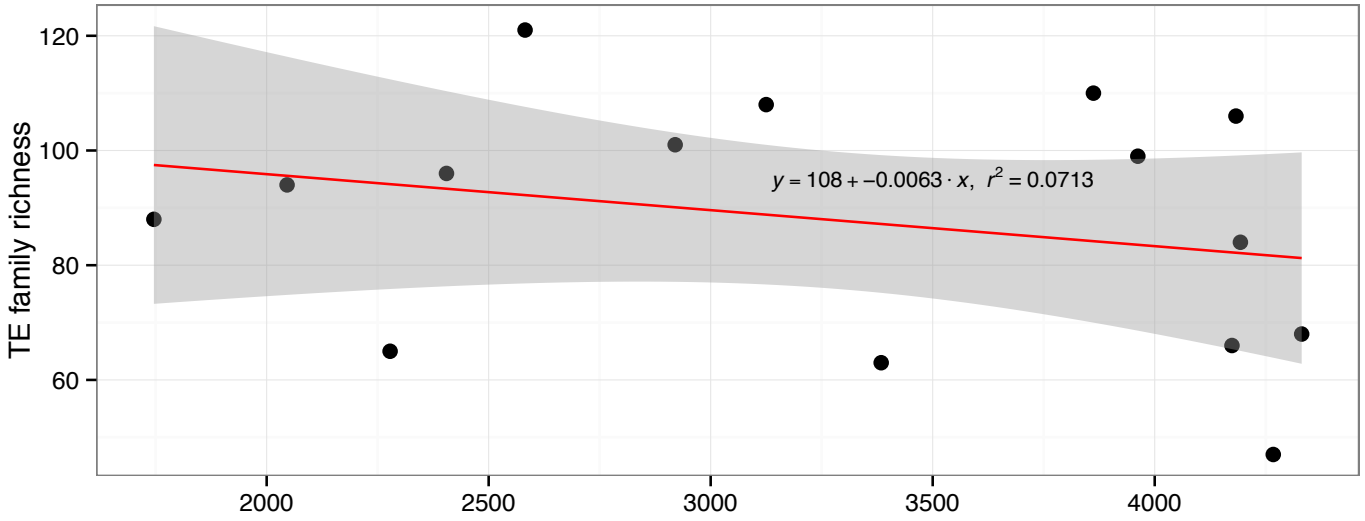

RLG_wily
RLG_iketas
RLG_rahi
RLG_rewu
RLG_wimu
RLG_kefe
RLC_jiliwu
RLC_amov
RLG_teda
RLC_ogaow
RLG_begi
H. argophyllus
H. annuus
H. niveus ssp. tephrodes
H. verticillatus
H. porteri
Phoebanthus tenuifolius
Family
5
10
15
20
Mean genomic abundance [%]
Family
RLC_amov
RLC_jiliwu
RLC_ogaow
RLG_begi
RLG_iketas
RLG_kefe
RLG_rahi
RLG_rewu
RLG_teda
RLG_wily
RLG_wimu

Figure 3.6. Rank abundance of TE families in the Heliantheae. Along the y-axis is the rank abundance of the top 2% of TE families in the Heliantheae, in decreasing order. Each panel depicts the rank abundance of TE families in phylogenetic order of the tribe from the base of the plot. The x-axis shows the percent genomic abundance of each TE family.

A)


TE family richness
120
100
80
60
2000
2500
3000
3500
4000
y = 108 + −0.0063 · x, r² = 0.0713

B)

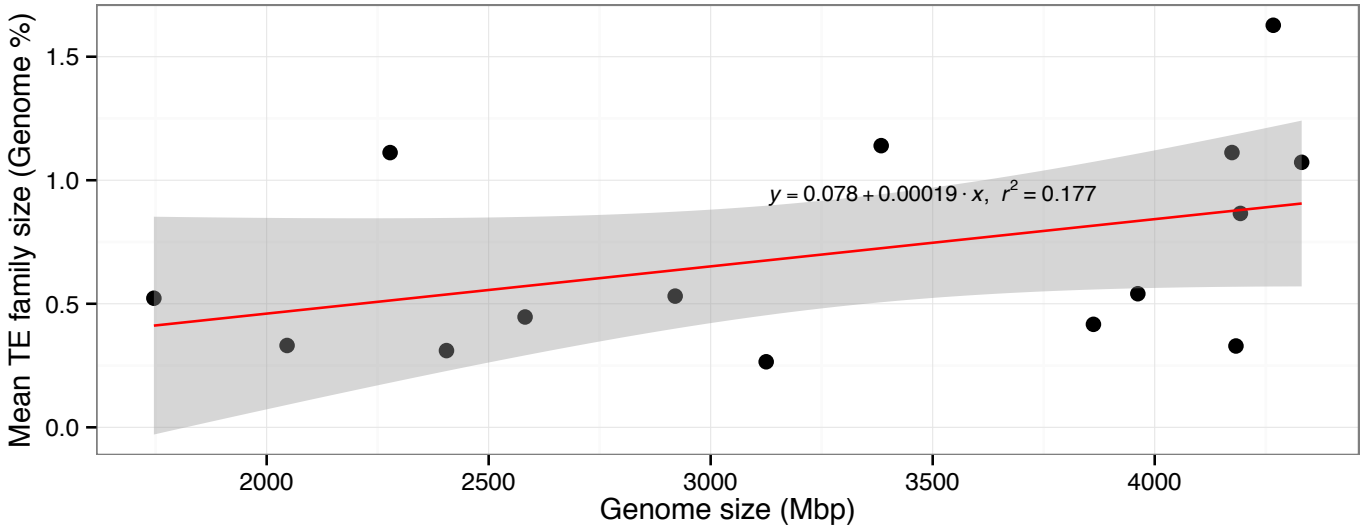

Mean TE family size (Genome %)
1.5
1.0
0.5
0.0
2000
2500
3000
3500
4000
Genome size (Mbp)
y = 0.078 + 0.00019 · x, r² = 0.177

Figure 3.7. Relationship between genome size and TE family size and richness. Along the x-axis is shown the genome size of each species in mega-base pairs. A) The TE richness, or total number of TE families seen, is shown along the y-axis. B) The mean TE family size as a percent of the genome is depicted on the y-axis.

B)

A)

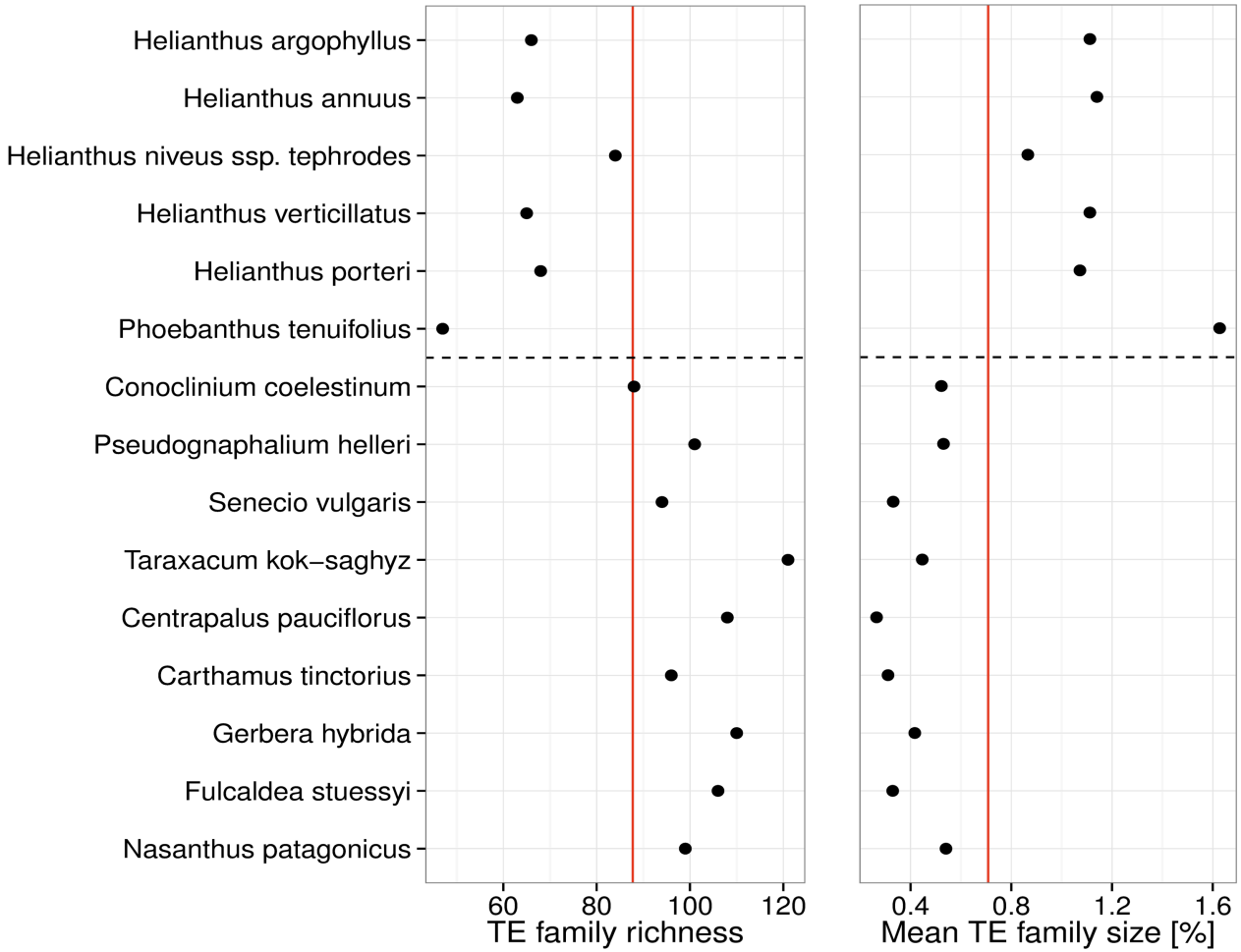

Helianthus argophyllus
Helianthus annuus
Helianthus niveus ssp. tephrodes
Helianthus verticillatus
Helianthus porteri
Phoebanthus tenuifolius
Conoclinium coelestinum
Pseudognaphalium helleri
Senecio vulgaris
Taraxacum kok−saghyz
Centrapalus pauciflorus
Carthamus tinctorius
Gerbera hybrida
Fulcaldea stuessyi
Nasanthus patagonicus
60
80
100
120
TE family richness
0.4
0.8
1.2
1.6
Mean TE family size [%]

Figure 3.8. Phylogenetic relationship between TE richness and TE family size. A) The TE family richness is shown along the x-axis for each species, which are depicted in phylogenetic order from the outgroup species at the base of the y-axis to the most derived lineages in the Asteraceae at the top of the y-axis. B) The mean TE family size as a percentage of the genome is shown along the x-axis. In both panels, the red vertical line indicates the mean and the horizontal dashed black line shows the base of the Heliantheae (with all species in the Heliantheae being shown above the line).

CHAPTER IV

TRANSPOSOME: INVESTIGATING TRANSPOSABLE ELEMENT FAMILIES FROM UNASSEMBLED SEQUENCE READS[1]

[1]Staton S.E., and Burke J.M. 2014. To be submitted to *Bioinformatics.*

SUMMARY

Transposable elements (TEs) make up the vast majority of DNA on the planet, and have had a profound impact on the evolution of Eukaryotes. Despite the broad taxonomic distribution of TEs, the evolutionary history of these sequences is largely unknown for many taxa due to a lack of genomic resources and identification methods for the vast majority of taxa. Advances in DNA sequencing have made it possible to rapidly generate large amounts of individual sequence reads; however, producing a genome assembly remains a challenging and costly procedure. Given that most TE annotation methods are designed to work on genome assemblies, we sought to develop a method to provide a fine-grained classification of TEs from DNA sequence reads. Here, we present a method for the efficient annotation of TE families from low-coverage whole genome shotgun data, enabling the rapid identification of TEs in a large number of taxa. Our software, Transposome, has been used to identify patterns of TE evolution in 16 plant species thus far, and is freely available (http://sestaton.github.io/Transposome).

INTRODUCTION

There is extreme variation in genome sizes in eukaryotes, from 9.78 Mbp in the unicellular algae *Cyanidium caldarium* to 124.59 Gbp in the angiosperm *Fritillaria assyriaca* (Gregory 2005). This variation was once considered a "paradox" due to the apparent lack of correlation between DNA content (i.e., genome size) and organismal complexity (Thomas 1971). However, this pattern is no longer considered a paradox because we know that genomes are not composed entirely of gene sequences (Britten and Kohne 1968), and it is likely that all of the major mechanisms contributing to genome size variation have already been described (Bennetzen *et al*., 2005; Gregory 2005; Bennetzen *et al*., 2007). The source of genome size variation may be attributed to environmental factors and biological attributes of species, though the major drivers of genome size increases, aside from polyploidy, are mobile DNA sequences called transposable elements (TEs).

Current approaches to identifying TEs involve using structural and similarity-based approaches with a genome assembly (e.g., Xu *et al*., 2007; Ellinghaus *et al*., 2008; Steinbiss *et al*., 2009), mathematical or k-mer based methods from genome assemblies or random sequence reads (e.g., Kurtz *et al*., 2008; Bao and Eddy 2002), signature-based methods with annotated TEs (Wheeler *et al*., 2013) and cluster-based approaches from unassembled sequence reads (Novak *et al*., 2013). Without question, the most accurate method for identifying TEs would be through a combination of the above methods. One caveat with the aforementioned approaches is that most require a genome sequence (i.e., assembled genome) as input. However, it is not practical to generate a genome assembly for every species of interest given that costs are currently still too high for non-model

systems and genome assembly algorithms are not currently able to resolve large and complex genomes, such as those of many plant species. There are other practical reasons to avoid draft genomes for repeat discovery. Specifically, draft genomes are typically not representative of the true nature of genomic repeats (Alkan *et al*., 2011), which is likely due to the fact that graph-based genome assembly algorithms are designed to avoid repetitive regions. The ideal solution to improving repeat annotation from understudied species would leverage high throughput short sequence read data, allowing thorough phylogenetic descriptions of TE properties. Implementing such a tool is technically challenging, as evidenced by the fact that there has been only one tool published that is able to provide biological descriptions of repeat types from millions of sequence reads (Novak *et al*., 2013). This repeat finding program, called RepeatExplorer, is implemented as a web-based Galaxy tool and is targeted at biologists with no programming knowledge. Unfortunately, this implementation is computationally inefficient, making analyses of multiple species impractical, and there this is no programmatic interface with this tool, making execution of the code disconnected from the data. Given the biological significance of repetitive sequences and the lack of tools for describing repeats in non-model species, we have developed a tool, called Transposome, that requires no programming experience but also offers an interface for experienced programmers to extend and modify the toolkit making it easy to incorporate into larger analysis pipelines.

RESULTS AND DISCUSSION

Genomic repeat annotation is a challenging task, in part because there are dozens of tools available and most have not been analyzed in terms of performance or accuracy (Lerat 2010). We analyzed both the performance and accuracy of Transposome using

whole-genome shotgun (WGS) data from two well-studied plant species, maize (*Zea mays* L.) and the common sunflower (*Helianthus annuus* L.). We were limited to only one program, RepeatExplorer, for comparison because all other repeat annotation programs that use WGS data as the substrate may take weeks to run on miniscule data sets (Saha *et al*., 2008), thus we restricted our comparisons to the relevant applications. Our findings clearly indicate that Tranposome is more accurate at identifying the total level of genomic repetitiveness and is more computationally efficient than RepeatExplorer (Table 4.1). Given this finding, we further evaluated Transposome with respect to accuracy at the individual TE family level. It should be noted that this type of analysis is not possible with the type of results generated by RepeatExplorer, and running a large number of simulations using RepeatExplorer would take many years.

Transposable element accumulation

How effectively are we sampling the genomic diversity of repetitive elements? In order to address this question, we sampled data at varying levels of genome coverage and performed an analysis with Transposome using default parameters on each read set. For this experiment, we only evaluated maize, as this species has the most comprehensive repeat library of any plant species. The repeat database used in this analysis was downloaded from the maize TE database (maizetedb.org). From the sampled data we constructed what is referred to in the ecological literature as an accumulation curve (Ugland *et al*., 2003), which allows us to assess the level of diversity (i.e., number of TE families) being captured. Because we know what fraction of the maize genome is comprised of TEs, we were also able to assess the total fraction of diversity being sampled at varying levels of genome coverage (i.e., the total percent of the genome).

The maize genome has recently been shown to be >75% TEs, with approximately 70% of the genome being composed of just 20 TE families (Baucom *et al*., 2009). With just 1 million randomly sampled paired-end sequences (see Experimental Procedures), Transposome was able to correctly identify 17 of the top 20 families, with the top 11 families being correctly identified as in Baucom et al. (2009), though not in the same order (Table 4.2). Many of the maize TE families are very similar in size; thus it is likely that there would be some inconsistencies with respect to rank order.

How many sequence reads are required to capture the majority of the TEs by genome coverage? Given that most applications would take months or years to analyze large data sets, it is not feasible to assess the effect for varying genome coverage or analysis parameters. However, it is clear that Transposome is able to accurately predict the genomic coverage of TEs and the number of TE families by rank that account for that genome coverage (Figure 4.1). This result clearly indicates that it is not necessary to have high genome coverage sequence data in order to infer >90% of the TEs by abundance, though it is not clear from this analysis what type of error there is in the estimation of each family abundance.

Estimating accuracy in statistical predictions of TE abundance

Is the prediction we obtain for a single analysis an accurate representation of the genomic abundance of a particular repeat type? This is a question that is rarely addressed in the literature, perhaps due to the fact that most analyses take months to run and multiple estimates are impractical. To address this question we simulated a much wider range of genome coverage values than for the above analysis (0.5 – 5.1% genome coverage), and we also took three random samples at each coverage level to assess the

magnitude of error in our calculations. We find that for TE families that have a high abundance in the genome (i.e., >5%), accurate estimates of abundance can be made with little error using very low genomic coverage data (i.e., less than 2% genome coverage for maize). It is also clear from this analysis that more accurate estimates of rare TE families can be made with higher genome coverage data (Figure 4.2A), and under sampling the genome will result in either overestimating abundance (Figure 4.2B), or underestimating abundance for some TE families (Figure 4.2C). Taken together, we were able to determine (at least for sunflower), that at least 3% of the genome would need to be sampled in order to identify TE families with genomic abundance <1% while keeping error estimates low (below 15% coefficient of variation, in this study). This result can help guide future studies, though the exact figures obtained will vary depending on the species.

Implementation and usage of Transposome

Transposome borrows the clustering approach implemented in RepeatExplorer (Novak *et al*., 2013), though it improves upon the performance and usability of RepeatExplorer in several ways. First, RepeatExplorer runs a full pipeline of analyses, which can take more than a month for large data sets. Since very long run times make redoing parts of analyses impractical, we developed Transposome in a modular fashion so that redoing analyses with varying clustering and annotation parameters is possible. Second, we found that RepeatExplorer generated highly partitioned clusters, erroneously separating individual TE families into multiple clusters. This behavior has been reported previously (Fortunato 2010), and is a result of the Louvain clustering algorithm. Despite this behavior, the Louvain clustering method outperforms other clustering algorithms in

terms of computation time, and is currently best available method for constructing very large graphs so we implemented an automated method to correct for the problem of over-clustering by using paired-end sequence read information. Third, we are unaware of any tools that would allow annotation of repeats to the TE family level, thus we developed methods to allow a fine-grained statistical evaluation of genomic composition. Lastly, Transposome can be set up very quickly and run on a local computer, removing the possibilities of data loss or long wait times due to high job volumes on a web server. Our methods are implemented in modern Object-Oriented Perl, making it possible to extend and modify Transposome with custom methods, though the primary usage of Transposome is through a command-line application and no programming experience is required to run any part of the analysis. Available through the Perl API are methods for randomly sampling and indexing sequence reads, permuting a wide range of clustering and annotation parameters, and many other utilities. The usage of Transposome as a toolkit is very similar to BioPerl (Stajich *et al*., 2002), though it is important to note that our sequence reading class is significantly faster than BioPerl for parsing high-throughput sequence data. The project is maintained with git for version control and Travis-CI for continuous integration and is hosted publicly on github (http://sestaton.github.io/Transposome) where users may explore the code and file issues. Documentation is available from the command-line for each Transposome class as well as the main application, and there is a wiki that covers the setup and usage of Transposome (https://github.com/sestaton/Transposome/wiki).

EXPERIMENTAL PROCEDURES

Algorithm description

Similarity-based annotation methods such as RepeatMasker (Smit *et al.*, 2010) are of little use for non-model systems because they are fundamentally biased to finding only sequences similar to those in a particular reference library. The approach we implemented consists of two steps. First, we developed a highly parallel all vs. all sequence comparison procedure to find shared similarity within the genome. This step makes use of mgblast (Pertea *et al*., 2003), a modified version of megablast that is very memory efficient. This enables us to execute many parallel processes, significantly speeding up run times. In terms of the thread model, the optimal solution to satisfy can be shown in the following equation:

$$\frac{n}{i} \geq t \times c$$

where *n* is the total number of sequences, *i* the subset size, *t* is the thread number, and *c* is the CPU number. If,

$$\frac{n}{i} < t \times c$$

then *c* CPUs will not be used, the user will be using

$$\frac{n}{i}$$

CPUs instead. Optimizing this equation will lead to the shortest run times by using all available computation resources. Second, we use a novel clustering algorithm called the Louvain Method that is able handle very large datasets (Blondel *et al*., 2008) efficiently and makes use of edge weight information (which corresponds to similar sequence pairs in this application). This algorithm is very fast, graphs grow like O(*n*), but one caveat is

that it over refines clusters leading to a separation of defined groups (Fortunato 2010). We modified this algorithm by using paired-end sequence information to find union in the graph constructed during the clustering processes, and this allows us to circumvent the issue of over clustering by using biological information. For this reason, Transposome is designed specifically to work with paired-end data only and has not been tested with single-end reads.

Analyzing the rank of cluster sizes alone is very informative for comparing genome structure between species. However, it is most useful to understand what types of repeats differ in abundance between species. Every sequence clustered by Transposome represents something repetitive in the genome, and these sequences are compared to a reference library using blastn from the BLAST+ package (Camacho *et al*., 2009). The particular reference library used by Transposome may be any sequence set, and the best reference set to use will vary depending on the species being investigated. Perhaps the most useful feature of Transposome is that results are translated directly into estimates of genomic composition of each repeat type, and the full taxonomic lineage of each repeat type is listed to the family level.

Sample data acquisition and treatment

Paired-end WGS reads for maize were downloaded from the Genbank Short-Read Archive (accession no. SRX142106), and WGS read data for sunflower were obtained from the Genbank BioProject site (BioProject ID: 236448; accession no. PRJNA236448). Both read sets consist of 100 bp Illumina paired-end reads with short inserts (median insert for maize: 300 bp; median insert for sunflower: 400 bp). Because the Genbank data format is unique and technically is not valid FASTQ (Cock *et al*., 2009), we used Pairfq

(version 0.09; https:/github.com/sestaton/Pairfq) to add the read pair information to the forward and reverse reads. Following this formatting step, all reads were quality trimmed with PRINSEQ (version 0.19.4; Schmieder and Edwards, 2011) with the parameters '-min_len 40 –noniupac –min_qual_mean 15 –lc_method entropy –lc_threshold 60 –trim_ns_right 10 –ns_max_p 20', and trimmed paired-end reads were re-synced with Pairfq prior to analysis.

For all simulations, a necessary procedure was to randomly sample the data at varying levels of genome coverage. This is computationally challenging for two reasons. First, modern sequence data sets consist of hundreds of millions of records. For example, the maize data described above consists of 302,177,385 FASTQ records, which is ~60.4 billion base pairs. In terms of the physical data size, these files totaled 1,208,709,540 lines (604,534,770 lines in each forward and reverse file) and take up approximately 35.9 GB of disk space. A naïve approach to handling a text file of over 1 billion lines will result in the loss of data, a large loss of time, and likely a system failure. Second, for paired-end sequence data, the forward and reverse reads cannot be out of sync, so an algorithm that generates truly random samples is of no use for this application.
The sampling algorithm we implemented is called Reservoir Sampling (algorthim R3.4.2; Knuth 1997), and requires only *n* records to be stored in memory at a time (where *n* is the desired number of samples to be taken), and guarantees even sampling of the entire data set. Furthermore, given the same random seed, the exact same records will be sampled for both forward and reverse files, which eliminates further processing of the data. This sampling method is called 'sample_seq' and is available through the Transposome::SeqUtil class. By default, the 'sample_seq' method writes to the standard

output stream, which allows data to be piped to other applications. It is also possible to write the sampled data to a file, as demonstrated in the Transposome wiki tutorial (https://github.com/sestaton/Transposome/wiki/Tutorial).

ACKNOWLEDGEMENTS

We thank Petr Novak for discussions about RepeatExplorer and advice on running this program.

LITERATURE CITED

Alkan C, Sajjadian S, Eichler E.E. 2011. Limitations of next-generation genome sequence assembly. *Nat. Methods* 8(1), 61-65.

Bao Z, and Eddy SR. 2002. Automated *de novo* Identification of Repeat Sequence Families in Sequenced Genomes. *Genome Res.* 12, 1269-1276.

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 or retroelements in the B73 maize genome. *PLOS Genetics* 5(11), 1-13.

Britten RJ and Kohne DE. 1968. Repeated sequences in DNA. *Science* 161, 529-540.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, and Madden, TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics,* 10, 421.

Cock PJA, Fields CJ, Goto N, Heuer ML, and Rice PM. 2009. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucl. Acids Res.* 38(6), 1767-1771.

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

Fortunato S. 2010. Community detection in graphs. *Physics Rep.* 486, 75-174.

Knuth D. 1997. The art of computer programming. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA.

Kurtz S, Narechania A, Stein JC, and Ware D. 2008. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics,* 9, 517.

Novak P, Neumann P, Pech J, Steinhaisl J, and Macas J. 2013. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* 29(6), 792-793.

Saha S, Bridges S, Magbanua ZV, and Peterson DG. 2008. Empirical comparison of *ab initio* repeat finding programs. *Nucl. Acids Res.* 36(7), 2284-2294.

Schmieder R. and Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27:863-864.

Smit AFA, Hubley R and Green P. 1996-2010. RepeatMasker Open-3.0. <http://www.repeatmasker.org>.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, et al. 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* 12(10), 1611-1618.

Steinbiss S, Willhoeft U, Gremme G, and Kurtz S. 2009. Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Res.* 37, 7002-7013.

Ugland KI, Gray JS, and Ellingsen KE. 2003. The species-accumulation curve and estimation of species richness. *J. Animal Ecol.* 72(5), 888-897.

Wheeler TJ, Clements J, Eddy SR, Hubley R, Jones TA, Jurka J, Smit AFA, and Finn, RD. Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucl. Acids Res.* 41 (D1), D70-D82.

Xu Z and and Wang H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 35, W265-W268.

TABLES FOR CHAPTER IV

Table 4.1. Basic performance metrics of programs for finding repeats from WGS reads. RepeatExplorer is executed on a remote web-server, which makes it not possible to get detailed computational resource usage.

Program	*Running time*	*Total genomic repeats*[1]	*Percent of repeats captured*[2]
RepeatExplorer	>30days	48	59.25
Transposome	4hr33m53s	72.02	88.91

[1]Figures represent the total repeat percentage of the genome.
[2]Based on estimates reported in Staton et al., 2012.

Table 4.2. Comparison of published maize TE family annotations with Transposome results. The RLX indicates ambiguity, which means these families contain no coding domains. The dashes indicate that Transposome did not identify that particular TE family.

Family	Superfamily	Family rank[1]	Family rank [2]	Genomic %[1]	Genomic %[2]
Huck	Gypsy	1	4	10.15	10.21
Ji	Copia	2	1	9.81	16.6
Cinful-zeon	Gypsy	3	3	8.18	11.23
Opie	Copia	4	2	7.74	12.11
Flip	Gypsy	5	5	4.18	5.15
Xilon-diguus	Gypsy	6	7	3.63	2.81
Prem1	Gypsy	7	6	3.34	3.61
Gyma	Gypsy	8	9	2.80	2.06
Grande	Gypsy	9	8	2.71	2.13
Doke	Gypsy	10	11	1.88	0.98
Giepum	Gypsy	11	10	1.20	1.56
Milt	RLX	12	--	0.93	--
Puck	Gypsy	13	17	0.90	0.26
Ruda	RLX	14	--	0.83	--
Tekay	Gypsy	15	15	0.69	0.58
Uwum	Gypsy	16	12	0.68	0.89
Dagaf	Gypsy	17	14	0.68	0.62
Iwik	RLX	18	--	0.36	--
Wiwa	Gypsy	19	16	0.29	0.45
CRM1	Gypsy	20	13	0.27	0.85

[1]Baucom et al., 2009
[2]This study

FIGURES FOR CHAPTER IV

A)

B)

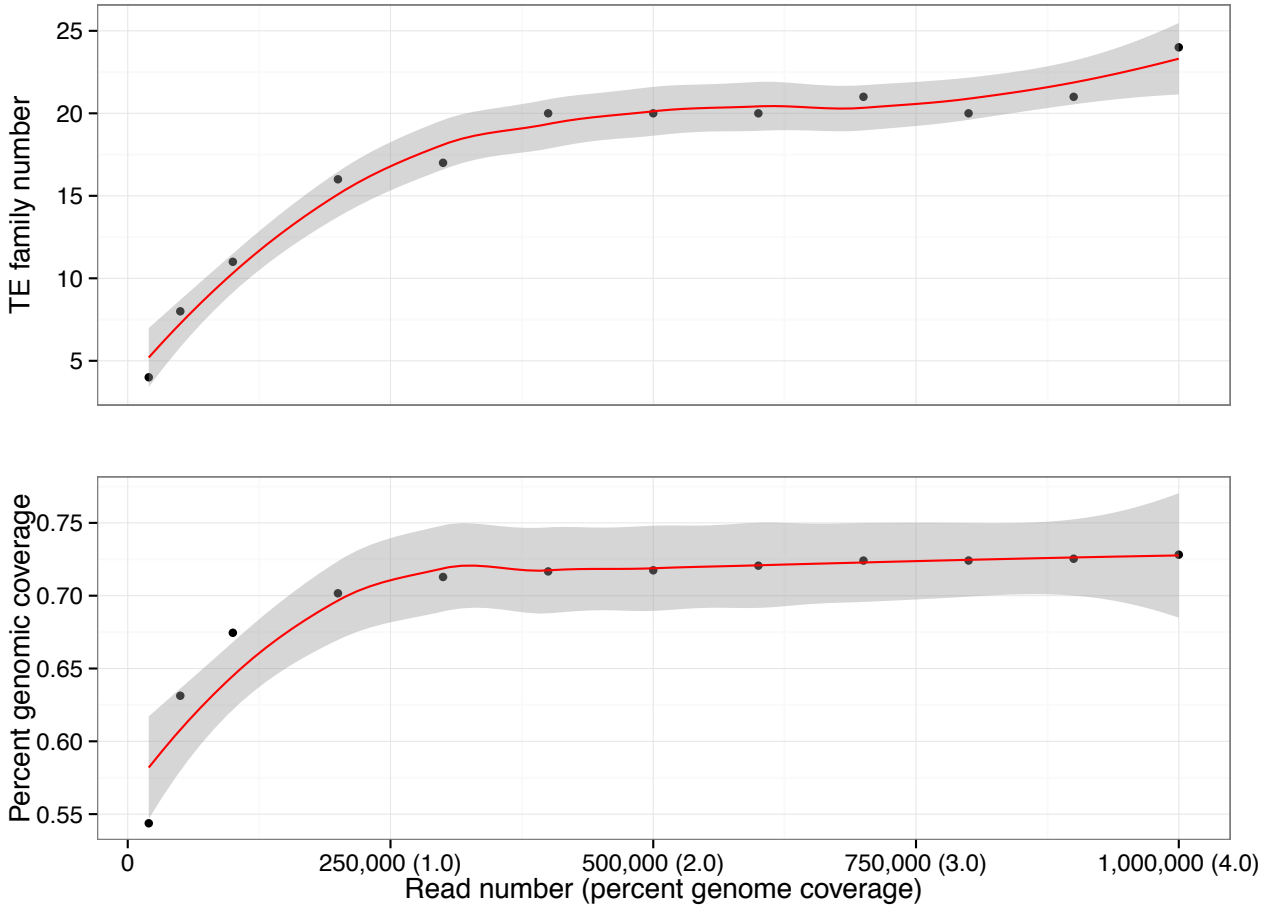
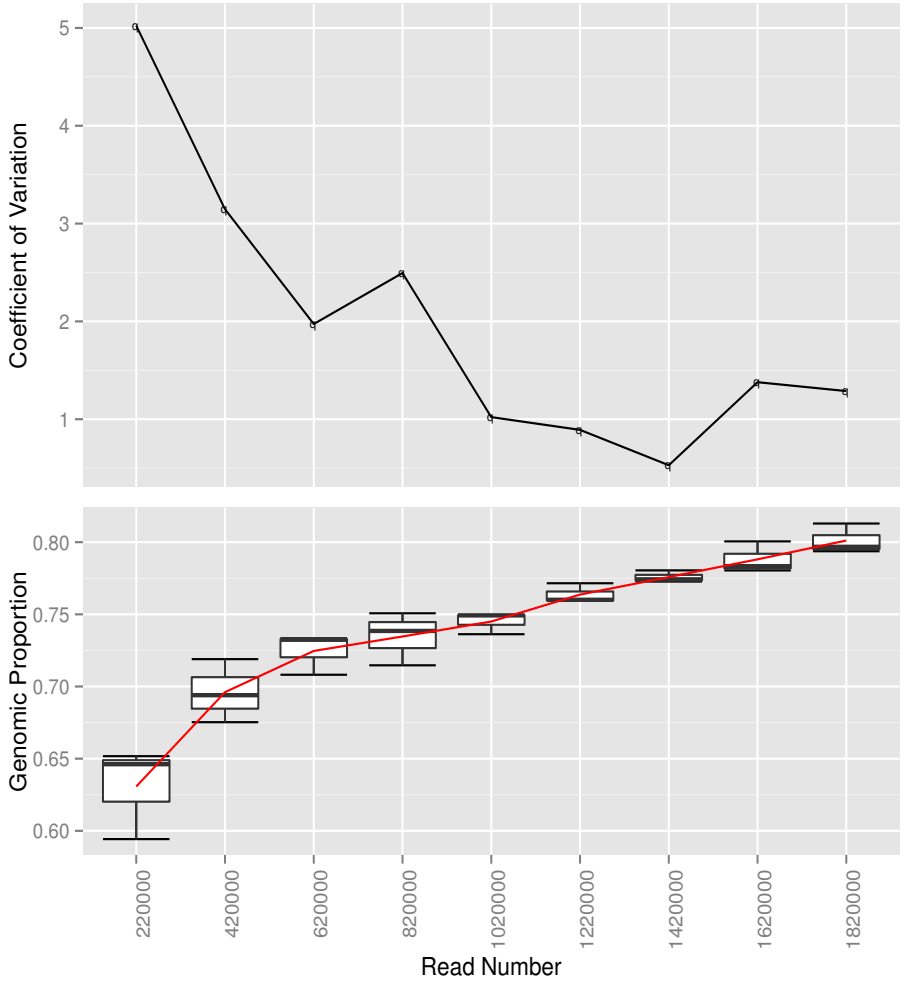
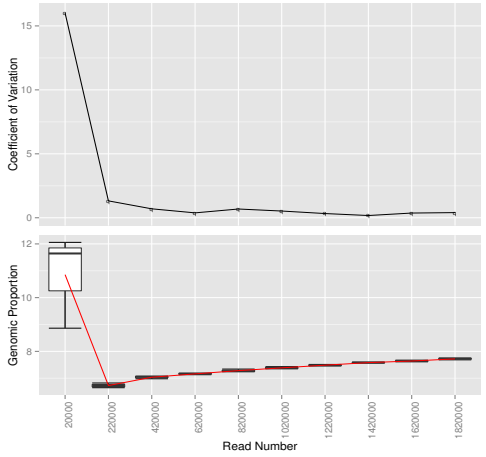


Figure 4.1. Maize TE family accumulation and percent TE accumulation with varying genome coverage data. The x-axis shows the number of reads used for each simulation, and the corresponding percent genome coverage is also displayed. A) The number of TE families identified at each level of genome coverage, and B) the percent of the genome described by the TEs identified at each level of genome coverage.

A)


Coefficient of Variation
5
4
3
2
1
Genomic Proportion
0.80
0.75
0.70
0.65
0.60
220000
420000
620000
820000
1020000
1220000
1420000
1620000
1820000
Read Number

B)


Coefficient of Variation
15
10
5
0
Genomic Proportion
12
10
8
20000
220000
420000
620000
820000
1020000
1220000
1420000
1620000
1820000
Read Number

C)

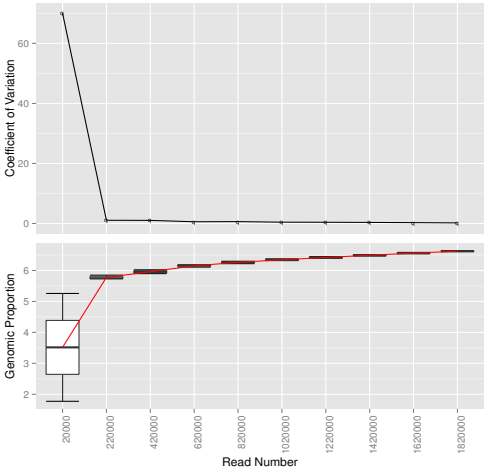

Coefficient of Variation
60
40
20
0
Genomic Proportion
6
5
4
3
2
20000
220000
420000
620000
820000
1020000
1220000
1420000
1620000
1820000
Read Number

Figure 4.2. Variation in the estimates of genome abundance for TE families using different levels of genome coverage. A) For rare TE families (here, RLG_suwi is shown), it is clear that high genome coverage data is required accurately estimate the genomic abundance with a coefficient of variation below 1%. Analyses from very low genome coverage data are likely to B) overestimate (in the case of RLG_wily) or C) underestimate (in the case of RLG_wimu) the genomic abundance of a TE family.

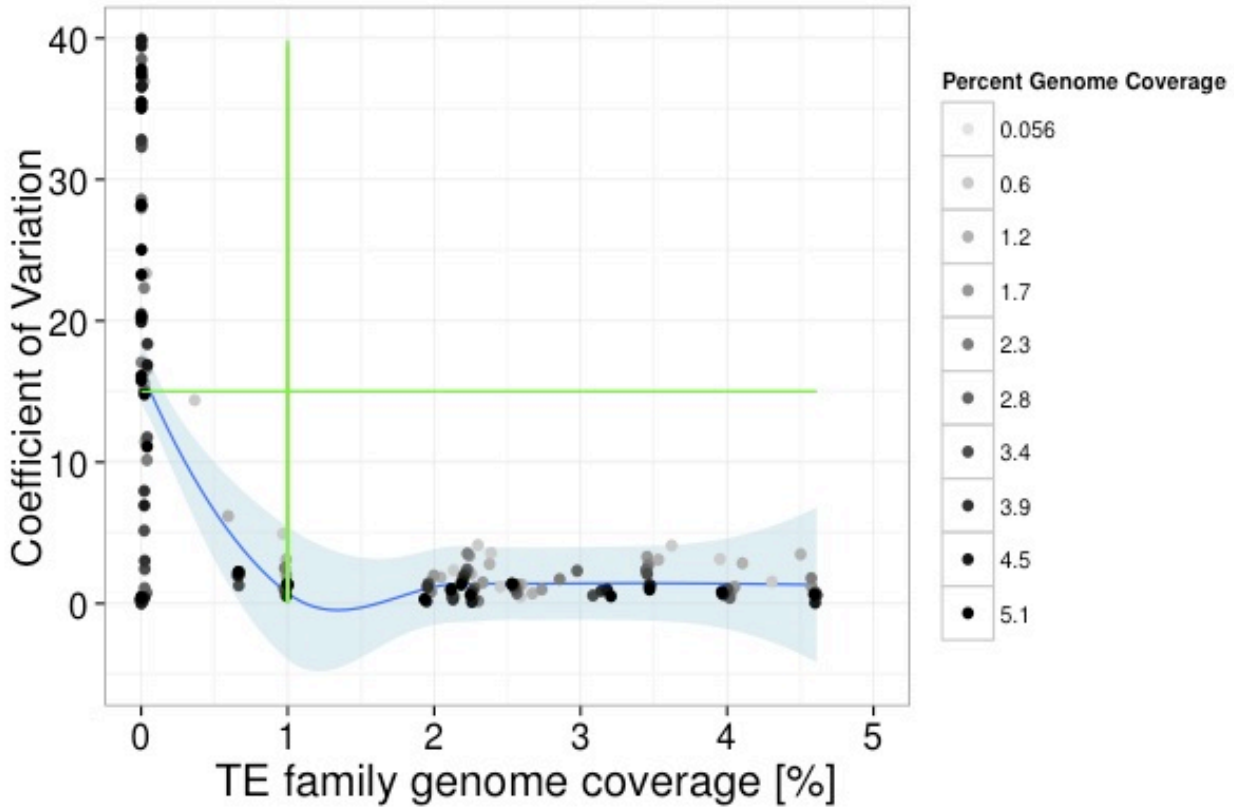

Coefficient of Variation
40
30
20
10
0
0
1
2
3
4
5
TE family genome coverage [%]
Percent Genome Coverage
0.056
0.6
1.2
1.7
2.3
2.8
3.4
3.9
4.5
5.1

Figure 4.3. Coefficient of variation for estimates of TE family abundance from a range of genome coverage simulations for sunflower. Each point in the graph depicts the percent genomic abundance of a TE family and the different shades of each point represent estimates using different levels of genome coverage data.

CHAPTER V

CONCLUSIONS

"*The rapid development as far as we can judge of all the higher plants within recent geological times is an abominable mystery.*" (Darwin 1887)

Charles Darwin was deeply intrigued by the extraordinary diversity of Angiosperms. It was not the diversity of plant species itself that was most puzzling but rather the short evolutionary timeframe over which the diversity had arisen (Darwin 1872; Crepet 1998; Friedman 2009). The idea of rapid speciation was controversial in Darwin's time because it was not clear what mechanisms could contribute to both ecological adaptations and reproductive isolation over such short timeframes (Berendse and Scheffer 2009; Friedman 2009). Though numerous mechanisms likely contribute to this diversity, it is now appreciated that eukaryotic transposable elements (TEs) have the potential to facilitate rapid genomic and phenotypic changes, potentially leading to both adaptations and species divergence (McClintock 1984; Kalendar *et al.*, 2000; Kobayashi *et al.*, 2004; Piegu *et al*., 2006; Ungerer *et al*., 2006; Hilbrict *et al.*, 2008; Xiao *et al*., 2008; Xie *et al.*, 2010). For example, TEs are known to contribute to the process of speciation (reviewed in Rebollo *et al*., 2010) and to the processes of epigenetic reprogramming and regulatory evolution (Feschotte 2008; Lisch 2009). These examples reflect the role of TEs at many different levels, from DNA transcription to chromosome behavior and shaping ecological distributions of species. It is clear that separating individual processes from phylogenetic and ecological patterns is difficult because there are interactions among all of these levels. The ongoing nature of TE evolution and their

complex arrangements in the genome have prohibited deciphering the importance of individual events in the past with marker-based, hybridization or mapping techniques. However, current technology makes it possible to investigate patterns of TE evolution in a large number of non-model species in a single study. This work represents the first demonstration, to our knowledge, of targeting a large number of previously undescribed species and providing a complete description of TE diversity and abundance in a phylogenetic framework. However, it remains somewhat of a mystery what processes are most important in shaping the long-term evolution of the genome community of TEs (Gregory 2005; Lynch 2007; Le Rouzic *et al*., 2007). The primary motivation for this work is to provide a better understanding of these processes through comparative genomic analyses of the plant family Asteraceae.

Transposable element contributions to sunflower genome evolution

The second chapter of this dissertation provided a fine-scale description of TEs in the genome of the common sunflower, *Helianthus annuus* (Staton *et al*., 2012). The most important findings from this work are that the sunflower genome has an extremely biased distribution of TE abundances with *Gypsy* LTR retrotransposons (LTR-RTs) accounting for more that 58% of the genome. The majority of *Gypsy* insertions appear to be unique to the sunflower lineage (i.e., having occurred within the past 1 MY), and we provide further evidence that *Gypsy* TEs are transcriptionally active in sunflower. Similar to findings in the maize genome (SanMiguel *et al*., 1996; Devos *et al*., 2002), we show that the majority of LTR-RT copies in the sunflower genome are intact rather than truncated, though we also demonstrate that there is a greater ratio of solo-LTRs to intact elements in *Copia* than is seen with *Gypsy*. This latter finding may be a result of the greater average

age of *Copia* elements, though it may also be an explanation for the biased composition of TEs in the sunflower genome (i.e., arising possibly through biased removal of *Copia* instead of simply biased amplification of *Gypsy*). Most interestingly perhaps, we found that over 55% of intact *Gypsy* TEs contain a chromodomain, which appears to be involved in integration of the element (Gao *et al*., 2008), and nearly all chromodomain-containing *Gypsy* elements had at least one tandem duplication of this domain. While chromodomain-containing *Gypsy* elements are widespread in the plant kingdom (Gorinsek *et al*., 2004), we have not found another report of duplicated chromodomains within *Gypsy* TEs. This finding is especially interesting given that chromodomain-containing *Gypsy* elements are also found in fungi and vertebrate lineages (Martin and Llorens 2000; Gorinsek *et al.,* 2004). We know that this pattern of duplication is not restricted to sunflower as we have also discovered duplicated chromodomains within *Gypsy* elements in lettuce and dandelion (also in the Asteraceae; *unpublished*). Importantly, the most abundant clade of chromodomain-containing *Gypsy* elements in sunflower, Tekay, was shown previously to exhibit signs of positive selection, and many clade-specific lineages of Tekay elements have been found in plants (Novikov *et al.,* 2012). In addition, it appears that transitions in plant *Gypsy* chromodomains have occurred coincident with transitions in the chromatin landscape in plant genomes. Given that Group II chromodomains, like the ones described here, are only found in plant *Gypsy* retrotransposons, we feel that this work may represent an important insight into the processes of plant genome organization and diversity.

Transposable element contributions to genome evolution in the Asteraceae

It is now clear that the sunflower genome is biased towards *Gypsy* TEs (e.g., Staton *et al*., 2012; Natali *et al*., 2013), and we have provided insight into the mechanisms of sunflower genome evolution. Yet, it was not clear if the patterns in *Helianthus annuus* were widespread in the genus *Helianthus* or across the entire Asteraceae. Through our analysis of 15 species from across the Asteraceae, including an outgroup species, we demonstrated that the genomic patterns described in my second chapter were not unique to the common sunflower, as similar patterns of genome composition and bias appear to be shared across the tribe Heliantheae. However, the basal lineages of the Asteraceae exhibit strikingly different patterns of genome diversity and relative abundance with respect to TEs as those species in the Heliantheae. With the exception of the Heliantheae, there is little variation in relative abundance and diversity of TEs in the Asteraceae, which suggests that neutral processes likely shape the genome community over large evolutionary timescales. However, in the Heliantheae we see a very different pattern, with a few highly abundant TE families and numerous rare TE families dominating the genome. The impact of these changes on the genome community has been an increase in the average TE family size, owing to the dominance of a few families and a decrease in TE diversity, perhaps due to the ongoing replacement of DNA that is pervasive in plant genomes (Ma and Bennetzen 2004), though in a biased manner in this case.

This work provides another example of TE amplifications coinciding with phylogenetic divergence, and provides further insight into the timing and mechanisms of genome evolution in the Asteraceae. Perhaps most importantly, we have developed an

efficient computational toolkit enabling large-scale genomic studies of TEs to be done in a repeatable fashion. It is our intention that continued development of this toolkit will provide a framework for phylogenetic studies of TEs and genome evolution in the plant kingdom, and help to uncover the mechanisms that govern the diversity of TEs within the genome.

Future directions

Our collective knowledge of TE properties in the plant kingdom has been largely shaped by the analysis of just a few model species. It is now clear that our 'model species' are not representative of the diversity in the plant kingdom; thus, a new focus has been placed on understanding TE dynamics in natural populations, and investigating the mechanistic role of these processes in a phylogenetic framework by simultaneously analyzing many species (as of this writing, "many species" realistically means that one could perform a *de novo* analysis of dozens of plant genomes in a single study using the methodology described herein). By taking a phylogenomic approach to analyzing TE dynamics, I believe we will gain a much deeper mechanistic understanding of TE properties than by making evolutionarily distant comparisons alone.

Given the ease with which whole-genome shotgun data sets can now be generated for a large number of species, I believe several important questions will be addressed in the near future. First, it is clear that whole genome duplication (WGD) has been an important process in the evolution of plants, and TE amplification and genome rearrangements often follow WGD events. How and why do WGD events contribute to TE amplification and genome restructuring? At this time, there do not seem to be general rules because WGD events may contribute to TE family-specific patterns of

amplification (Hawkins *et al.* 2006). Conversely, WGD events may lead to non-specific amplification of many TE families or no TE amplification at all (Hawkins *et al.*, 2008). Thus, it is important to further sample the plant kingdom to gain a better understanding of whether these few studies reflect that TE amplification is stochastic in nature following WGD, or if there are specific mechanisms that govern this process. Second, the most abundant form of TE in the plant kingdom is the chromodomain-containing *Gypsy* retrotransposons (i.e., the chromoviruses). These TEs integrate into the genome by recognizing specific epigenetic marks, and appear to have diversified in response to changes in the genomic environment (Novikov *et al.*, 2012). Currently, there have been no genome-wide studies of chromoviruses in plants to investigate the specific causes and consequences of diversification of these genomic elements. Given the importance of epigenetic modifications in regulating gene function, how have chromoviruses contributed to genome structure and function in the plant kingdom? I believe a major focus in the future will be in understanding how TEs shape the chromatin landscape. Third, it is likely we know all, or at least most, of the biological mechanisms that contribute to TE activity. However, what is the role of the environment in shaping TE-host evolution and which factors are more important? I began this dissertation with a discussion of Barbara McClintock's discovery of TEs because that is where investigations into TEs began. I believe it is fascinating to think that more that 60 years after her descriptions of "controlling elements" we have now returned to the point of asking whether TEs represent general mechanisms of genome response to environmental or biological stress. I sense in talking with colleagues that a major shift in the future will be away from genome-based analyses with model systems and towards broader sampling

of variation in natural systems. In particular, it is thought that more studies should be done to understand TE response to stresses and whether these represent specific adaptations. Of course, decades of work have been done in this area but in very few species.

Finally, one of the main questions posed in this dissertation relates to understanding if there are general mechanisms that determine the diversity of TEs in the genome. We see exceptional variation in TE diversity in the Asteraceae alone but what are the factors that determine TE richness and changes in the relative abundance of TEs? I have identified one likely cause of changes in abundance in my fourth chapter, which is whole genome duplication. It may also be true that clade age contributes to TE richness, as species at the base of the Asteraceae have a far greater number of TE families than those species in the crown lineages. Given the vast creative potential of TEs and their many contributions to plant genome evolution, understanding factors that contribute to TE diversification will reveal important mechanisms about both plant evolution and the factors that govern biodiversity.

LITERATURE CITED

Berendse F, and Scheffer M. 2009. The angiosperm radiation revisted, an explanation for Darwin's 'abominable mystery'. *Ecol Lett*. 12(9), 865–872.

Crepet WL. 1998. Botany: The abominable mystery. *Science* 282, 1653–1654

Darwin C. 1872. The origin of species by means of natural selection, or the preservation of favoured races in the struggle for life, 6th ed. John Murray, London, UK.

Darwin F. 1887. The life and letters of Charles Darwin, including an autobiographical chapter. John Murray, London, UK

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nature Rev. Gen.* 9, 397-405.

Flanagan JF, Mi L, Chruszcz M, Cymborowski M, Clines KL, Kim Y, Minor W, Rastinejad F, and Khorasanizadeh S. 2005. Double chromodomains cooperate to recognize the methylated histone H3 tail. *Nature,* 438, 1181-1185.

Flanagan JF, Blus BJ, Kim D, Clines KL, Rastinejad F, and Khorasanizadeh S. 2007. Molecular implications of evolutionary differences in CHD double chromodomains. *J. Mol. Biol.* 369, 334-342.

Friedman WE. 2009. The meaning of Darwin's "abominable mystery." *Am. J. Bot.* 96, 15-21.

Gao X, Hou Y, Ebina H, Levin HL, and Voytas DF. 2008. Chromodomains direct integration of retrotransposons to heterochromatin. *Genome Res.* 18, 359-369.

Gorinsek B, Gubensek F, and Kordis D. 2004. Evolutionary genomics of chromoviruses in eukaryotes. *Mol. Biol. Evol.* 21, 781-798.

Hawkins JS, Kim H, Nason JD, Wing RA, and Wendel JF. 2006. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. *Genome Res.* 16(10), 1252-1261.

Hawkins JS, Grover CE, Wendel JF. 2008. Repeated big bangs and the expanding universe: Directionality in plant genome size evolution. *Plant Science* 174, 557-562.

Hilbrict T, Varotto S, Sgaramella V, Bartels D, Salamini F, Furini A. 2008. Retrotransposons and siRNA have a role in the evolution of desiccation tolerance leading to resurrection of the plant Craterostigma plantagineum. *New Phyt.* 179, 877-887.

Kvikstad EM, Makova KD. 2010. The (r)evolution of SINE versus LINE distributions in primate genomes: Sex chromosomes are important. *Genome Res.* 20, 600-613.

Gregory, TR. 2005. Evolution of the genome. Elsevier, Inc.

Lisch D. 2009. Epigenetic regulation of transposable elements in plants. *Ann. Rev. Plant Biol.* 60, 43-66.

Ma J, and Bennetzen JL. 2004. Recent growth and divergence of rice nuclear genomes. *Proc. Natl Acad. Sci. U.S.A.* 101, 12404-12410.

Martın I and Llorens C. 2000. Ty3/Gypsy retrotransposons: description of new Arabidopsis thaliana elements and evolutionary perspectives derived from comparative genomic data. *Mol. Biol. Evol.* 17(7), 1040-1049.

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

Natali L, Cossu RM, Barghini E, Giordani T, Buti M, Mascagni F, Morgante M, Gill N, Kane NC, Rieseberg L, and Cavallini A. 2013. The repetitive component of the sunflower genome as shown by different procedures for assembling next generation sequencing reads. *BMC Genomics* 14, 686.

Novikov I, Smyshlyaev G, Novikova O. 2012. Evolutionary history of LTR retrotransposon chromodomains in plants. *Intl. Jour. Pl. Genomics* 874743.

Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, Collura K, et al. 2006. Doubling genome size without polyploidization: dynamics of retrotransposon-mediated genome expansions in Oryza australensis, a wild relative of rice. *Genome Res.* 16, 1262-1269.

Robello R, Horard B, Hubert B, and Vieira C. 2010. Jumping genes and epigenetics: Towards new species. *Gene* 454, 1-7.

Ungerer MC, Strakosh SC, and Zhen Y. 2006. Genome expansion in three hybrid sunflower species is associated with retrotransposon proliferation. *Curr. Biol.* 16, R872-R873.

Staton SE, Hartman Bakken B, Blackman B, Chapman M, Kane N, Tang S, Ungerer M, Knapp S, Rieseberg L, Burke J. 2012. The sunflower (Helianthus annuus L.) genome reflects a recent history of biased accumulation of transposable elements. *The Plant Jour.* 72, 142-153.

Xiao H, Jiang N, Schaffner E, Stockinger EJ, van der Knapp E. 2008. A retrotransposon mediated gene duplication underlies morphological variation of tomato fruit. *Science* 319, 1527-1530.

Xie D, Chen C, Ptaszek LM, Xiao S, Cao X, Fang F, Ng HH, Lewin HA, Cowan C, and Zhong S. 2010. Rewirable gene regulatory networks in the preimplantation embryonic development of three mammalian species. *Genome Res.* 20, 804-815.

APPENDIX FOR CHAPTER II

Text A1

BAC Assembly

For the four BAC clones sequenced on a Roche 454 sequencer at the U. of Georgia's Georgia Genomics Facility, sequences and corresponding quality scores from each flowgram produced by the Roche sequencer were sampled randomly with custom Perl scripts to obtain approximately 30x coverage which, based on our simulations, resulted in assemblies with the fewest contigs, highest N50 size, lowest per base error rate in assembled contigs, and the highest Q40 values. These BAC clones were first assembled with the gsAssembly software (v2.3) from Roche Life Sciences to assess the quality and coverage for each clone. MIRA (v3.0.3; Chevreux 1999) was used for final assemblies as it consistently produced fewer contigs and a greater total number of assembled bases while also allowing for a greater control of the assembly process by numerous command line switches. Each MIRA assembly was input to the minimus2 pipeline (Sommer *et al.*, 2007) to join fragmented assemblies, and the assemblies were viewed and edited with Tablet (Milne *et al.*, 2010) and Consed (Gordon *et al.*, 1998), respectively.

LTR retrotransposon (LTR-RT) and solo LTR identification

Primer binding site and domain identification is automated with LTR*digest* (version 1.3.4; Steinbiss *et al., 2*009) by searching each LTR-RT prediction against a database of sequences and running HMMER (version 2.3.2; Eddy 1998) on a set of Hidden Markov Models (HMMs) constructed from alignments of LTR-RT coding domains, respectively.

We used a set of 23 HMMs (both Pfam local models from the Pfam HMM_ls database and mulit-hit local models from the Pfam HMM_fs database for each domain were used for a total of 46 HMMs) and a custom database of 6120 plant tRNAs obtained from the genomic tRNA database (Chan and Lowe 2009) to identify primer binding sites.

Solo LTRs and truncated LTR-RTs were identified by masking the BAC sequences with intact (i.e. complete or full-length sequences) LTR-RTs with RepeatMasker (version Open-3.2.9; Smit *et al.,* 1996-2010), and then searching the masked BACs with HMMs created by aligning the two LTRs from a single retrotransposon and BLASTN with intact LTR-RTs with an e-value of 1e-5, respectively. Multi-hit local models representing the LTR sequence of each LTR-RT were used to identify fragments and multiple solo LTRs using HMMER version 2.3.2 (Eddy 1998), and only matches above 50% sequence identity to the entire model were retained as potential solo LTRs. Only fragments above 50 base pairs and with at least 80% sequence identity to an intact LTR-RT copy were considered as truncated copies.

The genome-wide frequency of solo LTRs and truncated elements vs. intact elements was determined by using similarity searches with BLAST. First, we used BLASTN with default parameters to search the WGS read set with both LTR sequences from the 79 intact *Gypsy* elements identified in the BAC clones. The same procedure was repeated for the 28 intact *Copia* elements (see Table S2 for more information about *Gypsy* and *Copia* families identified in this study). Next, the reverse transcriptase (RVT) protein sequence from each intact *Gypsy* element was used to search the WGS reads with TBLASTN using an e-value cutoff of 1e-5, and this procedure was repeated for each intact *Copia* element. We used a ratio of the total number of BLAST hits for LTR

sequences to the total number of BLAST hits for RVT sequences to assess the genome-wide percentage of intact LTR-RTs vs. truncated elements or solo LTRs. Given that the RVT and LTR sequences can be clearly differentiated between *Gypsy* and *Copia* elements, the ratio of LTR:RVT BLAST hits in the WGS reads provides an estimate of the relative genome-wide percentage of intact vs. truncated elements or solo LTRs for *Gypsy* and *Copia* elements.

Neighbor-joining analyses

The diversity and novelty of the sunflower *Gypsy* LTR retrotransposons that contained a chromodomain (i.e. the "chromoviruses") was assessed by comparing sequences identified in this study to previously identified chromoviruses from the Viridiplantae (Gorinsek *et al.*, 2004). The 19 reverse transcriptase sequences selected for comparison were chosen to represent all of the known clades of chromoviruses in plants. In addition, we identified chromodomain sequences for each putative chromovirus identified in this study using HMMER and analyzed the composition and structure of each domain. All sequences were aligned with MUSCLE (Edgar 2004) and alignments were edited manually with SeaView (Galtier *et al.*, 1996). The sequence composition for chromodomains was assessed using WebLogo (Crooks *et al.*, 2004) and displayed using Jalview (Waterhouse *et al.*, 2009). The putative secondary structure of the chromodomains was determined by aligning the sequences identified in this study to the chromodomain from Mouse Modifier Protein 1 (Ball *et al.,* 1997). Neighbor-Joining trees based on the reverse transcriptase sequences were constructed using the BioNJ algorithm (Gascuel 1997) with a Poisson model and 1000 bootstrap replicates. The chromovirus

Neighbor-Joining tree was rooted with the Osvaldo element (GenBank accession number AJ133521) from *Drosophila melanogaster*. All plots and neighbor-joining trees were rendered and edited with the R programming language (Paradis *et al.*, 2004; Wickman 2009; R Development Core Team 2011).

LITERATURE CITED

Chan PP, and Lowe TM. 2009. GtRNAdb: A database of transfer RNA genes identified in genomic sequences. *Nucleic Acids Res.* 37, D93-D97.

Chevreux B, Wetter T, and Suhai S. 1999. Genome sequence assembly using trace signals and additional sequence information. *Comp. Sci. Biol: Proc. German Conf. Bioinf. (GCB)* 99, 45-56.

Crooks GE, Hon G, Chandonia JM, and Brenner SE. 2004. WebLogo: A sequence logo generator. *Genome Res.* 14, 1188-1190.

Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics* 14, 755-763.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nuc. Acids Res.* 32, 1792–97.

Galtier N, Guoy M, and Gautier C. 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput. Appl. Biosci.* 12, 543-548.

Gascuel O. 1997. BioNJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* 14, 685-695.

Gordon D, Abajian C, and Green P. 1998. Consed: A graphical tool for sequence finishing. *Genome Res.* 8, 195-202.

Gorinse B, Gubensek F, and Kordis D. 2004. Evolutionary genomics of chromoviruses in eukaryotes. *Mol. Biol. Evol.* 21, 781-798.

Milne I, Bayer M, Cardle L, Shaw P, Stephan G, Wright F, and Marshall D. 2010. Tablet – next generation sequence assembly visualization. *Bioinformatics,* 26, 401-402.

Novak P, Neumann P, and Macas J. 2010. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinf.* 11, 378.

Paradis E, Claude J, and Strimmer K. 2004. APE: Analysis of Phylogenetics and

Evolution in R. *Bioinformatics,* 20, 289-290.

Sommer DD, Delcher AL, Salzberg SL, and Pop M. 2007. Minimus: a fast, lightweight genome assembler. *BMC Bioinf.* 8, 64.

Steinbiss S, Willhoeft U, Gremme G, and Kurtz S. 2009. Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Res.* 37, 7002-7013.

Waterhouse AM, Procter JB, Martin DBA, Clamp M, and Barton GJ. 2009. Jalview Version 2 – a multiple sequence alignment editor and workbench. *Bioinformatics* 25, 1189-1191.

Wickman, H. 2009. ggplot2: Elegant graphics for data analysis. *Springer Science*.

Table A2.1. Length statistics for BAC clone sequences.

BAC ID[1]	Length[2]
Contig2_c1666	45514
Contig4_c1666	30615
Contig1_c1236	13133
Contig5_c1236	13030
Contig1_c2588	38706
Contig2- a_c2588	28905
Contig3_c2588	29597
Contig1_c5898	10761
Contig4_c5898	6012
Contig6_c5898	8221
227-17	113591
P189P24	137647
P29D11	112328
P347K03	172652
P347L09	112473
Contig15_P35K12	55962
Contig29_P35K12	45458
P392C18	128671
P94O19	159960
127K04	112580
254L24	119871
Contig6_271N19	19133
Contig7_271N19	51993
P102A12	125640
Contig23_P245O15	16150
Contig25_P245O15	25104
Contig36_P245O15	71471
Contig37_P245O15	6023
P339N08	131960
P408L01	143555
P396I22	107161
07A15	105044
Total	2308222

1- The top eleven BAC IDs reflect the contig number and gene name (all gene names from Chapman *et al.,* 2008 Plant Cell) in the selected BAC clone. The remaining BAC IDs reflect the contig number, if there was more than one contig in the final assembly, and clone address for that BAC. 2- Lengths are represented in bp.

Table A2.2. Demography statistics for LTR retrotransposon families derived from BAC clone sequences. All families are indicated with the prefix RLG (for *Ty3*/*Gypsy*) or RLC (for *Ty1*/*Copia*) and RLG-X or RLC-X indicates the unclustered sequences (see Experimental Procedures).

Family	Count	Overall length[1]	LTR length[1,2]	Bases / percent of BACs	PBS motif	PPT motif	Domain organization	Solo:FL:TR[3]
RLC-X	11	8881	908	97687 / 3.79	Variable	Variable	Variable	1.36:1:0.09
RLC-ogaow	2	11154	612	22307 / 0.86	Variable	Variable	_____	0:1:0
RLC-suwi	2	7353	1527	14705 / 0.57	_____	aaggggggag(a)	_____	0:1:0
RLC-jiliwu	6	8701	563	52205 / 2.02	tatcagagcca(ag)	aaggggga(a|g)	INT-RT	0:1:0
RLC-amov	7	9542	579	66791 / 2.59	ta(t|a)cagagccg(g|t)(ag)	aaggggggag	INT-RT	0:1:0
All RLC	*28*	*9061*	*775*	*253695 / 9.86*				*0.53:1:0.03*
RLG-X	23	9741	1113	224048 / 8.71	Variable	Variable	Variable	0.52:1:0.4
RLG-ryse	2	12580	2407	25159 / 0.97	_____	aaggggggg	GAG-AP-RT-INT-Chromo	0:1:0
RLG-wimu	4	13652	2615	54608 / 2.12	gggcccaccgg	aaggggggtgagga	GAG-AP-RT-RH-INT-Chromo	0:1:0.5
RLG-kefe	4	10444	920	41774 / 1.62	tctgctaggaa	aaggataa	GAG-AP-RT-RH-INT-Chromo	0:1:0
RLG-begi	2	7856	1908	15712 / 0.61	_____	_____	GAG-RT-INT	0:1:0
RLG-tewuvu	3	9529	1417	28587 / 1.11	tatcagagcca	_____	GAG-AP-RT-RH-INT-Chromo	0:1:0
RLG-rewu	4	9283	604	37132 / 1.44	_____	ggacaagaaaaag	RT-INT	0:1:0
RLG-teda	2	13037	378	26073 / 1.01	tatcagagccagg	_____	_____	0:1:0
RLG-taoham	2	9719	2130	19437 / 0.75	tatcagagcca	_____	GAG-AP-RT-RH-INT-Chromo	0:1:0
RLG-wily	5	10567	2694	52835 / 2.05	tatcagagcca(a)	gggggggggg	GAG-AP-RT-RH-INT-Chromo	0:1:0.6
RLG-esuv	2	9125	1765	18249 / 0.70	tatcagagcca	_____	GAG-AP-RT-RH-INT-Chromo	0:1:0
RLG-rahi	4	10523	674	42093 / 1.63	cgcccaccgtggggc(ct)	_____	GAG-RT-RH-INT	0:1:0
RLG-iketas	22	8991	1831	197792 / 7.69	tatcagagcca(a|g)(g)	aaaggaggggaga	GAG-AP-RT-RH-INT-Chromo	0:1:0
All RLG	*79*	*9918*	*1551*	*783499 / 30.47*				*0.15:1:0.07*
Total	*107*	*9693*	*1346*	*1037194 / 40.33*				*0.14:1:0.06*

1 – Lengths are presented as averages.
2 – The average LTR lengths presented does not reflect solo LTRs, only LTRs from intact elements.
3 – Ratio of Solo LTRs (Solo) to full-length (FL) to truncated (TR) LTR retrotransposon copies.

Table A2.3. Statistics for putative events of illegitimate recombination for each LTR-RT family.

Family	Deletion number[1]	Deletion mean (stddev)	Deletion min	Deletion max	Deletion sum	HSP mean[2] (stddev)	HSP min	HSP max
RLC-amov	8	10.75 (0.71)	10	12	86	4.25 (0.46)	4	5
RLC-jiliwu	34	11.88 (1.39)	10	16	404	4.50 (0.46)	4	9
RLC-ogaow	13	11.62 (1.80)	10	15	151	4.08 (0.28)	4	5
RLC-suwi	10	11.30 (1.25)	10	14	113	4.60 (1.26)	4	8
RLC total	*65 | 16.25*	*11.38 (0.49)*	*10*	*14.25*	*188.50*	*4.36 (0.24)*	*4*	*6.75*
RLG-begi	5	12.60 (0.89)	11	13	63	4.49 (0.85)	4	6
RLG-esuv	1	13.00 (0.00)	13	13	13	4.00 (0.00)	4	4
RLG-iketas	136	12.01 (1.67)	10	16	1634	4.13 (0.44)	4	7
RLG-kefe	16	12.06 (1.53)	10	15	193	4.31 (0.60)	4	6
RLG-rahi	12	11.42 (1.62)	10	14	137	4.42 (0.67)	4	6
RLG-rewu	10	11.60 (1.26)	10	14	116	4.40 (0.52)	4	5
RLG-ryse	6	12.33 (1.97)	10	14	74	4.00 (0.00)	4	4
RLG-taoham	2	13.50 (2.12)	12	15	27	4.50 (0.71)	4	5
RLG-teda	0	NA	NA	NA	NA	NA	NA	NA
RLG-tewuvu	5	10.60 (1.34)	10	13	53	5.20 (1.64)	4	8
RLG-wily	16	11.88 (0.96)	10	14	190	4.25 (0.77)	4	7
RLG-wimu	4	11.50 (1.00)	10	12	46	4.00 (0.00)	4	4
RLG total	*213 | 17.75*	*11.04 (3.56)*	*9.67*	*12.75*	*212.17*	*4.34 (0.35)*	*4*	*5.17*
Total	*278 | 17.38*	*11.13 (3.06)*	*10.34*	*13.13*	*206.25*	4.07 (1.13)	4	5.56

[1]Totals reflect (sum| mean) for gap number. [2]HSP = High-scoring segment pairs. Statistics in the final three columns describe the nature of direct repeats flanking deletions.

APPENDIX FOR CHAPTER III

Table A3.1. Phylogenetic signal for TE families in the Asteraceae. Shown are TE families exhibiting significant phylogenetic signal as compared to a Brownian motion model of evolution along the phylogenetic tree. For each TE family, we demonstrate the observed PIC (phylogenetic independent contrast) scores and significance value, along with the value of the random PIC scores.

Order	Superfamily	Family	K	PIC variance (observed)	PIC variance (random mean)	PIC variance *P-value*	PIC variance *Z-value*
LTR-RT	Copia	COPIA2_LC	3.257875889	5.14E+15	3.67E+16	0.039	-0.909125888
LTR-RT	Copia	COPIA2_MT	3.30773527	1.17E+16	8.03E+16	0.008	-0.878074852
LTR-RT	Copia	Copia15_VV	1.302072147	7.08E+13	3.73E+14	0.018	-2.10973762
LTR-RT	Copia	RLC_X	2.663159151	2.05E+17	1.26E+18	0.019	-0.89629912
LTR-RT	Copia	RLC_amov	0.710946644	2.98E+16	7.74E+16	0.037	-1.907217855
LTR-RT	Copia	RLC_jiliwu	0.520839198	1.07E+17	3.06E+17	0.018	-1.210112268
LTR-RT	Copia	RLC_ogaow	0.938232015	7.84E+15	2.94E+16	0.022	-2.438721569
LTR-RT	Gypsy	RLG_X	0.703019737	2.05E+17	1.26E+18	0.019	-0.89629912
LTR-RT	Gypsy	RLG_kefe	1.076403493	2.52E+16	9.59E+16	0.002	-2.767755405
LTR-RT	Gypsy	RLG_rewu	0.726242862	1.10E+17	3.30E+17	0.049	-2.032258579
LTR-RT	Gypsy	RLG_ryse	0.661494372	1.95E+15	4.56E+15	0.025	-1.933088881
LTR-RT	Gypsy	RLG_teda	1.507538406	1.25E+16	7.55E+16	0.006	-3.239986584
LTR-RT	Gypsy	RLG_tewuvu	0.850900791	7.53E+16	3.50E+17	0.011	-0.931314665
LTR-RT	Gypsy	DM176	3.261186883	3.34E+11	2.18E+12	0.039	-0.836834688
LTR-RT	Gypsy	GYPSY16_AG	3.261186883	3.34E+11	2.18E+12	0.036	-0.848736222
LTR-RT	Gypsy	Gypsy123_DR	3.239593367	1.34E+12	9.04E+12	0.0495	-0.881410277
LTR-RT	Gypsy	Gypsy1_SM	0.870785834	9.46E+11	5.12E+12	0.043	-1.03495498
LTR-RT	ERV1	ERV1_N6_DR	3.210894013	3.43E+14	2.29E+15	0.0335	-0.875115482
Non-LTR-RT	L1	L1_11_DR	2.79895668	2.60E+13	1.63E+14	0.01	-1.042094132
Non-LTR-RT	L1	L1_12_DR	3.261186883	3.34E+11	2.41E+12	0.0355	-0.928489469
Non-LTR-RT	L1	L1_58_ACar	3.261186883	5.35E+12	3.61E+13	0.0395	-0.866658002
Non-LTR-RT	NeSL	LIN4b_SM	3.261186883	3.34E+11	2.15E+12	0.0365	-0.832815617
Non-LTR-RT	CR1	CR1_13_CQ	3.261186883	3.34E+11	2.23E+1	0.0385	-0.868749537
Non-LTR-RT	CR1	CR1_58_HM	3.261186883	3.34E+11	2.23E+12	0.036	-0.872823223
Non-LTR-RT	CR1	CR1_79_HM	3.261186883	2.14E+13	1.46E+14	0.0325	-0.870677374
Class II	hAT	P4_AG	3.261186883	3.34E+11	2.30E+12	0.0255	-0.892685921
Class II	Mariner/Tc1	SMAR15	3.261186883	3.01E+12	1.99E+13	0.0375	-0.854157409
Class II	Mariner/Tc1	ATHPOGON1	3.261186883	1.21E+14	7.69E+14	0.0355	-0.830930169
Class II	Mariner/Tc1	Helitron3_PPa	3.261186883	5.35E+12	3.45E+13	0.0405	-0.841984529

Table A3.2. Raw data statistics and genome size estimates. For each species in this study, we show the number of sequence reads generated and the corresponding genome coverage obtained from the genome size estimates (see Methods).

Subfamily	Tribe	Genus	Species	Genome size [1]	Num. reads/ Genome cov.
Calyceraceae	Calyceraceae	*Nastanthus*	*patagonicus*	3962340892	22733114/0.58
Barnadesioideae	Barnadesieae	*Fulcaldea*	*stuessyi*	4182557218	92343086/2.23
Mutisioideae	Mutisieae	*Gerbera*	*hybrida*	3861919879	19128428/0.50
Carduoideae	Cardueae	*Carthamus*	*tinctorius*	2405291468	18020913/0.76
Cichorioideae	Cichorieae	*Taraxacum*	*kok-saghyz*	2582325776	21388100/0.84
Cichorioideae	Vernonieae	*Centrapalus*	*pauciflorus*	3125365235	19627573/0.63
Cichorioideae	Senecioneae	*Senecio*	*vulgaris*	2045909989	15732065/0.78
Cichorioideae	Gnaphalieae	*Pseudognaphalium*	*obtusifolium*	2920317131	13066952/0.45
Asteroideae	Eupatorieae	*Conoclinium*	*coelestinum*	1746269472	20943700/1.21
Asteroideae	Heliantheae	*Phoebanthus*	*tenuifolius*	4267295897	148630586/3.52
Asteroideae	Heliantheae	*Helianthus*	*porteri*	4330738740	18192388/0.42
Asteroideae	Heliantheae	*Helianthus*	*verticillatus*	2278002736	18560744/0.82
Asteroideae	Heliantheae	*Helianthus*	*niveus ssp. tephrodes*	4192677026	23420666/0.56
Asteroideae	Heliantheae	*Helianthus*	*argophyllus*	4174346891	25568942/0.62
Asteroideae	Heliantheae	*Helianthus*	*annuus*	3384161947	22621880/0.68

[1]base pairs

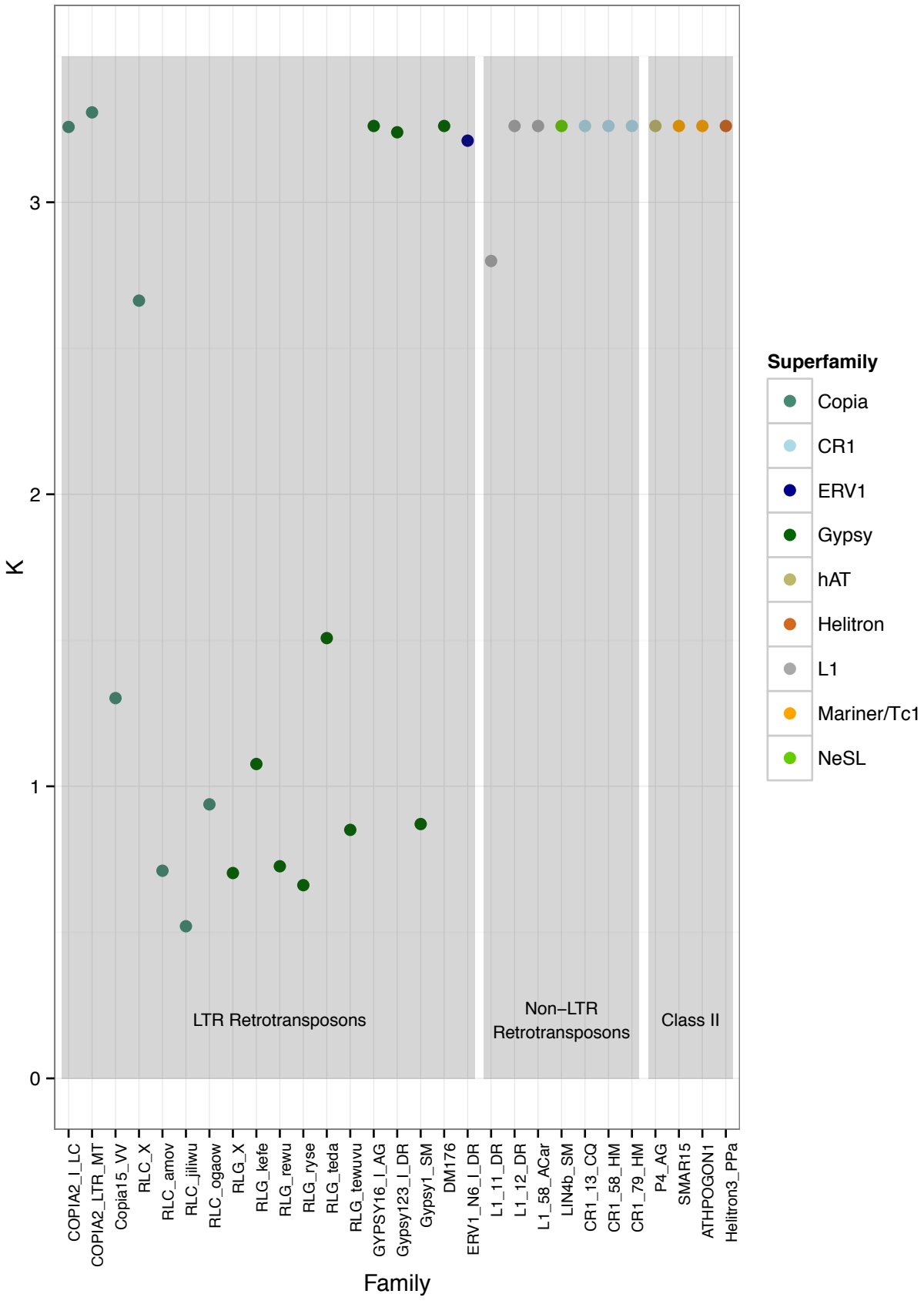

K
0
1
2
3
LTR Retrotransposons
Non−LTR Retrotransposons
Class II
COPIA2_I_LC
COPIA2_LTR_MT
Copia15_VV
RLC_X
RLC_amov
RLC_jiliwu
RLC_ogaow
RLG_X
RLG_kefe
RLG_rewu
RLG_ryse
RLG_teda
RLG_tewuvu
GYPSY16_I_AG
Gypsy123_I_DR
Gypsy1_SM
DM176
ERV1_N6_I_DR
L1_11_DR
L1_12_DR
L1_58_ACar
LIN4b_SM
CR1_13_CQ
CR1_58_HM
CR1_79_HM
P4_AG
SMAR15
ATHPOGON1
Helitron3_PPa
Family
Superfamily
Copia
CR1
ERV1
Gypsy
hAT
Helitron
L1
Mariner/Tc1
NeSL

Figure A3.1. TE families exhibiting significant phylogenetic signal. Along the x-axis are TE families in alphabetical order (divided by order, which is indicated by gray boxes) exhibiting significant phyologenetic signal (y-axis).

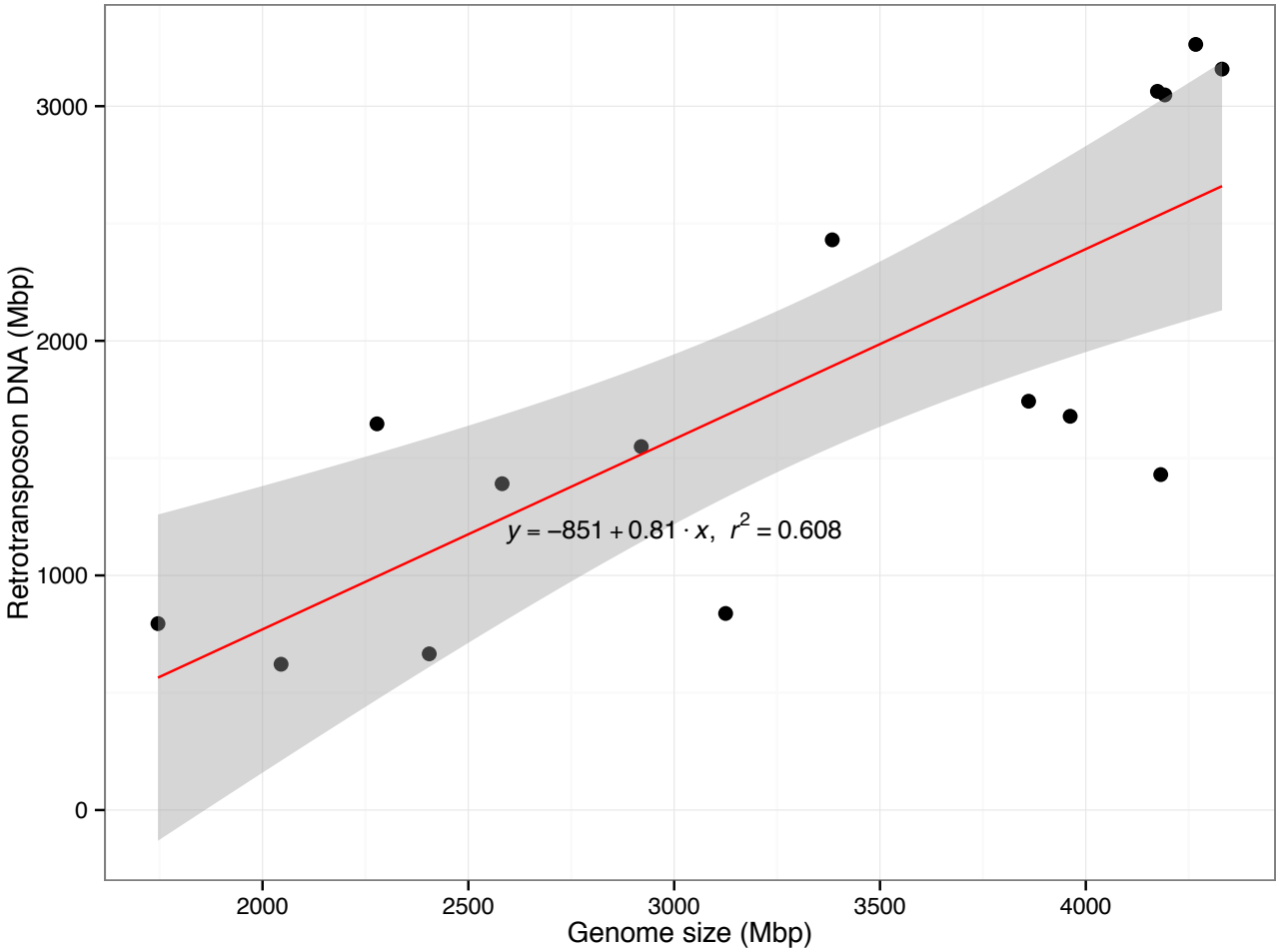
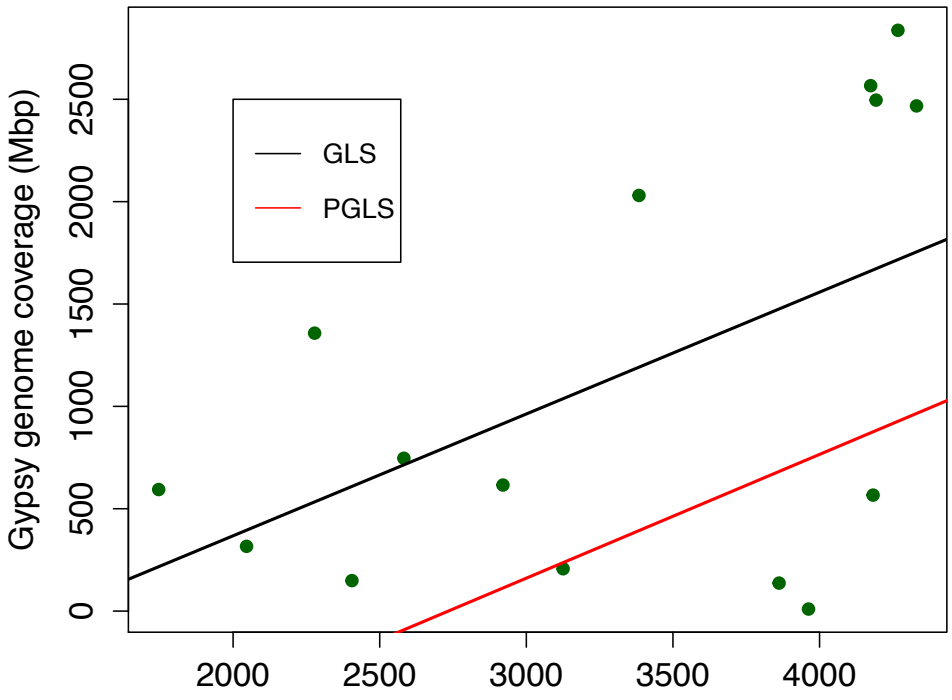


Figure A3.2. Relationship between retrotransposon DNA and genome size. The total amount of retrotransposon base pairs (y-axis) correlates very strongly with genome size (y-axis) in the Asteraceae.


Gypsy genome coverage (Mbp)
0
500
1000
1500
2000
2500
2000
2500
3000
3500
4000
GLS
PGLS

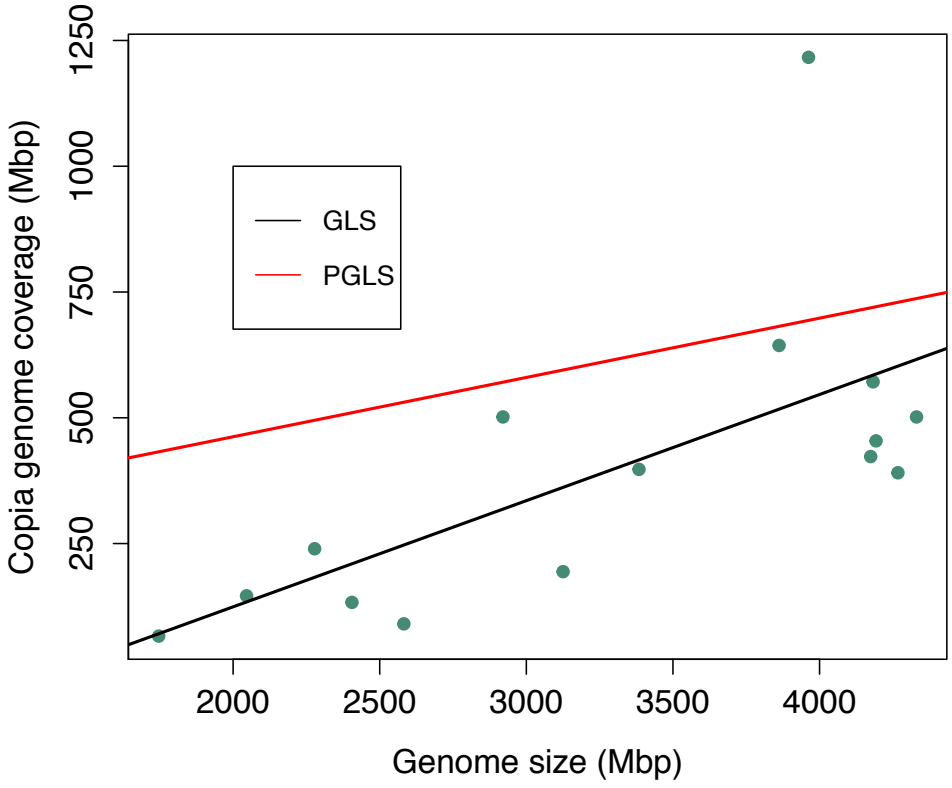

Copia genome coverage (Mbp)
250
500
750
1000
1250
2000
2500
3000
3500
4000
Genome size (Mbp)
GLS
PGLS

Figure A3.3. GLS and PGLS tests for the evolution of *Gypsy* and *Copia* composition. The genomic composition (y-axis) of A) *Gypsy* and B) *Copia* TEs correlates strongly with genome size (x-axis) as shown by the GLS fit (black line), even when considering the phylogenetic relatedness of the species with a PGLS test (red line).

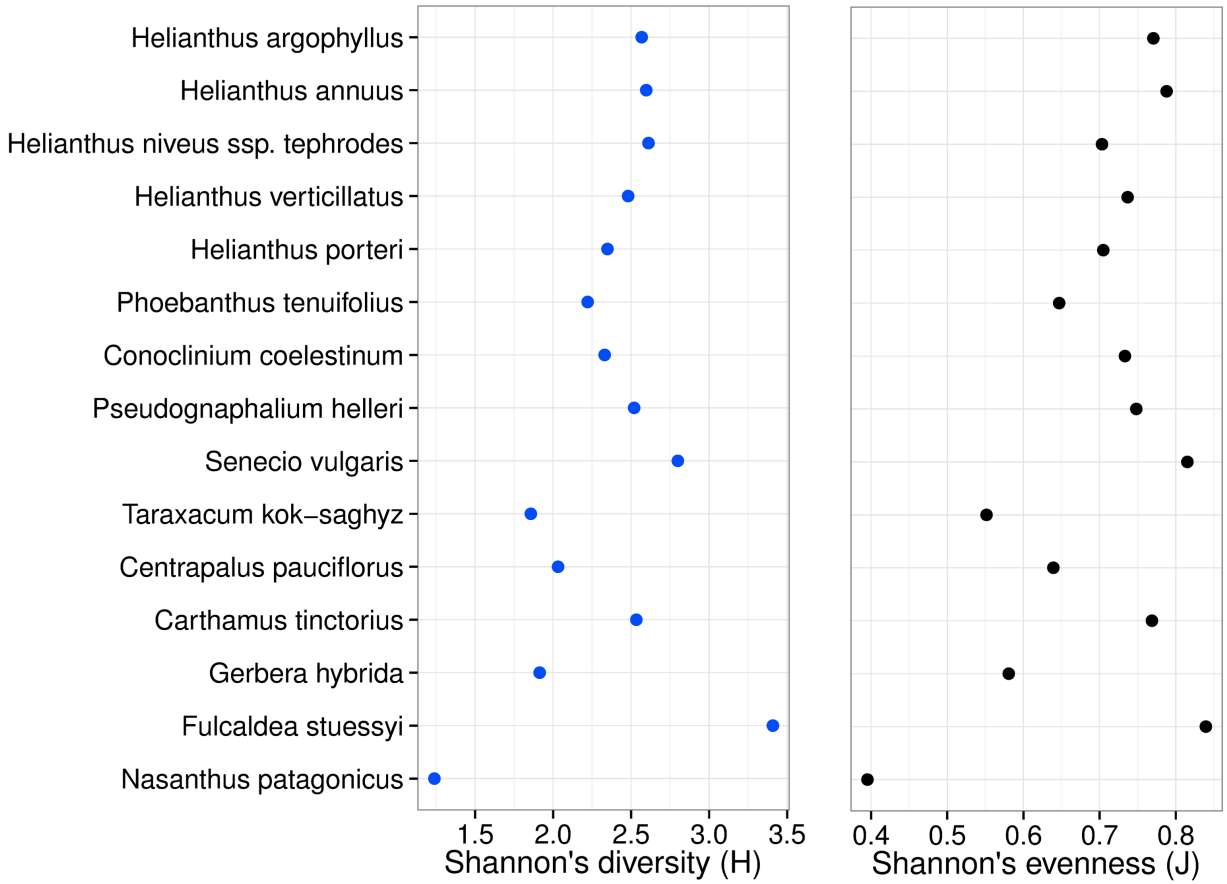

Helianthus argophyllus
Helianthus annuus
Helianthus niveus ssp. tephrodes
Helianthus verticillatus
Helianthus porteri
Phoebanthus tenuifolius
Conoclinium coelestinum
Pseudognaphalium helleri
Senecio vulgaris
Taraxacum kok−saghyz
Centrapalus pauciflorus
Carthamus tinctorius
Gerbera hybrida
Fulcaldea stuessyi
Nasanthus patagonicus
1.5
2.0
2.5
3.0
3.5
Shannon's diversity (H)
0.4
0.5
0.6
0.7
0.8
Shannon's evenness (J)

Figure A3.4. Genome diversity statistics for TE families. The species shown along on the y-axis are in phylogenetic order from the outgroup (base of the y-axis) to the most derived lineages of the Asteraceae (top of the y-axis). The filled blue points are Shannon's diversity (H), and the black points show Shannon's evenness (J).

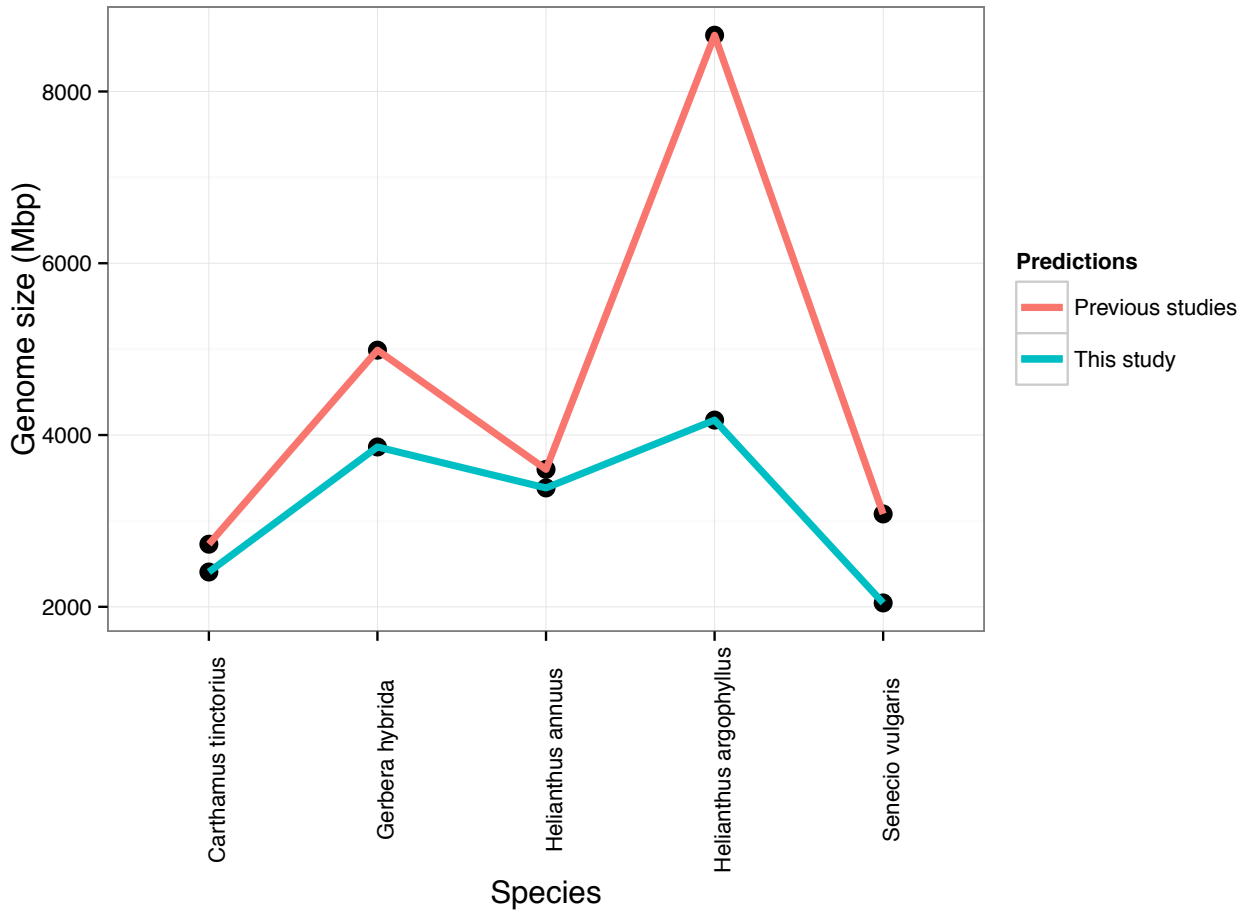


Figure A3.5. Published genome size estimates and genome size observations determined by the method described in this study. Along the x-axis are species for which published genome size estimates were available (obtained from: http://data.kew.org/cvalues). The y-axis shows genome size in mega-base pairs.