UC Davis
UC Davis Previously Published Works

Title

The genomic ecosystem of transposable elements in maize

Permalink

Journal

ISSN

Authors

Stitzer, Michelle C
Anderson, Sarah N
Springer, Nathan M
et al.

Publication Date

DOI

Peer reviewed

PLOS GENETICS

RESEARCH ARTICLE

The genomic ecosystem of transposable elements in maize

Michelle C. Stitzer[1]*, Sarah N. Anderson[2], Nathan M. Springer[2], Jeffrey Ross-Ibarra[1,3]

1 Center for Population Biology and Department of Evolution and Ecology, University of California, Davis, California, United States of America, **2** Department of Plant and Microbial Biology, University of Minnesota, Saint Paul, Minnesota, United States of America, **3** Genome Center, University of California, Davis, California, United States of America

* mcstitzer@ucdavis.edu

Abstract

Transposable elements (TEs) constitute the majority of flowering plant DNA, reflecting their tremendous success in subverting, avoiding, and surviving the defenses of their host genomes to ensure their selfish replication. More than 85% of the sequence of the maize genome can be ascribed to past transposition, providing a major contribution to the structure of the genome. Evidence from individual loci has informed our understanding of how transposition has shaped the genome, and a number of individual TE insertions have been causally linked to dramatic phenotypic changes. Genome-wide analyses in maize and other taxa have frequently represented TEs as a relatively homogeneous class of fragmentary relics of past transposition, obscuring their evolutionary history and interaction with their host genome. Using an updated annotation of structurally intact TEs in the maize reference genome, we investigate the family-level dynamics of TEs in maize. Integrating a variety of data, from descriptors of individual TEs like coding capacity, expression, and methylation, as well as similar features of the sequence they inserted into, we model the relationship between attributes of the genomic environment and the survival of TE copies and families. In contrast to the wholesale relegation of all TEs to a single category of junk DNA, these differences reveal a diversity of survival strategies of TE families. Together these generate a rich ecology of the genome, with each TE family representing the evolution of a distinct ecological niche. We conclude that while the impact of transposition is highly family- and context-dependent, a family-level understanding of the ecology of TEs in the genome can refine our ability to predict the role of TEs in generating genetic and phenotypic diversity.

Author summary

Transposable elements (TEs) are pieces of DNA that can jump to new positions in the genome. When they land at a new location, they generate a mutation. Such mutations in genes affecting kernel and plant pigmentation allowed the discovery of TEs in maize in the 1940's. Since, we have learned that TEs are a ubiquitous feature of eukaryotic genomes, and that TEs make up over 85% of all the DNA in a maize genome. Here, we investigate

the roles of individual TE copies and TE species interacting within the maize genome, and how these relationships are analogous to ecological communities. The community of transposable elements within a single genome represents a rich diversity of ecological strategies for survival in the complex, hostile environment of a host genome.

Introduction

'Lumping our beautiful collection of transposons into a single category is a crime'

-Michael R. Freeling, Mar. 10, 2017

Transposable elements (TEs) are pieces of DNA that can replicate or move themselves within a genome. The majority of DNA in plant genomes is TE derived, and their activity is the largest contributor to differences in genome size within and between taxa [1]. When they transpose, TEs also generate mutations as they insert into novel positions in the genome [2, 3]. These two linked processes—that of replication of the TE, and mutation suffered by the host genome—generate a conflict between individual lineages of TEs and their host genome. Individual TE lineages gain evolutionary advantage by increasing in copy number, while the host genome gains fitness if it can reduce deleterious mutations arising from transposition. As a result of this conflict, many genomes are littered with a bulk of TE-derived DNA that is often both transcriptionally and recombinationally inert [4]. While this conflict between TEs and their host has long been noted to shape general patterns of TE evolution [5–8], the details of how this conflict unfolds are tenuous and rarely well understood [9, 10].

The staggering diversity of TEs presents a major challenge in understanding their conflict with the host genome. For example, although they are united by their ability to move between positions in the host genome, the mechanisms by which TEs do so differ between the major TE classes. Class I retrotransposons, often the major contributor of TE DNA in plants [11], can be further divided into three orders—long terminal repeat (LTR), long interspersed nuclear element (LINE), and short interspersed nuclear element (SINE). All class I TEs are transcribed to mRNA by host polymerases, some are translated to produce reverse transcriptase and other enzymes, and all use TE encoded enzymes for reverse transcription of a cDNA copy that can be integrated at a new position in the host genome. In contrast, the two major orders of class II DNA TEs transpose in different ways. TIR elements are physically excised from one position on the chromosome and moved by TE-encoded transposase proteins that recognize short, diagnostic, terminal inverted repeats (TIRs). Helitron elements are thought to transpose via a rolling circle mechanism that generates a new copy after a single strand nick by an element-encoded protein and subsequent strand invasion and repair [12]. The process of transposition for most TEs (all LTR, TIR; some LINE, SINE) generates a target site duplication (TSD) in the host DNA at the integration site, and thus the identification of a TSD bordering a TE can confirm transposition. These well-described mechanisms of transposition generate predictable sequence organization that can be recognized computationally, but also generate differences in the genomic localization of these elements, via enzymatic site preference of TE encoded proteins [13, 14]. Most of these orders are further subdivided into TE superfamilies based on differences in the sequence, arrangement, and function of proteins encoded by the TE to ensure its transposition [15].

The process of transposition generates new TE copies within a genome, forming relationships between TEs that allow their systematic grouping into families. Many taxonomic

schemes for TEs exist [15–19], but the most widely-applied approach for genome-scale data [15] relies on sequence homology between copies. Although not entirely representative of TE evolutionary history [20, 21], such approaches nonetheless reflect to some degree the ability of TE encoded proteins to bind TE DNA and move other TE copies in *trans*, as recognition of specific nucleic acid sequences by TE-encoded proteins is a necessary step in the transposition process. The resulting TE families thus represent groups of related TEs that share both evolutionary history and transposition machinery, and are the groupings most naturally analogous to species in higher eukaryotes.

TE families differ from one another in many ways, including their total copy number, where they insert in the genome, which tissues they are expressed in, and how they are restricted epigenetically by the host genome. In the maize genome, some families are small, consisting of a few (e.g. *Bs* [22]) or tens (e.g. *Ds1* [23]) of copies, while others contain tens of thousands of copies (e.g. *huck*, *cinful-zeon* [24–28]). Some TE families are expressed in certain tissues (e.g. *Misfit* [29]), while others are expressed more broadly across many tissues (e.g. *cinful* [29]). Some families preferentially insert into genic regions (e.g. *Mu1* [30]), others in the centromere (e.g. *CRM1* [31]). And while some families lack DNA methylation, others are methylated across the entire body of the TE, and yet others act to spread methylation outwards into flanking sequences [32, 33].

Although the major classes of TEs are found across taxa, their relative abundances differ [34] and there is no clear consensus as to the factors that explain the diversity of TEs within a genome [35–38]. One approach to understand the diversity of TEs is to consider the genome as a community and apply principles of community ecology to understand their distribution and abundance [39]. Initially proposed in terms of a dichotomy between TEs that have specialized in heterochromatic or euchromatic niches [7], thoughts about the ecology of the genome have been refined into a continuum of space, with different TE lineages existing in different genomic niches [8, 39, 40]. Empirical descriptions of TEs in a community ecology context, however, have been limited to a few families [41–43]. The genome reflects the interface between ecological and evolutionary processes, as TEs alter their environment by inserting. This in turn affects how TEs evolve and adapt—making the genome a system to explore the interface between ecology and evolution [44–46]. Due to the long time scales over which TEs can persist in genomes, distinguishing whether processes occurring within the genome reflect ecological or evolutionary time scales can be difficult, although the two can be separated [47].

Here, we utilize the genomic ecosystem as a framework to describe patterns we observe in the extant maize genome. We take advantage of the diversity of TEs in maize, the record of past transposition still detectable in the genome, and the rich developmental and tissue-specific resources of maize to investigate the family-level ecological and evolutionary dynamics of TEs in maize. We integrate many metrics that can be measured at the level of TE family to present a natural history of TEs in the B73 maize genome, characterizing and describing the genomic features that differentiate superfamilies and families of TEs. We model survival of individual copies and families in the genome to facilitate an understanding of the complex and interactive strategies TEs use to associate with their host and each other, and identify suites of traits that act to define specific genomic niches and survival strategies. We conclude that understanding the diversity of TEs in the maize genome helps not only to describe TE function, but also that of the host genome.

Results

General features of TE orders and superfamilies

We identified members of each of the 13 superfamilies of transposable elements (TEs) previously identified in plants [15] in our structural annotation of the maize B73 reference genome.

This annotation resolves nested insertions of TEs within other elements, resulting in a total of 143,067 LTR retrotransposons (RLC (*Ty1/Copia*), RLG (*Ty3/Gypsy*), and RLX (*Unknown LTR*) superfamilies), 1,640 LINE and SINE (nonLTR) retrotransposons (RIL, RIT, and RST superfamilies), 171,570 TIR transposons (DTA (*hAT*), DTC (*CACTA*), DTH (*Pif/Harbinger*), DTM (*Mutator*), DTT (*Tc1/Mariner*), and DTX (*Unknown TIR*) superfamilies), and 22,234 Helitrons (DHH superfamily) (Table 1 and Fig 1A). We determined the number of families, median length, median age, distance to the nearest gene, and the number of base pairs each superfamily contributes to the genome (Fig 2; Interactive distributions per family: https://mcstitzer.shinyapps.io/maize_te_families/). For each family and superfamily, we determined the proportion of elements that are nested within another TE and the proportion of elements that are split into multiple pieces by other TE insertions.

Even at the broad taxonomic level of order, there are considerable differences among TEs. Because of their size, (median length 8.4 kb; Fig 2A and S1(C) Fig) LTR retrotransposons contribute more total base pairs to the genome (1,363 Mb; Fig 1B) and are commonly disrupted by another TE copy ($\approx \frac{2}{3}$ disrupted; S1(D) Fig). LTR retrotransposons are also typically far from genes (median distance 16.4 kb; Fig 2B, only 3.5% within a gene transcript; S1(A) Fig, median distance to a syntenic gene 31.9 kb; S1(B) Fig) and $\approx \frac{1}{2}$ of copies insert into a preexisting TE copy (Fig 2C). The median time since insertion of LTR retrotransposons is 315,000 years (Fig 2D). In contrast, despite having more copies (Table 1), TIR elements contribute fewer base pairs to the genome (74.1 Mb) and are rarely disrupted by the insertion of another TE copy ($< 5\%$ disrupted) (S1(D) Fig), presumably due to their much smaller size (median length 306 bp; Fig 2A and S1(C) Fig). TIR elements as a group are slightly further from genes (median distance 17.2 kb; Fig 2B, 1.7% within a gene transcript; S1(A) Fig, median distance to a syntenic gene 29.0 kb, S1(B) Fig), and commonly insert into preexisting TE copies ($\approx 70\%$ of copies; Fig 2C). They represent the most recent insertions, with a median age of 185,000 years (Fig 2D). Although Helitron elements are fewer in number than TIR elements, they contribute more base pairs to the genome (93.8 Mb) and are more commonly disrupted by the insertion of another TE ($\approx \frac{1}{4}$ of copies; S1(D) Fig) due to their increased length (median length 2.4 kb). Helitrons are also closer to genes than TIR elements (median distance 10.4 kb; Fig 2B, with 22.9% overlapping a gene transcript S1(A) Fig; median distance 25.4 kb from a syntenic gene; S1(B) Fig), and less frequently insert into a preexisting copy (50% of copies are found within another TE; Fig 2C). Helitrons are represented by relatively old copies, with a median age of 500,000 years (Fig 2D). NonLTR retrotransposons (LINEs and SINEs) contribute only 2.9 Mb, are relatively short (median length 548 bp), and only 5% of copies are disrupted by the insertion of another TE (S1(D) Fig). LINEs and SINEs are however often close to genes (median distance 2.3 kb; Fig 2B, 18.6% in a gene transcript; S1(A) Fig, median distance to a syntenic gene 10.1 kb; S1(B) Fig) and only 37% insert into another TE copy (Fig 2C). These nonLTR retroelements have a median age of 350,000 years (Fig 2D).

Within these orders, variation also exists among superfamilies (Fig 2). For example, TE superfamilies are found nonuniformly along chromosomes (Fig 3 and S2 Fig): while some superfamilies like RLG (Ty3/Gypsy) and DTC (CACTA) are enriched in centromeric and pericentromeric regions, others, like RLC (Ty1/Copia) and DTA (hAT) are found more commonly on chromosome arms. As maize genes are enriched on chromosome arms, this distribution is reflected in the distance each superfamily is found from genes (Fig 2B). Similarly, while most TIR superfamilies are found far from genes (median 17.2 kb), DTM (Mutator) elements are only a median distance of 2.4 kb away from genes (Fig 2B). And although TIR elements are often short (median 311 bp), DTC elements have a median length of 2886 base pairs (Fig 2A).

Table 1. TE superfamilies in the maize genome.

Class	Order	Superfamily	Common Name	Number Copies	Number Families
DNA transposon	Helitron	DHH	Helitron	22,339	1,722
DNA transposon	TIR	DTA	hAT	5,096	275
DNA transposon	TIR	DTC	CACTA	2,768	73
DNA transposon	TIR	DTH	Pif/Harbinger	63,216	458
DNA transposon	TIR	DTM	Mutator	928	67
DNA transposon	TIR	DTT	Tc1/Mariner	67,533	269
DNA transposon	TIR	DTX	Unknown TIR	34,778	76
Retrotransposon	LTR	RLC	Ty1/Copia	46,553	2,788
Retrotransposon	LTR	RLG	Ty3/Gypsy	75,761	7,719
Retrotransposon	LTR	RLX	Unknown LTR	20,789	13,290
Retrotransposon	nonLTR	RIT	RTE	296	2
Retrotransposon	nonLTR	RIL	L1	477	29
Retrotransposon	nonLTR	RST	SINE	892	533

Features of TE families

These descriptive statistics measured at the order and superfamily level are an aggregate across many TE families. TE families are defined based on sequence homology between copies [48], using a 80% sequence similarity cutoff described in Wicker et al. (2007) [15]. This results in thousands of families of LTR retrotransposon and Helitron elements, and hundreds of families of DNA TIR elements (Table 1). Although the majority of all TE families have fewer than ten copies (Fig 1A), the largest LTR retrotransposon and Helitron families in the genome consist of thousands of copies. Consistent with previous analyses built on subsets of maize bacterial
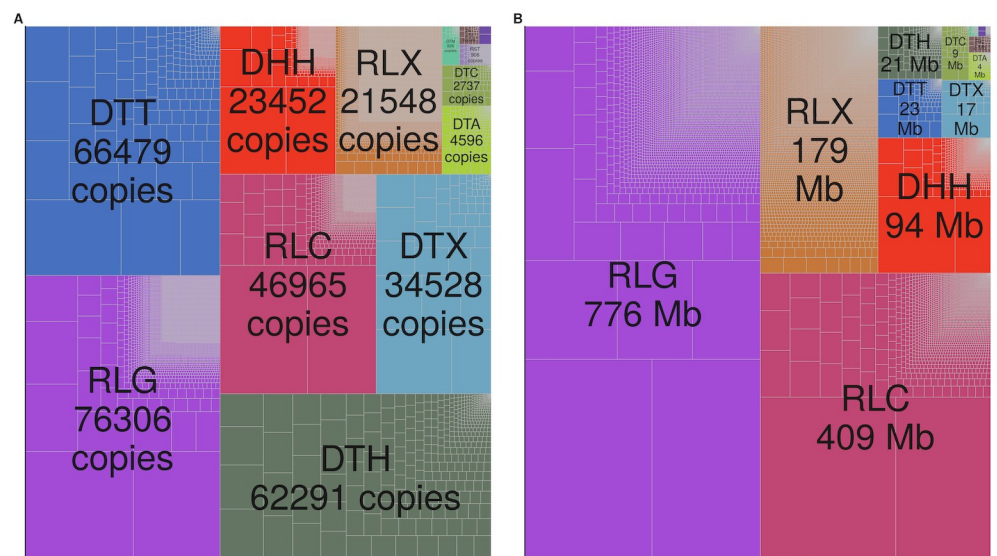
Fig 1. Abundance of TEs. The relative copy number (A) and size in million base pairs (Mb) (B) of families and superfamilies shown by the size of the rectangle. Superfamilies are denoted by color, and each family is bounded by gray lines within the superfamily. Superfamily names begin with a two letter code: 'DT' belong to the order of Terminal Inverted Repeat transposons, 'DH' refers to the order Helitron, 'RL' belong to the order Long Terminal Repeat retrotransposons, and 'RI' and 'RS' are nonLTR retrotransposons (LINEs and SINEs). Superfamily names beginning with 'D' are Class II DNA transposons, while those starting with 'R' are Class I retrotransposons.
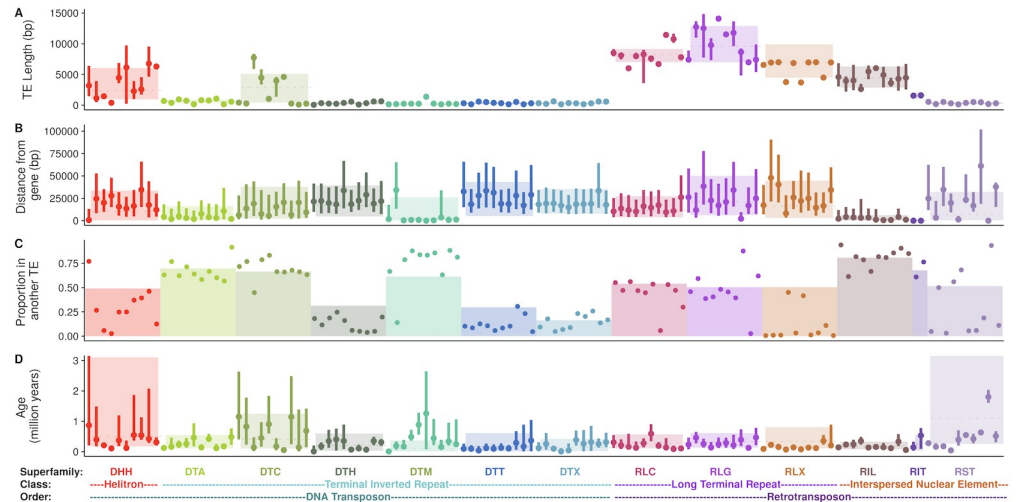
Fig 2. Characteristics of each superfamily of TE. Superfamilies are classified into orders and classes, as shown at the bottom of the plot. (A-D) Family characteristics of each of the most numerous 10 families (with ≥ 10 copies) of each superfamily. Family names are listed in S1 Table. (A) TE length, (B) Distance to the closest gene, (C) proportion of TE copies found within another TE, and (D) TE age. In (A, B, & D) family medians are shown as points, with lines representing upper to lower quartiles. Superfamilies are shown as colored rectangles, where the dotted line reflects the median and box boundaries reflect lower and upper quartiles. In (C), families are shown as points and superfamily proportions as a barplot.

https://doi.org/10.1371/journal.pgen.1009768.g002

artificial chromosomes (BACs) [26, 49], a majority (75%) of maize LTR retrotransposon families are present only as a single copy in the B73 genome. The average LTR family contains 6.1 copies, with this distribution ranging from 1 to 16,289 copies. In contrast, the family size distribution of TIR transposons is more uniform, with the average family containing 142 elements (range 1 to 9953) and only 10% of families are represented by a single copy. Helitron families are smaller, with 14 copies on average (66% represented by a single copy), and nonLTR retrotransposon families have an average of 3 copies (77% consist of a single copy).

Families are also found nonuniformly along chromosomes (Fig 3B, 3C, 3D and 3E and S3 Fig). Sometimes, the distribution of copies in the largest families in a superfamily match the pattern seen when summarized across all members of a superfamily, such as the five largest RLC families which all share an enrichment on chromosome arms (Fig 3D). There are also families that differ from the aggregate superfamily distribution. For example, the second largest RLG family (RLG00003) is enriched on chromosome arms, and the third largest RLG family (RLG00005) is more uniformly distributed along the chromosome (Fig 3E).

Further, the ages of different TE families vary greatly as well (Fig 4 and S4 Fig). We determine ages of individual copies based on terminal branch lengths of TE phylogenetic trees. For LTR retrotransposons, we additionally measure divergence between the two LTRs of each insertion (See Methods). Some families have not had a new insertion in the last 100 kya, while others have expanded rapidly in that time frame (Fig 4B, 4C, 4D and 4E). Some families display cyclical dynamics, readily generating new insertions that are retained, with pulses of stasis in between (e.g. DTA00073, Fig 4C). Others show sustained activity in the past (e.g. DHH00004, Fig 4B). In total, 70% of TIR families, 20% of LTR families (estimated with LTR-LTR divergence), 15% of nonLTR families and only 7% of Helitron families have been active in the last 100 kya.
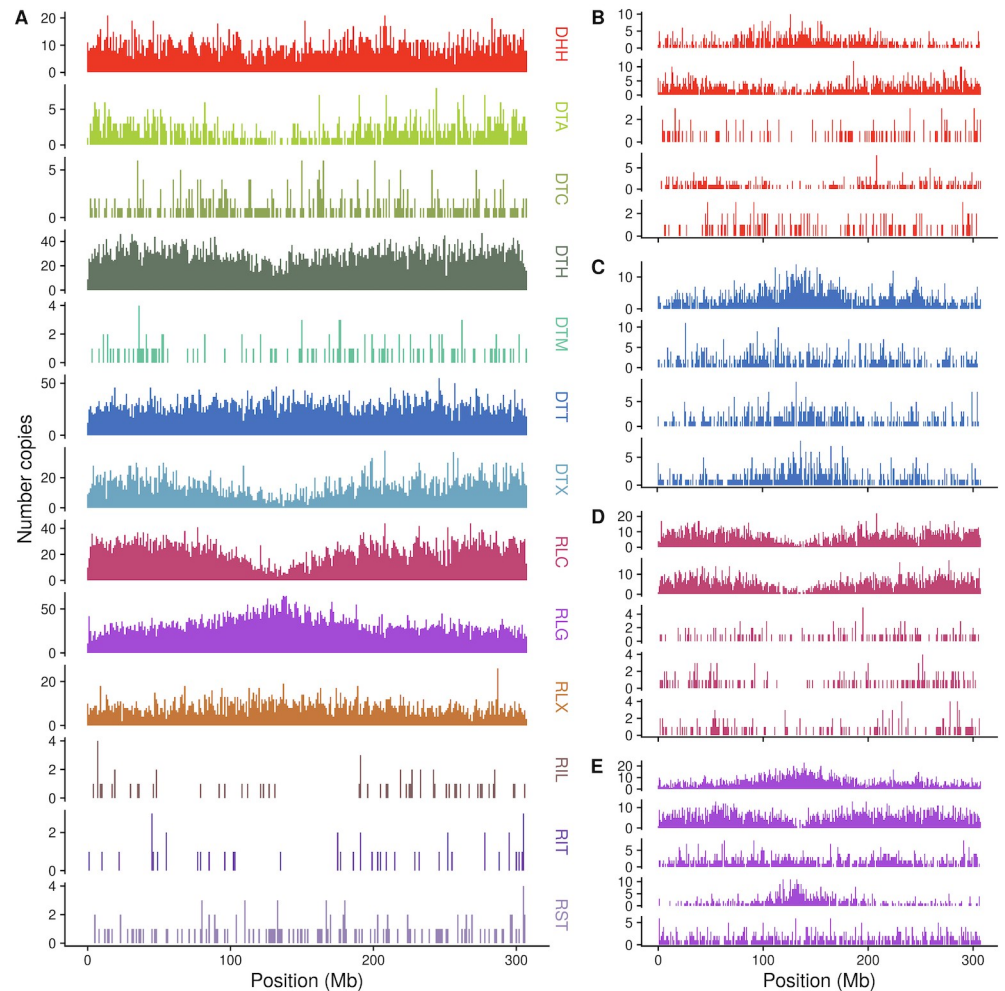
Fig 3. Chromosomal distribution of superfamilies and example families. Counts of number of insertions in 1 Mb bins across chromosome 1 for (A) TE superfamilies and (B-E) the 5 families with highest copy number in each of four superfamilies, DHH (B), DTT (C), RLC (D), and RLG (E). Family names are listed in S1 Table.

https://doi.org/10.1371/journal.pgen.1009768.g003

Features of the transposition process

Here, we address features that restrict and allow movement of TE copies, as well as influence their survival after insertion.

TE proteins. Numerous sequence features of the TE itself are required for the complex transposition process to occur, and these are best understood at the level of TE family. One requirement is the presence of TE encoded proteins that catalyze movement. Functional characterization of TE protein coding capacity is complicated by difficulty in identifying the effect of stop codons or nonsynonymous changes on transposition. Instead we measure homology to TE proteins (see Methods for details), although we recognize this does not fully reflect whether a TE copy can produce a transpositionally-competent protein product. Although TE-encoded proteins are often of similar length within a TE superfamily due to domain conservation and shared ancestry, the longest ORF in a TE varies by family (Fig 5A). Sometimes this is due to the presence of nonautonomous or noncoding copies. While nonautonomous copies rely on protein production in *trans* by other family members, autonomous TE copies encode their
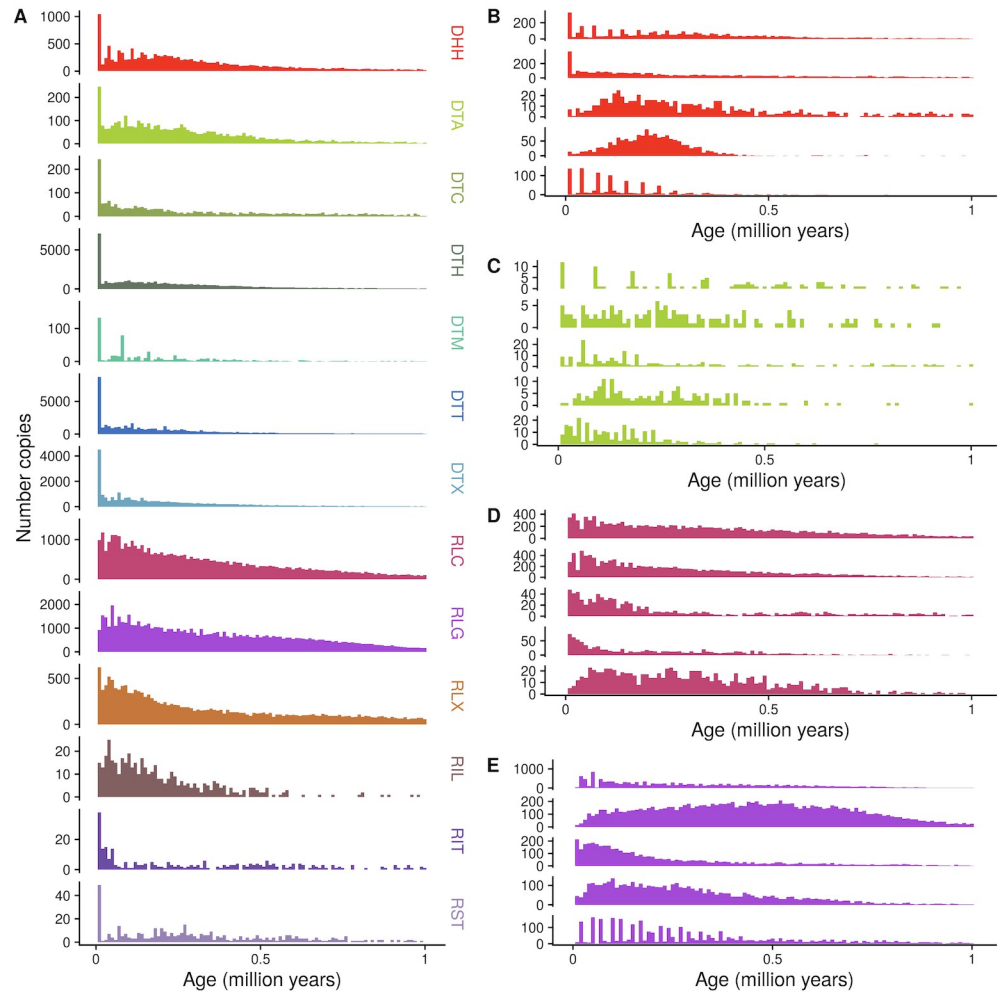
Fig 4. Age distribution of (A) superfamilies and (B-E) five largest families of (B) DHH, (C) DTA, (D) RLC, and (E) RLG. Family names are listed in S1 Table. Counts of number of insertions in 10,000 year bins are shown. As they are rare, TE copies older than 1.1 million years are not shown. Ages are calculated with terminal branch lengths for all TEs except LTR retrotransposons, which are calculated with LTR-LTR divergence. See S5 Fig for LTR retrotransposon plots with terminal branch length ages.

https://doi.org/10.1371/journal.pgen.1009768.g004

own transposition machinery in *cis*. The majority (52%) of LTR families have at least one family member that retains some remnant of coding capacity for all the TE proteins necessary for transposition. In contrast, only 0.6% of TIR families, 0.3% of helitron families, and 0.2% of nonLTR families have at least one family member with protein coding sequence that matches known TE proteins. For all TEs, coding capacity varies substantially between families (Fig 5B). Several LTR retrotransposon families have a small proportion of potentially autonomous copies (S12(C) Fig), and yet other families where coding potential for required proteins is split between between different TE copies (e.g. RLG00001, where 1.6% of copies code only for GAG and 13.5% of copies code for only POL, although both proteins are required for retrotransposition; S12(A), S12(B) and S12(C) Fig). Also, families range from having almost exclusively potentially autonomous coding copies ($\geq$ 75% of copies in 14 families of DTC, RLC, and RLG, S3 Table), to having exclusively nonautonomous noncoding copies (842 families, spanning all 13 superfamilies; S4 Table).
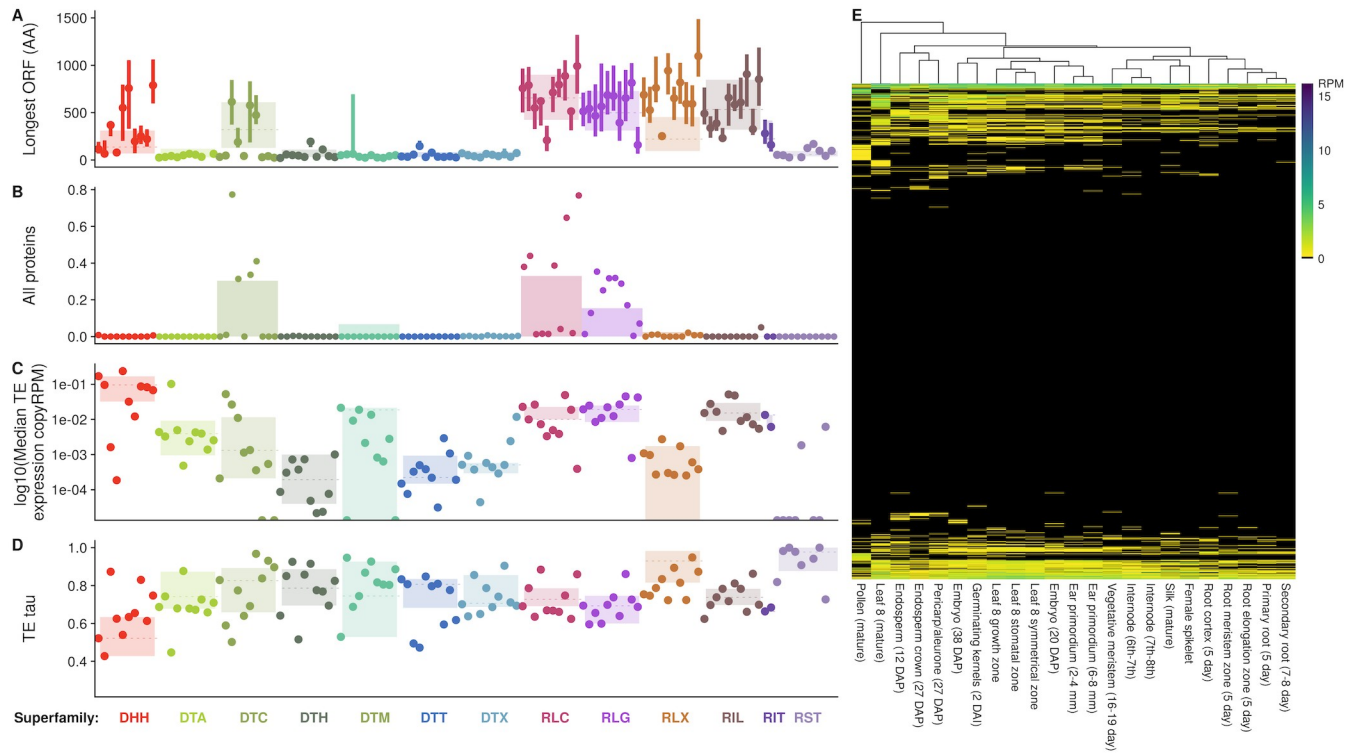
Fig 5. TEs code for proteins that are expressed, and expression varies by family across tissues. In A-D, families are in the same order as presented in Fig 2, and listed in S1 Table. (A) Length of longest open reading frame within the TE, measured in amino acids. (B) Proportion of family with all proteins required for transposition. (C) log10 median TE expression across tissues, per-TE copy. (D) Tissue specificity of TE expression τ, with low values representing constitutive expression, and high values representing tissue specificity. (E) Per copy TE expression across tissues (RPM, reads per million), clustered by expression level. Families with greater than 10 copies are shown in rows, and tissues in columns.

https://doi.org/10.1371/journal.pgen.1009768.g005

Coding capacity for TE proteins likely dictates the ability to generate new insertions, and as such is associated with TE age and the timing of activity of the family. Averaged across all orders, TEs that code for proteins are younger than those that do not code for proteins (median age of 198 kya vs. 263 kya; significant effect of protein coding in Wilcoxon rank sum test, $p < 2e - 16$). Further, noncoding copies from families that lack a coding member in B73 show an elevated median age (266 kya) when compared to noncoding copies in families with coding members (174 kya) (all pairwise Wilcoxon rank sum tests show a significant effect of coding status, all $p < 2.1e - 11$). For most superfamilies, coding members are older than non-coding copies from families with coding members (S7 Fig).

TE expression. Beyond simply coding for TE proteins, another requirement for TE transposition and transgenerational inheritance is expression of the TE itself such that the TE-encoded protein can be generated. Mapping of RNA-seq reads to repetitive TE families is a challenge, as it can be impossible to identify the exact copy that is expressed when a read maps equally well to multiple TE copies [50]. We choose to summarize multiply mapping reads and TE expression at the level of per-copy RPM of the family. This likely averages relevant variation in expression known to exist between copies within maize TE families [51, 52], but reflects patterns observed at the level of the family. Measured in this way, large families are generally transcriptionally repressed (Fig 5C), while small families show higher median per-copy expression levels. Most families are not expressed in any tissues surveyed (Fig 5E). While superfamily medians and median expression per copy of the ten largest families per superfamily show

below 0.1 RPM per copy (Fig 5C), per copy rates of expression can be higher for small families. For example, the 19 copies of RLC00184 (also known as *stonor*) show high median expression of 4.33 RPM per copy.

Tissue specificity can reflect different strategies for TE survival, like that a TE must jump in germline tissue to ensure its transgenerational inheritance at a new locus. Tissue specificity, measured as τ (see Methods), is highest when values of τ are equal to 1, and 0 when constitutively expressed at identical levels across all tissues. Helitrons and most LTR retrotransposon superfamilies (RLC and RLG) show lower median τ than TIR and nonLTR retrotransposon superfamilies (All pairwise Wilcoxon rank sum tests significant ($< 4.2e − 12$) except TIR and nonLTR comparison; Fig 5D). Tissue specificity can be extreme, with some families showing expression in only one tissue (Fig 5D)). For example, DTH00434 shows maximal per copy expression in mature pollen (4.3 RPM), with highly tissue specific expression ($\tau = 0.998$).

TE regulation. TE expression is likely limited by regulation of the TE by the host genome, which we measure via DNA methylation and MNase hypersensitivity in the TE and regions surrounding it. TEs on average are heavily regulated by their host genome: average cytosine methylation across TEs is high (averaged across five tissues, 82% of cytosines in a CG context in a TE are methylated, 67% in a CHG context, and 4% in a CHH context), although this varies across superfamilies (S5 Table) and families (Fig 6A, 6B and 6C). Only a small fraction of base pairs within TEs is in chromatin accessible to MNase, averaging 0.2% in shoot tissue, and 0.08% in root tissue (S9(K) and S9(M) Fig), both lower than genome-wide proportions (0.5% in shoot, 0.2% in root; significantly different from genome-wide values, one-sample Wilcoxon Signed Rank test p $< 2.2e − 16$ for both shoot and root). Despite this overall pattern of regulation, the host genome restricts some families of TEs differently. For example, the median CG methylation of the family DTM00796 is only 52% in anther tissue (Fig 6A), while most other families show higher methylation. There is even more extreme variation in CHG methylation across TE families (Fig 6B): though many TE families show low CHH methylation across the body of the TE, some families of DNA transposons show relatively high CHH methylation (Fig 6C). Although the numbers presented here are for anther tissue, these patterns are robust across tissues (S8 Fig).

Some TEs can preferentially insert to particular genomic locations, often based on local chromatin state [13, 14]. Others can modify the methylation patterns in flanking regions after insertion [32, 33]. Genome-wide, methylation levels in the region surrounding a TE insertion reflect these processes, and are variable both across families and DNA methylation contexts. Of the 1,243 TE families with ten or more copies, median methylation levels averaged across all tissues are elevated within the TE compared to 500 bp away for 734 TE families for CG methylation, 957 for CHG methylation, and 1086 families for CHH methylation. This pattern can be visualized as the decay of methylation moving away from the TE (Fig 6D, 6E and 6F and S8 Fig). The magnitude of reduction in local CG and CHG methylation moving away from the TE differs in extent and pattern, from families where methylation is reduced immediately adjacent to the TE to others with minimal reductions even 2 kb away from the TE (Fig 6D and 6E). In contrast, most families show rapid reductions in CHH methylation within 100 bp away from the edge of the TE (Fig 6F).

TE base composition. Observed DNA methylation levels may be impacted by the base composition of the TE, as cytosines must be present to be methylated. TE families differ in GC content (S9(A) Fig); with extremes ranging from 21% (DTT13542) to 84% (DTH14236) median GC content. This appears to be a consequence of bases carried by the TE itself and not of regional mutation pressure, as variation in GC content in the TE is greater than that of the flanking sequence (S9(B) Fig). For example, GC content in the 1kb flanking DTH14236 is over 30% lower than that in the TE (52% GC in the flanking region). Beyond the proportion of
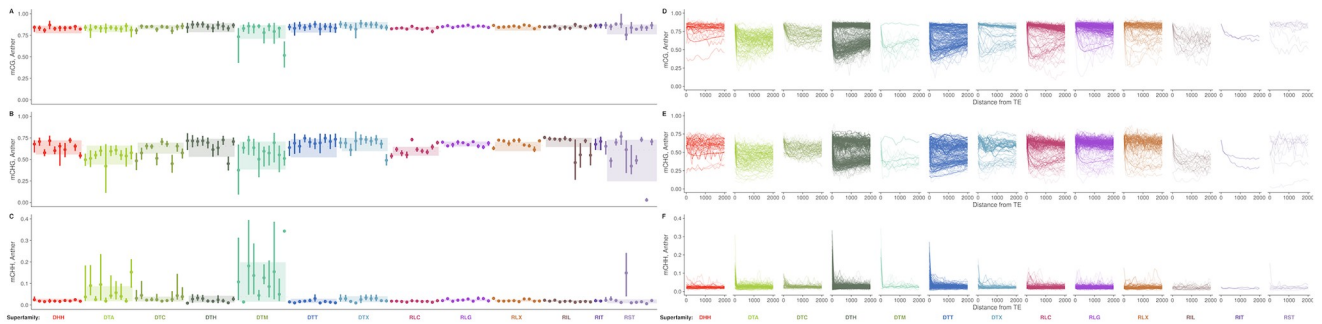
Fig 6. TEs and their flanking sequences are regulated by their host genome. Families are presented in the same order as in Fig 2, and listed in S1 Table. CG methylation in TE (A), CHG methylation in TE (B), and CHH methylation in TE (C). CG methylation in 2 kb flanking the TE (D), CHG methylation in 2 kb flanking the TE (E), and CHH methylation in 2 kb flanking the TE (F). All methylation data from anther tissue, other tissues shown in S8 Fig. In (A—C), superfamily median is shown as a dashed line with the interquartile range in the shaded box. In (D—E), median methylation for regions up to 2 kb up and downstream of the TE are plotted for each family, with family size denoted by line transparency (darker lines are larger families).

https://doi.org/10.1371/journal.pgen.1009768.g006

cytosines in the sequence, the context in which these cytosines are found can impact whether and how they are methylated. For example, 51 families have a median of 0 cytosines that can be methylated in either the CG or CHG context (S6 Table). And even with similar GC content, families differ in the contexts in which they have those cytosines, as families can have moderate GC proportions, but high proportions of these in a CG context (e.g. DTM00473; S9(A) and S9 (C) Fig). This is reflected in increased TG content, potentially a consequence of deamination of methylated cytosines (S9(I) and S10 Figs). Notably, TEs with high amounts of methylatable cytosines within the TE do not always share high methylatable cytosine proportions for the region flanking the TE (S9(C), S9(D), S9(E), S9(F), S9(G) and S9(H) Fig).

Although difficulty in mapping short reads to a highly repetitive genome precludes a comprehensive analysis of population frequencies of TEs across maize individuals, we use the proportion of segregating sites within TEs as a proxy for copy number. We measure segregating sites in a panel that includes 1,218 maize and teosinte individuals [53]. While as a whole TEs have fewer segregating sites per base pair (median 0.022) than the genome-wide proportion (0.039) (one-sided Wilcoxon signed rank test, p < 2.2e − 16; S9(O) Fig), some TE families show high numbers of segregating sites (e.g. DTH10060, 0.177 segregating sites per bp). In contrast to the sequence carried by the TE, variation in the region the TE is inserted into is considerably closer to genome-wide averages than that of the TE itself (median 0.034 segregating sites per bp; S9(P) Fig), but still significantly different (one-sided Wilcoxon signed rank test, p < 2.2e − 16).

Features structuring TE survival after insertion. The recombinational environment that a TE exists in can impact the efficacy of natural selection on the TE, as higher recombination can unlink deleterious variation from adaptive mutations [54], leading to a positive relationship between recombination and diversity. While LTR retrotransposons are more commonly found in low recombination regions (median 0.30 cM/Mb), Helitrons and TIR elements are more commonly found in higher recombination regions (both show a median 0.43 cM/Mb), and nonLTR retrotransposons are found in the highest recombination regions (median 0.57 cM/Mb) (significant effect of TE order on recombination rate, all pairwise Wilcoxon rank sum tests are significant at p < 1.4e − 06 except the Helitron-TIR comparison). This varies by family—the two largest families of DTT differ in median recombination regions from 0.14 cM/Mb to 0.53 cM/Mb (S11(A) Fig).

Additionally, selection can act on TEs if they have an impact on the expression of genes they land near. Although it is impossible to determine whether a TE insertion causes changes

in nearby gene expression using only the B73 genome, we observe differences among super-families and families of TEs in the expression levels of the closest gene. Across tissues, genes near TIR and nonLTR elements have higher median expression (1.37 RPKM for TIR and 1.83 RPKM for nonLTR) than genes near LTR (1.04 RPKM) and Helitrons (0 RPKM) (S11(C) Fig) (All pairwise Wilcoxon rank sum tests significant, p < 0.023). Notably, this pattern intensifies for genes within 1 kb of the TE, where median gene expression is over 4 RPKM for genes near TIR and nonLTR elements, but 0 RPKM for these genes close to LTR and Helitron elements (S11(D) Fig). Much of this signal is driven by non-syntenic genes—average expression is much higher for the closest syntenic genes (≈12 RPKM) but shows no significant difference amongst orders (Kruskal-Wallis rank sum test, p-value = 0.2353) (S11(F), S11(G) and S11(H) Fig). Some families are often found near highly expressed genes (e.g. DTA00133, median expression 22.38 RPKM), while median expression of the closest gene for $\approx \frac{1}{3}$ of families is close to zero. However, when genes near TEs are expressed, their expression is much more constitutive than that of TE families (S11(E) Fig and Fig 5E), with mean τ values of 0.75 for genes near TEs and 0.93 for TE families themselves. Tissue specificity varies by family and superfamily as well, and there is a weak correlation between tissue specificity of expression of TE families and expression of the genes they are closest to (Pearson's correlation 0.067, p = 4e-12).

The maize genome arose from an autopolyploidy event [55], and has been sorted into two extant subgenomes [56]. Subgenome A has retained more genes and base pairs than subge-nome B, accounting for 64.8% of sequence [48], and 64% of all TEs (S11(B) Fig). Additionally, the median age of TEs in subgenome B is lower (0.24 Mya) than those in subgenome A (0.26 Mya) (Wilcoxon Signed-rank test shows significant effect of subgenome; p < $2.2e − 16$). These differences are likely due to the effect of ongoing transposition erasing any signature of TE dif-ferences between parents of the allopolyploidy event, as genome-wide the family with the old-est median age (DTH16531) is only 8.5 million years old.

Modeling survival of TEs

To account for the myriad differences of these 341,426 TE copies in 27,444 families, we approach our understanding of the survival of TEs in the genome by modeling age as a response to the TE-level features and the genomic regions in which TEs exist today. Age reflects survival of TEs, measuring the amount of time since transposition that they have per-sisted at a genomic position without being lost to selection or drift. Hence, we measure the pre-dictive ability of features of the TE itself and the genomic region it inserted into on TE survival as measured by age.

Random forest regressions using age as a response variable and features that are measured at the level of the individual TE explain moderate amounts of variance (27.7%), and show low mean squared error (0.014). Across all TEs, information on the superfamily a TE belongs to contributes the most to prediction accuracy for age; after permuting their values, the square root of mean squared error (RMSE) increases by 162 kya (Fig 7B). Other features that increase RMSE by over 100 kya include the size of the family the TE comes from, the length of the TE (both in total bp and when including bases coming from copies nested within it), the TE fam-ily, the number of segregating sites per bp within the TE, and the median expression of the TE family across all sampled tissues. In aggregate, features of the region flanking each TE explain approximately as much variation in age as features of the TE itself, but there are more flanking features than those measured on the TE. On average, each feature of the TE contributes over 4 times more predictive power than that of a flanking feature (square root mean squared error of 39 kya for a TE feature, 8 kya for a flanking feature) (Fig 7A and 7B).
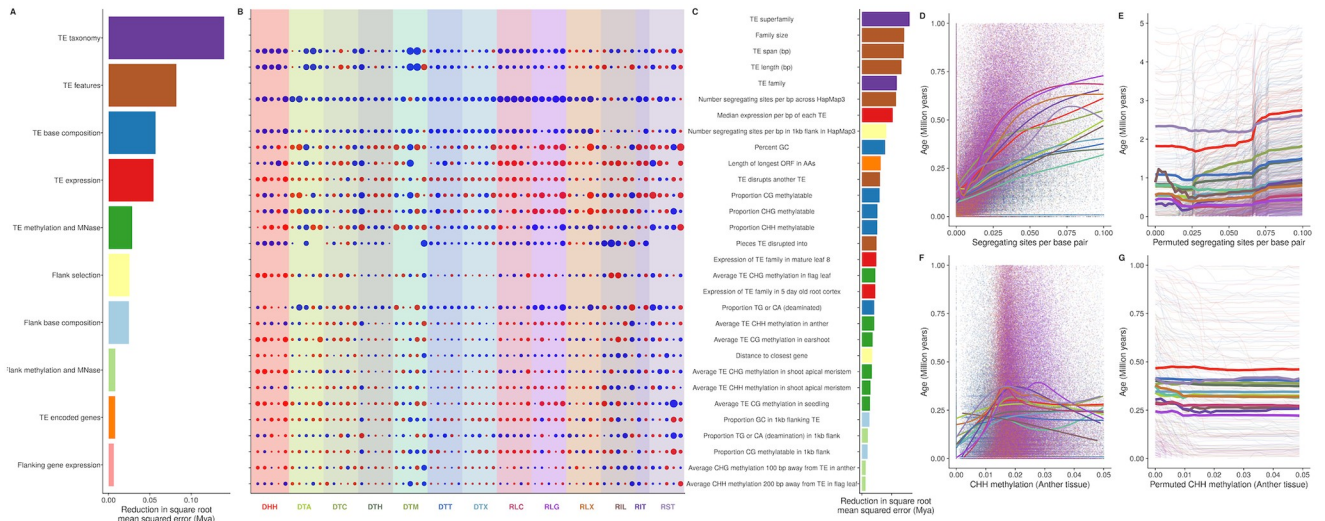
Fig 7. Features ranked by importance. (A) Reduction in mean squared error gained by including a feature in a model, summarized into categories. (B) Correlations of each of the top 30 features with age for the five largest families in each superfamily. Features labeled to the right in (C). Size of point is scaled by correlation coefficient, and color by whether the relationship is positive (blue) or negative (red). Rows without values are features that are fixed within a family, thus have no variance. (C) Reduction in mean squared error for top 30 individual features. Colors match categories in (A). (D) Raw correlations between age and segregating sites per base pair (E) Model predictions for the relationship between age and segregating sites per base pair (F) Raw correlations between age and anther CHH methylation of the TE (G) Model predictions for the relationship between age and anther CHH methylation of the TE.

https://doi.org/10.1371/journal.pgen.1009768.g007

These generalities reflect underlying nonlinearities in the relationships between individual features and age, which are often family-specific. Indeed, correlations of these top features with age differ not only in magnitude, but even in sign between individual families (Fig 7C). To provide additional insight into the local behavior of the relationship between a feature of interest and age, we use the fitted random forest models to predict age for TE copies as we vary the feature. For example, the number of segregating sites in the TE is positively correlated with age in the raw data (Fig 7D), and is confirmed via this permutation approach (Fig 7E). Yet despite this overall pattern, individual families vary in sign (Fig 7C) and slope (Fig 7E) of the relationship. Other features, like CHH methylation of the TE in anther tissue, show relationships that vary by superfamily, where RIT and DHH appear older with increasing CHH methylation of the TE in the anther, while other superfamilies show decreasing age (Fig 7G), a pattern less apparent in the raw correlations (Fig 7F). Across all features, there are largely family-specific combinations of both the direction and strength of correlation with age (Fig 7C). In total, while genomic and TE features contribute to prediction of age, interactions among these features make it difficult to predict the survival of any single family.

Discussion

General patterns

As 85% of the maize genome is repetitive sequence [26, 49], and 63% structurally recognizable TE sequence [48], TEs contribute more to the maize genome than sequence that is uniquely 'maize.' Like most plant genomes [11], retrotransposons contribute more base pairs to the maize genome than do DNA transposons (Table 1 and Fig 2B). This is a consequence of the high number of copies (Fig 2A) and the large size of individual retrotransposons (Fig 2C), likely due to a 'copy and paste' replication mode that leaves existing copies intact when

generating new copies. Also like other plant genomes [57, 58], several superfamilies of DNA transposon in the maize genome are found closer to genes than are retrotransposons (Fig 2B). This is likely due to targeted insertion into euchromatic sequences [59, 60], and differences in removal through natural selection after insertion [61, 62].

TE superfamilies

The bulk of TE sequence is often described at a finer scale, that of individual superfamilies of TEs. Each TE superfamily defined in the maize genome has representatives across the tree of life [63–65], suggesting an ancient origin of these genomic parasites. Some superfamilies have retained dramatic and consistent differences in their spatial patterning across chromosomes over hundreds of millions of years. For example, the superfamily RLG is enriched near centromeres in all plants [66–68] including maize (Fig 3A), highlighting a genomic niche that allows long-term survival near the centromere. Similar patterns exist at deep time scales for DNA transposon superfamilies, which preferentially insert near genes in both monocots and dicots [57, 58, 69, 70] and in maize are enriched on chromosome arms where genes are concentrated (Fig 3A).

These patterns likely reflect the evolution of different ecological strategies of TEs in the genome. Kidwell and Lisch (1997) [7] described two extremes to the 'ecology of the genome'—one, a TE that preferentially inserts far from genes, into low recombination heterochromatic regions, and a second, risky TE that inserts near low copy sequences, more likely to disrupt gene function. We observe these extremes at play in the maize genome, in that LTR retrotransposons dominate the heterochromatic space, with over half of all copies greater than 16 kb from a gene (Fig 2B), and most copies heavily methylated (Fig 6A, 6B and 6C). The alternate strategy also exists in the maize genome, with risky insertions near genes and transcribed regions seen for several TIR superfamilies. For example, over half of Mutator transposons (DTM) are found within 1 kb of a gene (and over one quarter of DTM within 100 bp of a gene) (Fig 2B). This likely results from the preferential insertion of DTM elements upstream of genes [30, 60, 71, 72]. We note that we find TIR copies are found further from genes (17.2 kb) than previously reported for grass genomes [49, 58]. We believe this may be due to previous analyses based on preferential assembly and identification of genic TEs—indeed subsetting to the 893 TIR families found in the Maize TE Database [49] results in a much reduced 1.6 kb median distance to genes. On a genome-wide scale at the level of all TEs, the spatial patterns we observe could result from either preferential insertion or differential removal by selection after insertion. Further characterization of these ecological strategies will be facilitated by investigating TE polymorphism across maize individuals [33, 73] and *de novo* recent insertions that selection has not yet acted on [74].

TE families

While superfamily level observations are useful for gaining an overview of the distribution and survival of TEs in a genome, more detailed study on a time scale relevant to the evolution of the genus *Zea* comes from studying TE families. Maize TE families are shared with closely related host species, but the number of shared families rapidly decreases with phylogenetic distance. Many families are shared with congeners *Zea diploperennis* [75–77] and *Zea luxurians* [78], but few families investigated are found in maize's sister genus *Tripsacum* (1 mya divergence; [79]) [75–77, 80, 81], and the only families shared between maize and *Sorghum* (12 mya; [55]) are shared only as a result of horizontal transfer events between the species [82]. This suggests that in order to understand TE evolution at a timescale relevant to maize as a

species, it is essential to understand families of TEs, rather than the aggregate properties of superfamilies or orders.

Indeed, our family-level analysis also reveals patterns obscured when TEs are averaged together at the level of superfamily. For example, despite the fact that the RLG superfamily is enriched in centromeric and pericentromeric domains (Fig 3A), the second largest family RLG00003 (homologous to the RLG family *huck* [83]) is predominantly found on chromosome arms (Fig 3D). While many RLG elements contain a chromodomain targeting domain in their polyprotein [84] allowing targeted insertion to centromeres, RLG00003 does not (S12 (G) Fig). This lack of a chromodomain may explain a proximal cause of the observed niche of RLG00003, although other factors are certainly at play, as other families with centromeric enrichment also lack chromodomains (S12(G) Fig). DNA transposons are also best described at the family level. While Mutator (DTM) elements are found a median distance of 2.5 kb from genes (Fig 2B) and have long been observed to target insertions near genes in maize [30, 60, 72], the second largest family, DTM13640, is found a median distance of 34 kb away from genes (Fig 2B). The mechanism for gene targeting seems to be mediated through recognition of open chromatin [60, 71], but precise details of the targeting are unknown. Further investigation into the families that insert near and far from genes may pinpoint how their molecular mechanisms of targeting may differ.

Furthermore, differences in the timing of transpositional activity vary extensively between families. Most TE families in maize have had most new insertions in the last 1 million years (Fig 4). Some TE families have bursts of activity, punctuated by a lack of surviving new insertions, while others appear to be headed towards extinction. All of these timings are much more recent than allopolyploidy in maize ($\approx$ 12 mya) [55] and families show little subgenome bias in their distribution (S9(B) Fig), suggesting that these represent lineages evolving within maize.

Maize was domesticated from teosinte (*Zea mays* subsp. *parviglumis*) 9,000 years ago [85, 86]. It is tempting to address the contribution of TEs to this major transition, especially given the contribution of TE insertions to maize domestication and improvement [87–89]. Although we caution that mutation rates and estimation can complicate ascertainment (see below), 46,949 TEs across all 13 superfamilies have an estimated age of less than 9,000 years, and 24,630 TEs have an estimated age of 0. This suggests that transposition has been ongoing since the divergence of maize from its wild ancestor, but we caution that we lack appropriate confidence intervals for these estimates, especially as non-zero age requires observing at least one nucleotide mutation.

The family-level ecology of the genome

It can be difficult to predict exactly why a particular TE family differs from other families. Community ecologists aim to understand the environmental factors that give rise to the observed diversity of organisms living in one place, including not just features of the environment but also interactions between species. TE families are analogous to species in the genomic ecosystem, and because the genomic environment a TE experiences is constrained to the cell, TEs are forced to interact in both time (Fig 4) and space (Fig 3). We predict each family of TE is adapted to its genomic ecological niche, where the genomic features we measure represent the environmental conditions and resources limiting a species' ecological niche [90]. TEs additionally can act as ecosystem engineers, modifying the environment they insert into, and generating new habitat for future colonization [10, 44].

In the genomic ecosystem, we can observe interactions between species much like we would see in a traditional ecosystem. We see a number of patterns, including cyclical dynamics

of TE activity through time for several families, sustained activity through time, and a reduction in new copies towards the present (Fig 4 and S4 Fig). This means that the genomic environment a newly inserted TE experiences is affected by the activity and abundance of all other TE families in the genome. At one extreme, members of the same family can even encode different proteins required for retrotransposition in different TE copies, where both types are required to be transcribed and translated for either to transpose. Such a system approaches a mutualism, where the success of one type depends on another. Previous knowledge of these systems was limited to the maize retrotransposon families Cinful, which codes for polyprotein domains, and Zeon, which codes for GAG [25] (represented here by a single family, RLG00001). This strategy has been successful in maize, and RLG00001 [48, 77] for example makes up 135Mb of sequence. *Sorghum*, in contrast, has a genome $\frac{1}{3}$ the size of maize [91] and lacks homologs to RLG00001. Such symbiotic relationships within a TE family have been thought of as remarkably rare [92, 93]; however we identify 25 LTR retrotransposon families where GAG and POL protein domains are found in separate TE copies but less than 1% of copies contain both, suggesting that this pattern is much more prevalent than previously described. These types of elements are best classified as subtypes of a single family, because the *cis* components of the LTR are recognized by protein domains of both GAG and POL proteins, leading to homogenization of sequence signals. As noted by Le Rouzic et al. (2007) [92], symbiotic TE families face a major barrier in being horizontally transferred, as both copies must be transmitted through an already rare process. Their prevalence in the maize genome thus supports instead a long term coevolution of the maize genome and the TEs that live within it, specializing and diversifying with different ecological strategies.

Unlike most contemporary ecological communities, which are censused when a researcher surveys them, the genomic ecosystem carries a record of past transposition. We can investigate this past 'fossil' record using the age of individual TE copies. This allows a robust analysis of the features that define TE survival across time. The TEs we see today are a readout of the joint processes of new transposition—which may not be uniform through time—and removal through selection, deletion, and drift [62]. Survival of a TE can be measured by its age or time since insertion, as our observation of a TE is conditioned on the fact it has not been removed by either neutral processes or selection. Changes in the TE community over time give rise to evolution.

Although relative age differences between TE insertions are limited only by our ability to count mutations, absolute age estimates can be shifted by mutation rate estimates. We use a maize-specific mutation rate [94], which leads to a five times younger estimated age of maize LTR retrotransposons than the 3–6 million years originally estimated by SanMiguel et al. (1998) [95]. Additionally, as nucleotide mutation rates in TEs may be higher than other parts of the genome (≈ 2 fold higher in TEs in *Arabidopsis thaliana*, [96]), we consider our age estimates to represent an upper bound of TE age. Nonetheless, age represents a comparable metric of survival in the genome, especially when summarized across multiple copies and families. Our random forest model predicting age of TEs thus relates the action of transposition to the processes that occur afterwards on an evolutionary time scale. The model shows that TE superfamily and family size provide best predictive power for age (Fig 7B).

Another TE feature with high predictive power for age and survival in the genome is the length of the TE, both of itself and its length including copies nested within it. For most families, there is a negative correlation between TE length and age (Fig 7C). However, we find that the relationship between TE length and age in maize is often nuanced, with some long TEs surviving over millions of years (Fig 7C and S13(A) and S3(B) Fig). In other taxa, selection is stronger on long TEs, mediated by a higher potential for nonhomologous (ectopic)

recombination [97–99]. Although a number of factors may contribute to these patterns, it also seems likely that a genome as repetitive and TE-rich as maize perhaps could not have evolved without mechanisms to prevent improper pairing of nonhomologous sequences with high nucleotide similarity [100].

Other predictors of age are expected. For example, we expect a new insertion to be younger if we show that the TE disrupts another TE, which we see for most families shown in Fig 7C. Additionally, we expect the proportion of segregating sites in the TE and the region flanking a TE insertion will be positively associated with TE age, as they reflects a count of the mutations that have accumulated on the haplotype carrying the TE. There is a positive relationship between age and segregating sites for most families shown in Fig 7C. We note that imprecise repair of a double stranded break after excision of a TIR element [70] could obscure this signal to some extent, increasing the number of flanking SNPs while decreasing the average frequency of the TE. Consistent with this mechanism, the superfamily DTT, which excises precisely without introducing nucleotide mutations [101] shows lower median flanking segregating sites per base pair (0.0295) than TIR elements from other superfamilies (0.0310) (Wilcoxon Signed-rank test shows significant effect of DTT superfamily; $p < 2.2e − 16$).

Elevated CHH methylation of TEs has been found in recently activated TEs in *Arabidopsis thaliana* [102] and in TEs near genes in maize [103, 104]. We find complicated, nonlinear relationships of CHH methylation with age (Fig 7F and 7G). These differences between and within families may reflect a natural senescence of TE copies. Young copies not yet silenced by the genome lack CHH methylation, intermediate age copies are effectively silenced with higher CHH methylation levels, and the oldest TEs with low CHH methylation are defunct copies incapable of transposition that are no longer silenced. More detailed study of recently active maize TE families will allow understanding of the temporal dynamics of transcriptional and post-transcriptional silencing of TEs.

In spite of previous predictions, distance to a gene and recombination are not found in the top 30 explanatory variables of age. Old TEs are underrepresented near genes in humans and *Arabidopsis thaliana* [105, 106], consistent with selection against such insertions. Recombination has been implicated in both the removal of TEs and in modifying their impact on fitness via ectopic recombination [6]. We believe that both distance to a gene and recombination rate reflect broad-level summaries of genomic regions, such that they are not predictive in our model once other local features are included. For example, regions with high recombination rate generally show low CG methylation in maize [107], but a subset of genes in such regions show CG methylation across the gene body. Since CG methylation plays a role in TE survival (Fig 7B), inclusion of this feature in our models will thus reduce the importance of recombination rate. Similarly, CHH methylation is most prominent in regions of the genome close to genes, presumably a result of RNA-directed DNA methylation reinforcing the boundary between heterochromatin and euchromatin [103, 104]. As this elevated CHH methylation is often over the TE closest to a gene [104], the distance of a TE to the closest gene may provide largely redundant information beyond what is captured by measurements of CHH methylation. Finally, despite many other features being correlated with either gene density or recombination rate, the two are inextricably linked, as recombination in maize primarily initiates in genes [108]. Together, these combine to reveal few patterns in the relationship between distance to gene, recombination rate, and age of TE copies (S13(C) and S13(D) Fig).

Finally, in spite of the fact our model includes more than 400 features of the genomic environment, TE taxonomy contributes substantially to prediction of TE age (Fig 7A and 7B). We have seen that the relevance and direction of effect of individual features can differ among families (Fig 7C), essentially generating family-specific niches in the genomic ecosystem. In fact, there is no genomic feature we measure which shows even the same direction of

correlation with age across all families. The importance of taxonomy in our model suggests that there are unmeasured latent variables that are best captured with superfamily and family labels. This further emphasizes that the analysis of TEs in maize should focus on family, as each family is surviving in a slightly different way, exploiting a unique genomic niche.

Conclusion

Genes in the maize genome are 'buried in non-genic DNA' [109] consisting predominately of TEs. The interaction between TEs and the genes of the host genome can structure and inform genome function. The diversity of TEs in an elaborate genome like maize generates a complex ecosystem with many interdependencies and nuances, limiting the ability to predict the functional consequences of a particular TE based only on superfamily or order. Instead, TE families represent a biologically relevant level on which to understand TE evolution, and the features most important for determining survival of individual copies represent dimensions of the ecological niche they inhabit. These observations suggest that the co-evolution between TE and host is ongoing, and inference of the impacts of transposons requires a multifaceted approach. The nuanced understanding generated from exhaustive analysis of genomic features and survival of individual families of TEs serves as a starting point to begin to understand not only TE evolution, but also the evolution of the host genomes they have coevolved with.

Methods

Scripts for generating summaries from data sources and links to summarized data are available at http://www.github.com/mcstitzer/maize_genomic_ecosystem. Interactive distributions per family can be found at https://mcstitzer.shinyapps.io/maize_te_families/.

TE sequence properties

We base our analysis on the TE annotation of the maize inbred line B73 [48], updated to more fully capture TIR elements (see S1 Text). TEs that are nested inside of other TEs are divided for further analyses, by assigning each TE base pair in the genome to a single copy by iteratively removing copies in order of arrival. We remove from analysis any TE for which less than 50 bp remains after resolving nested copies. We add the positions of retrotransposon long terminal repeats (LTRs) to these annotations as produced by LTRharvest [110], and delimit the internal protein coding genes of LTR TEs using LTRdigest [111] and GyDb 2.0 retrotransposon gene HMMs [112]. We additionally identify the longest open reading frame (ORF) in each TE model using transdecoder [113], and identify whether this longest ORF is homologous to known transposases, integrases, and replicases respectively for TIRs, nonLTR retrotransposons, and helitrons (JCVI GenProp1044 http://www.jcvi.org/cgi-bin/genome-properties/GenomePropDefinition.cgi?prop_acc=GenProp1044 and PFAM PF02689, PF14214, PF05970) using hmmscan [114] with default parameters. We characterize copies as autonomous based on the content of their protein coding domains, requiring evidence of all 5 proteins (GAG, AP, RT, RNaseH, INT) for LTR retrotransposons, a reverse transcriptase match for LINEs, a transposase profile match for TIR transposons, and a Rep/Hel profile match for Helitrons. This measure is lenient in defining coding content, as it does not penalize stop codons and frameshifts throughout these coding regions.

After insertion, TE copies accumulate nucleotide substitutions that can be used to estimate their age. To estimate age based on divergence of a TE copy from others in the genome, we generated phylogenies of TE copies by first aligning the entire TE sequence of each copy in each superfamily using Mafft [115] (allowing sequences to be reverse complemented with the option `--adjustdirection`) and then building an unrooted tree using FastTree [116].

To make tree building computationally efficient in spite of the high number of TE copies and large element size, we use a maximum of 1000 bp for tree building for the largest 5 superfamilies (3' terminal for Helitrons, 5' terminal for LTR retrotransposons and TIR elements). The terminal branch length of each copy is used as a measure of its age, representing nucleotide substitutions since divergence from the closest related copy in the B73 reference genome. This measure of age makes a number of assumptions about the tempo and mode of transposition—for example, we assume nucleotide mutations in a TE arose at its current location, which may not be true for TIR elements that excise and move to a new location. Nonetheless, it is the only approach to calculate ages of individual TIR and Helitron elements [117, 118] without relying on a consensus element generated from a multiple sequence alignment that can be biased towards recently transposed copies that have not yet been removed by natural selection or genetic drift [118, 119].

Because the 5' and 3' LTR of LTR retrotransposons are identical upon insertion [83], we also estimate their time since insertion using the number of substitutions that occur between the two LTRs. For each LTR retrotransposon copy, we align both LTRs with Mafft [115] and calculate nucleotide divergence with a K2P correction using dna.dist in the ape package of R [120, 121]. For all age measures, we relate nucleotide divergence to absolute time using a mutation rate of 3.3×10^{-8} substitutions per site per year [94]. These LTR-LTR estimates are generally in line with terminal branch length age estimates (Spearman's correlation 0.65), with LTR-LTR ages often older than terminal branch length ages (S6 Fig).

TE environment and regulation

We characterize the genomic environment of the TE and features that overlap the TE. For each TE, we characterize the distance to the closest gene (gene annotation AGPv4, Zm00001d.2, Ensembl Plants v40) irrespective of strand using GenomicRanges [122]. We additionally measure expression of these closest genes across a developmental atlas of the maize inbred line B73 [123] (accessed from MaizeGDB as walley_fpkm.txt using AGPv4 gene names). In order to estimate the overall dynamics and tissue-specificity of expression, we calculated both the median expression and τ [124] for each of these genes. τ is calculated as the summed deviance of each tissue from the tissue of maximal expression, divided by total number of tissues minus 1. τ values thus range from 0 to 1, with low values representing constitutive expression and high values indicating tissue-specific expression. We further characterize whether the closest gene is found in a syntenic position in *Sorghum bicolor* (typically indicative of higher conservation) using curated lists across grass genomes, excluding maize genes matched to multiple *Sorghum* orthologs [125, 126].

In addition to host genes, TEs themselves can be transcribed. Using RNAseq reads from the Walley et al. (2016) [123] expression atlas (NCBI SRP029238), we counted reads that align uniquely to a specific member of a TE family, as well as multiply mapped reads that align to a single family, as in Anderson et al. (2018) [51] and Anderson et al. (2019) [52]. This allows estimation of the expression level of a TE family, despite the repetitive nature of TEs that limits unique mapping of reads. Reads that map to TEs located within genic sequences (generally within introns) were excluded because their expression is indistinguishable from transcription from the gene promoter. We take the mean value of reads per million across the two to three replicates per tissue, and divide by the total family size to get a per-copy metric of expression. As with genes, we calculate median expression across tissues and tissue specificity using τ.

To identify the recombinational environment in which each TE exists, we use a 0.2 cM genetic map of maize generated from the Nested Association Mapping (NAM) panel [127]. We convert AGPv2 coordinates to AGPv4 coordinates using the Ensembl variant converter

[128]. To approximate the recombination rate in genomic regions, we fit a monotonic polynomial function to each chromosome [129]. Using this function and TE start and end positions, we calculate a cM value for each TE, and convert to cM/Mb values by dividing by the length of the TE in megabases.

The chromatin environment a TE exists in can impact transposition [60]. We converted data on MNase hypersensitive sites in roots and shoots [130] from the AGPv3 reference genome to AGPv4 coordinates using the Ensembl variant converter [128]. We counted how many hypersensitive sites exist in each TE, as well as the proportion of base pairs of the TE that are hypersensitive. We also calculate these metrics for the 1 kb region flanking the TE on both sides.

Regulation of TEs by the host genome is often mediated via epigenetic modifications. We use bisulfite sequencing reads from shoot apical meristem, anther, ear shoot, seedling leaf, and flag leaf [104, 131]. We trim adapters using TrimGalore, map using bsmap 2.7.4 with parameters (-v 5 -r 0 -q 20) [132], and summarize in 100 bp windows as in Li et al. (2015) [104], to characterize the local proportion of methylated cytosines in all three contexts (CG, CHG, CHH; where H is any base but G). We summarize the average levels of each measure over each TE copy and each of 20 100 bp windows of flanking sequence on either side, imputing missing data with the family mean.

To identify differences between TE copies in their base composition, we calculate GC content plus the number of di- and tri- nucleotide sites containing cytosines in a methylatable context (CG, CHG, CHH). We count these contexts in each TE using the bedtools nuc command [133] and divide by TE length to determine the proportion of the sequence that is methylatable for each context. We also calculate these measures of methylatability for the 1 kb flanking the TE on each side.

We also measure the number of segregating sites per TE base pair and the 1 kb flanking in the *Zea mays* Hapmap3.2.1 dataset [53] as well as the subgenome [48] each TE is found within.

As we cannot calculate accurate summaries of genomic features for families with a small number of TE copies, we include only those families with more than ten copies when presenting results in the text that identify specific outlier families, such as the family with highest GC content. When presenting summaries at the superfamily and order level or results modeling TE age, we include information from all TE copies, including those from smaller families.

Analysis and interpretation

We implement random forest regression models (in the R package 'randomForest' [134]) to understand the importance of different genomic features to TE survival in the genome as measured as the age of individual extant copies. We train models on 50% of copies, and summarize 1000 iterations of trees. The remaining TEs are retained as a test set to estimate model performance. Any missing data is assigned a value of -1, and the categorical variable of superfamily is considered as a factor. Because of limitations in converting numbers to binary, we limit the categorical variable of family to the 31 largest families, and code all others as 'smaller.' We summarize the overall importance of each feature in predicting age by permuting its values across individual TE copies and observing the change in mean squared error of the model prediction of the actual value, scaled by its standard deviation. We summarize features into categories reflecting TE taxonomy, TE base composition, TE methylation and chromatin accessibility, TE expression, TE-encoded proteins, nearest gene expression, regional base composition, regional methylation and chromatin accessibility, and regional recombination and selection. A full description of the individual measurements that go into each category is found in S2 Table.

In order to interpret family-specific relationships for the top predictors of age, we perform further analyses. We calculate the Pearson's correlation coefficient of each predictor with age, using samples from each family. To visualize the nonlinear relationships and interactions produced by such models, we calculate Individual Conditional Expectations (ICE plots [135], R package 'pdp' [136]), which summarize the contributions of permuted values of a variable of interest to the response, while conditioning on observed values at all other variables. We provide permuted values summarizing 95% of the observed data, to provide predictions in a region of parameter space the model is trained on. We summarize these responses as deviation of the predicted value generated with permuted data from the true value, and plot as individual lines and superfamily averages.

Supporting information

S1 Fig. Family characteristics of each of the largest 10 families of each superfamily with at least 10 copies. (A) Proportion of TEs within the transcript of a gene, including introns and UTRs. (B) TE span along the genome, summing both the base pairs of the TE and the base pairs of the TEs nested within it. (C) Proportion of TEs that are intact, that is, uninterrupted by the insertion of another TE. In (A and C), families are shown as points and superfamily proportions as a barplot, and in (B) families are shown with medians as points and lines representing ranges of upper to lower quartiles, with superfamilies shown as colored rectangles.
(TIF)

S2 Fig. Chromosomal distribution of superfamilies across all 10 maize chromosomes. Count of TE copies of each superfamily in 1 megabase bins across each chromosome.
(TIF)

S3 Fig. Distribution on chromosome 1 of five largest families with at least ten copies in each superfamily. Count of TE copies in 1 megabase bins along chromosome 1. (A) DHH, (B) DTA, (C) DTC, (D) DTH, (E) DTM, (F) DTT, (G) DTX, (H) RLC, (I) RLG, (J) RLX, (K) RIL, (L) RIT, (M) RST. Note that some families have no copies on chromosome 1, including DTT10880 and DTX10177. Additionally, the RIT superfamily only has two families.
(TIF)

S4 Fig. Ages in 10,000 year bins across each of the largest 10 families of each superfamily with at least 10 copies. (A) DHH, (B) DTA, (C) DTC, (D) DTH, (E) DTM, (F) DTT, (G) DTX, (H) RLC, (I) RLG, (J) RLX, (K) RIT, (L) RIL, (M) RST. The RIT superfamily only contains two families.
(TIF)

S5 Fig. LTR-LTR ages and terminal branch length ages for LTR retrotransposons. Ages in 10,000 year bins across each of the largest 10 families of each superfamily with at least 10 copies. Left plots (A-D) show LTR-LTR ages, right plots (E-H) show terminal branch length (TBL) ages. (A) all copies, LTR-LTR, (B) RLC families, LTR-LTR, (C) RLG families, LTR-LTR, (D) RLX families, LTR-LTR, (E) all copies, TBL, (F) RLC families, TBL, (G) RLG families, TBL, (H) RLX families, TBL.
(TIF)

S6 Fig. LTR-LTR ages vs. terminal branch length ages for LTR retrotransposon superfamilies. Spearman's correlation coefficient shown on plot for each superfamily.
(TIF)

S7 Fig. Age of TE copies split by coding potential of self and family. Violin plots with three lines, at median and 25th and 75th percentile. Only ages younger than 2 million years are shown. "Coding copy" refers to those copies that code for protein, "noncoding copy" refers to those copies that don't code for protein, but a family member does, and "noncoding family" refers to copies from families without a coding member in B73.
(TIF)

S8 Fig. Methylation in TE and flanking sequence, across tissues. A-J: mCG; K-T: mCHG; U-end mCHH. Tissues on y-axis, from top to bottom: Anther, SAM (shoot apical meristem), Ear-shoot, Flagleaf, Seedling leaf.
(TIF)

S9 Fig. Features of the TE and flanking sequences. GC content in the TE (A) and 1kb flanking sequence (B). Proportion of sites methylatable in the CG context in the TE (C) and 1kb flanking sequence (D), methylatable in the CHG context in the TE (E) and 1kb flanking sequence (F), proportion of sites methylatable in the CHH context in the TE (G) and 1kb flanking sequence (H). Proportion of sites containing a TG or CA dinucleotide in the TE (I) and 1kb flanking sequence (J). Proportion of sites in MNase hypersensitive regions in root in TE (K) and 1kb flank (L), and shoot in TE (M) and 1kb flank (N). Proportion of segregating sites in the TE (O) and 1kb flank (P).
(TIF)

S10 Fig. The proportion of methylatable cytosines is negatively correlated with the proportion of TG/CA dinucleotides. The x-axis reflects the proportion of cytosines in a CG context within the TE, and the y-axis reflects the proportion of dinucleotides in the TE that contain a TG or CA.
(TIF)

S11 Fig. Recombination, subgenome, and expression of closest gene. (A) Recombination rate across the TE, (B) proportion of TEs in subgenome A, (C) log10 median expression of the closest gene to each TE, (D) log10 median expression of genes within 1kb of the TE, (E) Tau of closest gene to each TE, (F) log10 median expression of the closest syntenic gene, (G) log10 median expression of closest syntenic genes within 1 kb, and (H) Tau of the closest syntenic gene.
(TIF)

S12 Fig. Protein coding gene presence of individual LTR GAG and POL domains. Shown are (A) the proportion of TEs with evidence of agglutination factor (GAG) domain present, (B) the proportion of TEs with evidence of all polyprotein domains present (aspartic protein-ase, integrase, reverse transcriptase, and RNaseH), (C) the proportion of TEs with both GAG and Polyprotein present in the same element. Families are shown as points and superfamily proportions as barplot.
(TIF)

S13 Fig. Predicted and observed relationship of age to TE length and distance to gene. Raw relationship (A & C) and predicted relationship (B & D) of TE length (A & B) and distance to gene (C & D).
(TIF)

S1 Table. Ten largest families in each superfamily, as shown left to right in plots.
(TXT)

S2 Table. Categories that each feature measured for each TE is classified into.
(TXT)

S3 Table. 14 families with at least 10 copies in the B73 genome, with at least 75% of copies coding for transposition related proteins.
(TXT)

S4 Table. 842 families with at least 10 copies in the B73 genome that lack coding representatives.
(TXT)

S5 Table. Mean methylation levels across superfamilies, averaged across all tissues, and averaged within a tissue.
(TXT)

S6 Table. TE families that lack methylatable cytosines (presented as family median values).
(TXT)

S1 Text. TIR annotation methods.
(PDF)

Acknowledgments

Author Contributions

Conceptualization: Michelle C. Stitzer, Sarah N. Anderson, Nathan M. Springer, Jeffrey Ross-Ibarra.

Data curation: Michelle C. Stitzer, Sarah N. Anderson.

Formal analysis: Michelle C. Stitzer, Sarah N. Anderson.

Funding acquisition: Michelle C. Stitzer, Jeffrey Ross-Ibarra.

Methodology: Michelle C. Stitzer, Sarah N. Anderson.

Software: Michelle C. Stitzer, Sarah N. Anderson.

Supervision: Nathan M. Springer, Jeffrey Ross-Ibarra.

Visualization: Michelle C. Stitzer.

Writing – original draft: Michelle C. Stitzer.

Writing – review & editing: Michelle C. Stitzer, Sarah N. Anderson, Nathan M. Springer, Jeffrey Ross-Ibarra.

References

1. Bennetzen JL, Kellogg EA. Do plants have a one-way ticket to genomic obesity? The Plant Cell. 1997; 9(9):1509. https://doi.org/10.1105/tpc.9.9.1509 PMID: 12237393

2. Lisch D. How important are transposons for plant evolution? Nature Reviews Genetics. 2013; 14 (1):49–61. PMID: 23247435

3. Oliver KR, McComb JA, Greene WK. Transposable elements: powerful contributors to angiosperm evolution and diversity. Genome Biology and Evolution. 2013; 5(10):1886–1901. https://doi.org/10.1093/gbe/evt141 PMID: 24065734

4. Fedoroff NV. Transposable elements, epigenetics, and genome evolution. Science. 2012; 338 (6108):758–767. https://doi.org/10.1126/science.338.6108.758 PMID: 23145453

5. Charlesworth B, Charlesworth D. The population dynamics of transposable elements. Genetical Research. 1983; 42(01):1–27.

6. Charlesworth B, Langley CH. The population genetics of Drosophila transposable elements. Annual Review of Genetics. 1989; 23(1):251–287. https://doi.org/10.1146/annurev.ge.23.120189.001343 PMID: 2559652

7. Kidwell MG, Lisch D. Transposable elements as sources of variation in animals and plants. Proceedings of the National Academy of Sciences. 1997; 94(15):7704–7711. https://doi.org/10.1073/pnas.94.15.7704 PMID: 9223252

8. Venner S, Feschotte C, Biémont C. Dynamics of transposable elements: towards a community ecology of the genome. Trends in Genetics. 2009; 25(7):317–323. https://doi.org/10.1016/j.tig.2009.05.003 PMID: 19540613

9. Linquist S, Cottenie K, Elliott TA, Saylor B, Kremer SC, Gregory TR. Applying ecological models to communities of genetic elements: the case of Neutral Theory. Molecular Ecology. 2015;. https://doi.org/10.1111/mec.13219 PMID: 25919906

10. Kremer SC, Linquist S, Saylor B, Elliott TA, Gregory TR, Cottenie K. Transposable element persistence via potential genome-level ecosystem engineering. BMC Genomics. 2020; 21:1–15. https://doi.org/10.1186/s12864-020-6763-1 PMID: 32429843

11. Bennetzen JL. Transposable element contributions to plant gene and genome evolution. Plant Molecular Biology. 2000; 42(1):251–269. https://doi.org/10.1023/A:1006344508454 PMID: 10688140

12. Thomas J, Pritham EJ. Helitrons, the eukaryotic rolling-circle transposable elements. In: Mobile DNA III. American Society of Microbiology; 2015. p. 893–926.

13. Labrador M, Corces VG. Interactions between transposable elements and the host genome. Mobile DNA II. 2002; p. 1008–1023.

14. Sultana T, Zamborlini A, Cristofari G, Lesage P. Integration site selection by retroviruses and transposable elements in eukaryotes. Nature Reviews Genetics. 2017; 18(5):292. https://doi.org/10.1038/nrg.2017.7 PMID: 28286338

15. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics. 2007; 8(12):973–982. https://doi.org/10.1038/nrg2165 PMID: 17984973

16. Finnegan DJ. Eukaryotic transposable elements and genome evolution. Trends in Genetics. 1989; 5:103–107. https://doi.org/10.1016/0168-9525(89)90039-5 PMID: 2543105

17. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and Genome Research. 2005; 110(1-4):462–467. https://doi.org/10.1159/000084979 PMID: 16093699

18. Kapitonov VV, Jurka J. A universal classification of eukaryotic transposable elements implemented in Repbase. Nature Reviews Genetics. 2008; 9(5):411–412. https://doi.org/10.1038/nrg2165-c1 PMID: 18421312

19. Piégu B, Bire S, Arensburger P, Bigot Y. A survey of transposable element classification systems–a call for a fundamental update to meet the challenge of their diversity and complexity. Molecular Phylogenetics and Evolution. 2015; 86:90–109. https://doi.org/10.1016/j.ympev.2015.03.009 PMID: 25797922

20. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. Reply: A unified classification system for eukaryotic transposable elements should reflect their phylogeny. Nature Reviews Genetics. 2009; 10(4):276. https://doi.org/10.1038/nrg2165-c3

21. Wicker T. So many repeats and so little time: how to classify transposable elements. In: Plant Transposable Elements. Springer; 2012. p. 1–15.

22. Johns M, Mottinger J, Freeling M. A low copy number, copia-like transposon in maize. The EMBO journal. 1985; 4(5):1093–1101. https://doi.org/10.1002/j.1460-2075.1985.tb03745.x PMID: 2988938

23. Sutton W, Gerlach W, Peacock W, Schwartz D. Molecular analysis of Ds controlling element mutations at the Adh1 locus of maize. Science. 1984; 223(4642):1265–1268. https://doi.org/10.1126/science.223.4642.1265 PMID: 17759347

24. Hake S, Walbot V. The genome of Zea mays, its organization and homology to related grasses. Chromosoma. 1980; 79(3):251–270. https://doi.org/10.1007/BF00327318

25. Sanz-Alferez S, SanMiguel P, Jin YK, Springer PS, Bennetzen JL. Structure and evolution of the Cinful retrotransposon family of maize. Genome. 2003; 46(5):745–752. https://doi.org/10.1139/g03-061 PMID: 14608391

26. Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, Deragon JM, et al. Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. PLoS Genetics. 2009; 5(11):e1000732. https://doi.org/10.1371/journal.pgen.1000732 PMID: 19936065

27. SanMiguel P, Vitte C. The LTR-retrotransposons of maize. In: Handbook of Maize. Springer; 2009. p. 307–327.

28. Diez CM, Meca E, Tenaillon MI, Gaut BS. Three Groups of Transposable Elements with Contrasting Copy Number Dynamics and Host Responses in the Maize (*Zea mays* ssp. *mays*) Genome. PLoS Genetics. 2014; 10(4):e1004298. https://doi.org/10.1371/journal.pgen.1004298 PMID: 24743518

29. Vicient CM. Transcriptional activity of transposable elements in maize. BMC Genomics. 2010; 11 (1):601. https://doi.org/10.1186/1471-2164-11-601 PMID: 20973992

30. Cresse AD, Hulbert SH, Brown WE, Lucas JR, Bennetzen JL. Mu1-related transposable elements of maize preferentially insert into low copy number DNA. Genetics. 1995; 140(1):315–324. https://doi.org/10.1093/genetics/140.1.315 PMID: 7635296

31. Zhong CX, Marshall JB, Topp C, Mroczek R, Kato A, Nagaki K, et al. Centromeric retroelements and satellites interact with maize kinetochore protein CENH3. The Plant Cell. 2002; 14(11):2825–2836. https://doi.org/10.1105/tpc.006106 PMID: 12417704

32. Eichten SR, Ellis NA, Makarevitch I, Yeh CT, Gent JI, Guo L, et al. Spreading of heterochromatin is limited to specific families of maize retrotransposons. PLoS Genetics. 2012; 8(12):e1003127. https://doi.org/10.1371/journal.pgen.1003127 PMID: 23271981

33. Noshay JM, Anderson SN, Zhou P, Ji L, Ricci W, Lu Z, et al. Monitoring the interplay between transposable element families and DNA methylation in maize. PLoS Genetics. 2019; 15(9):e1008291. https://doi.org/10.1371/journal.pgen.1008291 PMID: 31498837

34. Elliott TA, Gregory TR. Do larger genomes contain more diverse transposable elements? BMC Evolutionary Biology. 2015; 15(1):69. https://doi.org/10.1186/s12862-015-0339-8 PMID: 25896861

35. Ågren JA, Wright SI. Co-evolution between transposable elements and their hosts: a major factor in genome size evolution? Chromosome Research. 2011; 19(6):777.

36. Ågren JA, Greiner S, Johnson MT, Wright SI. No evidence that sex and transposable elements drive genome size variation in evening primroses. Evolution. 2015; 69(4):1053–1062. https://doi.org/10.1111/evo.12627 PMID: 25690700

37. Sotero-Caio CG, Platt RN, Suh A, Ray DA. Evolution and diversity of transposable elements in vertebrate genomes. Genome Biology and Evolution. 2017; 9(1):161–177. https://doi.org/10.1093/gbe/evw264 PMID: 28158585

38. Bast J, Jaron K, Schuseil D, Roze D, Schwander T. Asexual reproduction drives the reduction of transposable element load. eLife. 2019; 8:e48548. https://doi.org/10.7554/eLife.48548 PMID: 31486772

39. Brookfield JF. The ecology of the genome—mobile DNA elements and their hosts. Nature Reviews Genetics. 2005; 6(2):128–136. https://doi.org/10.1038/nrg1524 PMID: 15640810

40. Kidwell MG, Lisch DR. Transposable elements as sources of genomic variation. Mobile DNA II. 2002; p. 59–90.

41. Promislow DE, Jordan IK, McDonald J. Genomic demography: a life-history analysis of transposable element evolution. Proceedings of the Royal Society of London B: Biological Sciences. 1999; 266 (1428):1555–1560. https://doi.org/10.1098/rspb.1999.0815 PMID: 10467744

42. Abrusán G, Krambeck HJ. Competition may determine the diversity of transposable elements. Theoretical Population Biology. 2006; 70(3):364–375. https://doi.org/10.1016/j.tpb.2006.05.001 PMID: 16814337

43. Saylor B, Elliott TA, Linquist S, Kremer SC, Gregory TR, Cottenie K. A novel application of ecological analyses to assess transposable element distributions in the genome of the domestic cow, Bos taurus. Genome. 2013; 56(9):521–533. https://doi.org/10.1139/gen-2012-0162 PMID: 24168673

44. Jones CG, Lawton JH, Shachak M. Organisms as ecosystem engineers. In: Ecosystem management. Springer; 1994. p. 130–147.

45. Pelletier F, Garant D, Hendry A. Eco-evolutionary dynamics. Philosophical Transactions of the Royal Society B: Biological Sciences. 2009; 364(1523):1483. https://doi.org/10.1098/rstb.2009.0027

46. Post DM, Palkovacs EP. Eco-evolutionary feedbacks in community and ecosystem ecology: interactions between the ecological theatre and the evolutionary play. Philosophical Transactions of the Royal Society B: Biological Sciences. 2009; 364(1523):1629–1640. https://doi.org/10.1098/rstb.2009.0012 PMID: 19414476

47. Linquist S, Saylor B, Cottenie K, Elliott TA, Kremer SC, Ryan Gregory T. Distinguishing ecological from evolutionary approaches to transposable elements. Biological Reviews. 2013; 88(3):573–584. https://doi.org/10.1111/brv.12017 PMID: 23347261

48. Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, et al. Improved maize reference genome with single molecule technologies. Nature. 2017; 546(7659):524–527. https://doi.org/10.1038/nature22971 PMID: 28605751

49. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. Science. 2009; 326(5956):1112–1115. https://doi.org/10.1126/science.1178534 PMID: 19965430

50. Slotkin RK. The case for not masking away repetitive DNA. Mobile DNA. 2018; 9(1):15. https://doi.org/10.1186/s13100-018-0120-9 PMID: 29743957

51. Anderson SN, Zynda G, Song J, Han Z, Vaughn M, Li Q, et al. Subtle Perturbations of the Maize Methylome Reveal Genes and Transposons Silenced by Chromomethylase or RNA-Directed DNA Methylation Pathways. G3: Genes, Genomes, Genetics. 2018; p. g3–200284. https://doi.org/10.1534/g3.118.200284 PMID: 29618467

52. Anderson SN, Stitzer MC, Zhou P, Ross-Ibarra J, Hirsch CD, Springer NM. Dynamic Patterns of Transcript Abundance of Transposable Element Families in Maize. G3: Genes, Genomes, Genetics. 2019; https://doi.org/10.1534/g3.119.400431 PMID: 31506319

53. Bukowski R, Guo X, Lu Y, Zou C, He B, Rong Z, et al. Construction of the third-generation Zea mays haplotype map. GigaScience. 2018; 7(4):1. https://doi.org/10.1093/gigascience/gix134 PMID: 29300887

54. Hill WG, Robertson A. The effect of linkage on limits to artificial selection. Genetics Research. 1966; 8 (3):269–294. https://doi.org/10.1017/S0016672300010156 PMID: 5980116

55. Swigoňová Z, Lai J, Ma J, Ramakrishna W, Llaca V, Bennetzen JL, et al. Close split of sorghum and maize genome progenitors. Genome Research. 2004; 14(10a):1916–1923. https://doi.org/10.1101/gr.2332504 PMID: 15466289

56. Schnable JC, Springer NM, Freeling M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proceedings of the National Academy of Sciences. 2011; 108(10):4069–4074. https://doi.org/10.1073/pnas.1101368108 PMID: 21368132

57. Bennetzen J, Liu R, Ma J, Pontaroli A. Maize genome structure and rearrangement. Maydica. 2005; 50(3/4):387.

58. Han Y, Qin S, Wessler SR. Comparison of class 2 transposable elements at superfamily resolution reveals conserved and distinct features in cereal grass genomes. BMC Genomics. 2013; 14(1):71. https://doi.org/10.1186/1471-2164-14-71 PMID: 23369001

59. Jiang N, Wessler SR. Insertion preference of maize and rice miniature inverted repeat transposable elements as revealed by the analysis of nested elements. The Plant Cell. 2001; 13(11):2553–2564. https://doi.org/10.1105/tpc.010235 PMID: 11701888

60. Liu S, Yeh CT, Ji T, Ying K, Wu H. Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. PLoS Genetics. 2009; 5 (11):e1000733–e1000733. https://doi.org/10.1371/journal.pgen.1000733 PMID: 19936291

61. Wright SI, Le QH, Schoen DJ, Bureau TE. Population dynamics of an Ac-like transposable element in self-and cross-pollinating Arabidopsis. Genetics. 2001; 158(3):1279–1288. https://doi.org/10.1093/genetics/158.3.1279 PMID: 11454774

62. Tenaillon MI, Hollister JD, Gaut BS. A triptych of the evolution of plant transposable elements. Trends in Plant Science. 2010; 15(8):471–478. https://doi.org/10.1016/j.tplants.2010.05.003 PMID: 20541961

63. Eickbush TH, Jamburuthugoda VK. The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Research. 2008; 134(1-2):221–234. https://doi.org/10.1016/j.virusres.2007.12.010 PMID: 18261821

64. Yuan YW, Wessler SR. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. Proceedings of the National Academy of Sciences. 2011; 108(19):7884–7889. https://doi.org/10.1073/pnas.1104208108 PMID: 21518873

65. Kapitonov VV, Jurka J. Helitrons on a roll: eukaryotic rolling-circle transposons. TRENDS in Genetics. 2007; 23(10):521–529. https://doi.org/10.1016/j.tig.2007.08.004 PMID: 17850916

66. Du J, Tian Z, Hans CS, Laten HM, Cannon SB, Jackson SA, et al. Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison. The Plant Journal. 2010; 63(4):584–598. https://doi.org/10.1111/j.1365-313X.2010.04263.x PMID: 20525006

67. Neumann P, Navrátilová A, Koblížková A, Kejnovský E, Hřibová E, Hobza R, et al. Plant centromeric retrotransposons: a structural and cytogenetic perspective. Mobile DNA. 2011; 2(1):4. https://doi.org/10.1186/1759-8753-2-4 PMID: 21371312

68. Slotkin RK, Nuthikattu S, Jiang N. The impact of transposable elements on gene and genome evolution. In: Plant Genome Diversity Volume 1. Springer; 2012. p. 35–58.

69. Bureau TE, Wessler SR. Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. The Plant Cell. 1994; 6(6):907–916. https://doi.org/10.1105/tpc.6.6.907 PMID: 8061524

70. Wicker T, Yu Y, Haberer G, Mayer KF, Marri PR, Rounsley S, et al. DNA transposon activity is associated with increased mutation rates in genes of rice and other grasses. Nature Communications. 2016; 7:12790. https://doi.org/10.1038/ncomms12790 PMID: 27599761

71. Singer T, Yordan C, Martienssen RA. Robertson's Mutator transposons in A. thaliana are regulated by the chromatin-remodeling gene Decrease in DNA Methylation (DDM1). Genes & Development. 2001; 15(5):591–602. https://doi.org/10.1101/gad.193701

72. Jiang N, Ferguson AA, Slotkin RK, Lisch D. Pack-Mutator–like transposable elements (Pack-MULEs) induce directional modification of genes through biased insertion and DNA acquisition. Proceedings of the National Academy of Sciences. 2011; p. 201010814.

73. Anderson SN, Stitzer MC, Brohammer AB, Zhou P, Noshay JM, O'Connor CH, et al. Transposable elements contribute to dynamic genome content in maize. The Plant Journal. 2019;. https://doi.org/10.1111/tpj.14489 PMID: 31381222

74. Dooner HK, Wang Q, Huang JT, Li Y, He L, Xiong W, et al. Spontaneous mutations in maize pollen are frequent in some lines and arise mainly from retrotranspositions and deletions. Proceedings of the National Academy of Sciences. 2019; 116(22):10734–10743. https://doi.org/10.1073/pnas.1903809116 PMID: 30992374

75. Zhang Q, Arbuckle J, Wessler SR. Recent, extensive, and preferential insertion of members of the miniature inverted-repeat transposable element family Heartbreaker into genic regions of maize. Proceedings of the National Academy of Sciences. 2000; 97(3):1160–1165. https://doi.org/10.1073/pnas.97.3.1160 PMID: 10655501

76. Meyers BC, Tingey SV, Morgante M. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Research. 2001; 11(10):1660–1676. https://doi.org/10.1101/gr.188201 PMID: 11591643

77. Estep M, DeBarry J, Bennetzen J. The dynamics of LTR retrotransposon accumulation across 25 million years of panicoid grass evolution. Heredity. 2013; 110(2):194–204. https://doi.org/10.1038/hdy.2012.99 PMID: 23321774

78. Tenaillon MI, Hufford MB, Gaut BS, Ross-Ibarra J. Genome Size and Transposable Element Content as Determined by High-Throughput Sequencing in Maize and Zea luxurians. Genome Biology and Evolution. 2011; 3:219–229. https://doi.org/10.1093/gbe/evr008 PMID: 21296765

79. Ross-Ibarra J, Tenaillon M, Gaut BS. Historical divergence and gene flow in the genus Zea. Genetics. 2009;. https://doi.org/10.1534/genetics.108.097238 PMID: 19153259

80. Gerlach W, Dennis E, Peacock W, Clegg M. The Ds1 controlling element family in maize andTripsacum. Journal of Molecular Evolution. 1987; 26(4):329–334. https://doi.org/10.1007/BF02101151 PMID: 2836599

81. Purugganan MD, Wessler SR. Molecular evolution of magellan, a maize Ty3/gypsy-like retrotransposon. Proceedings of the National Academy of Sciences. 1994; 91(24):11674–11678. https://doi.org/10.1073/pnas.91.24.11674 PMID: 7972122

82. Roulin A, Piegu B, Fortune PM, Sabot F, D'hont A, Manicacci D, et al. Whole genome surveys of rice, maize and sorghum reveal multiple horizontal transfers of the LTR-retrotransposon Route66 in Poaceae. BMC Evolutionary Biology. 2009; 9(1):58. https://doi.org/10.1186/1471-2148-9-58 PMID: 19291296

83. SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, et al. Nested retrotransposons in the intergenic regions of the maize genome. Science. 1996; 274(5288):765. https://doi.org/10.1126/science.274.5288.765 PMID: 8864112

84. Malik HS, Eickbush TH. Modular evolution of the integrase domain in the Ty3/Gypsy class of LTR retrotransposons. Journal of Virology. 1999; 73(6):5186–5190. https://doi.org/10.1128/JVI.73.6.5186-5190.1999 PMID: 10233986

85. Matsuoka Y, Vigouroux Y, Goodman MM, Sanchez J, Buckler E, Doebley J. A single domestication for maize shown by multilocus microsatellite genotyping. Proceedings of the National Academy of Sciences. 2002; 99(9):6080–6084. https://doi.org/10.1073/pnas.052125199 PMID: 11983901

86. Piperno DR, Ranere AJ, Holst I, Iriarte J, Dickau R. Starch grain and phytolith evidence for early ninth millennium BP maize from the Central Balsas River Valley, Mexico. Proceedings of the National Academy of Sciences. 2009; 106(13):5019–5024. https://doi.org/10.1073/pnas.0812525106

87. Studer A, Zhao Q, Ross-Ibarra J, Doebley J. Identification of a functional transposon insertion in the maize domestication gene tb1. Nature Genetics. 2011; 43(11):1160–1163. https://doi.org/10.1038/ng. 942 PMID: 21946354

88. Yang Q, Li Z, Li W, Ku L, Wang C, Ye J, et al. CACTA-like transposable element in ZmCCT attenuated photoperiod sensitivity and accelerated the postdomestication spread of maize. Proceedings of the National Academy of Sciences. 2013; 110(42):16969–16974. https://doi.org/10.1073/pnas. 1310949110 PMID: 24089449

89. Huang C, Sun H, Xu D, Chen Q, Liang Y, Wang X, et al. ZmCCT9 enhances maize adaptation to higher latitudes. Proceedings of the National Academy of Sciences. 2018; 115(2):E334–E341. https:// doi.org/10.1073/pnas.1718058115 PMID: 29279404

90. Hutchinson GE. Concluding Remarks. Cold Spring Harbor Symposia on Quantitative Biology. 1957; 22:415–427. https://doi.org/10.1101/SQB.1957.022.01.039

91. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, et al. The Sorghum bicolor genome and the diversification of grasses. Nature. 2009; 457(7229):551. https://doi.org/10. 1038/nature07723 PMID: 19189423

92. Le Rouzic A, Dupas S, Capy P. Genome ecosystem and transposable elements species. Gene. 2007; 390(1):214–220. https://doi.org/10.1016/j.gene.2006.09.023 PMID: 17188821

93. Wagner A. Cooperation is fleeting in the world of transposable elements. PLoS Computational Biology. 2006; 2(12):e162. https://doi.org/10.1371/journal.pcbi.0020162 PMID: 17140282

94. Clark RM, Tavaré S, Doebley J. Estimating a nucleotide substitution rate for maize from polymorphism at a major domestication locus. Molecular Biology and Evolution. 2005; 22(11):2304–2312. https://doi. org/10.1093/molbev/msi228 PMID: 16079248

95. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. The paleontology of intergene retrotransposons of maize. Nature Genetics. 1998; 20(1):43–45. https://doi.org/10.1038/1695 PMID: 9731528

96. Weng ML, Becker C, Hildebrandt J, Rutter MT, Shaw RG, Weigel D, et al. Fine-Grained Analysis of Spontaneous Mutation Spectrum and Frequency in Arabidopsis thaliana. Genetics. 2018; p. genetics–301721. https://doi.org/10.1534/genetics.118.301721 PMID: 30514707

97. Petrov DA, Aminetzach YT, Davis JC, Bensasson D, Hirsh AE. Size matters: non-LTR retrotransposable elements and ectopic recombination in Drosophila. Molecular Biology and Evolution. 2003; 20 (6):880–892. https://doi.org/10.1093/molbev/msg102 PMID: 12716993

98. Shen JJ, Dushoff J, Bewick AJ, Chain FJ, Evans BJ. Genomic dynamics of transposable elements in the western clawed frog (Silurana tropicalis). Genome Biology and Evolution. 2013; 5(5):998–1009. https://doi.org/10.1093/gbe/evt065 PMID: 23645600

99. Hollister JD, Gaut BS. Population and evolutionary dynamics of Helitron transposable elements in Arabidopsis thaliana. Molecular Biology and Evolution. 2007; 24(11):2515–2524. https://doi.org/10.1093/ molbev/msm197 PMID: 17890239

100. Fu H, Zheng Z, Dooner HK. Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. Proceedings of the National Academy of Sciences. 2002; 99 (2):1082–1087. https://doi.org/10.1073/pnas.022635499 PMID: 11792865

101. Gilbert DM, Bridges MC, Strother AE, Burckhalter CE, Burnette JM, Hancock CN. Precise repair of mPing excision sites is facilitated by target site duplication derived microhomology. Mobile DNA. 2015; 6(1):15. https://doi.org/10.1186/s13100-015-0046-4 PMID: 26347803

102. Cavrak VV, Lettner N, Jamge S, Kosarewicz A, Bayer LM, Scheid OM. How a retrotransposon exploits the plant's heat stress response for its activation. PLoS Genetics. 2014; 10(1):e1004115. https://doi. org/10.1371/journal.pgen.1004115 PMID: 24497839

103. Gent JI, Ellis NA, Guo L, Harkess AE, Yao Y, Zhang X, et al. CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize. Genome Research. 2013; 23(4):628–637. https://doi.org/ 10.1101/gr.146985.112 PMID: 23269663

104. Li Q, Gent JI, Zynda G, Song J, Makarevitch I, Hirsch CD, et al. RNA-directed DNA methylation enforces boundaries between heterochromatin and euchromatin in the maize genome. Proceedings of the National Academy of Sciences. 2015; 112(47):14728–14733. https://doi.org/10.1073/pnas. 1514680112 PMID: 26553984

105. Medstrand P, Van De Lagemaat LN, Mager DL. Retroelement distributions in the human genome: variations associated with age and proximity to genes. Genome Research. 2002; 12(10):1483–1495. https://doi.org/10.1101/gr.388902 PMID: 12368240

106. Hollister JD, Gaut BS. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Research. 2009; 19 (8):1419–1428. https://doi.org/10.1101/gr.091678.109 PMID: 19478138

107. Rodgers-Melnick E, Bradbury PJ, Elshire RJ, Glaubitz JC, Acharya CB, Mitchell SE, et al. Recombination in diverse maize is stable, predictable, and associated with genetic load. Proceedings of the National Academy of Sciences. 2015; p. 201413864. https://doi.org/10.1073/pnas.1413864112 PMID: 25775595

108. Dooner HK, He L. Polarized gene conversion at the bz locus of maize. Proceedings of the National Academy of Sciences. 2014; 111(38):13918–13923. https://doi.org/10.1073/pnas.1415482111

109. Bennetzen J. In: Maize Genome Structure and Evolution; 2009. p. 179–199.

110. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics. 2008; 9(1):1. https://doi.org/10.1186/1471-2105-9-18

111. Steinbiss S, Willhoeft U, Gremme G, Kurtz S. Fine-grained annotation and classification of de novo predicted LTR retrotransposons. Nucleic Acids Research. 2009; p. gkp759. https://doi.org/10.1093/nar/gkp759 PMID: 19786494

112. Llorens C, Futami R, Covelli L, Domínguez-Escribá L, Viu JM, Tamarit D, et al. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Research. 2010; p. gkq1061. https://doi.org/10.1093/nar/gkq1061 PMID: 21036865

113. Brian H, Papanicolaou A. Transdecoder (Find Coding Regions Within Transcripts)—http://transdecoder.github.io; 2018.

114. Eddy S. Hmmer 3.1—http://hmmer.org/; 2018.

115. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution. 2013; 30(4):772–780. https://doi.org/10.1093/molbev/mst010 PMID: 23329690

116. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PloS One. 2010; 5(3):e9490. https://doi.org/10.1371/journal.pone.0009490 PMID: 20224823

117. Bergman CM, Bensasson D. Recent LTR retrotransposon insertion contrasts with waves of non-LTR insertion since speciation in Drosophila melanogaster. Proceedings of the National Academy of Sciences. 2007; 104(27):11340–11345. https://doi.org/10.1073/pnas.0702552104 PMID: 17592135

118. Fiston-Lavier AS, Vejnar CE, Quesneville H. Transposable element sequence evolution is influenced by gene context. arXiv preprint arXiv:12090176. 2012;.

119. Brookfield JF, Johnson LJ. The evolution of mobile DNAs: when will transposons create phylogenies that look as if there is a master gene? Genetics. 2006; 173(2):1115–1123. https://doi.org/10.1534/genetics.104.027219 PMID: 16790583

120. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2018; xx:xxx–xxx.

121. R Core Team. R: A Language and Environment for Statistical Computing; 2018. Available from: https://www.R-project.org/.

122. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Computational Biology. 2013; 9(8):e1003118. https://doi.org/10.1371/journal.pcbi.1003118 PMID: 23950696

123. Walley JW, Sartor RC, Shen Z, Schmitz RJ, Wu KJ, Urich MA, et al. Integration of omic networks in a developmental atlas of maize. Science. 2016; 353(6301):814–818. https://doi.org/10.1126/science.aag1125 PMID: 27540173

124. Kryuchkova-Mostacci N, Robinson-Rechavi M. A benchmark of gene expression tissue-specificity metrics. Briefings in Bioinformatics. 2016; 18(2):205–214.

125. Zhang Y, Ngu DW, Carvalho D, Liang Z, Qiu Y, Roston RL, et al. Differentially regulated orthologs in sorghum and the subgenomes of maize. The Plant Cell. 2017; 29(8):1938–1951. https://doi.org/10.1105/tpc.17.00354 PMID: 28733421

126. Schnable J. Grass Syntenic Gene List sorghum v3 maize v3/4 with teff and oropetium v2. 2019;

127. Ogut F, Bian Y, Bradbury PJ, Holland JB. Joint-multiple family linkage analysis predicts within-family variation better than single-family analysis of the maize nested association mapping population. Heredity. 2015; 114(6):552–563. https://doi.org/10.1038/hdy.2014.123 PMID: 25585918

128. Monaco MK, Stein J, Naithani S, Wei S, Dharmawardhana P, Kumari S, et al. Gramene 2013: Comparative plant genomics resources. Nucleic Acids Research. 2014; 42(D1):D1193–D1199. https://doi.org/10.1093/nar/gkt1110 PMID: 24217918

129. Murray K, Müller S, Turlach B. Fast and flexible methods for monotone polynomial fitting. Journal of Statistical Computation and Simulation. 2016; 86(15):2946–2966. https://doi.org/10.1080/00949655.2016.1139582

130. Rodgers-Melnick E, Vera DL, Bass HW, Buckler ES. Open chromatin reveals the functional maize genome. Proceedings of the National Academy of Sciences. 2016; 113(22):E3177–E3184. https://doi.org/10.1073/pnas.1525244113 PMID: 27185945

131. Eichten SR, Vaughn MW, Hermanson PJ, Springer NM. Variation in DNA methylation patterns is more common among maize inbreds than among tissues. The Plant Genome. 2013; 6(2). https://doi.org/10.3835/plantgenome2012.06.0009

132. Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics. 2009; 10(1):232. https://doi.org/10.1186/1471-2105-10-232 PMID: 19635165

133. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26(6):841–842. https://doi.org/10.1093/bioinformatics/btq033 PMID: 20110278

134. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002; 2(3):18–22.

135. Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. Journal of Computational and Graphical Statistics. 2015; 24(1):44–65. https://doi.org/10.1080/10618600.2014.907095

136. Greenwell BM. pdp: An R Package for Constructing Partial Dependence Plots. The R Journal. 2017; 9 (1):421–436. https://doi.org/10.32614/RJ-2017-016