Title

Birth and Death of LTR-Retrotransposons in Aegilops tauschii

Permalink

https://escholarship.org/uc/item/5s0463cx

Journal

Genetics, 210(3)

ISSN

0016-6731

Authors

Dai, Xiongtao
Wang, Hao
Zhou, Hongye
et al.

Publication Date

2018-11-01

DOI

10.1534/genetics.118.301198

Peer reviewed

Birth and Death of LTR-Retrotransposons in *Aegilops tauschii*

Xiongtao Dai ,*,[1] Hao Wang,† Hongye Zhou ,† Le Wang,‡ Jan Dvořák,‡ Jeffrey L. Bennetzen,† and Hans-Georg Müller *

*Department of Statistics, University of California, Davis, California 95616, †Department of Genetics, University of Georgia, Athens, Georgia 30602, and ‡Department of Plant Sciences, University of California, Davis, California 95616

ORCID ID: 0000-0002-6996-5930 (X.D.)

ABSTRACT Long terminal repeat-retrotransposons (LTR-RTs) are a major component of all flowering plant genomes. To analyze the time dynamics of LTR-RTs, we modeled the insertion rates of the 35 most abundant LTR-RT families in the genome of *Aegilops tauschii*, one of the progenitors of wheat. Our model of insertion rate (birth) takes into account random variation in LTR divergence and the deletion rate (death) of LTR-RTs. Modeling the death rate is crucial because ignoring it would underestimate insertion rates in the distant past. We rejected the hypothesis of constancy of insertion rates for all 35 families and showed by simulations that our hypothesis test controlled the false-positive rate. LTR-RT insertions peaked from 0.064 to 2.39 MYA across the 35 families. Among other effects, the average age of elements within a family was negatively associated with recombination rate along a chromosome, with proximity to the closest gene, and weakly associated with the proximity to its 5′ end. Elements within a family that were near genes colinear with genes in the genome of tetraploid emmer wheat tended to be younger than those near noncolinear genes. We discuss these associations in the context of genome evolution and stability of genome sizes in the tribe Triticeae. We demonstrate the general utility of our models by analyzing the two most abundant LTR-RT families in *Arabidopsis lyrata*, and show that these families differed in their insertion dynamics. Our estimation methods are available in the R package TE on CRAN.

KEYWORDS transposable elements; insertion rates; demography; population dynamics

LONG terminal repeat-retrotransposons (LTR-RTs) are present in virtually all studied eukaryotes, and make up the majority of the nuclear genomes in many flowering plants (Bennetzen and Wang 2014). LTR-RTs are classified into five superfamilies: *Copia*, *Gypsy*, *Bel-Pao*, retrovirus, and endogenous retrovirus (ERV), and among them, *Copia* and *Gypsy* are predominant in plant genomes, which each contain hundreds of different LTR-RT families that are operationally distinguished by their different LTR sequences (Wicker *et al.* 2007). Any single plant will routinely contain several hundred different LTR-RT families, of which a few will be highly abundant (contributing hundreds to thousands of copies) but most will have only one to five intact members (Baucom *et al.* 2009a,b). Variation in the copy numbers of LTR-RTs is the major factor responsible for the huge (>3000-fold) genome size variation in flowering plants (Leitch and Leitch 2013). Because LTR-RTs transpose via integration of a reverse-transcribed transcript, while leaving the donor element in place, they can rapidly increase their number in a genome. The most dramatic case of this amplification has been observed in the *Zea* lineage, where the massive transposition of several different LTR-RT families in the ancestors of *Zea luxurians* led to more than a doubling of genome size in <2 MY, resulting in the addition of >2400 Mb of new LTR-RT DNA in that short time period (Estep *et al.* 2013).

The transposition of different LTR-RT families exhibits episodic and apparently stochastic activation over evolutionary time (Wicker and Keller 2007; Baucom *et al.* 2009b). Because the two LTRs of a single LTR-RT are usually identical at the time of insertion, insertion dates can be estimated by investigating the degree of LTR divergence within a single LTR-RT (SanMiguel *et al.* 1998). Such analyses indicate that individual LTR-RT families exhibit different histories of

doi: https://doi.org/10.1534/genetics.118.301198

Manuscript received May 30, 2018; accepted for publication August 20, 2018; published Early Online August 27, 2018.

Supplemental material available at Figshare: https://doi.org/10.25386/genetics.6988631.

[1]Corresponding author: Department of Statistics, Iowa State University, Ames, IA 50011. E-mail: xdai@iastate.edu

"amplification bursts" in any given lineage, and that this accounts for the great variation in the structure of even closely related plant genomes. Even in small plant genomes, like that of rice (*Oryza sativa*, ~400 Mb), LTR-RTs can add hundreds of megabases of new LTR-RTs per MY. However, this process does not always lead to genome size expansion over evolutionary time, because there are also processes for the rapid removal of DNA from flowering plant genomes (Devos *et al.* 2002; Ma *et al.* 2004; Vitte *et al.* 2007; Hawkins *et al.* 2009). Unequal homologous recombination (HR) between the LTRs of a single LTR-RT leads to the loss of all internal sequences and the generation of a solo-LTR, which is a process that attenuates transposition-driven genome growth. DNA loss by accumulated deletions caused by illegitimate recombination can slow or even reverse genome growth. The mechanism(s) of illegitimate recombination responsible for genome shrinkage has not been proven, but deletions resulting from the repair of double-strand breaks or adjacent single-strand nicks appear to be the most important drivers (Kirik *et al.* 2000; Devos *et al.* 2002; Vaughn and Bennetzen 2014; Schiml *et al.* 2016).

The relative rates of amplification and removal of LTR-RTs and other unnecessary DNA varies across plant lineages (Vitte and Bennetzen 2006), and may also be quite variable across regions in the plant genome (Ma and Bennetzen 2006) and over evolutionary time within a lineage (Estep *et al.* 2013). This dynamic state of plant genomes creates the raw material for natural selection, especially when one considers that a high percentage of transposable element (TE) insertions of all types can lead to altered regulation, both genetic and epigenetic, of nearby genes (Lisch and Bennetzen 2011). Understanding the significance of genome dynamics created by TE activities and rates of genome change will require more accurate parametrization and modeling than any of the isolated observations published to date. This study, which focused on modeling the dynamics of the LTR-RT families during the evolution of the *Aegilops tauschii* genome, provides an important step in that direction.

Ae. tauschii is one of the three diploid progenitors of hexaploid bread wheat. It has a large genome, ~4.3 Gb, that contains at least 66% LTR-RTs (Luo *et al.* 2017), mostly present as nested arrays of TEs between tiny gene islands (Gottlieb *et al.* 2013). These intergenic arrays are entirely replaced in a span of 3–4 MY driven by the deletion of old elements and the insertion of new ones (Dubcovsky and Dvorak 2007).

The dynamic nature of the *Ae. tauschii* LTR-RTs is employed here in modeling their biodemography. The insertion rates of LTR-RTs have been analyzed previously in many species, including *O. sativa* (Vitte *et al.* 2007; Wang *et al.* 2008; Baucom *et al.* 2009b), Triticeae (Wicker and Keller 2007), wheat chromosome 3B (Choulet *et al.* 2010), maize (SanMiguel *et al.* 1998), and *Arabidopsis* (Wicker and Keller 2007), but an explicit statistical modeling approach was not used. Statistical models have been proposed for analyzing the dynamics of TEs in *Drosophila* (Charlesworth and Langley 1989), *Saccharomyces cerevisiae* (Promislow *et al.* 1999), *Arabidopsis thaliana* (Hollister and Gaut 2007), *Hylobates* (Wacholder *et al.* 2014), and hominids (Marchani *et al.* 2009; Levy *et al.* 2017).

In this work, we formulate the insertion/deletion dynamics of LTR-RTs in terms of birth/death processes that change the age distribution over time, building on a model from biodemography (Müller *et al.* 2007), and recover the insertion rates for each LTR-RT family in the *Ae. tauschii* genome with ≥ 50 elements. A key difference between the age distribution and the insertion rate is that the former describes the ages of only the intact elements that survived the deletion process to the present, while the latter is the rate of insertion for all surviving and deleted LTR-RTs. We reject the hypothesis that LTR-RTs were inserted into the *Ae. tauschii* genome at a uniform rate with high significance. Since the removal of LTR-RTs from a genome cannot be easily dated, we conduct a sensitivity analysis to investigate different scenarios of death rates and their influence on insertion rates. In a regression analysis, we find that death rates, as proxied by the average age of intact elements, were associated with several genomic factors.

Materials and Methods

LTR-RTs

Intact LTR-RTs with a target site duplication (TSD) were identified by using LTR_FINDER (Xu and Wang 2007) and LTRharvest (Ellinghaus *et al.* 2008) to scan the *Ae. tauschii* genome sequence v4.0 (Luo *et al.* 2017), and by combining nonredundant predictions of the two program tools. An intact LTR-RT element was identified if the element showed all of the following characteristics: (1) highly similar 5′ and 3′ LTRs, (2) TG-CA termini of the LTRs, and (3) exact TSD [*e.g.*, see Ma *et al.* (2004)]. Artificial predictions were excluded by manual inspection, with more details included in the Supplemental Materials. A group of elements were classified into a family if their 25-bp TE ends exhibited ≥80% identity.

A total of 18,024 intact copies of 390 LTR-RT families were identified, and we performed the demographic analysis on 15,781 elements in the 35 largest LTR-RT families, all with ≥ 50 copies. The 35 families consisted of 9 *Copia* and 26 *Gypsy* families (Supplemental Material, Table S1). The divergence of an LTR-RT was defined as the number of mismatches in the two LTRs divided by the LTR length. Indels were not included in this analysis.

Additionally, 14,481 solo-LTRs were identified by applying RepeatMasker (Smit 2004) to search the *Ae. tauschii* genome with intact LTR-RT masked, using a solo-LTR sequence library built from the 5′ LTRs of the intact elements. TSDs of 4–6 bp at both ends of a solo-LTR were required.

To demonstrate that our modeling framework is applicable to other species, we annotated LTR-RTs in the *A. lyrata* genome. We found 397 intact elements in 38 LTR-RT families (Table S2) and analyzed the two largest families, namely a

Gypsy and a *Copia* family with 183 and 58 copies, respectively. No counterparts of these two *A. lyrata* LTR-RT families were found among those annotated in *A. thaliana* (Lamesch *et al.* 2011) using the Basic Local Alignment Search Tool and the family allocation system proposed by Wicker *et al.* (2007). An element in the largest *Gypsy* family with an outlying number of mismatches (> 7.5 SD above the mean) was removed from our analysis.

Statistical modeling for LTR-RT insertion activities

For each LTR-RT family, we model its population demographics as follows. Throughout, any time $t \geq 0$ refers to time in years in the past relative to the current calendar time, *i.e.*, t years before the current calendar time, which is set to 0. The age distribution at any time t in the past is defined as the distribution of the ages (*i.e.*, time since insertion) of all intact LTR-RTs within the family at that time. We use the probability density function $g(a,t)$ to represent the age (a) distribution at time t, so $g(a,0)$ is the age distribution at present. We let $\gamma(t)$ denote the birth rate or insertion rate (insertions per million year) at time t in the past, and assume that $\gamma(t)$ corresponds to the intensity of an inhomogeneous Poisson point process; then $\gamma(t)$ is proportional to the expected number of elements inserted into the genome within period $[t, t+\Delta]$, for an infinitesimal time interval Δ.

The insertion rate $\gamma(t)$ is assumed to be changing over time to reflect periods with changing insertion activities, in contrast to the assumption of constant insertion rate (Promislow *et al.* 1999; Marchani *et al.* 2009). A key difference between the age distribution $g(a,0)$ at present-time $(t=0)$ as a function of age a, and the insertion rate $\gamma(t)$, as a function of time t, is that the $g(a,0)$ describes the ages of only the intact elements that survived the deletion process to the present day, while the $\gamma(t)$ is the rate of birth for all elements at some time t in the past, regardless of whether they have been deleted or not at present. The insertion rate $\gamma(t)$ corresponds to the underlying genome dynamics, while the age distribution $g(a,0)$ does not directly reflect $\gamma(t)$ because even if $\gamma(t)$ has been constant, $g(a,0)$ will be decreasing because elements inserted in more distant past are less likely to survive.

Since LTR-RTs are subject to rapid deletion (Devos *et al.* 2002; Ma *et al.* 2004), one must take into account the deletion process when estimating the insertion rate, instead of simply regarding the age distribution as solely indicative of the insertion rate and effectively making a zero-deletion assumption. Assume each newly inserted LTR-RT has probability $\bar{F}(a) = P(X > a)$ to survive the deletion process to age a, where X is the life span of an LTR-RT, and that the survival function $\bar{F}(a)$ does not depend on the calendar time t. This assumption means that the intensity of deletion activities depends only on the age of the elements but not on calendar time, which is likely to hold if the overall genetic and epigenetic environment that affects retrotransposon deletion has been constant in the past. At time t, the density of intact elements of age a (those born at $t+a$ years in the past) is proportional to the product of $\gamma(t+a)\bar{F}(a)$, where $\gamma(t+a)$ is the birth intensity at time $t+a$ years before present, and $\bar{F}(a)$ is the fraction of elements surviving past age a. By normalizing the product into a density function, we obtain the age distribution

$$g(a,t) = \frac{\gamma(t+a)\bar{F}(a)}{\int_0^\infty \gamma(t+s)\bar{F}(s)ds}. \tag{1}$$

The integral in the previous display is finite if $\gamma(t)$ is bounded and $E(X)$ is finite. By fixing time t at $t=0$, the current calendar time, and by reordering (1), we obtain the insertion rate a years ago as

$$\gamma(a) = \frac{g(a,0)}{\bar{F}(a)} \int_0^\infty \gamma(s)\bar{F}(s)ds \propto \frac{g(a,0)}{\bar{F}(a)}, \tag{2}$$

where $\propto$ denotes a proportional relationship, since the integral does not depend on a. The ratio $g(a,0)/\bar{F}(a)$ can be interpreted as the shape of the insertion rate function $\gamma(a)$, which contains information for peak insertion periods and the time-dynamic change in the rate of insertion activities, and thus is the target of investigation.

We next estimate the survival function $\bar{F}(a)$. In the literature, it is generally assumed that the distribution of the life span of TEs is exponential, which means that the rate of removal of TEs is constant and the distribution is characterized by half-life. The half-life for rice LTR-RTs was estimated to be < 3 MY (Ma *et al.* 2004; Vitte *et al.* 2007) and that for rice *Copia* elements $\sim$796,000 years (Wicker and Keller 2007). Throughout our analysis, we adopt this commonly made assumption that life span X follows an exponential distribution, and estimate its half-life through Maximum Likelihood Estimation (MLE).

Estimating age distribution

In the current literature, the age distribution $g(a,0)$ is generally estimated by the histogram of the insertion date estimates (Ma *et al.* 2004; Vitte *et al.* 2007; Wicker and Keller 2007; Wang *et al.* 2008; Nystedt *et al.* 2013), which are in turn estimated using LTR divergence $d = N/l$, where N is the number of mismatches in the aligned LTRs of a retroelement and l is the length of the alignment. This estimate is only a proxy for the true age due to the randomness of mutations in the LTRs of an element, and the accuracy is lower for elements with shorter LTRs. Due to the variability in the individual estimates, their pooling within a family is subject to increased statistical error, which provides the motivation for the improved methodology introduced here.

Assume that the number of mutations in a single LTR with length l inserted x years ago follows a Poisson distribution with rate rlx [the same assumption as in Marchani *et al.* (2009)], where $r = 1.3 \times 10^{-8}$ substitutions/(year $\cdot$ site), as proposed by Ma and Bennetzen (2004). Then, the number of mismatches N on a pair of LTRs follows a Poisson distribution with rate $2rlx$. The conventional age estimate $d/(2r) = N/(2lr)$ will vary around age x, the center of its distribution.

To demonstrate the variability of the estimates, assume that each of the LTR-RTs within a single family has LTR length $l = 500$ bp, is inserted at $x = 1$ MYA, and the number of mismatches N between the two LTRs follows the Poisson distribution specified above. The distribution of the number of mismatches N shows considerable variability (Figure 1), even in this case where all elements are inserted into the genome at the same time, with a large coefficient of variation (0.277), defined as the ratio of SD over the mean. The histogram estimate of the age distribution by pooling the individual age estimates will have the same coefficient of variation as N rather than concentrate at 1 MYA, regardless of how many elements are in the family. Therefore, an approach based on the raw divergence is inadequate.

We approach this problem by modeling the number of mismatches N directly without estimating the age of individual elements, where we find that the distributions of N within most of the LTR-RT families are well approximated by negative binomial distributions [see, for example, the solid and dashed lines in the fitting of the number of mismatches in *Gypsy 1* (*Fatima*), Figure 3B]. Therefore, we use a negative binomial distribution to approximate the marginal distribution of N. For each family, we assume that the length l of each LTR is the same and is well approximated by the alignment length. This is reasonable since 97% of the elements had alignment lengths within ±10% around their corresponding family mean. Let random variable A be the age or insertion date of an element, which is assumed to be an independent and identical realization from the age distribution of its family. Then, the conditional distribution of the number of mismatches for a given insertion date is $N|A = a \sim \text{Poisson}\ (2rla)$. By a known probabilistic relationship (Leemis and McQueston 2008), the distribution of A follows a gamma distribution, which is flexible to model exponentially decreasing distributions and many unimodal age distributions. Denote the negative binomial distribution for N as $\text{NB}(n, p)$, with size n and success probability p, and the gamma distribution for A as $\Gamma(\alpha, \beta)$ with shape $\alpha = n$ and rate $\beta = 2prl/(1-p)$. We obtain estimates $(\hat{n}, \hat{p})$ for (n, p) by MLE, and then use

$$\hat{\alpha} = \hat{n}, \hat{\beta} = 2\hat{p}rl/(1-\hat{p}) \tag{3}$$

as the parameter estimates for the gamma distribution of A. The estimated age distribution $g(a, 0)$ is set to be the density of $\Gamma(\hat{\alpha}, \hat{\beta})$. The probability distributions and the MLE algorithms used are described in the Supplemental Materials.

In the special case where the size parameter of the negative binomial is $n = 1$, the negative binomial distribution for N reduces to a geometric distribution with probability p, and the age distribution will follow an exponential distribution with rate $2prl/(1-p)$. Under the assumption that the age distribution is exponential, as a special case of the gamma distribution, the rate of the exponential distribution can be estimated by

$$\hat{\lambda} = 2\hat{p}rl\Big/(1-\hat{p}), \tag{4}$$

where $\hat{p}$ is the MLE for the probability p.

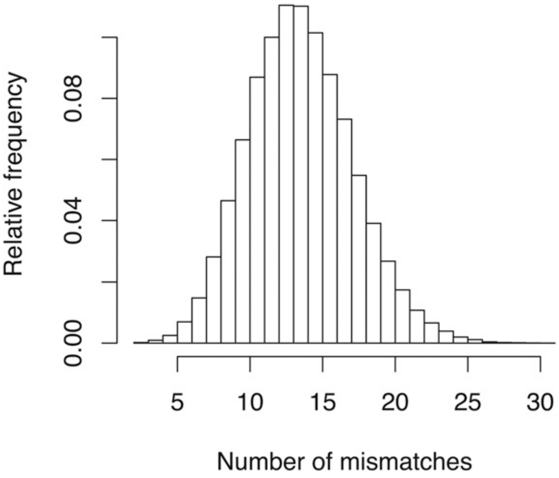

Figure 1 The distribution of the number of mismatches in a simulation when all elements are of length 500 bp, inserted 1 MYA, and have the same rate of mutation (and zero death rate). Under random mutations, the histogram of raw mismatches (or divergence) is seen to be inadequate for representing the age distribution.

Alternatively, the inaccuracy in the individual age estimates may be handled by nonparametrically deconvoluting the histogram of age estimates in estimating the age distribution. However, upon implementing this approach, we found that nonparametric deconvolution was unstable as it requires extensive tuning, which diminishes its practical value.

Inference

It is of biological interest to test for a given LTR-RT family whether the insertion rate $\gamma(t)$, and thus transposition activity, is constant/homogeneous over time. Formally, the null hypothesis is $H_0 : \gamma(t) = c$ for some constant c *vs.* the alternative $H_1 : \gamma(t) \neq c$ for all c. By (1) we find that under H_0 for any time z

$$g(a, z) = c\bar{F}(a)\Big/\int_0^\infty c\bar{F}(s)ds = \bar{F}(a)/E(X) = f(a),$$

where the second equality is due to a probabilistic equivalence, the third equality is due to a property of exponential distributions, and $f(a)$ is the exponential density function of the survival time X. This implies that $g(a, 0)$ is exponential and that the distribution of N is geometric, a special case of the negative binomial distribution (Leemis and McQueston 2008). Then, rejecting the null hypothesis H_0 of a constant insertion rate is implied by rejecting that N follows a geometric distribution. We carried out this test by embedding the geometric distribution into the negative binomial family, and tested for

$H_0 : N$ follows a geometric distribution vs
$H_1 : N$ follows a negative binomial distribution.

We are free to choose the alternative hypothesis, which does not affect the size (type I error rate) of the test, but could limit the power (type II error rate) if the true alternative is inadvertently omitted.

Sensitivity analysis and death rate

The birth rate can be obtained from Equation 2 after estimating the age distribution if one knows the survival function $\bar{F}(a)$, which corresponds to the death rate. However, even with the exponential life span assumption, the death rate is difficult to estimate precisely from the data because deletion mechanisms could remove TEs completely (see the *Discussion* section), leaving no trace that the deletion has occurred. Therefore, we compare a range of death rates and conduct a sensitivity analysis.

The exponential death rate parameter $\hat{\lambda}$ for the distribution of survival times X is estimated by fitting a geometric distribution to the mismatch data and then recovering the exponential rate, as in Equation 4. Since a single estimate of λ may not be accurate because there is no guarantee of a constant birth rate, we investigated three scenarios: Baseline death rates $\lambda = \hat{\lambda}$, low death rates $\lambda = \hat{\lambda}/2$, and high death rates $\lambda = 2\hat{\lambda}$, where $\hat{\lambda}$ is the MLE for the death rate under a constant birth rate model obtained according to Equation 4. As per Equation 2, we only estimate the birth rate up to a constant multiplier, and we normalized all birth rates into density functions that have area under the curve equal to 1.

To justify the plausibility of the baseline death rate $\hat{\lambda}$ for *Ae. tauschii*, we constructed an additional death rate estimate $\hat{\lambda}_s$ from the solo-LTRs, which is expected to serve as a lower bound for the true death rate due to various other factors causing or affecting TE removal from the genome, such as insertions of other TEs into the LTR-RT, deletions via illegitimate recombination, and purifying selection. Since the survival function of an intact element past age a is $\bar{F}(a) = e^{-\lambda a}$ under constant death rate λ, using a_i and y_i to denote the age and survival status (intact = 1, solo = 0) of an LTR element, we obtain the likelihood function for the ith LTR element as

$$L_i(\lambda) = y_i e^{-\lambda a_i} + (1 - y_i)\left(1 - e^{-\lambda a_i}\right), \text{ for } \lambda > 0.$$

Approximating a_i using $d_i/2r$ where d_i is the divergence in the ith element, we obtain the death rate estimate $\hat{\lambda}_s$ based on solo-LTRs as the MLE of the joint likelihood $L(\lambda) = \prod_{i=1}^{n} L_i(\lambda)$.

Goodness-of-fit of negative binomial fit

For some of the families, negative binomial distributions showed a lack of fit for the mismatch data, which may result in unreliable age distribution estimates. We used the Kullback–Leibler (Kullback and Leibler 1951) (KL) divergence as a criterion to evaluate the goodness-of-fit of our negative binomial models. For discrete probability distributions P and Q, the KL divergence from P to Q is defined to be

$$D_{KL}(P\|Q) = \sum_{i=0}^{\infty} P(i) \log \frac{P(i)}{Q(i)},$$

where we use the kernel density estimate (KDE) as P, representing the underlying "true" distribution, and the negative binomial distributions as Q. By inspecting the difference between the KDE and the negative binomial fit, we found that for families with $D_{KL} \leq 0.025$ a negative binomial distribution provided a reasonably good fit, while this was not the case for families with $D_{KL} > 0.025$ (*Gypsy* families 24, 35, 36, 40, and 44 and *Copia* families 27, 38, and 45, which have relatively small copy numbers). For those families, a mixture of two negative binomial distributions was fitted to the mismatch data by MLE (see the Supplemental Materials), with 1000 random starting points to search for the global maximizer of the likelihood function. That resulted in a recovered age distribution which was a mixture of two gamma distributions.

Regression analysis of TE ages

Given a population of previously inserted LTR-RTs, the mean age of the surviving intact elements is negatively correlated with the death rate in this population. Therefore, the death rate of LTR-RTs under different genomic variables can be proxied by the mean age of intact elements. We investigated through a linear mixed model the relationship between response insertion date of a TE, as approximated by $d/2r$, and its other attributes, including the chromosome, local meiotic HR rate (cM/Mb), log distance (base pair) to the nearest gene, superfamily membership (either *Gypsy* or *Copia*), the synteny of the closest gene (yes or no) with the wild emmer wheat (*Triticum turgidum* ssp. *dicoccoides*) genome (Dvorak *et al.* 2018), the closest codon (start or stop), and an LTR-RT family random effect. All intact elements with known chromosomal membership ($n = 17834$) were included. The local meiotic HR rates were estimated by the first derivative of a local quadratic smoother applied on genetic linkage data in cM (Fan and Gijbels 1996), with Gaussian kernel and bandwidth equal to 5 Mb. To calculate the distances to the nearest gene, we used only high-confidence genes (Luo *et al.* 2017).

Since we found that the log distance to the nearest gene had the strongest effect on the mean age of a TE, we performed an additional nonparametric regression to scrutinize the pattern of change of this effect in dependence on the distance to nearest gene. First, we fitted a linear mixed effects model that was the same as the model described in the last paragraph, except that the log distance to the nearest gene was not included as a predictor, and then extracted the residuals (of age) from this regression model. Second, a nonparametric regression using cubic regression splines was fitted to the age residuals as response and log distance to the nearest gene as predictor. This nonparametric regression assesses a possibly nonlinear relationship between the log distance to the nearest gene and the mean element age.

Implementation

User-friendly and fast algorithms that implement the proposed analysis are made available in the R package TE on CRAN. The estimation methods for the insertion rate, age

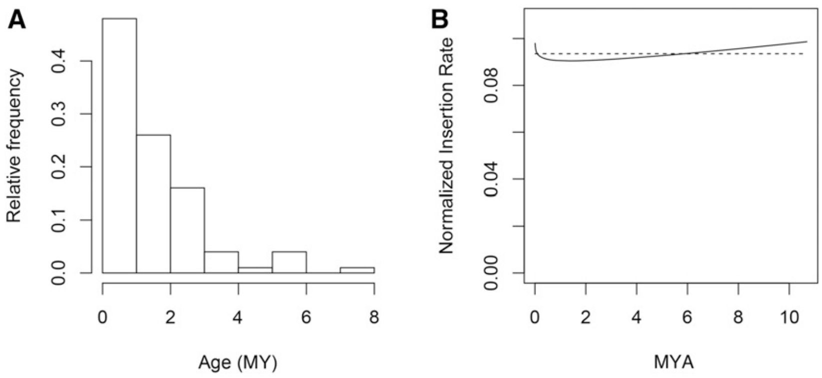

Figure 2 Simulated distribution of the number of mismatches, where each element is inserted into the genome uniformly over the past 10 MY, and has a half-life of 1 MY and LTR length equal to 500 bp. (A) Histogram for the mismatches in a random selection of 100 such elements that survive to the current time. An exponential decay pattern is apparent while the true insertion rate is constant. (B) The estimated insertion rate using a negative binomial fit (solid, our proposed method) and a geometric fit (dashed), correctly accounting for the death process and thus producing a constant insertion rate. MY, million years.

distribution, deletion rate, the hypothesis test of a constant insertion rate, and the sensitivity analysis produced by a single distribution fit were implemented in the function EstDynamics and the estimation for the mixture model in EstDynamics2. For the ease of comparison with other approaches for dating insertion dynamics (Promislow *et al.* 1999; Marchani *et al.* 2009), we also implemented functions MasterGene and MatrixModel. Helper functions such as PlotFamilies and SensitivityPlot for generating additional plots that display multiple families are also provided.

Data availability

LTR-RT data and code are included in the R package TE, available on CRAN (https://cran.r-project.org/). Supplemental material available at Figshare: https://doi.org/10.25386/genetics.6988631.

Results

LTR-RT demographics in the Ae. tauschii genome

To compare our approach with age histogram-based methods and to assess the false-positive rate of our method, we conducted a simulation with a data set generated under H_0, where each element was inserted uniformly over the past 10 MY, had a half-life of 1 MY, and an LTR length equal to 500 bp. Under these conditions, the distribution of ages computed from mismatches in LTRs followed an exponential decay (Figure 2A). This could lead, based on previous approaches, to inference of an exponential insertion rate, when the true insertion rate is uniform. This simulation thus illustrates that the age histogram of LTR-RTs may lead to incorrect assessments of the insertion rates. This is because the insertion rate bears out the history of insertions for all LTR-RTs, including those that have been removed from the genome at present time, while the age distribution reflects only the surviving intact LTR-RTs without adjusting for the survival bias caused by the removal of elements from the genome. In contrast to the histogram method, the insertion rate estimate based on our method (solid curve, Figure 2B) closely approximated the true uniform insertion rate (dashed line, Figure 2B) used in this simulation by correctly accounting for the death rate. We tested H_0 at the 0.05 significance level in 2000 simulations under the same setting as Figure 2, and the proportion of times rejecting H_0 was 0.051, showing that our test accurately controls the false-positive rate.

The age distribution of the largest *Gypsy* family, *Fatima* (Figure 3A, in the mismatch scale rather than timescale), was well fitted by a negative binomial distribution to the number of mismatches, since the negative binomial distribution fit was close to the kernel density estimate with $D_{KL} = 0.001$. The age distribution based on our method had a more salient peak at 1.28 MYA in the timescale (transformed from a peak at 15.6 mismatches) than that produced by the histogram method, which significantly underestimated the age distribution near the peak period, suffering from the convolution with the Poisson error. In some LTR-RT families, the negative binomial distribution showed a certain degree of lack of fit, as defined by $D_{KL} > 0.025$, and this is illustrated by *Gypsy* family 24 (*Nusif*) (Figure 3B). The lack of fit was remedied by using a mixture of two negative binomial distributions ($D_{KL} = 0.009$). This mixture model improved the fit and reduced D_{KL} to < 0.025 for all families that formerly showed lack-of-fit ($D_{KL} > 0.025$) when a single negative binomial fit was applied.

The null hypothesis that the insertion rate is constant was rejected with very small *P*-values for each of the 35 *Ae. tauschii* LTR-RT families (Tables S3 and S4). Compared to the peaks in the age distributions (Figure 4A), earlier peaks of the normalized insertion rates (Figure 4B) were attenuated and later peaks were amplified. This is because old elements are less likely to survive the deletion process as compared to young elements, and thus later peaks in the age distribution require more adjustment to yield the insertion rates than earlier peaks. Here, the insertion rates were estimated by adjusting the age distribution with the baseline death rate $\hat{\lambda}$. Each LTR-RT family was active during a different time range (Figure 4B) with peak activity ranging from 0.064 to 2.39 MYA. The most recent insertions were two *Copia* elements in family 27 (*Maximus*) that occurred 0.064 MYA; these two elements had only one and four mismatches in their LTRs and were vastly different from other elements in the same family that had an average of 40 mismatches. These two elements could be products of gene conversions between their LTRs. More work is needed to confirm that.

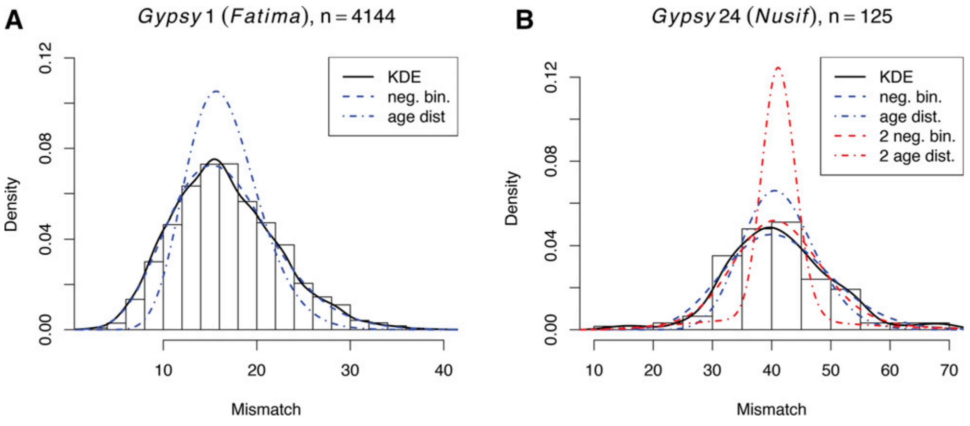

Figure 3 Distributional fits and recovered age distribution of *Gypsy* family 1, *Fatima* (A), produced by function EstDynamics and *Gypsy* family 24, *Nusif* (B), produced by EstDynamics2. We display the proportions of long terminal repeat-retrotransposons falling into each bin of mismatches (histogram bars), kernel density estimate (black solid), the negative binomial fit by maximum likelihood estimation (blue dashed), and the recovered age distribution expressed in mismatch timescale (blue dash-dot). For *Gypsy* family *Nusif* (B), a negative binomial fit shows lack of fit as measured by Kullback–Leibler divergence (see subsection *Goodness-of-fit of negative binomial fit*). Therefore, we used a mixture of two negative binomial distributions (red dashed) to improve the fit, for which the recovered age distribution is a mixture of gamma distributions (red dash-dot).

A nearly universal destiny of LTR-RTs is to be removed from the genome. That can take place by deleting the entire element or converting it into a solo element by HR, which ultimately is obliterated by small deletions, insertions, and substitutions. The abundance and divergence of solo elements can therefore be used as a lower bound of LTR-RT death rate. The ratios of solo to intact elements in each family varied extensively (Figure S1) from 0.01 (*Copia* 16) to 7.11 (*Gypsy* 34), and were significantly positively associated with the mean LTR length ($P < 10^{-12}$, Figure S2). Within most families, the ratio increased steadily as elements aged. The death rates $\hat{\lambda}_s$ estimated from the solo-LTRs (Tables S3 and S4) were consistently smaller than the baseline death rate $\hat{\lambda}$ (see *Materials and Methods*) except for *Gypsy* 34, which had the largest solo-to-intact element ratio.

Some age histograms, such as those for families *Copia 3*, *Gypsy 31*, and *Gypsy 40* (Figure S3A), show a peak in ages of complete elements and an additional peak of solo elements. Analyzing these three families with our model shows that insertion rates in these three LTR-RT families experienced two bursts of insertions and, subsequently, silencing (Figure S3B).

The age distribution obtained with the matrix population model (Promislow *et al.* 1999) and the master gene model (Marchani *et al.* 2009) for the largest *Gypsy* and *Copia* families (1 and 4, respectively) (Figure S4 and Table S5) was an exponential and a uniform distribution, respectively. Neither distribution reflected the existence of the peaks in the age distributions and the uniform fit produced by the master gene model omitted the oldest elements. The death rates q estimated by the matrix population model were smaller than $\hat{\lambda}_s$ for both families.

To demonstrate the sensitivity of our results to the estimates of death rates, we studied the insertion rates corresponding to three death rate scenarios for the first, third, and fifth largest *Copia* and *Gypsy* families (Figure 5), which were based on family-specific baselines according to Equation 4 (see *Materials and Methods*). The varying death rates between families are supported by their varying abundance of solo-LTRs (Figure S1). An important outcome of exploring the three death rate scenarios is that, while the precise date of the peaks of insertion times may move in time, the sequence of peaks across families is not much affected by varying assumptions on death rates. Salient peaks in the insertion rates were evident in each family, meaning that these families all underwent periods of rapid amplification. In a scenario assuming a higher death rate, peaks were shifted back in time, meaning that the peak of insertion activity was calculated to be at a more ancient date; this is a consequence of Equation 2.

Relationships between Ae. tauschii LTR-RT ages and biological predictors

To investigate factors affecting the death rates of LTR-RTs in the *Ae. tauschii* genome, we computed regression coefficients (Table 1) between element age (MYA) approximated by LTR divergence (response) and the following potential predictors: (1) chromosome membership (1D to 7D), (2) local meiotic HR rate (cM/Mb), (3) proximity to the 5′ (start codon) or 3′ (stop codon) end of the nearest gene, (4) colinearity of the nearest gene with genes on a homeologous pseudomolecule in wild emmer wheat (yes or no) (Dvorak *et al.* 2018), (5) log distance to the nearest gene (bp), and (6) superfamily membership (*Copia* or *Gypsy*).

All predictors had a significant effect on the age of LTR-RTs (response) at the $P = 0.05$ level. (1) Taking chromosome 1D arbitrarily as a baseline in the regression analysis, LTR-RTs on chromosomes 2D, 4D, 6D, and 7D were on average older than those on chromosome 1D, whereas those on chromosomes 3D and 5D did not significantly differ. (2) An increase by 1 cM/Mb in meiotic HR rate was associated with a decrease of 0.021 MY in age. (3) If an element was near the 5′ end of the nearest gene, it was on average 0.012 MY younger than if it was near the 3′ of the nearest gene, but this effect was weak and barely significant ($P = 0.049$). (4) If a gene nearest to an element was in a colinear location with a homologous gene in wild emmer wheat (Dvorak *et al.* 2018), then the element

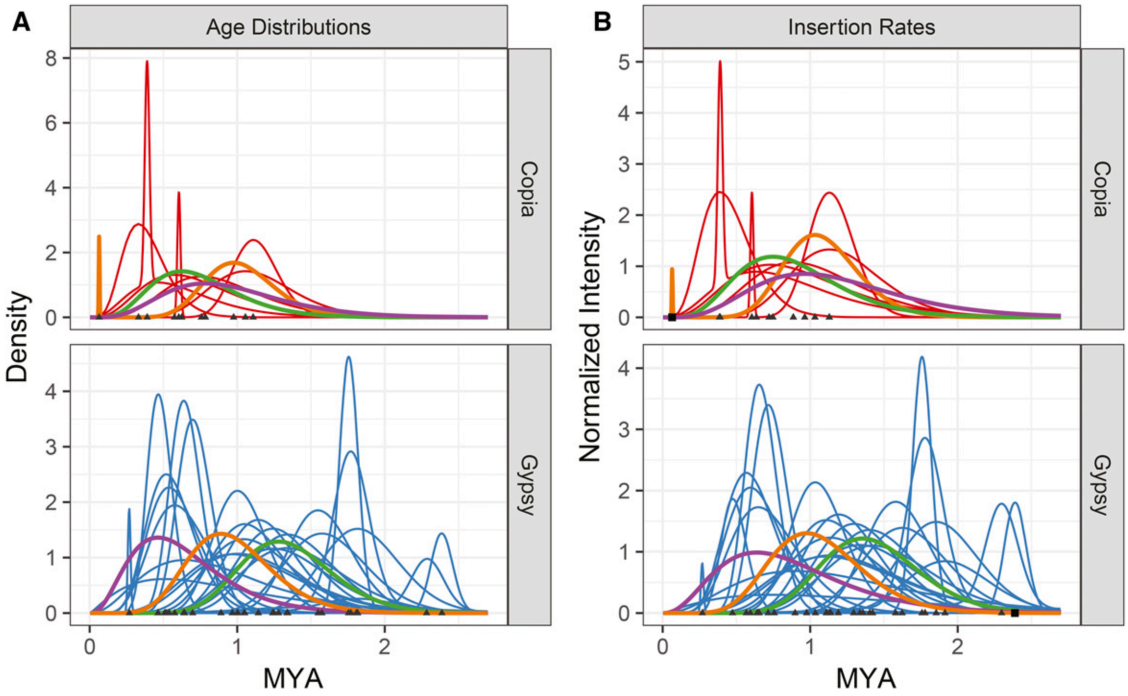

Figure 4 (A) Age distributions and (B) normalized insertion rates in the 35 largest *Ae. tauschii* long terminal repeat-retrotransposon families. Each curve represents the estimated age distribution (A) or insertion rate as normalized into a probability density function (B) of a single family. *Copia* families are shown in red and *Gypsy* families in blue, and the first, third, and fifth largest *Copia* and *Gypsy* families are highlighted in green, purple, and orange, respectively. Gray triangles on the *x*-axis indicate the peak locations. The peak insertion activities ranged from 0.064 to 2.39 MYA, marked by black squares (right).

was on average 0.050 MY younger than an element that was near a gene in a noncolinear location. (5) A unit increase in the log distance to the nearest gene was associated with an increase in age of 0.070 MY, and this effect had the largest *t*-value. In an additional nonparametric regression (Figure S5), we demonstrated that the mean LTR-RT age did not vary as the distance to the nearest gene increased up to 22 kb, after which point the mean element age was linearly associated with the log distance to the nearest gene. (6) *Gypsy* families were on average 0.094 MY older than *Copia* families.

LTR-RT demographics in the A. lyrata genome

To compare LTR-RT demographics in the large genome of *Ae. tauschii*, containing a great abundance of TEs, with a genome containing far fewer TEs, we applied our model to two *A. lyrata* LTR-RT families, *Gypsy 1* and *Copia 2*. For both families, a single negative binomial distribution provided a close fit to the mismatches with $D_{KL} < 0.025$ (Figure S6). The recovered age distribution for *Gypsy* 1 had a peak that translated to 0.125 MYA, while that for *Copia* 2 peaked at present. The constant insertion rate hypothesis was rejected for the *Gypsy* 1 family ($P = 1.3 \times 10^{-5}$) but not for the *Copia* 2 family ($P = 0.259$). The two families have distinctly different demographics and distinctly different profiles of the rates during the past 3 MY (Figure S7). The *Gypsy* 1 profile showed a recent burst of insertion activities, while that of *Copia* 2 showed a nearly constant amplification during the past 3 MY.

Discussion

LTR-RT demographics modeling

LTR-RTs are a large component of plant genomes. In the tribe Triticeae, which includes wheat and its ancestors, they represent >60% of the genome (Avni *et al.* 2017; Luo *et al.* 2017; Mascher *et al.* 2017; Zhao *et al.* 2017). It has been stated many times that this large and dynamic component of plant genomes has a profound effect on the evolution of plant genome structure and gene expression. To advance our understanding of these effects, we developed and deployed statistical models to study the rates of birth and death of LTR-RTs in *Ae. tauschii*, one of the three progenitors of bread wheat (Luo *et al.* 2017). By inferring insertion rates for each LTR-RT family from the age distribution of its members, inferred from the divergence of LTRs and the death rate, we arrived at a more realistic portrait of amplification of LTR-RTs during genome evolution than the current approaches based on an age histogram.

The advantages of our method are twofold. First, our model takes into account the deletion process, producing more realistic estimates that put more weight on the older and thus harder to observe elements. Using our model, one can formally test the hypothesis that the insertion rate is constant over time. In this case, the distribution of divergence is exponentially decaying due to the death of old elements. Neglecting the death rate and equating the age distribution with the insertion rate, a feature common to other approaches,

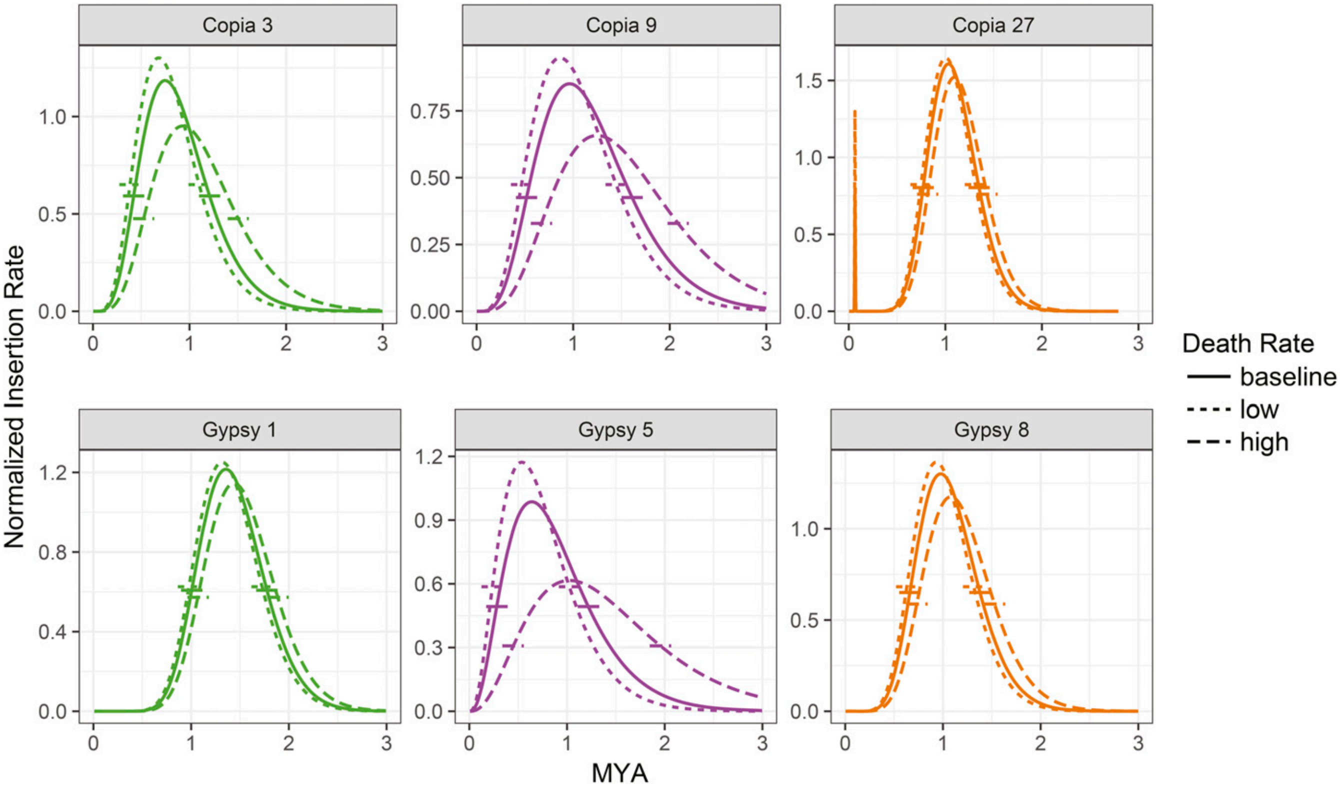

Figure 5 Sensitivity analysis for the first, third, and fifth largest *Copia* (top row) and *Gypsy* (bottom row) families, respectively. For each family, three death rate scenarios are shown: baseline death rates $\lambda = \hat{\lambda}$ (solid), low death rates $\lambda = \hat{\lambda}/2$ (dotted), and high death rates $\lambda = 2\hat{\lambda}$ (dashed). Short horizontal lines on each curve mark the times when the insertion activities are half as strong as the peak intensity in each scenario.

can lead to an imprecise estimate of the insertion rate, especially if the insertion rate is constant or nearly constant, a situation that we found to be true for *Copia* family 2 of *A. lyrata*. Our method showed a constant insertion rate for this family, whereas a histogram of age distribution suggested an exponential decay that would lead to a wrong impression of a recent burst of insertions (compare right panels of Figures S6 and S7). Second, the relationship between the number of mutations and the age of an LTR-RT is subject to random variation. Our model takes that into account, which results in more pronounced peaks in the age distribution estimates.

Our model accommodates varying insertion rates of LTR-RT families over time, which is appropriate for dynamic transposition activity as demonstrated in simulations (Le Rouzic *et al.* 2007). Our hypothesis test rejected constancy of insertion rates for all 35 *Ae. tauschii* LTR-RT families and *Gypsy* family 1 in *A. lyrata*, with the biological implication that LTR-RT families go through bursts of various intensity of amplification, followed by decline and eventual removal. LTR-RT families such as *Copia 27* (Figure 5) and *Copia 9*, *Gypsy 31*, and *Gypsy 40* (Figure S1) had at least two past activations, as shown by their bimodal insertion rates and/or age distributions. This demonstrates that the activation and silencing in each LTR-RT family can be cyclic.

Our approach is applicable to modeling the TE demographics in species with LTR-RT dynamics that may differ from that in *Ae. tauschii*. This is illustrated by the analyses of LTR-RT families in *A. lyrata*, a much smaller genome with a much lower content of LTR-RTs. The analyses of two major *A. lyrata* LTR-RT families with our approach revealed two contrasting patterns, even though a single negative binomial distribution provided a close fit to the LTR mismatches for both families. The constant insertion rate hypothesis was rejected for the *Gypsy 1* family but not for the *Copia 2* family. The two families exhibit distinctly different profiles of the insertion rates during the past 3 MY. The *Gypsy 1* profile showed a recent burst of insertions peaking at 0.125 MYA, while that of *Copia 2* showed a nearly constant amplification during the past 3 MY. *Copia 2* was the only family we observed in our analysis of *Ae. tauschii* and *A. lyrata* that showed peak amplification at the present time, and may therefore provide a window into the insertion rates prior to attenuation and decline.

The matrix population model of Promislow *et al.* (1999) provides easy-to-calculate insertion and death rate estimates under constant insertion and exponential age distribution assumption. In contrast, our approach accommodates greater variation in age distribution by allowing a more flexible gamma distribution fit. Marchani *et al.* (2009) proposed a master gene model, which is largely applicable to TE families with nearly constant insertion rate and without element removal, but the model failed to capture peaks in the age distributions in *Ae. tauschii* families. Previous analyses of TE

Table 1 Regression coefficient estimates for the age of an LTR-RT as response and various genomic factors as predictors

Predictor[a]	Regression[b] coefficient	SE	*t*-value	*P*-value
Intercept	0.252	0.043	5.796	0.000
Chr2	0.024	0.011	2.207	0.027
Chr3	0.009	0.011	0.765	0.444
Chr4	0.022	0.012	1.876	0.061
Chr5	0.011	0.011	0.999	0.318
Chr6	0.023	0.012	1.904	0.057
Chr7	0.042	0.011	3.741	0.000
HR	−0.021	0.007	−3.062	0.002
Near 5′ end	−0.012	0.006	−1.971	0.049
Near colinear gene	−0.050	0.006	−8.335	0.000
Log distance to a gene	0.070	0.003	25.434	0.000
Gypsy superfamily	0.094	0.045	2.089	0.037

Chr, chromosome; HR, homologous recombination.

[a] The following predictors are considered: chromosome membership (Chr2–7, with Chr1 as baseline), adjacent HR rate, proximity to the 5′ end of a gene (with proximity to the 3′ end as baseline), colinearity of the closest gene with the homologous gene on a homeologous pseudomolecule in wild emmer (Dvorak *et al.* 2018) (with baseline that the gene is next to a noncolinear gene), log distance (bp) to the nearest gene, and long terminal repeat-retrotransposon superfamily membership (with Copia as baseline).

[b] The regression coefficients for the various predictors correspond to: for Chr2, the difference between the mean age on Chr2 and that on Chr1; analogously for Chr3–7; for "Near 5′ end," the difference between the mean age of long terminal repeat-retrotransposons (LTR-RTs) near the 5′ end of a gene and those near the 3′ end; for "Near colinear gene," the difference between the mean age of LTR-RTs next to colinear genes and those next to noncolinear genes; for *Gypsy* superfamily, the difference in mean age between *Gypsy* and *Copia* elements; and for "Intercept," the mean age in the reference level, consisting of *Copia* elements on Chr1 closest to a noncolinear gene and its 3′ end.

dynamics in *A. thaliana* employed multiple genome lineages (Hollister and Gaut 2007) to reveal a positive relationship between *Helitron* age and its distance to the nearest gene. Our analysis revealed the same relationship employing a single lineage (accession) of the much larger *Ae. tauschii* genome with abundant LTR-RTs.

For pragmatic reasons, we employed the exponential life span assumption followed by other authors (Ma and Bennetzen 2004; Vitte *et al.* 2007; Wicker and Keller 2007), which amounts to a constant hazard rate. The resulting baseline death rate estimates $\hat{\lambda}$ were larger than $\hat{\lambda}_s$ estimated from the solo-LTRs except for one LTR-RT family, in accordance with the fact that the deletions of an entire element are not reflected in solo-LTRs. As the ratio of solo to intact element increased steadily within most families for older elements (Figure S1), the death rate due to forming solo-LTRs does not seem to vary over time, which partially supports our constant death rate assumption.

If the death rates would vary significantly for TEs inserted at different times, the death rates could be a factor shaping the age distributions, in addition to the insertion rates. For example, if the insertion rate were constant but a removal process that applies to only young LTR-RTs started at time t, then one would observe a sharp decrease in abundance of elements younger than age t. This could occur if the fitness of an individual bearing such elements would decrease steeply as the copy number of the LTR-RTs increased (Charlesworth and Langley 1989), or if the species acquired a new excision mechanism at time t that targets only newly inserted elements. However, the estimation of time- or age-dependent hazard rates requires the observation of historical TE removal events. Further elucidation is left for future work because quality data on deletion are unavailable at this stage.

The LTR-RT age distribution estimated from a randomly selected focal individual will be an unbiased estimate of the population (average) age distribution because we consider the total number of inserted copies, regardless of whether each copy is fixed in the population. Nonetheless, since younger TEs not fixed in the genome are more likely to vary in copy numbers between different individuals, sampling variation for these younger elements will increase and thus result in larger variance in the estimates corresponding to the younger age range.

Gene conversion was not considered in our modeling of LTR-RT demographics. Gene conversion between paralogous LTR-RTs would lead to homogenization of dispersed copies and probably minimally affect the results since our model is concerned with a single-genome sequence, not with allelic frequencies in a population, unlike the case considered for example in Blumenstiel *et al.* (2014). Gene conversion between LTRs within a single element would homogenize them and lead to underestimation of time of insertion as suggested for the two exceptional elements of Copia family 27. Modeling such gene conversions needs future work, and this factor should be included into our model if found significant. However, it is of interest that we did not find elements with identical LTRs, which we would have interpreted as current LTR-RT insertions, among the 35 *Ae. tauschii* LTR-RT families we studied. This remarkable absence of currently inserted elements is contrary to what is expected if gene conversions within LTR-RTs were common.

Element ages and genome evolution

The balance between the insertion and removal rates of LTR-RT elements determines the global rate of sequence turnover in a genome, and shapes the structure of its chromosomes (Devos *et al.* 2002; Ma and Bennetzen 2006; Vitte and Bennetzen 2006). Our regression analysis revealed that the age of LTR-RTs within *Ae. tauschii* LTR-RT families was negatively associated with recombination rates along the chromosome, and positively with proximity to genes and their 5′ ends. Since genes and their 5′ ends frequently act as recombination hotspots (Schnable *et al.* 1998), the three associations seem to indicate that the local recombination rate is a major factor determining the local rate of the removal of elements from the genome (Ma and Bennetzen 2006). However, the same associations can be cited in support of purifying selection as the major factor. One reason for that is the Hill–Robertson effect, which is the positive relationship between the effectiveness of purifying selection and the local recombination rate (Hill and Robertson 1966). The other reason lies in the inverse relationship between gene expression and degree of methylation of nearby LTR-RTs

(Hollister and Gaut 2009; Hollister *et al.* 2011). In *Arabidopsis*, methylated LTR-RTs near genes are associated with low gene expression and are therefore subjected to purifying selection, compared to those that are far away from genes or those that are not methylated. Likewise, either recombination or purifying selection could be cited as major factors accounting for our observation that elements near genes colinear with homologous genes in wild emmer, thereby enriched for the conserved gene repertoire of the *Ae. tauschii* genome, tend to be younger than those near noncolinear genes. This gene repertoire effect was also captured (Table S6) if colinearity of the *Ae. tauschii* genes was assessed against the *Brachypodium distachyon* genome (Luo *et al.* 2017). A significant portion of the *Ae. tauschii* genes that are in noncolinear locations may be pseudogenes or may not be expressed, and be free of natural selection but also not acting as recombination hotspots.

Methylated LTR-RTs were estimated to detrimentally affect gene expression at distances of 1.0 kb in *A. lyrata* and 2.5 kb in *A. thaliana* (Hollister and Gaut 2009; Hollister *et al.* 2011). These estimates can be turned around and used as estimates of the distance from a gene for which purifying selection would act against a methylated LTR-RT. In the *Ae. tauschii* genome with much greater content of LTR-RTs compared to the *Arabidopsis* genome, genes affected the mean element age for >22 kb, with no actual distance limit. This seems to argue for strong purifying selection acting against LTR-RTs and against recombination as a major factor causing this association, since it seems unlikely that the initiation of recombination near a gene can cause recombination leading to a deletion of an element that far away. This long-distance effect of genes on the age, and hence the rate of turnover, of LTR-RTs in the intergenic regions of the *Ae. tauschii* genome highlights the potential importance of epigenetic effects of LTR-RTs on gene expression in the Triticeae genomes and the need for more work in this area.

LTR-RT insertion rates and genome sizes in Triticeae

The tribe Triticeae includes >300 species (Love 1984), most of them polyploid, that radiated over a period of ~10 MY (Huang *et al.* 2002; Ramakrishna *et al.* 2002; Dvorak and Akhunov 2005). In that time period, the arrays of LTR-RTs between genes have been turned over more than once in individual Triticeae lineages (Dubcovsky and Dvorak 2007). Based on this massive turnover and the fact that LTR-RTs account for large portions of the Triticeae genomes, one would expect to find great stochastic variation in genome sizes in the tribe. However, the sizes of genomes in the diploid Triticeae species vary by a factor of two, from 3.9 to 8.1 Gb (Dvořák 2009). An impressive property of the insertion rate profiles of the 35 most abundant *Ae. tauschii* LTR-RT families over the past 3 MY is their similarity and the absence of extremes. One by one, families go through an amplification burst, which reaches a peak of height similar among the families, and then the burst is quenched. Several families, *e.g.*, *Copia 3*, *Copia 27*, *Gypsy 31*, and *Gypsy 40*, show the cyclic nature of this process. Work in *Arabidopsis* highlighted the importance of epigenetic silencing for LTR-RT transposition by methylation and small interfering RNAs (Tsukahara *et al.* 2009; Hollister *et al.* 2011). The logical conclusion of these observations is that the constancy of genome sizes in Triticeae, despite the unprecedented rate of LTR-RT turnover, resides in the constancy of epigenetic control over LTR-RT amplification over the past 10 MY.

Acknowledgments

We thank Matthew Dawson for providing the estimates of recombination rates, Patrick McGuire for critiquing the manuscript, and Jiale Xu for helpful discussions. The study was supported by the National Science Foundation (NSF-ISO-1238231, NSF-DMS-1407852, and NSF-DMS-1712864). We have no competing interests.

Author contributions: J.D. conceived the birth and death analysis, and this conceptualization was refined through contributions from all authors. H.W. annotated the intact LTR-RTs and provided the data. H.Z. and L.W. provided the solo-LTR data and gene orientation data, respectively. X.D. and H.-G.M. proposed the statistical model and analyzed the data. X.D. developed the R package TE. X.D., J.D., and J.L.B. wrote the manuscript. All authors contributed to revising the manuscript.

Literature Cited

Avni, R., M. Nave, O. Barad, K. Baruch, S. O. Twardziok *et al.*, 2017 Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science 357: 93–97. https://doi.org/10.1126/science.aan0032

Baucom, R. S., J. C. Estill, C. Chaparro, N. Upshaw, A. Jogi *et al.*, 2009a Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. PLoS Genet. 5: e1000732. https://doi.org/10.1371/journal.pgen.1000732

Baucom, R. S., J. C. Estill, J. Leebens-Mack, and J. L. Bennetzen, 2009b Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome. Genome Res. 19: 243–254. https://doi.org/10.1101/gr.083360.108

Bennetzen, J. L., and H. Wang, 2014 The contributions of transposable elements to the structure, function, and evolution of plant genomes. Annu. Rev. Plant Biol. 65: 505–530. https://doi.org/10.1146/annurev-arplant-050213-035811

Blumenstiel, J. P., X. Chen, M. He, and C. M. Bergman, 2014 An age-of-allele test of neutrality for transposable element insertions. Genetics 196: 523–538. https://doi.org/10.1534/genetics.113.158147

Charlesworth, B., and C. H. Langley, 1989 The population genetics of *Drosophila* transposable elements. Annu. Rev. Genet. 23: 251–287. https://doi.org/10.1146/annurev.ge.23.120189.001343

Choulet, F., T. Wicker, C. Rustenholz, E. Paux, J. Salse *et al.*, 2010 Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. Plant Cell 22: 1686–1701. https://doi.org/10.1105/tpc.110.074187

Devos, K. M., J. K. Brown, and J. L. Bennetzen, 2002 Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome Res. 12: 1075–1079. https://doi.org/10.1101/gr.132102

Dubcovsky, J., and J. Dvorak, 2007 Genome plasticity a key factor in the success of polyploid wheat under domestication. Science 316: 1862–1866. https://doi.org/10.1126/science.1143986

Dvořák, J., 2009 Triticeae genome structure and evolution, pp. 685–711 in Genetics and Genomics of the Triticeae. Plant Genetics and Genomics: Crops and Models, Vol 7, edited by G. Muehlbauer and C. Feuillet. Springer-Verlag, New York. https://doi.org/10.1007/978-0-387-77489-3_23

Dvorak, J., and E. D. Akhunov, 2005 Tempos of gene locus deletions and duplications and their relationship to recombination rate during diploid and polyploid evolution in the *Aegilops-Triticum* alliance. Genetics 171: 323–332. https://doi.org/10.1534/genetics.105.041632

Dvorak, J., L. Wang, T. Zhu, C. M. Jorgensen, K. R. Deal *et al.*, 2018 Structural variation and rates of genome evolution in the grass family seen through comparison of sequences of genomes greatly differing in size. Plant J. 95: 487–503. https://doi.org/10.1111/tpj.13964

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

Estep, M. C., J. D. DeBarry, and J. L. Bennetzen, 2013 The dynamics of LTR retrotransposon accumulation across 25 million years of panicoid grass evolution. Heredity 110: 194–204. https://doi.org/10.1038/hdy.2012.99

Fan, J., and I. Gijbels, 1996 *Local Polynomial Modelling and its Applications: Monographs on Statistics and Applied Probability*, Vol. 66. CRC Press, Boca Raton, FL.

Gottlieb, A., H.-G. Müller, A. N. Massa, H. Wanjugi, K. R. Deal *et al.*, 2013 Insular organization of gene space in grass genomes. PLoS One 8: e54101. https://doi.org/10.1371/journal.pone.0054101

Hawkins, J. S., S. R. Proulx, R. A. Rapp, and J. F. Wendel, 2009 Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. Proc. Natl. Acad. Sci. USA 106: 17811–17816. https://doi.org/10.1073/pnas.0904339106

Hill, W. G., and A. Robertson, 1966 The effect of linkage on limits to artificial selection. Genet. Res. 8: 269–294. https://doi.org/10.1017/S0016672300010156

Hollister, J. D., and B. S. Gaut, 2007 Population and evolutionary dynamics of helitron transposable elements in *Arabidopsis thaliana*. Mol. Biol. Evol. 24: 2515–2524. https://doi.org/10.1093/molbev/msm197

Hollister, J. D., and B. S. Gaut, 2009 Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res. 19: 1419–1428. https://doi.org/10.1101/gr.091678.109

Hollister, J. D., L. M. Smith, Y.-L. Guo, F. Ott, D. Weigel *et al.*, 2011 Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. Proc. Natl. Acad. Sci. USA 108: 2322–2327. https://doi.org/10.1073/pnas.1018222108

Huang, S., A. Sirikhachornkit, X. Su, J. Faris, B. S. Gill *et al.*, 2002 Genes encoding plastid acetyl-CoA carboxylase and 3-phopshoglycerate kinase of the *Triticum/Aegilops* complex and the evolutionary history of polyploid wheat. Proc. Natl. Acad. Sci. USA 99: 8133–8138. https://doi.org/10.1073/pnas.072223799

Kirik, A., S. Salomon, and H. Puchta, 2000 Species-specific double-strand break repair and genome evolution in plants. EMBO J. 19: 5562–5566. https://doi.org/10.1093/emboj/19.20.5562

Kullback, S., and R. A. Leibler, 1951 On information and sufficiency. Ann. Math. Stat. 22: 79–86. https://doi.org/10.1214/aoms/1177729694

Lamesch, P., T. Z. Berardini, D. Li, D. Swarbreck, C. Wilks *et al.*, 2011 The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 40: D1202–D1210. https://doi.org/10.1093/nar/gkr1090

Leemis, L. M., and J. T. McQueston, 2008 Univariate distribution relationships. Am. Stat. 62: 45–53. https://doi.org/10.1198/000313008X270448

Leitch, I. J., and A. R. Leitch, 2013 Genome size diversity and evolution in land plants, pp. 307–322 in *Plant Genome Diversity: Physical Structure*, *Behaviour and Evolution of Plant Genomes*, Vol. 2, edited by J. Greilhuber, J. Dolezel, and J. F. Wendel. Springer-Verlag, Berlin.

Le Rouzic, A., T. S. Boutin, and P. Capy, 2007 Long-term evolution of transposable elements. Proc. Natl. Acad. Sci. USA 104: 19375–19380. https://doi.org/10.1073/pnas.0705238104

Levy, O., B. A. Knisbacher, E. Y. Levanon, and S. Havlin, 2017 Integrating networks and comparative genomics reveals retroelement proliferation dynamics in hominid genomes. Sci. Adv. 3: e1701256. https://doi.org/10.1126/sciadv.1701256

Lisch, D., and J. L. Bennetzen, 2011 Transposable element origins of epigenetic gene regulation. Curr. Opin. Plant Biol. 14: 156–161. https://doi.org/10.1016/j.pbi.2011.01.003

Love, A., 1984 Conspectus of the *Triticeae*. Feddes Repert. 95: 425–521.

Luo, M. C., Y. Q. Gu, D. Puiu, H. Wang, S. O. Twardziok *et al.*, 2017 Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. Nature 551: 498–502. https://doi.org/10.1038/nature24486

Ma, J., and J. L. Bennetzen, 2004 Rapid recent growth and divergence of rice nuclear genomes. Proc. Natl. Acad. Sci. USA 101: 12404–12410. https://doi.org/10.1073/pnas.0403715101

Ma, J., and J. L. Bennetzen, 2006 Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. Proc. Natl. Acad. Sci. USA 103: 383–388. https://doi.org/10.1073/pnas.0509810102

Ma, J., K. M. Devos, and J. L. Bennetzen, 2004 Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res. 14: 860–869. https://doi.org/10.1101/gr.1466204

Marchani, E. E., J. Xing, D. J. Witherspoon, L. B. Jorde, and A. R. Rogers, 2009 Estimating the age of retrotransposon subfamilies using maximum likelihood. Genomics 94: 78–82. https://doi.org/10.1016/j.ygeno.2009.04.002

Mascher, M., H. Gundlach, A. Himmelbach, S. Beier, S. O. Twardziok *et al.*, 2017 A chromosome conformation capture ordered sequence of the barley genome. Nature 544: 427–433. https://doi.org/10.1038/nature22043

Müller, H.-G., J.-L. Wang, W. Yu, A. Delaigle, and J. R. Carey, 2007 Survival and aging in the wild via residual demography. Theor. Popul. Biol. 72: 513–522. https://doi.org/10.1016/j.tpb.2007.07.003

Nystedt, B., N. R. Street, A. Wetterbom, A. Zuccolo, Y.-C. Lin *et al.*, 2013 The Norway spruce genome sequence and conifer genome evolution. Nature 497: 579–584. https://doi.org/10.1038/nature12211

Promislow, D. E. L., I. K. Jordan, and J. E. McDonald, 1999 Genomic demography: a life-history analysis of transposable element evolution. Proc. Biol. Sci. 266: 1555–1560. https://doi.org/10.1098/rspb.1999.0815

Ramakrishna, W., J. Dubcovsky, Y. J. Park, C. Busso, J. Embereton *et al.*, 2002 Different types and rates of genome evolution detected by comparative sequence analysis of orthologus segments from four cereal genomes. Genetics 162: 1389–1400.

SanMiguel, P., B. S. Gaut, A. Tikhonov, Y. Nakajima, and J. L. Bennetzen, 1998 The paleontology of intergene retrotransposons of maize. Nat. Genet. 20: 43–45. https://doi.org/10.1038/1695

Schiml, S., F. Fauser, and H. Puchta, 2016 Repair of adjacent single-strand breaks is often accompanied by the formation of tandem sequence duplications in plant genomes. Proc. Natl. Acad. Sci. USA 113: 7266–7271. https://doi.org/10.1073/pnas.1603823113

Schnable, P. S., A. P. Hsia, and B. J. Nikolau, 1998 Genetic recombination in plants. Curr. Opin. Plant Biol. 1: 123–129. https://doi.org/10.1016/S1369-5266(98)80013-7

Smit, A. F., 2004 Repeat-Masker Open-3.0. Available at: http://www. repeatmasker.org. Accessed November 2, 2015.

Tsukahara, S., A. Kobayashi, A. Kawabe, O. Mathieu, A. Miura *et al.*, 2009 Bursts of retrotransposition reproduced in *Arabidopsis*. Nature 461: 423–426. https://doi.org/10.1038/nature08351

Vaughn, J. N., and J. L. Bennetzen, 2014 Natural insertions in rice commonly form tandem duplications indicative of patch-mediated double-strand break induction and repair. Proc. Natl. Acad. Sci. USA 111: 6684–6689. https://doi.org/10.1073/pnas.1321854111

Vitte, C., and J. L. Bennetzen, 2006 Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proc. Natl. Acad. Sci. USA 103: 17638–17643. https://doi.org/10.1073/pnas.0605618103

Vitte, C., O. Panaud, and H. Quesneville, 2007 LTR retrotransposons in rice (Oryza sativa, L.): recent burst amplifications followed by rapid DNA loss. BMC Genomics 8: 218. https://doi.org/10.1186/1471-2164-8-218

Wacholder, A. C., C. Cox, T. J. Meyer, R. P. Ruggiero, V. Vemulapalli *et al.*, 2014 Inference of transposable element ancestry. PLoS Genet. 10: e1004482. https://doi.org/10.1371/journal.pgen.1004482

Wang, L., L. D. Brown, T. T. Cai, and M. Levine, 2008 Effect of mean on variance function estimation in nonparametric regression. Ann. Stat. 36: 646–664. https://doi.org/10.1214/009053607000000901

Wicker, T., and B. Keller, 2007 Genome-wide comparative analysis of *copia* retrotransposons in *Triticeae*, rice, and *Arabidopsis* reveals conserved ancient evolutionary lineages and distinct dynamics of individual *copia* families. Genome Res. 17: 1072–1081. https://doi.org/10.1101/gr.6214107

Wicker, T., F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy *et al.*, 2007 A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. 8: 973–982. https://doi.org/10.1038/nrg2165

Xu, Z., and H. Wang, 2007 LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 35: W265–W268. https://doi.org/10.1093/nar/gkm286

Zhao, G. Y., C. Zou, K. Li, K. Wang, T. B. Li *et al.*, 2017 The *Aegilops tauschii* genome reveals multiple impacts of transposons. Nat. Plants 3: 946–955. https://doi.org/10.1038/s41477-017-0067-8

Communicating editor: S. Wright