Birth and Death of LTR Retrotransposons in *Aegilops tauschii*

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

[*]Department of Statistics, University of California, Davis CA USA
[†]Department of Plant Sciences, University of California, Davis CA USA
[‡]Department of Genetics, University of Georgia, Athens GA USA

Running Title: Birth and Death of LTR-RTNs

Keywords: Transposable elements; insertion rates; demography; population dynamics

Correspondence to: Xiongtao Dai (dai@ucdavis.edu)
Department of Statistics
University of California, Davis
Davis, CA, 95616
1-530-574-9114
ORCID ID: 0000-0002-6996-5930

Abstract

Long Terminal Repeat (LTR) retrotransposons are the majority component of most flowering plant genomes, in particular for *Aegilops tauschii*, a progenitor of bread wheat. This study develops novel estimates for the *time-dynamic* insertion rates of the LTR retrotransposon families in *Ae. tauschii*. For each LTR retrotransposon family, the estimation of insertion rate (birth) consists of an improved estimate of the age distribution that takes into account random mutations, and an adjustment by the deletion rate (death) of LTR retrotransposons. This adjustment is crucial because older elements are more likely to be deleted and thus less observable. Our analyses reject the hypothesis that the LTR retrotransposons were inserted into the *Ae. tauschii* genome at a uniform rate, and find that peak insertion activities range from 0.064 to 2.39 million years ago across different families. Through simulations, we demonstrate the proposed hypothesis test is specific under the null hypothesis of uniform insertion activities, when a histogram of divergence would otherwise suggest a decreasing insertion rate. Finally, we confirm sites near genes tend to lose LTR retrotransposons more rapidly. The proposed estimation methods are available in R package `TE` available on CRAN.

Introduction

Long Terminal Repeat (LTR) retrotransposons are present in virtually all studied eukaryotes, and make up the majority of the nuclear genomes in most flowering plants [1]. LTR retrotransposons are classified into five subfamilies: *Copia*, *Gypsy*, *Bel-Pao*, Retrovirus and ERV, and among them, *Copia* and *Gypsy* are predominant in plant genomes, which each contains hundreds of different LTR retrotransposon families that are operationally distinguished by their different LTR sequences [2]. Any single plant will routinely contain several hundred different LTR retrotransposon families, of which a few will be highly abundant (contributing hundreds to thousands of copies), but with most families having intact element copy numbers of only 1-5 [3, 4]. Variation in the copy numbers of these LTR retrotransposons is the major factor responsible for the huge (>3000 fold) genome size variation in flowering plants. Because LTR retrotransposons transpose via integration of a reverse transcribed transcript, without any donor element excision, they can very rapidly increase their copy 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 retrotransposon families in the ancestors of *Zea luxurians* led to more than a doubling of that genome size in <2 million years, requiring the addition of >2400 Mb of new LTR retrotransposon DNA in that short time period [5].

The transposition of these different LTR retrotransposon families exhibits episodic and apparently stochastic activation over evolutionary time [4, 6]. Because the two LTRs of a single LTR retrotransposon are usually identical at the time of insertion, insertion dates can be estimated by investigating the degree of LTR divergence within a single LTR retrotransposon [7]. Such analyses indicate that individual LTR retrotransposon families exhibit different histories of "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 retrotransposons can add hundreds of Mb of new LTR retrotransposons per million years. However, this process does not always lead to genome size expansion over evolutionary time, because there are also very rapid processes for the removal of DNA from flowering plant genomes [8-11]. Unequal homologous recombination between the LTRs of a single LTR retrotransposon routinely leads to the loss of all internal sequences and the generation of a solo LTR. This attenuates transposition-driven genome growth, but does not reverse it. However, DNA loss by accumulated deletions caused by illegitimate recombination can slow or even reverse genome growth. The mechanism(s) of illegitimate recombination responsible for the process of genome shrinkage has not been proven, but deletion outcomes of the repair of double-strand breaks or adjacent single-strand nicks appear to be the most important driver [8, 12-14].

The relative rates of amplification and removal of LTR retrotransposons and other unnecessary DNA varies across plant lineages [15], and may also be quite variable

101 across regions in the plant genome [16] and over evolutionary time within a lineage
102 [5]. This genome dynamism creates the raw material for natural selection to derive
103 superior individuals, especially when one considers that a high percentage of
104 transposable element (TE) insertions of all types can lead to altered regulation, both
105 genetic and epigenetic, of nearby genes [17]. Understanding the significance of
106 genome dynamism created by TE activities and rates of genome change will require
107 more accurate quantitation and modelling than any of the isolated observations
108 published to date. This study provides an important step in that direction.

109 The focus of this study is modelling the dynamism of the LTR retrotransposon
110 families during the evolution of the *Aegilops tauschii* genome. *Ae. tauschii* is one of
111 the three diploid progenitors of bread wheat. It has a large genome, about 4.3 Gbp,
112 that is at least 66% LTR retrotransposons [18], mostly present as nested arrays of
113 TEs between tiny gene islands [19]. These intergenic arrays are entirely replaced in
114 a span of three to four million years, because of the deletions of old elements and
115 insertions of new elements [20].

116 This dynamic nature of the *Ae. tauschii* LTR retroelements is employed here in
117 modelling their biodemography. The insertion rates of LTR retrotransposons have
118 been analysed previously in *Oryza sativa* [4, 10, 21], Triticeae [6], and *Arabidopsis*
119 [6], but a principled statistical modelling approach was not used. Statistical models
120 have been proposed for analysing the dynamics of retrotransposons in some
121 species, including *Drosophila* [22], *Saccharomyces cerevisiae* [23], *Arabidopsis*
122 *thaliana* [24], and *Homo sapiens* [25].

123 Here, we frame the insertion/deletion dynamics of LTR retrotransposons in terms
124 of birth/death processes that change the age composition over time, with the goal to
125 recover the insertion rate for *Ae. tauschii* LTR retrotransposon families with ≥ 50
126 elements. We model the relationship between LTR retrotransposon insertion rates,
127 deletion rates, and age distributions, building on a model from biodemography [26],
128 and demonstrate the utility of these models to infer insertion rates. A key difference
129 between the age distribution and the insertion rate is that the former describes the
130 ages of only the intact elements that survived the deletion process to the present
131 day, while the latter is the rate of insertion activities for all LTR retrotransposon.
132 For an LTR retrotransposon family, the insertion rate is estimated by the ratio of the
133 age distribution and the deletion rate, adjusting for the fact that older elements are
134 more likely to be deleted and thus less observable. We also propose a new estimate
135 for the age distribution by fitting a negative binomial distribution to the distribution
136 of the number of mismatches in each pair of LTRs of the same LTR family, and then
137 transforming to a gamma age distribution by a probability identity.

138 Our results reject, with high significance, the hypothesis that LTR retrotransposons
139 were inserted into the *Ae. tauschii* genome at a uniform rate. The death rates of LTR
140 retrotransposons are difficult to obtain because deletion events cannot be easily
141 dated, so a sensitivity analysis is conducted to investigate different scenarios of
142 death rates and the resulting insertion rate estimates. We also investigate the
143 associations between the age of LTRs and other genomic variables including
144 recombination rates, distance to the nearest gene, membership in LTR

retrotransposon superfamilies, and chromosome location, using a regression analysis. Proposed analysis algorithms are included in a user-friendly R package that we have named `TE`, which is available on the Comprehensive R Archive Network (CRAN).

Methods and Materials

LTR Retrotransposons

Intact LTR retrotransposons with a target site duplication were identified by using `LTR_FINDER` [27] and `LTRharvest` [28] scanning of the *Ae. tauschii* genome assembly [18] and combining non-redundant predictions of the two program tools. An intact LTR 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 target site duplication (TSD); see for example [9]. Artificial predictions were excluded by manual inspection; see the Supplementary Materials for more details. A group of elements were classified into a family if their 25 bp TE ends exhibited at least 80% identity.

A total of 18,024 copies of 390 LTR retrotransposon families were identified, and we performed the demographics analysis on 15,781 copies, which were in the 35 largest LTR retrotransposon families, all with $\geq$ 50 copies each, consisting of 9

164 *Copia* families and 26 *Gypsy* families (Table S1). The divergence of an LTR
165 retrotransposon is defined as the number of mismatches in the two LTRs divided by
166 the LTR length. Indels were not included in this analysis.

167 Statistical Modelling for LTR Retrotransposon Insertion Dates

168 For each LTR retrotransposon family, we model its population demographics as
169 follows. Throughout, any time $t \geq 0$ refers to time in years in the past relative to the
170 current calendar time, i.e., t years before the current calendar time, which is set to 0.
171 The age distribution at any time t in the past is defined as the distribution of the
172 ages (i.e., time since insertion) of all intact LTR retrotransposons within the family
173 at that time. We use the probability density function $g(a, t)$ to represent the age (a)
174 distribution at time t. Then $g(a, 0)$ is the age distribution or the distribution of the
175 true insertion dates at present. We let $\gamma(t)$ denote the birth rate or insertion rate
176 (insertions per myr) at time t in the past, and assume that $\gamma(t)$ corresponds to the
177 intensity of an inhomogeneous Poisson point process; then $\gamma(t)$ is proportional to
178 the expected number of elements inserted into the genome within period $[t, t + \Delta]$,
179 for an infinitesimal time interval Δ.

180 The insertion rate $\gamma(t)$ is assumed to be changing over time to reflect periods with
181 changing insertion activities, in contrast to the assumption of constant insertion rate
182 of [23, 25]. A key difference between the age distribution $g(a, 0)$ at present-time
183 $t = 0$, as a function of age a, and the insertion rate $\gamma(t)$, as a function of time t, is
184 that the former describes the ages of only the intact elements that survived the

deletion process to the present day, while the latter is the rate of birth for all elements at some time t in the past, regardless of whether they are 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 the $\gamma(t)$ because even if $\gamma(t)$ has been constant throughout, $g(a, 0)$ will be decreasing, since older elements are more likely to be deleted and thus less observable.

Since LTR retrotransposons are subject to rapid deletion [8, 9], 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 retrotransposon has probability $\overline{F}(a) = P(X > a)$ to survive the deletion process to age a, where X is the life span of an LTR retrotransposon, and that the survival function $\overline{F}(a)$ does not depend on the calendar time t. This assumption means 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 remained relatively 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 $\gamma(t + a)\overline{F}(a)$, where $\gamma(t + a)$ is the birth intensity at time $t + a$ years before present, and $\overline{F}(a)$ is the fraction of elements surviving past age a. By normalizing the product into a density function, we obtain the age distribution

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

The integral in the previous display is finite as long as $\gamma(t)$ 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

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

where $\propto$ denotes a proportional relationship, since the integral does not depend on a. The ratio $g(a,0)/\overline{F}(a)$ 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 over the millennia, and thus is the target of investigation.

We next estimate the survival function $\overline{F}(a)$. In the literature it is generally assumed that the distribution of the life span of TEs is exponential, which means the hazard rate for removal of a TA is constant and the distribution is characterized by half-life. The half-life for rice LTR retrotransposons was estimated to be less than 3 myr [9, 10], and that for rice *Copia* elements around 796,000 yr [6]. 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 substituting the histogram of the insertion date estimates [6, 9, 10, 21, 29], 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. However, we note that this estimate is only a proxy for the true age due to randomness of mutations, and the accuracy is lower for elements with shorter LTRs. Due to the variability in the individual estimates, pooling estimates within a family is subject to increased statistical error, which provides the motivation for the improved methodology introduced here.

Assume 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 [25]), where $r = 1.3\times10^{-8}$ substitutions/(year · site), as proposed by Ma and Bennetzen [30]. Then, the number of mismatches N on a pair of LTRs follows a Poisson distribution with rate $2rlx$. Then 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 elements within a single family has LTR length $l = 500$ bp, is inserted $x = 1$ Mya (million years) ago, and the number of mismatches N between the two LTRs follows the Poisson distribution specified above. The distribution of N is shown in the left panel of Figure 1. There is considerable variability in the number of mismatches even in

this case where all elements are inserted into the genome at the same time, with a large coefficient of variation, defined as the ratio of standard deviation over mean (0.277). The histogram estimate of the age distribution by pooling the individual age estimates will have the same coefficient of variation rather than concentrate at 1 Mya, regardless of how many elements are in the family. Therefore this direct approach based on the raw divergence needs to be improved.

We approach this problem by modelling the number of mismatches directly to estimate the age distribution, or the insertion date distribution. We observe that the distributions of the number of mismatches within most of the LTR retrotransposon families are well approximated by negative binomial distributions (see for example the solid and dashed lines in the left panel of Figure 3), so we use this distribution to approximate the marginal distribution of N. For each family, we assume the length l of each LTR is the same and is well approximated by the alignment length. This is a reasonable assumption, since 97% of the elements had alignment length within $\pm 10\%$ 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 relation [31], the distribution of A follows a gamma distribution, which is flexible enough to model exponentially decreasing 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 maximum likelihood estimation (MLE), and then use

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

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 online Supplementary 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

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

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

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

Inference

It is of biological interest to test for a given LTR retrotransposon 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 versus the alternative H_1: $\gamma(t) \neq c$ for all c. By (1) we find that under H_0 for any time z

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

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 density function of the survival time X which is exponential. This implies $g(a, 0)$ is exponential and the distribution of N is geometric, a special case of the negative binomial distribution [31]. 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.

Note that 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) of the test if the true alternative is inadvertently omitted.

We show as example a simulated dataset under H_0 in Figure 2, where each element is inserted uniformly over the past 10 myr, and has a half-life of 1 myr and LTR length equal to 500 bp. In this scenario, although the true insertion rate is uniform, the distribution of mismatches would show an exponential decay, as demonstrated in the left panel of Figure 2, so that the age distribution and the insertion rates are vastly different, and a histogram of divergence leads to an incorrect assessment of the insertion rate. Our proposed method, however, is able to recover the uniform insertion rate in this case, as displayed in the right panel of Figure 2. Testing the null hypothesis at 0.05 significance level in 2,000 simulations under the same setting as Figure 2, the proportion of times H_0 was rejected was 0.051, showing our test has the correct size.

Sensitivity Analysis

We can estimate the birth rate by equation (2) after estimating the age distribution if we know the survival function $\overline{F}(a)$, which corresponds to the death rate. However, even with the exponential life span assumption, the death rate is hard to estimate from the data because the deletion events are not observed, so we compare a range of death rates and conduct a sensitivity analysis.

The exponential 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). As a single estimate may not be accurate because there is no guarantee of a good fit for the geometric distribution,

we investigated three scenarios: Baseline death rates $\lambda = \hat{\lambda}$, low death rates $\lambda = \hat{\lambda}/2$, and high death rates $\lambda = 2\hat{\lambda}$. Note that, as in (2), we can only estimate the birth rate up to a constant multiplier, so we normalized all birth rates into density functions that have area under the curve equal to one.

Goodness-of-fit of Negative Binomial Fit

For some of the families, negative binomial distributions showed a lack of fit for the mismatch data. Lack of fit may result in unreliable age distribution estimates. We used the Kullback–Leibler [32] (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 of Q from P is defined to be

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

where we use the kernel density estimate (KDE) as P, representing the underlying "true" distribution, and the negative binomial distributions as Q. For families with $D_{KL} > 0.025$ (*Gypsy* families 24, 35, 36, 40, 44 and *Copia* families 27, 38, 45; they have relatively small copy numbers), we use a mixture of two negative binomial distributions to fit the mismatch data, which provided good fits in all such cases, where the threshold 0.025 was set by visually inspecting the goodness-of-fit. When $D_{KL} > 0.025$, the recovered age distribution using the mixture approach is a mixture

of two Gamma distributions. The estimates were obtained by MLE, with 1000 random starting points to search for the global maximizer.

Regression Analysis of TE Ages

We fitted a linear mixed effects model to investigate the relationship between response LTR divergence d of a TE, as a proxy for its insertion date, and its other attributes, including the chromosome number, local recombination rate, log distance (in bp) to the nearest gene, superfamily membership (either *Gypsy* or *Copia*), and a LTR family random effect. The local recombination rates were estimated by the first derivative of a local kernel quadratic smoother applied on genetic linkage data in centimorgans [33], with Gaussian kernel and bandwidth equal to 5Mb. To calculate the distances to the nearest gene we used only high confidence genes [34].

Results

An example of recovered age distribution for the largest *Gypsy* family *Fatima* (in the mismatch scale rather than time scale) is shown in the left panel of Figure 3. The histogram of N is shown with the fitted distributions overlaid. The fitted negative binomial distribution is very close to the kernel density estimate, showing a good fit. The recovered age distribution has a more salient peak at 1.28 mya in the time scale (transformed from a peak of 15.6 in the mismatch scale) than that produced by the histogram method, where the latter significantly underestimates the age

356 distribution near its peak period, suffering from the convolution with the Poisson
357 error. *Gypsy* family 24 (*Nusif*) in the right panel of Figure 3 shows a lack-of-fit to a
358 negative binomial distribution, which is remedied by a mixture of two negative
359 binomial distributions.

360 The constant insertion rate hypotheses were rejected for all LTR transposon
361 families with very small *p*-values (Table S2 and S3), indicating that the insertion
362 rates are not constant over time. We show the estimated age distributions and
363 insertion rates of all families in the left and the right panels of Figure 4, respectively,
364 where the insertion rates were estimated with the baseline death rate $\hat{\lambda}$, and then
365 normalized into probability densities. Since older elements are less likely to survive
366 the deletion process, the insertion rates as compared to age distributions
367 compensated for this effect by attenuating earlier peaks and amplifying later peaks.
368 Each family was active during a different time range, while the peak insertion
369 activities for most families tended to occur around 1 mya, ranging from 0.064 mya
370 to 2.39 mya. The most recent sharp insertion rate spikes at 0.064 mya are due to
371 two *Copia* elements in family 27 (*Maximus*) that have only 1 and 4 mismatches,
372 vastly different from other elements in the same family that have an average of 40
373 mismatches. This shows that *Copia* family 27 had an ancient burst of activity,
374 followed by a recent amplification that may be on-going.

375 To demonstrate the sensitivity of our results to the assumption on death rates, we
376 studied and show three death rate scenarios for the top five *Copia* families and the
377 top five *Gypsy* families in Figure 5, which are based on family-specific baseline

scenarios. An important consequence of exploring the three death rate scenarios is that, while the precise location of the peaks of insertion times may move in time, the sequence of peaks is not much affected by varying assumptions on elimination rates, therefore validating the location of the peaks. Salient peaks are evident in each family, meaning that these families all underwent periods of rapid amplification. In a scenario assuming a higher death rate, peaks are shifted back in time; this is a consequence of equation (2).

The results of a regression analysis for the association between LTR divergence and TE attributes are reported in Table 1. LTR retrotransposons on chromosomes 2D, 4D, and 7D have significantly larger divergence (and thus are older) as compared to those on chromosome 1D. The recombination rate has a significant negative effect, while the log distance to the nearest gene has a significant positive effect on insertion dates. The distance to the nearest gene may be a proxy for higher recombination rates near genes, which leads to more unequal homologous recombination events, and thus more frequent removal of complete elements [16] and younger TEs. The predictor recombination rate is a smooth average of local recombination rates in finer scales. On average, *Gypsy* families tended to be older than *Copia* families.

Discussion

TEs drive the evolution of genome structure, both by their insertion activities and by their subsequent contributions as sites of chromosome breakage and ectopic

homologous recombination [1]. In flowering plants, it is commonly observed that even closely related lineages can have dramatically different histories of TE activity [5]. Beyond restructuring genomes, TEs also provide the raw material for epigenetic changes and most other changes in gene regulation [17], as well as possibly contributing to the function of structural components like centromeres [35]. Hence, a detailed and quantitatively robust analysis of TE activity is warranted to permit the understanding of TE contributions to the evolution of both structure and function in any genome.

LTR retrotransposons are uniquely well suited for the study of genome dynamics for several reasons. First, the identity of the two LTRs at the time of any insertion event allows the subsequent determination of the insertion date by quantifying LTR divergences within a single element [7]. Second, the transposition mechanism for LTR retrotransposons does not involve element deletion from the donor site, so that each insertion can be viewed as a simple, one element, amplification. Third, most LTR retrotransposons avoid inserting near or into genes [3, 36], so the effects of natural selection on LTR retrotransposon retention are minimized, although not fully neutralized. Fourth, the processes for LTR retrotransposon sequence removal (unequal homologous recombination to generate solo LTRs and illegitimate recombination) have been identified [8, 9], so they can be factored into any analysis of LTR retrotransposon dynamics.

The advantages of our proposed models over an estimate based on the previously utilized histogram of divergence are twofold. First, our insertion rate takes into

account the deletion process, producing more realistic estimates that puts more weight on the older and thus harder to observe elements (Figure 2). Modeling the death rate has a significant impact in the insertion rate if it is constant or near constant, in which case a histogram of divergence would show an exponential decay, as demonstrated in a simulated scenario (Figure 4) and as empirically shown in other species, e.g. rice [10, 21]. Using our model, one can formally test the hypothesis that the insertion rates are constant over time, which is in doubt especially if the distribution of divergence is exponentially decaying. Second, the randomness of mutations are taken into account, which results in more pronounced peaks in the age distribution estimates (Figure 3), indicating the insertion rates are more concentrated around bursts of activities than what appears in a histogram of divergence.

User-friendly and fast algorithms for the proposed analysis are conveniently available in the R package `TE` on CRAN, enabling easy comparisons with classical approaches. The package `TE` includes `EstDynamics` and `EstDynamics2` for estimating insertion rates and age distributions, where the former also tests the hypothesis of a constant insertion rate, and `PlotFamilies` and `SensitivityPlot` for generating additional plots.

For a pragmatic estimation of the death rate of TEs, we employed an exponential life span assumption [6, 10, 30] that amounts to a constant hazard rate. With this we produce more realistic insertion rate estimates than those obtained from previous methods. Time- or age-varying hazard rate estimates, however, require the

observation of historical TE removal events, for example by comparing multiple species. This is left for future work because high quality data informing deletion events is unavailable at this stage. Our current framework modelled the insertion rate and age distributions as parametric, allowing for fast computation without tuning parameters, while a possible alternative Bayesian framework modelling the insertion activities as a latent process was not considered here.

Our proposed models allow for time-varying insertion rates that are appropriate for dynamic transposition activity, which is a realistic scenario as demonstrated in simulations [37] and by the LTR retrotransposons of *Ae. tauschii*, where recent insertions are near absent for reasons currently unknown. Our time-dynamic modelling approach is in contrast to Promislow et al. [23], who modelled the insertion activity as constant over time or two-stage, and Marchani et al. [25], who targeted the age of the master gene for a retrotransposon subfamily. Previous work [22, 24] studied the TE dynamics of species with small genomes and multiple available linages, where the latter associated Helitron element ages with occupation frequency, while we study a single accession of *Ae. tauschii* with a large genome size (4.3 Gbp) and abundant repeated elements (65.9%), and found that the age of an LTR retrotransposon was associated with variables such as distance to the nearest genes and recombination rates.

The results of these studies indicate that a robust statistical analysis of LTR retrotransposon dynamics is feasible with the appropriate computational strategy and statistical models. As predicted, but never confirmed rigorously, our analyses

indicate bursts of LTR retrotransposon activity that are family specific, and we show multiple peaks of activity even within a single family history. Survival is also modelled, and confirms predictions that sites near genes (where negative selection is more likely to act) lose LTR retrotransposons more rapidly. Similarly, LTR retrotransposons within regions, such as genic areas, that exhibit high levels of meiotic recombination (where solo LTR generation should be more frequent) also were substantiated as sites of relatively rapid LTR retrotransposon loss. Taken in their entirety, these studies support a rigorous approach to analysing LTR retrotransposon histories across plant lineages, thus creating the opportunity to investigate these dynamics from a phylogenetically powerful perspective.

Data availability. Data and code are included in the R package `TE`, available on CRAN.

Authors' contributions. JD conceived the birth and death analysis, and this conceptualization was refined through contributions from all authors. HW annotated the LTR retrotransposons and provided the data. XD and HGM proposed the statistical model and analysed the data. XD developed the R package `TE`. XD, JD, and JB wrote the manuscript. All authors contributed to revising the manuscript.

Competing interests. We have no competing interests.

Funding. The study was supported by National Science Foundation (NSF-ISO-1238231).

Acknowledgements. We thank Matthew Dawson for providing the estimates for recombination rates, and Patrick McGuire for checking the manuscript.

References

[1] Bennetzen, J.L. & Wang, H. 2014 The Contributions of Transposable Elements to the Structure, Function, and Evolution of Plant Genomes. *Annual Review of Plant Biology* **65**, 505-530. (doi:10.1146/annurev-arplant-050213-035811).

[2] Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., et al. 2007 A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics* **8**, 973-982. (doi:10.1038/nrg2165).

[3] Baucom, R.S., Estill, J.C., Chaparro, C., Upshaw, N., Jogi, A., Deragon, J.-M., Westerman, R.P., SanMiguel, P.J. & Bennetzen, J.L. 2009 Exceptional Diversity, Non-Random Distribution, and Rapid Evolution of Retroelements in the B73 Maize Genome. *PLoS Genetics* **5**, e1000732. (doi:10.1371/journal.pgen.1000732).

[4] Baucom, R.S., Estill, J.C., Leebens-Mack, J. & Bennetzen, J.L. 2009 Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome. *Genome Research* **19**, 243-254. (doi:10.1101/gr.083360.108).

[5] Estep, M.C., DeBarry, J.D. & Bennetzen, J.L. 2013 The dynamics of LTR retrotransposon accumulation across 25 million years of panicoid grass evolution. *Heredity* **110**, 194-204. (doi:10.1038/hdy.2012.99).

[6] Wicker, T. & Keller, B. 2007 Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient

509 evolutionary lineages and distinct dynamics of individual copia families. *Genome*
510 *Research* **17**, 1072-1081. (doi:10.1101/gr.6214107).
511 [7] SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y. & Bennetzen, J.L. 1998 The
512 paleontology of intergene retrotransposons of maize. *Nature genetics* **20**, 43-45.
513 [8] Devos, K.M., Brown, J.K. & Bennetzen, J.L. 2002 Genome size reduction through
514 illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome*
515 *research* **12**, 1075-1079.
516 [9] Ma, J., Devos, K.M. & Bennetzen, J.L. 2004 Analyses of LTR-retrotransposon
517 structures reveal recent and rapid genomic DNA loss in rice. *Genome Research* **14**,
518 860-869.
519 [10] Vitte, C., Panaud, O. & Quesneville, H. 2007 LTR retrotransposons in rice (Oryza
520 sativa, L.): recent burst amplifications followed by rapid DNA loss. *BMC Genomics* **8**,
521 218. (doi:10.1186/1471-2164-8-218).
522 [11] Hawkins, J.S., Proulx, S.R., Rapp, R.A. & Wendel, J.F. 2009 Rapid DNA loss as a
523 counterbalance to genome expansion through retrotransposon proliferation in
524 plants. *Proceedings of the National Academy of Sciences* **106**, 17811-17816.
525 [12] Kirik, A., Salomon, S. & Puchta, H. 2000 Species-specific double-strand break
526 repair and genome evolution in plants. *The EMBO Journal* **19**, 5562-5566.
527 [13] Vaughn, J.N. & Bennetzen, J.L. 2014 Natural insertions in rice commonly form
528 tandem duplications indicative of patch-mediated double-strand break induction
529 and repair. *Proceedings of the National Academy of Sciences* **111**, 6684-6689.
530 (doi:10.1073/pnas.1321854111).

[14] Schiml, S., Fauser, F. & Puchta, H. 2016 Repair of adjacent single-strand breaks is often accompanied by the formation of tandem sequence duplications in plant genomes. *Proceedings of the National Academy of Sciences* **113**, 7266-7271. (doi:10.1073/pnas.1603823113).

[15] Vitte, C. & Bennetzen, J.L. 2006 Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. *Proceedings of the National Academy of Sciences* **103**, 17638-17643. (doi:10.1073/pnas.0605618103).

[16] Ma, J. & Bennetzen, J.L. 2006 Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. *Proceedings of the National Academy of Sciences* **103**, 383-388. (doi:10.1073/pnas.0509810102).

[17] Lisch, D. & Bennetzen, J.L. 2011 Transposable element origins of epigenetic gene regulation. *Current Opinion in Plant Biology* **14**, 156-161. (doi:10.1016/j.pbi.2011.01.003).

[18] Luo MC, Gu Y-G, Puiu D, Wang H, Twardziok S, Deal KR, Huo N, Zhu T, Wang L & al., W.Y.e. 2017 Reference-quality sequence of the genome of Aegilops tauschii, the progenitor of the wheat D genome, suggests cause of rapid genome evolution. *Nature* **Submitted**.

[19] Gottlieb, A., Müller, H.-G., Massa, A.N., Wanjugi, H., Deal, K.R., You, F.M., Xu, X., Gu, Y.Q., Luo, M.-C., Anderson, O.D., et al. 2013 Insular Organization of Gene Space in Grass Genomes. *PLoS ONE* **8**, e54101. (doi:10.1371/journal.pone.0054101).

552 [20] Dubcovsky, J. & Dvorak, J. 2007 Genome Plasticity a Key Factor in the Success of
553 Polyploid Wheat Under Domestication. *Science* **316**, 1862-1866.
554 (doi:10.1126/science.1143986).
555 [21] Wang, L., Brown, L.D., Cai, T.T. & Levine, M. 2008 Effect of mean on variance
556 function estimation in nonparametric regression. *The Annals of Statistics* **36**, 646-
557 664. (doi:10.1214/009053607000000901).
558 [22] Charlesworth, B. & Langley, C.H. 1989 The Population Genetics of Drosophila
559 Transposable Elements. *Annual Review of Genetics* **23**, 251-287.
560 (doi:10.1146/annurev.ge.23.120189.001343).
561 [23] Promislow, D.E.L., Jordan, I.K. & McDonald, J.E. 1999 Genomic demography: a
562 life-history analysis of transposable element evolution. *Proceedings of the Royal*
563 *Society B: Biological Sciences* **266**, 1555-1560. (doi:10.1098/rspb.1999.0815).
564 [24] Hollister, J.D. & Gaut, B.S. 2007 Population and Evolutionary Dynamics of
565 Helitron Transposable Elements in Arabidopsis thaliana. *Molecular Biology and*
566 *Evolution* **24**, 2515-2524. (doi:10.1093/molbev/msm197).
567 [25] Marchani, E.E., Xing, J., Witherspoon, D.J., Jorde, L.B. & Rogers, A.R. 2009
568 Estimating the age of retrotransposon subfamilies using maximum likelihood.
569 *Genomics* **94**, 78-82. (doi:10.1016/j.ygeno.2009.04.002).
570 [26] Müller, H.-G., Wang, J.-L., Yu, W., Delaigle, A. & Carey, J.R. 2007 Survival and
571 aging in the wild via residual demography. *Theoretical Population Biology* **72**, 513-
572 522. (doi:10.1016/j.tpb.2007.07.003).

573 [27] Xu, Z. & Wang, H. 2007 LTR_FINDER: an efficient tool for the prediction of full-
574 length LTR retrotransposons. *Nucleic Acids Research* **35**, W265-W268.
575 (doi:10.1093/nar/gkm286).
576 [28] Ellinghaus, D., Kurtz, S. & Willhoeft, U. 2008 LTRharvest, an efficient and
577 flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*
578 **9**, 18. (doi:10.1186/1471-2105-9-18).
579 [29] Nystedt, B., Street, N.R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D.G.,
580 Vezzi, F., Delhomme, N., Giacomello, S., Alexeyenko, A., et al. 2013 The Norway
581 spruce genome sequence and conifer genome evolution. *Nature* **497**, 579-584.
582 (doi:10.1038/nature12211).
583 [30] Ma, J. & Bennetzen, J.L. 2004 Rapid recent growth and divergence of rice
584 nuclear genomes. *Proceedings of the National Academy of Sciences* **101**, 12404-
585 12410. (doi:10.1073/pnas.0403715101).
586 [31] Leemis, L.M. & McQueston, J.T. 2008 Univariate Distribution Relationships. *The*
587 *American Statistician* **62**, 45-53. (doi:10.1198/000313008x270448).
588 [32] Kullback, S. & Leibler, R.A. 1951 On Information and Sufficiency. *The Annals of*
589 *Mathematical Statistics* **22**, 79-86. (doi:10.1214/aoms/1177729694).
590 [33] Fan, J. & Gijbels, I. 1996 *Local polynomial modelling and its applications:*
591 *monographs on statistics and applied probability 66*, CRC Press.
592 [34] Luo, M.-C., Gu, Y.Q., Puiu, D., Wang, H., Twardziok, S.O., Deal, K.R., Huo, N., Zhu,
593 T., Wang, L., Wang, Y., et al. 2017 Genome sequence of the progenitor of the wheat D
594 genome Aegilops tauschii. *Nature*. (doi:10.1038/nature24486
595 https://www.nature.com/articles/nature24486 - supplementary-information).

596 [35] Nagaki, K., Song, J., Stupar, R.M., Parokonny, A.S., Yuan, Q., Ouyang, S., Liu, J.,
597 Hsiao, J., Jones, K.M. & Dawe, R.K. 2003 Molecular and cytological analyses of large
598 tracks of centromeric DNA reveal the structure and evolutionary dynamics of maize
599 centromeres. *Genetics* **163**, 759-770.
600 [36] SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-
601 Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., et al. 1996 Nested
602 Retrotransposons in the Intergenic Regions of the Maize Genome. *Science* **274**, 765-
603 768. (doi:10.1126/science.274.5288.765).
604 [37] Le Rouzic, A., Boutin, T.S. & Capy, P. 2007 Long-term evolution of transposable
605 elements. *Proceedings of the National Academy of Sciences* **104**, 19375-19380.

606

607

608 # Tables
609

610 **Table 1**. Regression coefficient estimates.

	Value	Std. error	t-value	p-value
Intercept	0.0066	0.0011	5.91	0.0000
Chr2	0.0006	0.0003	2.21	0.0270
Chr3	0.0002	0.0003	0.81	0.4160
Chr4	0.0007	0.0003	2.38	0.0173
Chr5	0.0004	0.0003	1.21	0.2267
Chr6	0.0007	0.0003	2.16	0.0304
Chr7	0.0012	0.0003	4.02	0.0001
Recombination rate	-0.0007	0.0002	-3.84	0.0001
Log distance	0.0017	0.0001	24.22	0.0000
Gypsy superfamily	0.0031	0.0012	2.64	0.0086

611

Figure captions

Figure 1. The distribution of the number of mismatches, when all elements are of length 500 bp and inserted 1 mya.

Figure 2. Simulated distributions of the number of mismatches, where each element is inserted into the genome uniformly over the past 10 myr and has a half-life of 1 myr and LTR length equal to 500 bp. Left: A random selection of 100 such elements that survive to the current time. Right: The estimated insertion rate using our proposed method.

Figure 3. Distributional fits and recovered age distribution of *Gypsy* family 1, *Fatima* (left), produced by function `EstDynamics`, and *Gypsy* family 24, *Nusif* (right), produced by `EstDynamics2`. The black lines show the kernel density estimate (KDE, solid), the negative binomial fit by MLE (dashed), and the recovered age distribution expressed in mismatch time scale (dash-dot). For *Gypsy* family *Nusif*, a negative binomial fit shows lack of fit as measured by Kullback–-Leibler (KL) divergence (see Subsection Lack-of-fit of Negative Binomial Fit). Thus, 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).

Figure 4. Age distributions (left panels) and normalized insertion rates (right panels) in the 35 largest families. Each curve represents the estimated age distribution (left) or insertion rate as normalized into a probability density function (right) of a single family. *Copia* families are shown in red and *Gypsy* families in blue. Grey triangles on the x-axis indicate the peak locations. The peak insertion activities for most families occur around 1 mya, ranging from 0.064 mya to 2.39 mya, marked by black squares.

Figure 5. Sensitivity analysis for the 1st, 3rd, and 5th largest *Copia* (left) and *Gypsy* (right) families, respectively. For each family, three death rate scenarios are shown: Baseline death rates $\lambda = \hat{\lambda}$ (solid), low death rates $\lambda = \hat{\lambda}/2$ (dashed), and high death rates $\lambda = 2\hat{\lambda}$ (dotted). Short horizontal lines on each curve mark the times when the insertion activities are half as strong as the peak intensity in each scenario.