# Title:

# TESS: A Forward Simulation Framework for Studying the Role of Transposable Elements in Genome Expansion and Contraction

## Authors:

Christopher W. Benson[1†], Ting-Hsuan Chen[2†*], Tianyu Lu[1], Olivia Angelin-Bonnet[3], Susan J. Thomson[2], Cecilia H. Deng[4], Shujun Ou[1*]
1 Department of Molecular Genetics, Ohio State University, Columbus, OH, 43210, USA
2 The New Zealand Institute for Plant and Food Research Limited, Lincoln 7608, New Zealand
3 The New Zealand Institute for Plant and Food Research Limited, Palmerston North 4442, New Zealand
4 The New Zealand Institute for Plant and Food Research Limited, Auckland 1025, New Zealand

† Christopher W. Benson and Ting-Hsuan Chen contributed equally to this work.
* Correspondence: Ting-Hsuan.chen@plantandfood.co.nz; ou.195@osu.edu

## Abstract

Genome size varies by several magnitudes among eukaryotes. Expansion and contraction are primarily driven by transposable element activity, but the underlying processes remain enigmatic due to a lack of historical records tracing these changes. Here, we present the TE Evolution Simulation Suite, or TESS, for versatile simulation of whole-genome sequences with varying transposon dynamics. We analyzed the pangenome of *Pucciniomycotina*, a subdivision of fungi containing the myrtle rust (*Austropuccinia psidii*) genome, which drastically expanded to 1018 Mb. Using TESS, we simulated digital replicas of *Pucciniomycotina* genomes and replicated the genome size evolution of *A. psidii* driven by LTR retrotransposons.

Keywords: Transposable Element, Genome Size, Forward Simulation, InDels, Solo LTR, Structural Variation, Speciation, Rust Fungus

# Background

Transposable elements (TEs), or “jumping genes”, are dynamic DNA sequences that replicate and relocate within genomes. TEs constitute 46% of the human (*Homo sapiens*) genome [1], 85% of the maize (*Zea mays*) genome [2], and over 90% of the genome of the fungal pathogen *Austropuccinia psidii* that causes the myrtle rust disease [3]. The prevalence of TEs across the Tree of Life reflects their dual roles as both genomic parasites and drivers of evolutionary innovation. Beyond inflating genome sizes, TEs drive structural variation, modulate gene expression, and facilitate chromosomal rearrangements [4]. In both plant and mammalian genomes, bursts of TE activity correlate with speciation events and adaptive radiations, enabling rapid responses to environmental stressors like pathogens or climate shifts [5].

Recent studies highlight TE’s contribution to the collective genetic repertoire of a species. Core genomes, conserved across individuals, are often gene-rich and TE-poor [2], while "variable" or "dispensable" regions are TE-dense and enriched for stress-response genes, pathogen resistance loci, and structural variants [6]. In *Brassica oleracea*, resistance gene analogs cluster in TE-rich variable regions [7]. Similarly, tandem clusters of leucine-rich repeat (LRR) genes in maize are found in genomic regions rich with young long terminal repeat (LTR) retrotransposons [8], suggesting that TE-driven structural plasticity contributes to adaptive traits. Related lineages can become karyotypically disparate over generations due to differences in their ability to repress TEs [9]. The global impact of TEs on macro-chromosomal structure plays an important role in evolutionary fates by influencing meiotic stability and similarity between cytotypes [10]. These findings underscore TEs as dynamic regulators of genome evolution, influencing both microevolutionary adaptation and macroevolutionary diversification.

Genome size varies considerably across taxa. Mammalian and bird genomes appear more tightly constrained in genome size than those of both fungi and plants [11]. Numerous studies across the Mammalia class indicate this size constraint is not through a lack of TE radiation bursts [12] but through mechanistic interplay with purifying selection, where bursts are followed by TE silencing [13] and DNA loss [12]. Genome size variation in plants facilitates adaptation to specific environments, primarily carried out by whole genome duplication (WGD) and TE expansion [14]. Extremes of genome size can be observed from the compact genome of Arabidopsis (~135 Mb) through to the giant genomes of the Chinese pine (*Pinus tabuliformis*) at ~25 Gb [15], lily (*Lilium sargentiae*) at ~36 Gb [16], and the New Caledonian fork fern (*Tmesipteris oblanceolata*) at ~160 Gb [17]. These are examples of unconstrained genome expansion, with extremely high TE content within intergenic regions and also within introns. Conversely, there are plenty of signs that genome size expansion is not left unchecked in plants: Arabidopsis has undergone several WGDs, yet remains compact partially through subsequent gene loss that purges redundant paralogs [18]. The dynamic of genome size change has been difficult to accurately evaluate primarily because of the lack of high-fidelity tools for simulation and modeling studies.

Advances in long-read sequencing and computational methods have empowered efforts in resolving TE dynamics [19], yet reconstructing their evolutionary history remains challenging due to TEs often being fragmented, nested, and highly identical to each other. Simulations of TE evolution are thus indispensable for testing hypotheses about genome diversification, evaluating TE annotation tools, and modeling the interplay between TE activity and selection. By

generating synthetic genomes with known TE architectures, including nested insertions, sequence divergence, and polymorphism, researchers can disentangle the stochastic and deterministic forces shaping genome evolution [20].

Existing TE simulators vary widely in scope and range of biological scenarios that they mimic. Early models, such as those by Clough et al. (1996), introduced foundational concepts by simulating varying modes of mammalian transposition [21]. Modern tools address specific niches. SimulaTE [22] generates complex TE landscapes through predefined TE architectures, limiting its utility for exploring stochastic evolutionary scenarios. Population-genetic frameworks like SLiM [23] excel at modeling selection and demography but compress TEs into individual units, precluding modeling of TE sequence mutagenesis, nesting insertion, and the consequential genome size evolution. TESD [24] incorporates spatial population structure and selection against deleterious insertions, yet its inaccessibility hinders further application. replicaTE [25] engineers synthetic genomes by deleting and reinserting TEs into real assemblies but lack dynamic iterations through evolutionary epochs. Meanwhile, pMEsim [26] focuses on simulating polymorphic TE variants for pan-genomes, sacrificing long-term evolutionary dynamics, and simuG [27] serves as a generic genome simulator that can simulate TE insertions. GARLIC [28] specializes in generating realistic artificial genomic sequences by modeling intergenic sequences based on kmer profiles, GC contents and local repeat landscape, which serve as high-quality negative controls for evaluating annotation tools.

While existing simulation tools collectively address diverse needs, few allow flexibility in genome customization, TE sequence mutagenesis, and dynamic modulation of insertion/deletion rates across evolutionary epochs; none provides granular family-specific controls with iterative simulations for hypothesis testing; and most prioritize synthetic data generation over empirical parameter inference.

# Results

## Introduction of the TE Evolution Simulation Suite (TESS)

Here, we present the TE Evolution Simulation Suite (TESS) that allows dynamic forward simulation of TE insertions, deletions, truncations, and divergence and speciation events. TESS is a simulation toolbox comprising three novel components, RandSeqInsert, TEGenomeSimulator, and PrinTE (**Fig. 1; Table S1**). Each component is capable of executing independently and interconnecting with the other, allowing for versatile simulation of complex genome evolution scenarios.

1. RandSeqInsert specializes in the simulation of TIR DNA transposon insertions. It uses a tree-based algorithm for efficient sequence insertions at random positions, allowing for benchmarking of TE annotation and structural variant detection tools;
2. TEgenomeSimulator uses single time-point simulation of TE insertion to produce a genomic evolutionary snapshot. It can model a breadth of TE mobilization scenarios with deference to the timing and duration of TE burst, and accumulated mutation. These features allow for flexible approximation of the TE composition of real genomes;

3. PrinTE enables iterative forward simulations to explore the effects of TE insertion, mutation, and deletion on genome size evolution, genomic divergence, and the impact of selection on TE turnover.

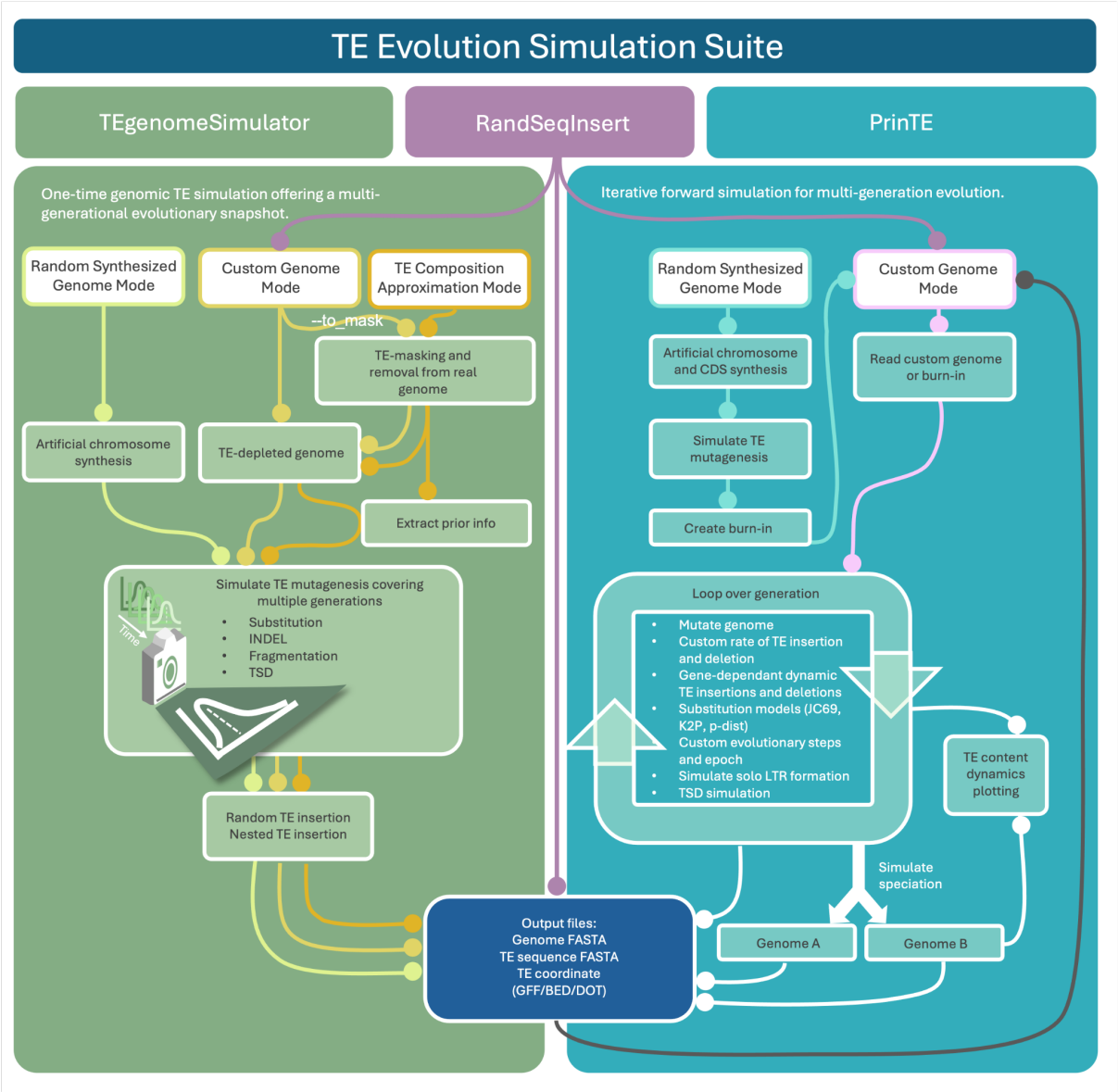


**Fig. 1. Programmatic components and functionality of the TE Evolution Simulation Suite (TESS).** TESS is composed of three modules: TEgenomeSimulator, RandSeqInsert, and PrinTE, represented by distinct colour blocks. The high-level computational workflows of TEgenomeSimulator and PrinTE are illustrated with rectangles and solid lines. The output from RandSeqInsert can serve as input for TEgenomeSimulator and PrinTE. See the main text for detailed descriptions of each module's functionality.

## *RandSeqInsert:* a light-weight simulator for TIR insertions and structural variants

RandSeqInsert was developed to simulate Terminal Inverted Repeat (TIR) transposable element insertions in genomic sequences. This high-performance Python tool implements an

AVL (named after inventors Adelson-Velsky and Landis) tree-based algorithm for efficient sequence insertions at random positions in target sequences, enabling precise control over insertion frequency (**Fig. S1A**). With support for nested insertion, the tool accurately represents how TIR transposons accumulate in genomes over time. RandSeqInsert can track all insertions, generating FASTA, BED, and DOT graph files to support further analysis of simulated transposition patterns (**Fig. S1**).

RandSeqInsert simulates target site duplications (TSDs) as biological insertion signatures with adjustable length and mutation rates, which allow for more accurate detection by structure-based *de novo* annotation tools (**Fig. S2**). We demonstrated the simulation of TIR insertions using RandSeqInsert and the detection of these events using minigraph [29], showcasing the simulation of biologically relevant structural variants and how this simulation framework can be extended to model TE variation at the population or pan-genome levels (**Fig. S1B**). Accurate simulation of TE insertions allows for the creation of ground truth sets for benchmarking and method development purposes (**Fig. S1; Fig. S2**).

### ***TEgenomeSimulator:*** **simulating diverse real-world TE evolution**

TEgenomeSimulator is a versatile tool that builds on and extends functions from denovoTE-eval [30], allowing users to define and test a range of TE mutagenesis scenarios that can be reflective of an evolutionary time point, followed by random and nested insertion into a randomly synthesized or user-provided genome (**Fig. S3**). It tracks TE locations, sequence identity, integrity, and breakpoints from nested insertion, providing flexibility to explore individual TE components and their simulated decay within synthetic genomes (**Table S2**). Key developments include support for multi-chromosome genome synthesis, flexible TE mutagenesis parameters incorporating informed TSD length adjustments, and the ability to analyze and reuse annotated TE composition profiles. Additionally, it streamlines the synthesis of artificial genomes, enabling efficient simulation of digital replicas of real genomes (**Table S2**).

Multiple simulation modes are implemented within TEgenomeSimulator (**Fig. 1**). The Random Synthesized Genome Mode is used to generate a fully synthesized genome adhering to user-defined parameters reflecting genome size and GC content. Users can then modify TE mutation parameters such as TE sequence identity and integrity, as well as copy number.

The simulation of TEs within synthetic genomes is often substantially different from what is observed in real genomes in terms of composition, length, and sequence diversity. To address this, TEgenomeSimulator implemented the TE Composition Approximation Mode, whereby TE composition of a provided genome can be used to build a model for inference of TE mutagenesis parameters. This genome-informed approach can produce a simulated genome whose TE content closely mirrors that of the original genome, in terms of copy number, genomic occupancy, sequence divergence, and length ratios relative to consensus sequences. As a result, it provides a synthetic snapshot of the genome's current TE evolutionary landscape, closely resembling the observed state of the input genome.

Moreover, the Custom Genome Mode implemented in TEgenomeSimulator gives users the flexibility to utilize a hybrid strategy, combining genome backbones with custom TE libraries. For example, users can first generate a digital replica of a real genome using the TE Composition Approximation Mode, then use this simulated genome as an input for the Custom Genome Mode along with a TE library that contains only LTR retrotransposon families from a

different species. By adjusting the range of mean identity, SD, and TE copy numbers, users can simulate a TE burst within the digital replica, modelling scenarios such as horizontal TE transfer.

The resulting genome can be used not only to benchmark computational methods but also as a burn-in phase for forward-time simulations in genome evolution studies, as such simulated genomes and their TE sequences can be used very effectively in combination with PrinTE.

## *PrinTE:* forward simulations of TE insertions, deletions, and genome divergence

PrinTE is a forward-time genome simulation tool designed to model the dynamics of TEs within eukaryotic genomes across evolutionary time and was developed to incorporate as many biologically relevant parameters into the model as possible (**Fig. S4**). It uses a modular approach similar to TEgenomeSimulator, beginning with a user-provided (Custom Genome Mode) or synthetic genome (Random Synthesized Genome Mode) that represents a present-day or theoretical ancestral genome. Users may generate these inputs from prior PrinTE runs, import them from TEgenomeSimulator or RandSeqInsert, or construct them entirely from scratch with PrinTE's built-in parameterization tools.

PrinTE implements two methods for simulation. The variable-rate method aims to incorporate consequences from some of the biological and selective forces that guide TE evolution. We established several rules in this method: 1) The host genome survival and TE proliferation are conflicting forces pitted against each other; 2) TE superfamilies compete with each other since they occupy the same limited space; 3) The rate of TE insertion is positively correlated with the number of intact TEs in the previous iteration; 4) TE deletions are guided by the number of TEs that are deleterious to the host's survival. For this purpose, the simulated genome consists of three components, all of which are receptive to new TE insertions: existing TEs, genic space, and intergenic space. If a TE inserts into another TE or intergenic space, there is no cost to the host and no impact on the TE deletion rate. In contrast, TE insertions into genic space are costly to the host, resulting in elevated TE deletion rates in the next generation. Thus, the outcome of the variable-rate method is less predictable, which may lead to generations of genome size increase followed by subsequent generations of decrease. It is useful for those who wish to explore the regulation of TEs through time and to understand the forces that affect TEs' bursts and purges.

The fixed-rate method is designed to generate more controlled and predictable outputs. In this method, per-generation TE insertion and deletion rates are predefined by the user, allowing the modeling of known evolutionary trajectories and the generation of results that more closely match expectations (**Fig. S4**). In practice, we find it most helpful to either start from a simulated genome that mirrors a present-day version and forecast a few million generations of TE evolution. Alternatively, it can start from a genome as we expect it may have looked a few million generations ago and simulate it to the present day. This approach allows users to test hypotheses relating to TE bloat and purge and to potentially superimpose TE evolutionary narratives onto the species of interest.

PrinTE can simulate as briefly as a single generation, which can generate subtle changes to the simulated genome. To allow for simulating millions of generations, PrinTE employs a stepwise approach in which it can simulate TE evolution of a number of generations in steps until it reaches the total number of generations specified by the user. To approximate

TE mutation in multiple generations, PrinTE uses an exponential decrease distribution to mimic DNA substitution of TE sequences over time, in which most TEs have few mutations.

In summary, TESS represents a state-of-the-art simulation of TE dynamics in eukaryotic genomes, enabling the modeling of TE contributions to genome evolution and the estimation of past and future genome evolution trajectories.

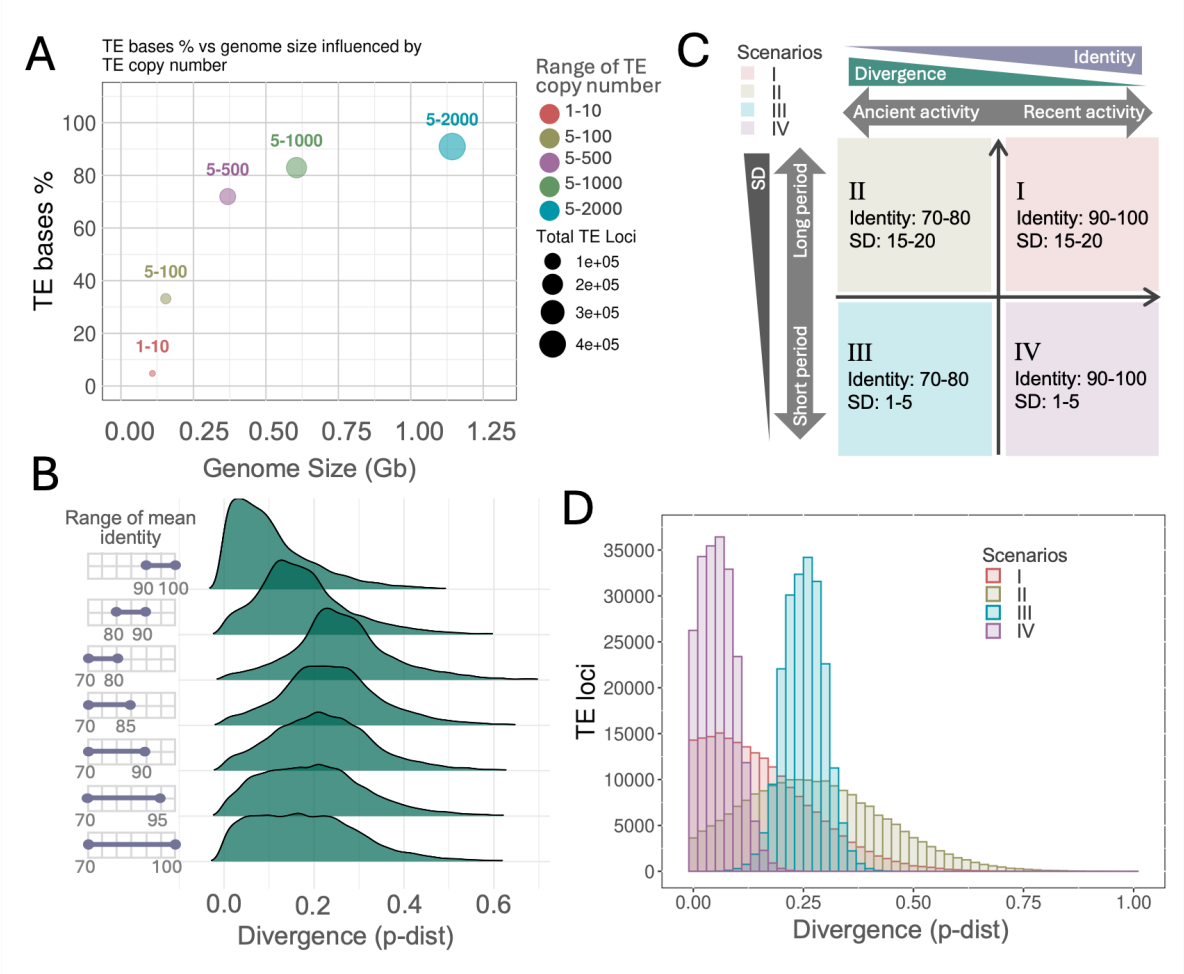


**Fig. 2. Modeling TE insertion and composition for various evolutionary scenarios using TEgenomeSimulator.** (**A**) Impact of altering TE copy number ranges for each TE family on simulated genome size. (**B**) Distribution of TE sequence divergence under variable ranges of mean identity across all TE families. A fixed TE copy number range of 5-100 was used in these simulations. (**C**) Four scenarios of historical TE activity that can be modelled by controlling the range of mean identity that represents the timeline of TE activities and the standard deviation (SD) of mean identity that represents the duration of TE activities. (**D**) Various distribution patterns of TE sequence divergence in association with the settings of mean identity and SD in **C**.

## Modeling the Time and Duration of TE Activity

The composition of TEs within a genome is shaped by a dynamic interplay of TE bursts, host-mediated removal processes, latency of TE activity, accumulation of single-nucleotide variants

227 (SNVs), insertions, deletions, and varying degrees of sequence truncation. The biological
228 insights from various TE profiles are often found associated with evolutionary milestones [31].
229 Accurately modeling TE composition under different amplification and duration scenarios is,
230 therefore, crucial for understanding genome evolution and implementing forward evolutionary
231 simulations.
232 The observed TE copy number in a genome represents the net result of TE amplification
233 and subsequent removal. A history of active TE mobilization, combined with inefficient removal
234 mechanisms, can lead to a substantial accumulation of TE sequences, often resulting in
235 genome expansion. TEgenomeSimulator encapsulates this concept through the modification of
236 TE copy number introduced into simulated genomes. Using TEgenomeSimulator's Custom
237 Genome Mode, in combination with the TE-depleted *Arabidopsis thaliana* (TAIR10) genome and
238 the curated *Arabidopsis* TE library (see Methods), we simulated a range of genome size
239 expansions from under 200 Mb to over 1 Gb. This was achieved by increasing the maximum
240 copy number per TE family from 10 to 2000 (**Fig. 2A**).
241 The observed TE divergence distribution is usually the result of multiple TE family bursts
242 at different times. To simulate this, we can vary the range of mean sequence identity in
243 TEgenomeSimulator. For simplicity, the minimum and maximum copy number of simulated TE
244 families was set to 5 and 100 respectively, and we applied a range of identity varying in
245 distribution sizes (10% - 30%) and sliding between 70% - 100% of mean sequence identity (**Fig.**
246 **2B**). As the identity range increases, the peak of the divergence distribution becomes
247 progressively flattened. This broad range of TE identity indicates continuous TE bursts and can
248 lead to a more gradual and flattened divergence profile.
249 The duration of TE bursts, sometimes an indicator of the lifespan of a TE family, which
250 can be measured by the standard deviation (SD) of sequence identity, also influences
251 divergence distribution. A short-lived TE burst results in a rapid deposition of nearly identical
252 insertions. These insertions then gradually diverge owing to accumulated mutations. In contrast,
253 a long-lasting burst over evolutionary time leads to the continuous deposition of similar TE
254 sequences, causing a less synchronized divergence pattern owing to overlapping degradation
255 timelines. Moreover, ancient TE bursts tend to show greater sequence deterioration due to
256 prolonged exposure to mutational decay. We can simulate genomes that represent different
257 evolutionary phases of TE activity (**Fig. 2C**), offering genomic canvases with distinct base-level
258 TE composition that can serve as different burn-in phases for more advanced simulations. As
259 shown in **Fig. 2C**, the representation of TE burst timeline and duration on the x-axis and y-axis,
260 respectively, allows four distinct TE divergence landscapes to emerge: I) high sequence identity
261 with high SD, modelling recent TE bursts with ongoing or prolonged activity over time; II) low
262 sequence identity with high SD, modeling ancient and prolonged activity, or multiple
263 independent bursts occurring over an extended period; III) low sequence identity with low SD,
264 reflecting ancient TE bursts that were quickly inactivated, with TE activity remaining silenced
265 since the initial event; IV) high sequence identity with low SD, inferring very recent and rapid TE
266 bursts, resulting in a synchronized insertion of highly similar sequences.
267 We modeled these four scenarios using the Custom Genome mode of
268 TEgenomeSimulator. The identity ranges and SD ranges were implemented as shown in **Fig.**
269 **2C**, with identity ranges of 90% to 100% and 70% to 80%, and SD ranges of 15 to 20 and 1 to
270 5. The TE copy number range was fixed between 5 and 1000. The resulting sequence

divergence profiles demonstrated a narrowing of the peak as the SD range decreased from 15–20 to 1–5, indicating more instantaneous bursts (**Fig. 2D**). Furthermore, shifting the identity range from 70–80% to 90–100% led to a leftward shift of the divergence peak (from ~0.25 to ~0.1) and an increase in peak height. These results demonstrate that TEgenomeSimulator is well-suited for modeling TE composition under diverse evolutionary scenarios, which can be considered the base-level scenarios before introducing additional evolutionary TE activity, by supplying the simulated genome as a new input to TEgenomeSimulator or PrinTE.

## Genome Size Evolution Through TE Expansion and Contraction

LTR retrotransposons (LTR-RTs) are the primary contributing TEs to genome size in many plant, fungi, and animal genomes [32]. To isolate the effect of TEs on genome size across temporal gradients, we focused on modeling LTR-RT evolution using PrinTE's fixed-rate method to model predictable patterns of TE expansion and contraction. Here, we simulated two dynamic scenarios subsequently on the same genome. The first scenario used a TE deletion rate fourfold higher than that of insertions. We observed a 45% decrease in genome size after five million generations of forward simulation (**Fig. 3A**). The decrease was initially driven by a swift reduction in the number of intact TEs (**Fig. S5A**). The number of fragmented TEs increased as new TE insertions created nesting and fragmentation of TEs in subsequent generations. The number of intact TEs became limited through generations of purging, and because of the constant deletion rate, we observed an elevated purging of fragmented TEs in later generations (**Fig. S5B**).

Following genome size decrease, we continued the simulation but reversed the TE insertion rate to fourfold higher than that of deletions. The elevated insertion rate caused rapid genome bloating (**Fig. 3A**) powered by an increase in the numbers of both intact and fragmented TEs (**Fig. S5C,D**). The sequence lengths of solo, fragmented, and intact LTR-RTs shifted in response to purging pressures from the host (**Fig. S6**). During periods of genome expansion, TEs (fragmented, intact, or solo) tend to be longer in length.

Because the two LTRs of an intact LTR-RT were identical at transposition, DNA mutations in this region are used to estimate the age of the LTR-RT [33,34]. During simulations of genome size decrease, when the rate of TE deletions exceeds the rate of insertion, we observed unbiased removal of LTR-RTs through generations. Reduced LTR-RT insertions and accumulation of DNA substitutions led to a shift in the distribution of LTR-RT ages across time (**Fig. 3B**). There were two distinct density peaks as simulation proceeds, even though the rate of TE insertions and deletions remains constant (**Fig. 3B**). The smaller peak represents younger LTR-RTs that were inserted recently and have had less time to be purged, and the larger peak that shifts to the right and decreases in density through time was retained from the standing body of LTR-RTs in the ancestral genome. In cases where the rate of TE insertions exceeds deletions, young LTR-RTs quickly proliferate in high abundance and dominate the density distribution of LTR-RTs in the simulated genome, masking the signal of remaining older LTR-RTs (**Fig. 3C**).

When species become reproductively isolated from each other, their TEs are also isolated and evolve independently. The two genomes share TEs that derive from their common ancestor and also contain novel insertions that are unique to the other. With PrinTE, we simulated a speciation event by evolving a genome for a period of time, then diverted the

genome into two isolated evolutionary paths with shifted TE dynamics to create divergent evolutionary trajectories of genome expansion or contraction (**Fig. 3D,E**). In the contracting genome, there are relatively few unique TEs created after speciation (**Fig. 3D; Fig. S7**). In contrast, the expanding genome had a higher abundance of young and unique TEs (**Fig. 3E; Fig. S7**). In the fixed-rate method, the rate of TE insertion and deletion remains relatively stable, only deviating by the stochastic variability introduced by sampling a Poisson distribution, while the number of insertions and deletions introduced scales to the size of the genome (**Fig. 3F**).

Deletion of TEs is mainly achieved by unequal or illegitimate recombination, which both operate through sequence homology between TE sequences. In PrinTE, we allow users to set TE deletion bias exponentially to the TE length using the `-k` parameter. Larger `k` has a stronger bias in deleting longer TEs and thus has more constraints on genome size (**Fig. S8**). When `k` is set to 0, TE deletion has no bias to TE length, allowing genome size growth through the accumulation of long TEs (**Fig. S8**). Illegitimate recombination between LTR regions of individual LTR retrotransposons will partially remove the element and create solo LTRs as remnants. Given a fixed deletion rate, PrinTE can model the proportion of LTR deletions that goes through this process by implementing the "`--solo-rate`" parameter. Increased solo rate can increase the number of solo LTRs in the genome and slow down the purging of LTR retrotransposons, resulting in a larger genome size (**Fig. S9**, **Fig. S10**). By adjusting parameters of deletion length bias and LTR solo rate that interplay with TE insertion and deletion rates, we can create complex scenarios for modeling genome size evolution in different systems.

Recognizing that the rates of TE insertion and deletion can be dynamic through time and may be modified by varying biological processes, we implemented the variable-rate method in PrinTE, which can be used to explore specific molecular features that constrain TE evolution and its impact on genome size. Here, rather than setting fixed TE insertion and deletion rates, we start with an initial rate and allow the tool to dynamically determine subsequent insertion and deletion rates based on the genetic makeup of the most recently simulated genome (**Fig. 3J**). We simulated the genome size and TE evolution of the present-day *Arabidopsis thaliana* and introduced theoretical parameter conditions to forecast one million generations of TE evolution. With the tested conditions, there is an initial period of genome bloating followed by a longer period of TE purging and genome size contraction (**Fig. 3G**). Expectedly, genome size increased through the accumulation of both intact and fragmented TEs and decreased due to the reduction of intact TEs (**Fig. 3H**). The divergence of intact LTR-RTs reflects the genome size expansion and contraction (**Fig. 3I**). These features allow us to simulate TE dynamics close to what is observed in eukaryotic genomes.

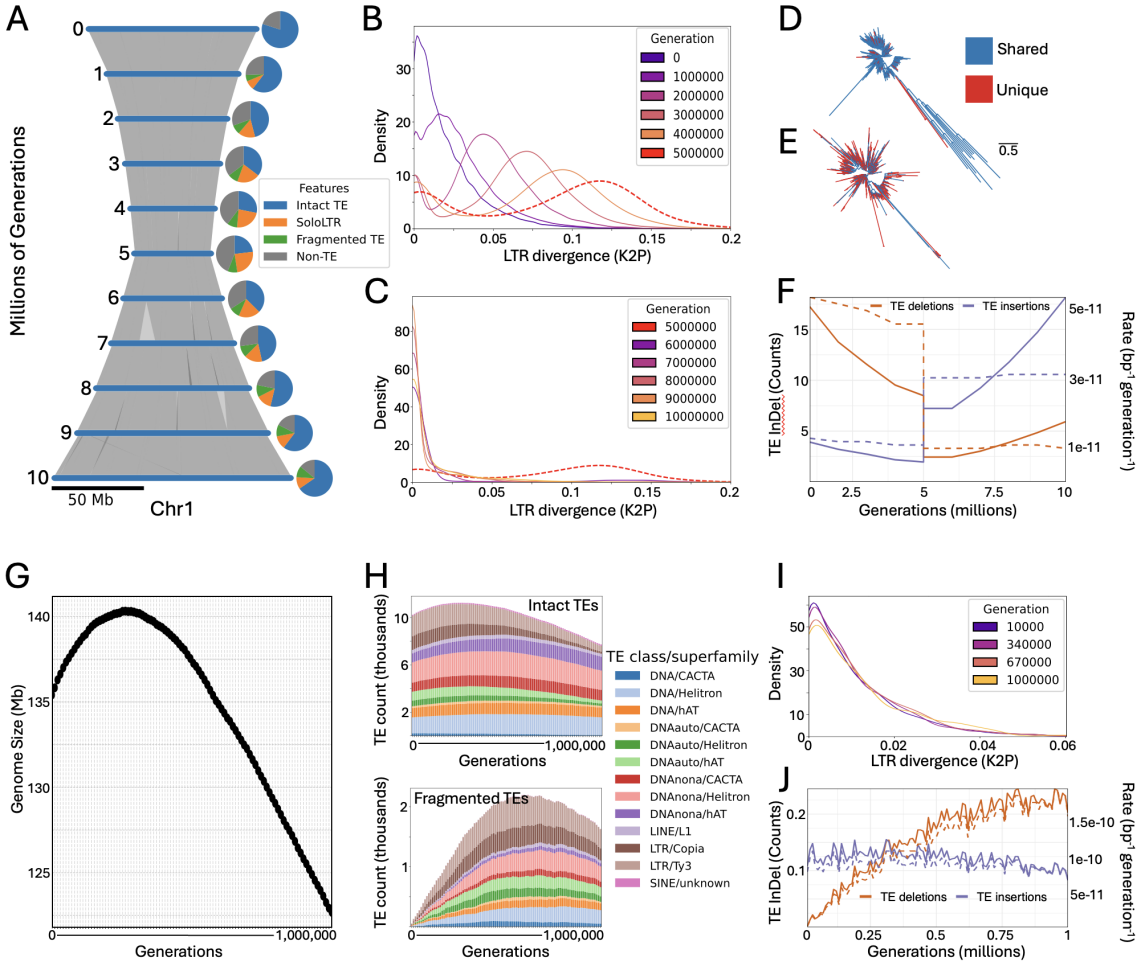


**Fig. 3. PrinTE models TE contribution to genome size evolution using fixed (A-F) and variable (G-J) rate methods.** (**A-F**) Ten million generations of LTR-RT evolution explores their long-term impact on genome size using the fixed rate method in PrinTE. For demonstration, we set the first five million generations of evolution using a rate of TE deletion that exceeds insertion by fourfold. Over the second five million generations, the rates were reversed, and insertion exceeded deletion by fourfold. (**A**) A syntenic ribbon plot illustrates the change in chromosome size. Pie charts corresponding to each time point show the relative contribution of each feature to genome size. (**B,C**) Density distributions of LTR-RT divergence estimated using the Kimura two-parameter (K2P) model during the first five million generations (**B;** genome size decrease) and the second five million generations (**C;** genome size increase). Five million generations is set as a transition time point between genome size decrease and increase and is shown in both plots as dashed red lines. Phylogenetic trees show the evolutionary relationship of LTR-RTs from the *Tekay* family after four million generations of genome size decrease (**D**) and four million geenrations of genome size increase (**E**). (**F**) LTR-RT insertion and deletion counts (solid line) and rates (dashed line) across 10 million generations. (**G,H**) A simulated genome replica of *Arabidopsis thaliana* is forward-evolved one million generations using

PrinTE's variable rates method. (**H**) Intact (top) and fragmented (bottom) TE counts per superfamily across one million generations. (**I**) The distribution of LTR-RT divergence estimated using the K2P model. (**J**) TE insertion and deletion counts (solid line) and rates (dashed line) illustrate the self-adjusting functionality of PrinTE's variable rate method.

## Simulating Digital Replicas of Eukaryotic Genomes

TE composition in real genomes varies substantially across TE families and species because of their complex and diverse evolutionary histories. This variability makes it challenging to simulate realistic genome-wide TE landscapes. To address this, we used TEgenomeSimulator's TE Composition Approximation Mode to model the TE profiles of *A. thaliana* (TAIR10) [35] and the myrtle rust pathogen, *A. psidii* [3]. The sizes of the simulated genomes, 118.2 Mb for *A. thaliana* and 0.98 Gb for *A. psidii*, closely matched the actual genome sizes (**Fig. 4A**). Similarly, the proportion of the genome occupied by TEs closely approximated the observed values in both organisms (**Fig. 4B**). In terms of TE loci and TE bases per superfamily, the simulated TE composition also aligned well with the real data (**Fig. S11**).

We estimated the sequence divergence distribution of each TE family in real genomes for simulations of digital replicas. While the simulated divergence profiles resembled the real data, they were not identical: the fine-scale fluctuations present in real genomes were smoothed because of the use of normal distributions in the simulator. This approach provides a generalized distribution pattern that captures the major trends in TE sequence evolution (**Fig. 4C**).

TE fragmentation patterns also vary widely across TE families and genomes. The Random Synthesized Genome and Custom Genome modes of TEgenomeSimulator use a beta distribution to model TE integrity across all TE families. When alpha and beta parameters are set between 0 and 1, with alpha equal to or smaller than beta (**Fig. S12A**), the distribution takes on a U-shape that approximates general fragmentation patterns previously observed in grapevine data [36]. However, because of the unique TE evolutionary histories of different genomes, this approach may fit well for one species but can overestimate the proportion of TE loci with intermediate degrees of fragmentation in others (**Fig. S12B**). Moreover, fragmentation patterns can differ substantially between TE families, in particular those identified by homology-based methods, which often lack members with high structural integrity (**Fig. S12C**). To improve accuracy, the TE Composition Approximation Mode directly extracts the empirical distribution of TE length ratios (relative to consensus sequences) for each TE family. This enables the simulator to reproduce the integrity distributions more authentically in the resulting simulated genomes (**Fig. 4D**).

These digital replicas of real genomes provide a robust foundation for simulating future TE-related events, such as new TE bloating or purging episodes, effectively serving as burn-in phases for modelling genome evolution from the current state.

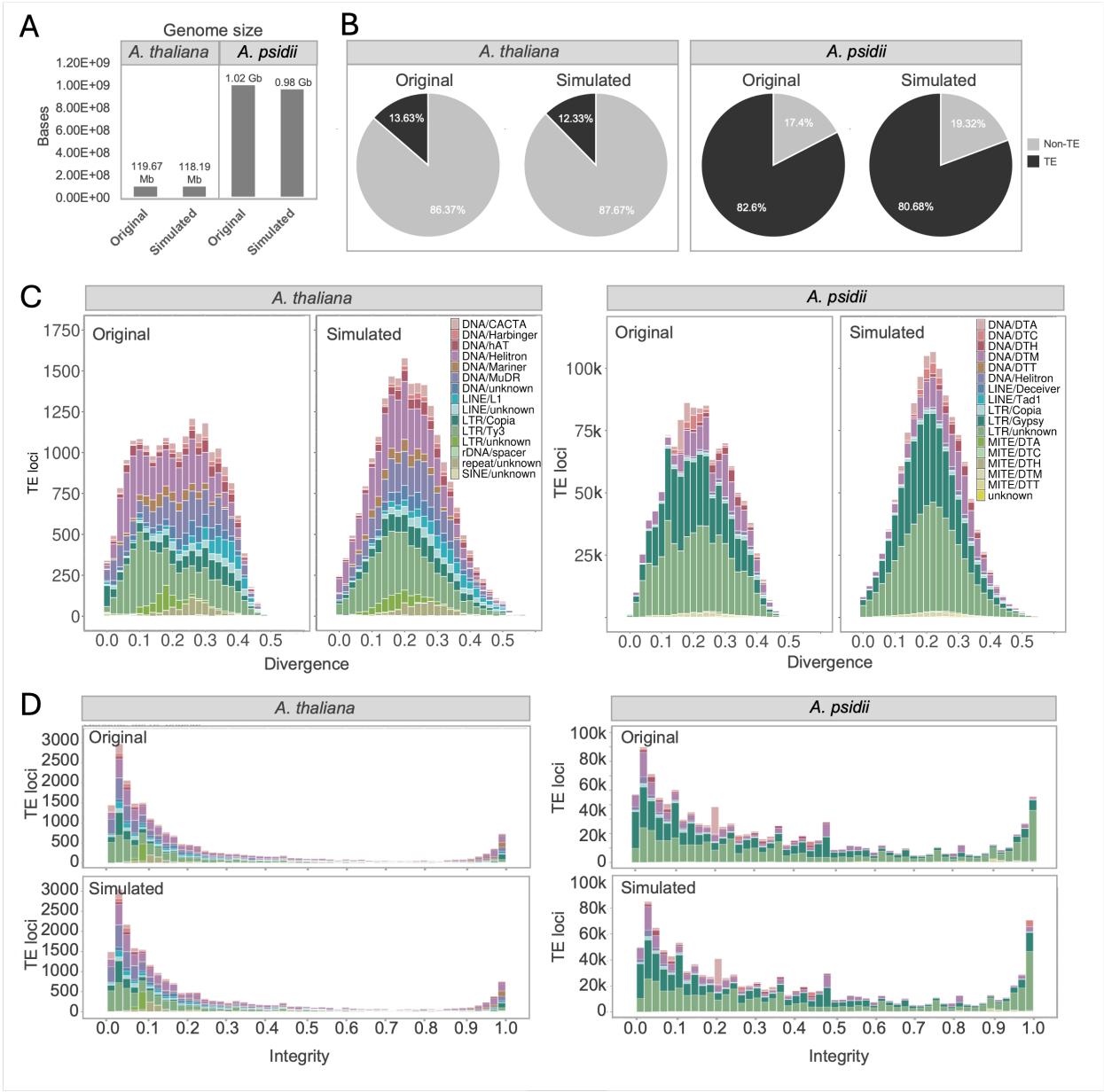


**Fig. 4**. **TE Composition Approximation for the *Arabidopsis thaliana* and *Austropuccinia psidii* genomes.** (**A**) Size of the original and simulated genomes. (**B**) The proportion of original and simulated genomes occupied by TE and non-TE bases. (**C**) TE sequence diversity distribution colored by TE superfamily. (**D**) The distribution of TE sequence integrity, which is the length ratio relative to the consensus sequence, is colored by TE superfamily as the same in (**C**).

## Genome Size Evolution of Rust Fungus

We further studied the impact of TE dynamics on genome size evolution using rust fungal genomes for their wide range of genome sizes. We downloaded 58 *Pucciniomycotina* fungal

genomes from NCBI and JGI MycoCosm [37] to investigate the expansion and contraction of genome size (**Table S3**). *Pucciniomycotina* is a subdivision of fungus that contains 10 classes and over 8400 species [38]. We identified and removed alternative haplotypes from these genomes using Redundans [39]. The resulting haploid genome size varies from 13.6 Mb (*Mixia osmundae*) to 1018.1 Mb (*Austropuccinia psidii*), representing over 70-fold differences (**Table S4**). We performed Benchmarking Universal Single-Copy Orthologs (BUSCOs) evaluations using compleasm [40] and the odb12 basidiomycota database. The complete BUSCOs (single and duplicated) averaged 90.7% (SD = 5.4) and contig N50 averaged 846 kbp (ranging from 111 kbp to 3.8 Mb), suggesting high assembly quality of haploid genomes (**Table S4**).

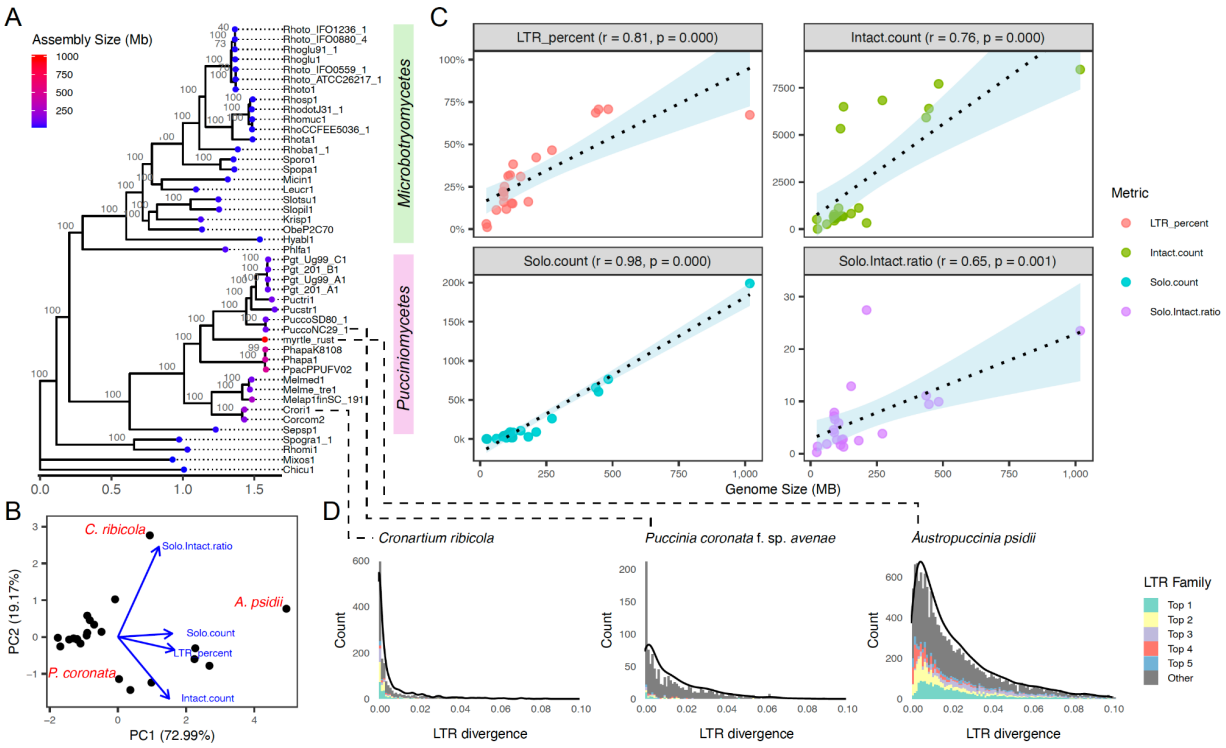


**Fig. 5.** (**A**) The phylogeny of *Pucciniomycotina* fungal genomes used in this study. Protein sequences of shared single-copy BUSCO were used for tree inference with 1000 bootstrap replications. Bootstrap support percentage is shown on each branch. The two classes, *Pucciniomycotina* and *Microbotryomycetes,* are labeled beside the phylogeny. The genome size of each species is indicated by color dots. (**B**) Principal Component Analysis (PCA) on LTR metrics in fungus genomes. Each dot represents one genome. PCA loadings of each metric are shown. Genomes selected for further studies are labeled. (**C**) Pearson's correlation (*r*) between genome size and LTR sequence content (LTR_percent), intact LTR-RT count (Intact.count), solo LTR count (Solo.count), and solo:intact ratio. Each dot represents one genome. The correlation *r* and p-value are shown on the panel title. (**D**) Divergence of intact LTR retrotransposons in *Cronartium ribicola* (white pine blister rust), *Puccinia coronata* f. sp. *avenae* (oat crown rust), and *Austropuccinia psidii* (myrtle rust) genomes. The top five most abundant LTR families in each genome are colored, and they are not shared between genomes. A density line is shown as the contour in each plot. The locations of these species on the phylogeny are indicated by dotted lines.

We identified the shared single-copy BUSCOs and used iqtree [41] to construct the phylogeny of *Pucciniomycotina*. The resulting phylogeny was over-represented by two classes, *Microbotryomycetes* and *Pucciniomycetes* (**Fig. 5A**). *Microbotryomycetes* genomes of *Microbotryomycetes* are small, ranging from 17.5 Mb to 50.3 Mb. Correspondingly, on average, only 0.4% of the genome was represented by LTR sequences with a range of 0% to 3.0% (**Table S4**). This observation suggests that the best strategy for controlling genome size is to avoid the invasion of LTR retrotransposons. *Pucciniomycetes* have widely distributed genome sizes, ranging from 26.8 Mb to 1018.1 Mb, with an LTR sequence content corresponding to 0.4% to 67.4% (**Table S4**).

To further study the dynamic of LTR retrotransposons in these fungal genomes, we *de novo* identified intact LTR elements and solo LTR sequences, and calculated whole-genome solo:intact ratio using LTR_retriever [33]. By using genomes with LTR sequence content larger than 1% (n = 22), we conducted Principal Component Analysis (PCA) on LTR metrics and genome size, showing clustering of smaller genomes and larger genomes as outliers, including *Cronartium ribicola* (white pine blister rust), *Puccinia coronata* f. sp. *avenae* (oat crown rust), and *Austropuccinia psidii* (myrtle rust; **Fig. 5B**). PCA loadings suggest that solo LTR number, intact LTR number, and their ratio were all drivers to the outliers of large genomes (**Fig. 5B**). Including total LTR content, these LTR metrics showed strong positive correlations with genome size (Pearson's $r \geq 0.65$, $p \leq 0.001$; **Fig. 5C**).

Not surprisingly, intact LTR elements as the product of LTR proliferation were strongly correlated with genome size ($r = 0.76$; **Fig. 5C**). However, solo LTR sequences as the product of LTR removal [33] showed an even stronger correlation, almost linearly correlated with genome size ($r = 0.98$; **Fig. 5C**). Another counter-intuitive observation is the positive correlation between genome size and solo:intact LTR ratio, which is a metric used to measure the intensity of LTR purging [34]. In the traditional understanding, the higher the solo:intact LTR ratio, the more intense purging of LTR sequences, and thus the smaller the genome [42]. The positive correlation between solo:intact ratio and genome size suggests that illegitimate recombination may not be an effective way of escaping genome obesity.

We selected *C. ribicola*, *P. coronata* f. sp. *avenae*, and *A. psidii* as representatives of three distinct roadmaps to genome size evolution. *C. ribicola* has a genome size of 211.6 Mb with the LTR divergence distribution showing an extremely quick turnaround time for LTR element insertions (**Fig. 5D; Fig. S13**). The solo:intact ratio of *C. ribicola* is 27.4 (**Table S4**), the highest among all genomes, suggesting a quick insertion–quick removal mode of action, and a relatively stable or slowly increasing genome size. *P. coronata* f. sp. *avenae* has a genome size of 113.2 Mb and a solo:intact ratio of 1.7 (**Table S4**). The LTR divergence distribution showed a much slower turnaround time for LTR insertions (**Fig. 5D; Fig. S13**), suggesting a slow insertion–slow removal mode of action, and a relatively stable genome size. *A. psidii* has an extremely large genome size of 1018.1 Mb and a solo:intact ratio of 23.5, the second highest ratio among our data (**Table S4**). However, the LTR divergence of *A. psidii* shows the peak of LTR insertion passed the current time (**Fig. 5D; Fig. S13**), suggesting a slowdown of insertion activities [8] and a decreasing genome size.

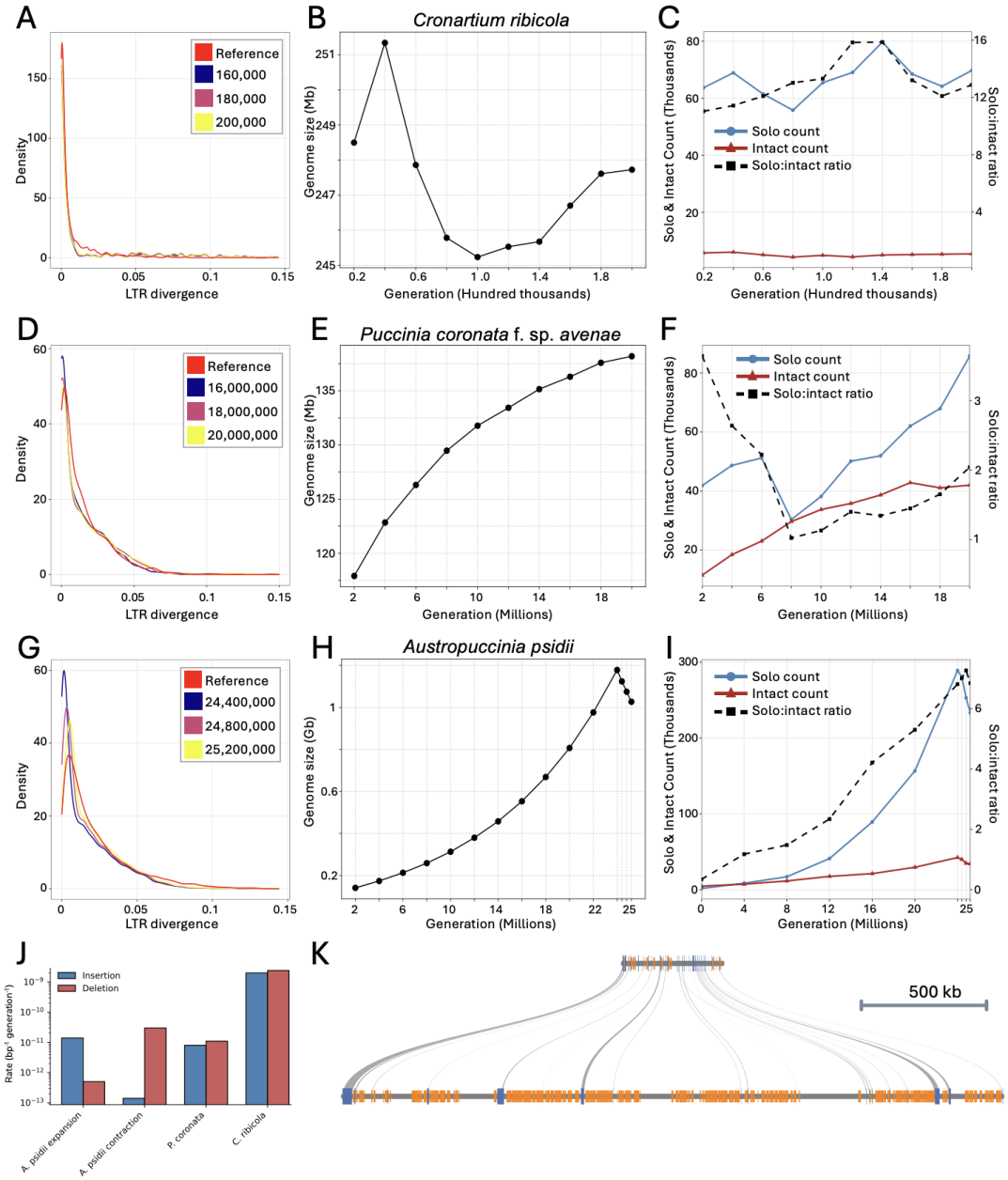


**Fig. 6. Simulating genome size evolution in three *Pucciniomycetes* genomes.** (**A-C**) *Cronartium ribicola* (white pine blister rust), (**D-F**) *Puccinia coronata* f. sp. *avenae* (oat crown rust), and (**G-I**) *Austropuccinia psidii* (myrtle rust). (**A, D, G**) Sequence divergence compares the empirical divergence distribution of LTR-RTs (Reference) to simulated genomes. LTR divergence was estimated using LTR_retriever. Three timepoints were plotted for *C. ribicola* (**A**) and *P. coronata* f. sp. *avenae* (**D**) to showcase the relatively stable TE dynamic. The three timepoints of *A. psidii* (**G**) highlight the rapid purging of LTR-RTs during the period of genome size contraction. (**B, E, H**) Genome size changes across generations. (**C, F, I**) The empirical solo:intact ratio estimated using LTR_retriever. (**J**) The estimated LTR insertion and deletion rates of fungal genomes. The expansion and contraction phase of the *A. psidii* simulation was

493 plotted separately. (**K**) Microsynteny between the simulated *P. coronata* f. sp. *avenae* before
494 (top) and after (bottom) the TE evolution that led to genome size increase in the simulated *A.*
495 *psidii* genome. Collinear genes are blue, and LTR-RTs are orange.

## 496 Accessing Genome Evolution with Forward Simulations of LTR Activities

497 Here, we demonstrate the utility of TESS in modeling genome size evolution under the impact of
498 TE bloat and purge. Despite their shared ancestry (**Fig. 5A**), the three selected fungal species
499 (*C. ribicola*, *P. coronata* f. sp. *avenae*, and *A. psidii* ) show immense variation in genome size
500 and LTR-RT evolution. To test our hypothesis that *C. ribicola* has a stable genome size that it
501 achieved by rapid but balanced LTR-RT insertion and deletion, we used TEgenomeSimulator to
502 create a digital replica of the *C. ribicola* genome and forward evolved it using PrinTE with a
503 variety of insertion and deletion rates in search of the observed LTR-RT divergence distribution.
504 At each trial, we compared the LTR-RT age distribution with that of the real genome (**Fig. S14)**.
505 We found that an LTR-RT insertion rate of $2\times10^{-9}$ per bp per generation and a deletion rate of
506 $2.4\times10^{-9}$ per bp per generation provided a simulated genome that closely matched the *C.*
507 *ribicola* genome with the genome size increasing moderately through time (**Fig. 6A,J**). Forward-
508 projecting 200,000 generations of TE evolution indicated that the density distribution of LTR-RT
509 ages in the simulation closely matched the distribution of the real genome (**Fig. 6A**, **Fig. S14**).
510 The solo:intact ratio of the simulated genome was 13 (**Fig. 6C**) in comparison to 27.4 in the real
511 genome (**Table S4**). Although the forward-simulated genome does not share an identical LTR-
512 RT footprint to that of *C. ribicola*, its proximity to the real data suggests that PrinTE can
513 effectively infer TE dynamics through forward simulations. Simulation of the *C. ribicola* genome
514 suggests that the species might have inserted and deleted (either intact or fragmented) one
515 effective LTR-RT every two generations during the latest 200,000 generations.
516 We repeated the same process using a digital replica of *P. coronata* f. sp. *avenae*a
517 simulated by TEgenomeSimulator to test our hypothesis that the species has slow TE insertion
518 and slow removal, leading to a relatively stable genome size. An LTR-RT insertion rate of $8\times10^{-}$
519 $^{12}$ per bp per generation and a deletion rate of $1.10\times10^{-11}$ per bp per generation were found to be
520 optimally adapted for maintaining the LTR-RT characteristics of the real genome for over 20
521 million generations (**Fig. 6B,J; Fig. S15**). Dating LTR-RTs in forward-simulated and the real
522 genome showed that the distributions were nearly identical (**Fig. 6D**). The solo:intact ratio in the
523 forward-simulated genome mimics that of the real genome, at 2.0 and 1.7, respectively (**Fig. 6F**,
524 **Table S4**). With these rates, we see a gradual increase in average genome size, adding
525 approximately 1.25 Mb per million generations (**Fig. 6E**). These results suggest that, in contrast
526 to *C. ribicola, P. coronata* f. sp. *avenae*a appears to have a mostly inactive LTR-RT footprint
527 with an average of 1090 generations between each effective LTR-RT insertion and an effective
528 deletion every 790 generations.
529 For our final simulation, we began with a relatively small genome that is typical in the
530 *Pucciniomycetes,* and simulated the TE evolution that led to the enormous genome bloating of
531 *A. psidii*. Based on the genome size and LTR-RT divergence distribution (**Fig. 5A,D**), we
532 expected that TE insertions exceeded deletions for a long period of time, leading to genome
533 size increase, followed by a shorter period of rapid LTR-RT deletions. Through evaluation of
534 various dynamic parameters, we found that an insertion rate of $1.4\times10^{-11}$ per bp per generation
535 and a deletion rate of $5\times10^{-13}$ per bp per generation for 24 million generations was optimally

536 tuned to mimic the observed LTR-RT age distribution of the *A. psidii* genome during the long
537 period of genome size increase (**Fig. 6G,J**; **Fig. S16A,B**). During this period, we aimed to
538 achieve a genome size that was marginally larger than the real genome so that we could tune
539 up the rate of deletion to reflect the more recent decline in young LTR-RTs observed in *A. psidii*
540 (**Fig. 5D**). At peak LTR-RT invasion, the simulated genome was 1,178 Mb in size (**Fig. 6H**).
541 The drop of recent LTR-RT insertions in *A. psidii* suggests a reduced insertion rate and
542 an increased deletion rate. Using a substitution rate of $2.0\times10^{-9}$ per bp per generation [43], we
543 estimated that the insertion reduction phase of *A. psidii* had lasted for 1.2 million generations.
544 We then forward-simulated an additional 1.2 million generations using elevated deletion rates.
545 Systematic evaluation of TE dynamic parameters revealed that an insertion rate of $1.4\times10^{-13}$ per
546 bp per generation and a deletion rate of $3\times10^{-11}$ per bp per generation can achieve a genome
547 size that is near *A. psidii*, at 1,027 Mb (**Fig. 6H,J**; **Fig. S16C**), and the LTR-RT divergence
548 distribution mimics the real genome (**Fig. 6G**; **Fig. S16D**). The solo:intact ratio of the simulated
549 *A. psidii* genome terminated at 6.8 (**Fig. 6I**), which is lower than the observed value at 23.5
550 (**Table S4**), suggesting other evolutionary processes may play roles in shaping the modern-day
551 *A. psidii* genome.
552 Simulation of *A. psidii* suggests two major phases that contributed to the present-day
553 genome size: a long period of genome size increase in which LTR-RT insertions exceeded
554 deletions by 28-fold and a short period where deletions exceeded insertions by 214-fold. Our
555 results support a recent contraction model for genome size evolution in *A. psidii* (**Fig. 6J**). It is
556 important to note that the estimated parameters are effective insertion and deletion rates
557 averaged from a long period of evolutionary time, which may not capture short periods of TE
558 expansion and contraction.
559 Because these three species were selected to represent the extremes of LTR-RT
560 dynamics in the sampled *Pucciniomycetes* genomes, we suggest that the amplification and
561 degradation of LTR-RT can vary widely within the fungal kingdom. In future studies, it will be
562 interesting to see if such a wide variety of LTR-RT dynamics exists at narrower phylogenetic
563 levels.

# 564 Discussion

565 The TE composition profile of a genome serves as an evolutionary archive, preserving
566 signatures of TE dynamics and host genome responses that collectively shape genome
567 evolution. Deciphering these patterns from empirical data is challenging. Whether to test
568 evolutionary hypotheses or to predict future trajectories, simulation is often necessary. To
569 address this, we built TESS, a simulation framework composed of RandSeqInsert,
570 TEgenomeSimulator, and PrinTE. RandSeqInsert is a light-weighted simulator that excels in
571 insertions of TIR DNA transposons with the AVL tree architecture. TEgenomeSimulator and
572 PrinTE serve as the two central pillars of TESS, each serving distinct but complementary roles.
573 Used together, they significantly expand the capacity for simulating complex TE dynamics and
574 testing evolutionary hypotheses.

## TEgenomeSimulator enables flexible modeling of TE composition and evolutionary dynamics

TEgenomeSimulator is a versatile tool for simulating TE composition across diverse evolutionary trajectories, with strong potential for broader applications. It encapsulates the dynamic balance between TE insertions and host-mediated TE removal through its TE copy number parameter, which directly influences the final size of the simulated genome (**Fig. 2A**). It can also simulate genomes with different histories of TE activities by varying the ranges of mean sequence density and its standard deviation representing the duration of the TE burst, allowing flexible modelling of different evolutionary scenarios (**Fig. 2B–D**).

Simulating TE compositions to closely resemble real genomes is a rare capability among existing tools. TEgenomeSimulator stands out by enabling realistic modelling of TE content based on empirical patterns (**Fig. 4**). It integrates three major features of TE composition, a) TE family-specific copy number, b) sequence identity (modeled using normal distribution), and c) TE integrity profiles, into a coherent mutagenesis framework. These features allow the replication of the evolutionary patterns observed in real genomes with additional critical information, enabling a more precise manipulation of TE compositions for hypothesis testing.

In comparison, GARLIC [28] is a genome-wide simulation program prioritized for benchmarking purposes. It begins by extracting coding and non-coding genes and repeats from a user-provided genome, then profiles repeat diversity and analyzes the intergenic space using k-mer and GC content profiles to model genomic context. To simulate TEs, GARLIC first mutates the consensus repeat sequences, then inserts them into a non-repetitive genomic background.

While GARLIC provides more controls in simulating genes and intergenic architecture, TEgenomeSimulator remains more TE-centric, offering flexible, user-driven control of TE-specific parameters without extensive genome preprocessing. This streamlined design lowers the barrier for simulating TE dynamics under various evolutionary assumptions. As such, GARLIC is well-suited for studies focused on genes and their annotations, whereas TEgenomeSimulator is ideal for hypothesis-driven investigations into TE-specific evolutionary scenarios with customizable control over TE dynamic parameters.

## PrinTE allows forward simulation on TE dynamics

Transposable elements are a major genomic component and subject to all evolutionary forces, including purifying selection, drift, and mutation. The ability to trace TEs' responses to these events will greatly facilitate the studies of TE dynamics. Forward simulation is a key approach that allows the artificial evolution of TEs and other genomic features, which provides opportunities to sample the intermediate state of the evolutionary process.

SLiM is a powerful forward simulation program that can simulate the majority of evolutionary scenarios [23]. However, its operation on single nucleotide variants limits the application on TEs with varying lengths (**Table S2**). SLiM provides a work-around, by first compressing the whole TE locus into a presence and absence unit and performing simulations on these units. Sequence content is then filled back to replace the units to generate genomic sequences. This approach cannot model the variation accumulated through generations on the simulated TE sequences, including the deletion and formation of solo LTRs, which are critical

617 for genome size studies. On the contrary, PrinTE operates on whole-genome sequences,
618 including invariant sites that allow intuitive modeling of complex scenarios and interpretation of
619 results. In particular, PrinTE models the insertion and deletion rates of TEs and generates solo
620 LTRs, which empowers the study of genome size evolution driven by LTR bursts and purges.
621 The input files of PrinTE are highly flexible (**Fig. 1**). It has a built-in engine to simulate
622 highly customizable starting genomes, in which users can specify genome size, TE content, TE
623 superfamily abundance, and TE divergence in their simulations. PrinTE integrates CDS
624 sequences in the simulation, allowing for modeling purifying selections, tracing large structural
625 variants, and studying the evolution of gene space. PrinTE also takes the simulated genomes
626 from TEgenomeSimulator, allowing forward simulation from different diverse backgrounds of
627 genomes. Most importantly, PrinTE can also take its own output as input, allowing for simulation
628 of varying TE dynamics and studying the host-TE arms race. This feature also enables the
629 simulation of genomic divergence and speciation driven by differences in TE activities.

## 630 TE contributes to genome size expansion

631 The contribution of TE contents to genome size expansion has been confirmed across many
632 taxa and in our data (**Fig. 5C**). Over time, genomic data have come faster and of a higher
633 quality and yet, as noted by Osmanski *et al.* (2023) [44], the role of TEs in genome architecture
634 and evolution is often an afterthought. More in-depth analyses become reliant on domain
635 specialists for a meaningful evolutionary perspective.
636 In order to put forward mechanistic hypotheses, authors have developed specialized
637 computational methods to detect signals of TE radiation and impacts on speciation and genome
638 size. For example, Ricci *et al.* (2018) [13] developed a measure of TE Density of Insertion and
639 Relative Rate of Speciation to model TE radiation bursts and silencing. Kapusta *et al.* (2017) [12]
640 estimated TE-driven DNA gains and losses and proposed the 'accordion' model of TE radiation
641 followed by DNA loss for tight genome size maintenance in mammals and birds. Osmanski *et al.*
642 (2023) [44] applied a Redundancy Analysis to study TE composition across mammalian taxa
643 generated through the Zoonomia genomes [45], and found evidence of multiple recent TE
644 expansions driving larger genome sizes and wide-spreading TE-silencing events across the
645 mammalian tree. These authors hypothesize that variations in TE defense mechanisms may
646 have played a role in whether specific lineages were overcome by a TE burst at a given
647 evolutionary time.
648 However, we have limited knowledge about the maintenance of compact genomes
649 despite evidence of TE radiation bursts and the maintenance and regulation of extremely large
650 genomes. The Chinese pine attributes 69.4% of its 25.4 Gb genome to TEs, 86.5% of which are
651 made up of LTR retrotransposons [15]. The Chinese pine genome showed no sign of recent
652 whole-genome duplication, but 91.2% of its genes were undergoing discrete duplication or
653 family expansion, strongly associated with TEs. Our results revealed two different genome
654 dynamics in the *Pucciniomycotina* subdivision of fungi, including the *Microbotryomycetes* class
655 that harbours compact genome size (18-50 Mb) and the *Pucciniomycetes* class with drastic
656 differences in genome size (27-1018 Mb) (**Fig. 5A**; **Table S4**). The *Microbotryomycetes* class is
657 almost devoid of LTR sequences (**Table S4**), suggesting that the most effective way of
658 maintaining compact genome sizes is to be immune to the invasion of LTR retrotransposons.
659 Genomes in the *Pucciniomycetes* class have varying richness of LTR retrotransposons (**Table**

**S4**), and the PCA reveals several types of LTR dynamics (**Fig. 5B**) that are likely major drivers of genome size variation.

The myrtle rust fungus expanded its genome size from an approximate ancestral size of 110 Mb to the current 1,018 Mb, experiencing a ~10 times increase. Our simulation studies revealed that the genome is experiencing a downsizing effect, with LTR deletions faster than insertions (**Fig. 6**). Using PrinTE's built-in genome simulation utility, we were able to simulate a genome that mimic the TE landscape of the ancestral genome of myrtle rust, and further used PrinTE to simulate two phases of evolution, including a long expansion period and a short contraction period. This model fits well with the observed composition of the myrtle rust genome, suggesting a close approximation of its trajectory of genome size evolution.

We recognize that TE activity is tightly linked to host genome regulation, particularly when it impacts gene function and organismal fitness. It is also important to account for more complex patterns in sequence identity distributions, which may go beyond simple normal distributions. In our study, TEgenomeSimulator helped to shape TE composition and sequence divergence under various evolutionary scenarios, while PrinTE simulated the real-time mobilization and insertion behavior of TEs. Crucially, the combined use of these two tools greatly enhanced the simulation capacity, enabling richer and more flexible hypothesis testing across both evolutionary and mechanistic contexts. Future development would benefit from incorporating features such as customizable identity distributions, more functional genomic context, epigenetic regulation mechanisms, and insertion site preferences.

# Conclusions

In conclusion, the TE Evolution Simulation Suite (TESS) is a versatile collection of tools that allows flexible forward simulation of whole-genome sequences with fine-grain user control of genome size, TE content, superfamily abundance, TE divergence profile, TE insertion, TE deletion, TE mutation, solo LTR formation, genome divergence, and speciation. We demonstrated the use of TESS in resolving the genome size evolution of *Pucciniomycotina* genomes. In particular, we simulated the genome size evolution of the myrtle rust (*Austropuccinia psidii*) and other large fungal genomes. Our work transforms the study of genome size evolution into a highly traceable and explorative simulation framework, and allows testing of data-driven hypotheses.

# Methods

## Implementation of RandSeqInsert

RandSeqInsert employs an effective method for randomly inserting genomic fragments into sequences (either user-provided or randomly generated). The implementation is built around a balanced binary tree data structure that enables efficient sequence manipulation while maintaining performance even for long sequences. Curated rice and maize TIR transposon libraries are included with the tool to serve as donor sequences during the simulation.

***Balanced Sequence Tree***

The backbone of RandSeqInsert is the "SequenceNode" class, which implements an AVL tree for sequence representation. An AVL tree is a balanced binary tree where the height difference between left and right branches of any node never exceeds 1 [46]. When this balance is disrupted during insertions, automatic tree rotations help to restore the balance.

The tree structure represents sequences as connected nodes that can be rearranged without extensive copying. Each node stores the unique identifier (UID), the actual sequence fragment (these can be non-TE backbones or inserted TEs), length data, total length of all sequences in its subtree (essential for position mapping), height of the node, 5' and 3' TSD sequences, and additional information. After each insertion operation, the balance factor is recalculated at each node along the path from the insertion site to the root, and rotations are performed automatically. Eventually, the complete sequence can be retrieved through an in-order traversal of the tree, which is a depth-first search algorithm that visits nodes in a binary tree following the Left-Root-Right order.

This structure offers major advantages over alternative approaches. A simple direct sequence implementation would require rebuilding the entire structure after each operation, resulting in $O(n)$ time complexity per operation and $O(n^2)$ complexity for $n$ operations. Using regular binary tree would risk performance degradation to $O(n)$ per operation when insertions occur in patterns (such as sorted order) that transform the tree structure into essentially a linked list (which is a linear data structure where each node contains both data and a reference to the next node), eliminating the logarithmic performance advantage. The AVL tree maintains $O(log\ n)$ per-operation time across all operation patterns through a self-balancing mechanism via rotational operations [46].

***Event Journal for Nested Insertions***

RandSeqInsert implements a sophisticated event sourcing architecture through the “NestingEventJournal” class to track complex nested insertion scenarios. This journal selectively records only nested insertions (TE-to-TE insertions), captures critical information for each nested insertion event, including the donor node UID, target node UID, left and right fragment node UIDs, and TSD information. This selective recording strategy enables the reconstruction of donor sequences that have been fragmented by subsequent insertions.

***Insertion Algorithms***

RandSeqInsert offers both recursive and iterative insertion algorithms, with the iterative method working better for large sequences by avoiding recursion limits. The insertion process identifies the appropriate position in the tree based on absolute sequence position, splits the target node if insertion occurs within it, creates new nodes for the inserted donor sequence and split fragments, and updates balance information while rebalancing the tree.

Position mapping is crucial to the algorithm. Given an absolute position in the full sequence, the algorithm determines which node contains this position, the relative position within that node, and how to split and reorganize nodes to incorporate the new sequence. When insertion occurs within a node, that node is split into three nodes: a left node containing the sequence before the insertion site, the current node repurposed for the inserted donor sequence, and a right node containing the sequence after the insertion site. When a donor sequence is inserted into another donor sequence (nested insertion), the system automatically records this event, preserving the original target sequence information for future reconstruction.

***Target Site Duplication (TSD) Modeling***

The implementation includes a detailed modeling of Target Site Duplications, a biological process where short sequences at insertion sites are duplicated during TE integration. The TSD generation process takes the source TSD sequence at the insertion site, creates both 5' and 3' TSD sequences with configurable lengths, and applies mutations to simulate biological variation according to configurable probabilities. TSD sequences are integrated into the node structure and are handled separately from the core donor sequence data.

Two types of mutations are supported: SNP mutations and InDel mutations. SNP mutations replace a single nucleotide at a random position with a different randomly selected base. InDel mutations represent a single random nucleotide insertion or deletion.

***Donor Sequence Tracking***

RandSeqInsert effectively tracks and records information about inserted donor sequences and their positions. The system identifies donor nodes and calculates their exact positions in the final sequence through a complete tree traversal. Following this process, it generates a FASTA file that records each donor sequence with its precise start and end coordinates, and a BED file that is compliant with PrinTE.

***Sequence Tree Visualization***

For research and analysis purposes, RandSeqInsert can generate two visualizations in Graphviz DOT format [47]. The sequence tree visualization displays the hierarchical structure of the AVL tree, with each node color-coded by type and annotated with position, length, and UID information. The event graph visualization demonstrates the insertion event relationships, showing the impacts of the nested insertions on sequence integrity. This dual visualization approach enables researchers to understand both the structural organization of the sequence data and the history of insertion events.

## Implementation of TEgenomeSimulator

TEgenomeSimulator was built on denovoTE-eval [30] and further developed to incorporate new functionalities. It comprises three modes: Random Synthesized Genome Mode (`-m 0 or --mode 0`), Custom Genome Mode (`-m 1 or --mode 1`), and TE Composition Approximation mode (`-m 2 or --mode 2`). These modes all perform random TE insertion and nested TE insertion sequentially. Depending on the mode chosen, the genome and TE sequences undergo distinct preparations prior to TE insertion. The implementation and input requirements for each mode are described in the following section.

***The Three Simulation Modes***

The Random Synthesized Genome mode will synthesize chromosome sequences using a set of user-defined parameters, including chromosome length and GC content. A TE library can be provided using the '`--repeat`' flag, which is used to simulate a population of TE sequences through random mutation, including nucleotide substitutions, 1-base InDel, target site duplication, and fragmentation. Simulated TE sequences are then randomly inserted into the synthesized chromosome sequences.

The Custom Genome mode requires users to provide a genome in which TE sequences have been removed. For convenience, this mode provides an option for users to generate such a genome using the '`--to_mask`' flag. RepeatMasker [48] is used to mask and remove repeats from the genome based on a TE library supplied by users. TE sequence mutation and random

nested insertions are introduced in the same way as the first mode. Users must also specify a second TE library with '`--repeat2`', which is used for TE masking and removal. Note that the simulator allows users to provide two different TE libraries for masking repeats and for TE sequence mutation, respectively.

The TE Composition Approximation mode utilizes the '`--to_mask`' flag implemented in the Custom Genome Mode, with the option of using two different TE libraries. By default, RepeatMasker output is utilized in the configuration of downstream TE mutations. In brief, the simulated TE composition is modeled on that of the original genome in terms of nucleotide substitution rate, InDel rate, integrity, and the absolute and relative abundance of TE families. The TE insertions and nested insertions in this mode are independent of the TE distribution in the original genome.

***Simulating TE Mutagenesis***

TE libraries provided by the user must adhere to the RepeatMasker recommended format (">repeat_family_name#class/subclass"), e.g, ">ATCOPIA10#LTR/Copia". It is recommended that the full sequence of each TE family be stored in the TE library file. For example, the LTR and internal domain sequences of an LTR retrotransposon are stitched in the conjunction form, LTR-INT-LTR, rather than stored separately. TEgenomeSimulator will check for split sequences of LTR retrotransposons in the user-provided library and assemble full sequences for the user.

TEgenomeSimulator extracts classification information from the TE library for incorporation into a final GFF file. Metadata, including TE family name, superfamily, subclass, and sequence length, are also extracted from the TE library headers and compiled into the output file 'TElib_sim_list.table'. This table also defines the mutagenesis parameters for each TE family, including TE family copy number, mean sequence identity ($I$), standard deviation of mean identity, and the proportion of sequence divergence ($D = 100 - I$) contributed by InDel ($i$). The remaining proportion of divergence is attributed to nucleotide substitution ($s$), such that $i + s = 100$. The table also includes values for target site duplication (TSD) length and proportions of fragmented TE loci and nested TE insertions, each relative to the total TE copy number. The assignment of copy number, sequence identity, and sequence integrity are further described below.

***Copy number assignment.*** In both Random Synthesized Genome and Custom Genome modes, the user is required to provide a value range for the random assignment of a copy number to each TE family. The TE Composition Approximation mode, however, uses the copy number information of each family obtained from RepeatMasker as the exact copy number value.

***Sequence identity***. Both Random Synthesised and Custom Genome modes assign sample sequence identity values from predefined normal distributions. The mean value and the standard deviation of the mean are, by default, randomly selected from ranges between 80 to 95 (%) and 1 to 20 (%), respectively. For the Genome Approximation mode, the mean and standard deviation of sequence identity distributions are derived from the RepeatMasker output for each TE family member. For each TE insertion, an integer is then randomly drawn from the generated distribution, regarded as the original identity of the family sequence. This value is further adjusted, following the approach implemented in denovoTE-eval [37], by adding 50% of the difference between the sampled identity value and 100. It is calculated as $I' = I + (100 - I)*0.5$, where $I$ is the original identity sampled from the normal distribution, and $I'$ is the adjusted

identity value. This adjustment causes a skew of the identity distribution to the right (i.e., closer to 100%), reflecting that TE sequences with high divergence from the family sequence are less likely to be recognized as family members.

***Target Site Duplication (TSD)***. TSD length is implemented based on literature findings [49], while the sequence of TSD is randomly synthesized. The simulator then randomly assigns the TE member to forward or reverse strands before performing sequence fragmentation and integration.

***Sequence integrity***. Both Random Synthesized Genome and Custom Genome modes implement a beta distribution to model the integrity of TE sequences. The default alpha and beta values are 0.5 and 0.7, respectively, to create an asymmetric U-shape distribution in which the probability of low sequence integrity (e.g., <10%) is higher than that of high integrity (e.g.,>90%). Alpha and beta values can be adjusted using the `--alpha` (or `-a`) and `--beta` (or `-b`) parameters. The integrity value for a TE member is sampled from the established beta distribution and then converted to a number representing the bases to be removed from the 5' end of that sequence. For example, if a TE member is 1 kbp at its full length and is assigned with 70% integrity, the simulator would remove 300 bases, which accounts for 30% of the intact length, from its 5' end**.** By default, only 0.1% of TE copies in each TE family remain untouched, ensuring the representation of intact TEs in the simulation. For sequence integrity in the TE Composition Approximation mode, TEgenomeSimulator utilizes a randomly sampled value from a sequence integrity distribution extracted from the RepeatMasker output.

***Random and Nested TE Insertion***

TEgenomeSimulator randomly samples a value between 0 and 30% as the default proportion of TE copy number to form nested insertions. This is only applied to LTR retrotransposons and is implemented across all simulation modes. The simulator performs simple TE insertion before executing nested TE insertion. Target sites for simple insertion and target TEs for nested insertion are randomly chosen.

***Output Files of TEgenomeSimulator***

TEgenomeSimulator generates a series of output files that are useful for downstream analysis, including the fasta file of the simulated genome, the fasta file of all TE sequences inserted in the simulated genome, and the TE annotation file in GFF format that records the sequence identity, integrity, and information to link the nested insertion to disrupted TEs.

## Implementation of PrinTE

### Burn-in

PrinTE simulates the forward evolution of genomes. The starting point for simulation is a burn-in genome in FASTA format and a corresponding PrinTE-formatted BED file. These will serve as the starting point (0 generation) for which no forward evolution has yet taken place. The burn-in FASTA and BED can be generated by TEgenomeSimulator, RandSeqInsert, *de novo* by PrinTE, or previous PrinTE outputs feeding back into PrinTE for additional forward simulations. PrinTE *de novo* burn-in allows many customizations, including genome size, chromosome number, percentage of the genome to be genic and TEs, and the ratio of each TE superfamily in the genome. Each TE insertion creates a TSD with a length determined by the TE superfamily, which is also customizable. In some instances, it may be beneficial to apply baseline mutation to

TEs in the burn-in such that they do not exactly match elements in the TE library. This is particularly relevant for self-dating elements, like LTR-RTs, where the default TE library only contains very young LTR-RTs, with exactly identical LTRs. For baseline mutations in the burn-in genome, we apply a normalized exponential decay function to skew mutations heavily toward lower mutation values:

The probability density function is aimed at flexibly and accurately modeling TE age distributions as they are observed in real data, where aggressive TE purging leads to higher probability of observing young TEs (**Fig. 5D**). The parameters $M_{max}$ (maximum mutation percentage; '`--TE_mut_Mmax`') and $k$ (decay constant; '`--TE_mut_k`') are customizable, allowing users to adjust the shape of the distribution to reflect a wide range of evolutionary and biological scenarios and to match the observed TE half-life of any species (see **Fig. S4A,B**).

***Two Simulation Modules***

$$f(M) = \frac{k}{M_{max} \cdot (1 - e^{-k})} \cdot e^{-k \cdot \frac{M}{M_{max}}}$$

Forward simulation begins once the burn-in genome is established. Users can specify the total number of generations ('`--generation_end`') and a smaller number of generations for each iteration ('`--step`'). There are three sequential modules for forward simulation: 1) genome mutation, 2) TE insertion, and 3) TE deletion. The pipeline will iterate over these processes for each step until the target number of generations is simulated. PrinTE can iterate through modules using constant and **fixed** TE insertion and deletion rates, or it can iterate using self-adjusting **variable** TE insertion and deletion rates. The fixed and variable methods begin identically, starting with the mutation of the genome. The expected number of DNA mutations per sequence ($\lambda m$) is calculated as:

$$\lambda m = DNA\ mutation\ rate \times genome\ size \times generations$$

Because TE insertions can occur at any generation of a defined evolutionary window (*i.e.*, a simulation step), they could accumulate DNA mutations less than $\lambda m$. To capture this dynamic, we used an exponential decay function similar to the decay function used for the burn-in genome to weight TE mutations toward lower values. We anchored the mutation distribution here to approximate the complicated evolutionary dynamics of TE activities. The actual number of mutations for each newly simulated TE is sampled from the probability density function, $f(mTE)$, to help weight their observed mutations toward lower frequencies:

$$f(m_{\mathrm{TE}}) = \frac{1.4 \times 10^{-6} \cdot e^{-1.4\times10^{-6} \cdot m_{\mathrm{TE}}}}{1 - e^{-1.4\times10^{-6} \cdot 3{,}000{,}000}}, \quad 0 \le m_{\mathrm{TE}} \le 3{,}000{,}000$$

***Simulation with Fixed Rates***

***TE insertion.*** After DNA mutations, TEs are inserted into the genome at random locations. They may insert into intergenic regions, into other TEs, resulting in nested TE insertions, or into

genes, resulting in fragmented genes. Information about nesting and fragmentation is recorded in the BED file. In the fixed rate method, the number of TEs to be inserted ($\lambda_{i-fix}$) is calculated as:

$$\lambda_{i-fix} = TE\ insertion\ rate \times genome\ size \times generations$$

***TE deletion and solo LTR.*** Donor TEs are randomly selected from the TE library based on user-defined abundance ratios per superfamily. A nested insertion will fragment the acceptor TE into two fragments. For example, if TE1 inserts into TE2, it will generate three sequences for further operations (e.g., deletion), including two resulting fragments of TE2 and the complete sequence of TE1. This differs from simple insertions into intergenic regions, where each insertion results in only one sequence for further operations. During the deletion process, full-length TEs and TE fragments are all eligible for deletion. The number of TE deletions ($\lambda_{d-fix}$) is calculated as follows:

$$\lambda_{d-fix} = TE\ deletion\ rate \times genome\ size \times generations$$

Illegitimate recombination is a prevalent process for TE deletions and is facilitated by sequence similarity between nearby homologous sequences [34]. Longer sequences have a higher probability of causing illegitimate recombination than shorter sequences because there are more similar base pairs to initiate pairing. As a result of this phenomenon, longer TEs have a higher probability of facilitating illegitimate recombination and a higher likelihood of being deleted than shorter TEs. We used an exponential decay function to model an exponentially increasing probability of TE deletion as TE length increases:

$$f(L) = e^{-k\cdot\left(1-\frac{L}{L_{95}}\right)}$$

Here, $L_{95}$ represents the TE length at the 95$^{th}$ percentile of the length distribution in the genome, $L$ represents the length of a TE (whether intact or fragmented), and $k$ is a user-defined parameter ('`--k`') that controls the rate at which deletion probability increases with sequence length. This modification allows us to favor the deletion of longer TE features by assigning higher deletion weights to sequences that are closer in length to the longest observed TE. Except for LTR-RTs, TE deletions span the full length of the TE feature. If a nested TE is selected for deletion, the adjacent TE fragments it disrupts are reconstituted into an intact TE. LTR-RTs are an exception to full-length deletions because they can undergo recombination with themselves owing to the sequence homology between their long terminal repeats, resulting in the formation of a solo LTR. PrinTE supports deletions that lead to solo LTR formation. If an LTR is selected for deletion, it will be converted to a solo LTR at a user-specified frequency, which is set by default to 95% ('`--solo_rate`').

***Simulation with Variable Rates***

In addition to the fixed-rate TE insertion/deletion method outlined above, in which the number of insertions and deletions per generation is constant and modeled by user-selected rates, PrinTE also offers a variable method that self-adjusts the number of insertions and deletions per

generation based on genomic landscapes. This aims to replicate a subset of the biological forces that affect TE insertions and deletions. The core processes, 1) genome mutation, 2) TE insertion, and 3) TE deletion, are identical between methods. The genome mutation step is identical between methods and is described above.

***TE insertions.*** In real genomes, the rate of TE transposition is influenced by the number of intact TEs, since only intact TEs are capable of initiating new insertions. Genomes with a high number of intact TEs tend to experience more transpositions per generation than those with few. Fragmented TEs and solo LTRs are no longer viable for transposition. In the variable method, the number of TE insertions ($\lambda_{i-var}$) is modeled as follows:

$$\lambda_{i-var} = TE\ insertion\ rate \times number\ of\ intact\ TEs \times generations$$

This simulates insertions via transposition, but some TEs will be acquired vertically and drift into the genome. We set that the frequency of vertically acquired transpositions is influenced by TE prevalence within the species. We set that species with TE-rich genomes are more likely to acquire additional TEs vertically, while species with TE-poor genomes are less likely to do so. For this calculation, the TE composition of the genome is surveyed before simulations and used as a proxy for TE abundance in the species. Vertically acquired TEs ($\lambda_{i-ver}$) are then simulated accordingly:

$$\lambda_{i-ver} = TE\ vertical\ rate \times TE\ composition\ of\ the\ species \times generations$$

***Insertion bias.*** The distribution of TEs across chromosomes is uneven and can be influenced by chromatin states, with euchromatin generally more accessible for TE insertion. In the variable method, we model this by treating genes and their surrounding regions as euchromatin, and the gene-poor genomic backbone as heterochromatin. We can modulate insertion location probability by chromatin type using a bias factor ('`--chromatin_bias_insert`'). We assign insertion probability weights to euchromatic and heterochromatic sequences as follows:

$$Euchromatin\ weight = length \times bias\ factor; \quad Heterochromatin\ weight = length$$

It is important to note that PrinTE does not use genes as functional units, but rather as representative segments of the genome to which cellular and selective attributes can be assigned. This enables the modeling of global genome features. Unlike the fixed-rate method, where TEs are inserted at random in proportions defined by the user, the variable-rate method inserts TEs in proportions that reflect the current genomic composition.

***Fitness-dependent TE deletions.*** TEs can trigger extensive and rapid proliferation, resulting in a sudden increase in their genomic abundance. If left unchecked, this proliferation can destabilize the genome and impose a selective disadvantage on the host organism. Under the variable-rate method, the number of TE deletions is dynamically adjusted according to the extent of gene disruption caused by TE insertions. In this context, genes are modeled as sequences subject to purifying selection, whereby insertions of TEs into these regions are deleterious and thus reduce the observed number of TE insertions. The number of TE deletions ($\lambda_{d-var}$) is then determined by the following expression:

In actual genomes, not all genes that confer a selective advantage exert the same influence on fitness. While most TE insertions into genes are deleterious, some have more severe effects. To capture this variability, we model the distribution of TE selection pressures across genes using a log-normal distribution (**Fig. S4C**):

$$\lambda_{d\text{-}var} = TE\ deletion\ rate \times \sum (disrupted\ gene\ weights) \times generations$$

$$f(x;\sigma) = \frac{1}{x\sigma\sqrt{2\pi}}\, e^{-\frac{(\ln x - \sigma^2)^2}{2\sigma^2}}, \quad x > 0$$

In this framework, each disrupted gene contributes a positive value to the total disrupted gene weight. The mode of the distribution is fixed at 1 to gain better control of the spread. Because of the long tail of the distribution, a small subset of genes exerts a disproportionately large influence on fitness. Consequently, TE insertions in these regions lead to a substantial increase in deletion events. The parameter σ, which is set via the command line ('`--sigma`'), determines the spread of the distribution. Larger values of σ produce longer tails and increase the range of selective effects among genes. In reality, TE insertions in some genomic regions can be lethal. PrinTE does not simulate lethality explicitly. Instead, we assume that simulating many generations using a stepwise approach effectively filters out lethal insertions. Over time, such deleterious types are eliminated from the population and do not persist in the simulated genomic landscape.

***TE deletion bias.*** Once the number of deletions is calculated, the specific TEs to delete are determined by the same exponential growth function ($f(L)$) used in the fixed-rate method multiplied by additional genomic attributes that might affect the probability of TE deletion. These genomic attributes include the selection coefficient parameter ('`--sel_coeff`') and the chromatin bias parameter ('`--chromatin_bias_delete`'). The selection coefficient parameter is used to assign additional deletion weight toward the TEs that disrupt genes. Similarly, we can assign additional weight toward deletions in simulated euchromatin regions. Longer TEs and TEs nearer to or within impacted genes are more likely to be deleted based on user-specified thresholds.

## Searching for TE Dynamic Parameters using PrinTE

To identify optimal TE dynamics in a given genome, we employed a systematic parameter sweep approach using PrinTE. This method involved iteratively testing combinations of key parameters to approximate empirical TE behavior and genome size observed in the genome of interest. The '`pipeline.report`' file contains an updating log of TE insertion and deletion counts per simulated iteration and can also be useful in monitoring PrinTE runs.

### *Parameter Space Exploration*

We conducted a grid search over two critical parameters, insertion and deletion rates, governing TE insertion dynamics. Users can specify parameter ranges (upper and lower limits) and increments of parameters. Based on the hypothesis of genome size dynamics, constraints on the deletion rate can be applied to narrow down the search space. For example, if a genome

evolves with a stable size, a deletion rate between 1.2 to 1.5 folds of the insertion rate can be used. We also monitor the size of simulated genomes and abort simulations when over or under the user-set size to ensure strict resource consumption. For example, if the target genome size is 150 Mb, simulation parameters that create genomes over 180 Mb in any steps are deemed unfit for the hypothesis and thus aborted to reduce the computation load. PrinTE's built-in '`--max_size`' parameter can be used to set the upper limit of genome size and stop the simulation when exceeding the specified size.

Additional parameters included length bias for deletions ('`--k`') and solo LTR rates ('`--solo_rate`'). In practice, we found k values of 0, 3, or 6 and solo rates from 70% to 95% (with 5% increments) could provide sufficient dynamics on top of the insertion and deletion rates. Higher k value will lead to slower genome size increase due to the preference for deleting longer TEs (**Fig. S8**). Higher solo rate will lead to faster genome size increase due to the creation of solo LTRs instead of full deletions (**Figs. S9, S10**) and increased solo:intact ratio.

***High-Throughput Simulation Framework***

PrinTE simulations were executed through an HPC cluster using array jobs (SLURM workload manager), enabling parallel processing of parameter sets. Users can set the total number of generations ('`-ge`') with sampling steps ('`-st`') for all simulations during parameter sweeps.

The species-specific substitution rate ('`--mutation_rate`') is also critical to the simulation outcome. In cases where an accurate substitution rate is not available for the species, users can try rates in different magnitudes (e.g., $1\times10^{-8}$, $1\times10^{-9}$, or $1\times10^{-10}$ per bp per generation) in the simulation, then compare the LTR divergence of simulated genomes to the real genome. Higher mutation rates will introduce more mutations to TEs in a given timeframe, thus creating gentler slopes of LTR divergence. In reverse, lower mutation rates will create steeper slopes of LTR divergence.

***Empirical Estimation of LTR Divergence and solo:intact Ratio***

Outputs were processed through a standardized pipeline to assess TE dynamics, including genome size change, LTR divergence distributions, and solo:intact ratio. Intact LTR elements were extracted from each simulation using the "`extract_intact_LTR.py`" script; their divergences were estimated with the K2P model and a mutation rate of $2.0\times10^{-9}$ per bp per generation using the "`seq_divergence.py`" script, and divergences were plotted using the "`ltr_dens.py`" script, all available in PrinTE.

Although the BED file output of PrinTE contains the annotation of all intact and solo LTR sequences, we perform a *de novo* scan of intact LTR-RTs using LTR_retriever (v3.0.1) [33] on the simulated genomes, so that these results are comparable to those obtained from the real genome. The resulting LTR library was used to annotate whole-genome LTR sequences with RepeatMasker (v4.1.7-p1) [48]. Solo LTRs were then identified by the "`solo_intact_ratio_wrap.sh`" script available in the LTR_retriever package, which also calculated the overall solo:intact ratio of each genome. For the same reason, LTR divergence distributions of simulated genomes were also plotted using intact LTR-RTs identified by

LTR_retriever. The "`plot_divergence.R`" script available in PrinTE was used to plot the density of LTR divergence in simulated and real genomes. Parameter sets were ranked by their ability to reproduce three empirical features: 1) Simulated genome sizes are consistent with hypothesis predictions; 2) Genome-wide LTR divergence distributions are consistent with empirical whole-genome data; 3) Solo:intact LTR ratios are close to the empirical value. This approach identified optimal parameters where simulated TE dynamics showed convergence with multiple empirical metrics while maintaining computational feasibility.

## Syntenic Ribbon Plot and LTR-RT Phylogeny

To simulate genomes presented in **Fig. 3A**, 19,621 *Arabidopsis thaliana* cds sequences were inserted into the simulated genome using PrinTE with "`--cds_num 19621`" to guide collinear anchoring and enable syntenic analysis. The simulated genomes were processed pairwise using "`jcvi.compara.catalog ortholog --cscore=.99`" in JCVI (v1.5.4) [50]. Syntenic anchors were visualized using "`matplotlib.pyplot`" in Matplotlib (v3.10.3) to create the ribbon plot (**Fig. 3A**). The LTR-RT phylogeny (**Fig. 3D,E**) was created using only elements in the *Tekay* family, which were classified as such by TEsorter (v1.4.7) [51]. *Tekay* LTR-RT headers were tagged with "`_shared`" and "`_uniq`" to facilitate post-PrinTE parsing. Shared LTR-RTs were inserted using PrinTE with parameters "`-p 80 -cn 1 -sz 10Mb -tmx 10 --TE_lib Tekay_shared.fa -tk 15`". The outputs ("`burnin.fa`" and "`burnin.bed`") were duplicated and allowed to continue simulation using either genome expansion parameter ( "`--generation_end 5000000 --step 500000 -F 2e-10,1e-10 --TE_lib Tekay_uniq.fa`") or genome contraction parameters ( "`--generation_end 5000000 --step 500000 -F -F 2e-11,1e-11 --TE_lib Tekay_uniq.fa`"). Intact LTR-RTs were extracted using "`extract_intact_LTR.py`" and classified using TEsorter with "`-db rexdb-plant -dp2`". Reverse transcriptase domains identified from *Tekay* LTR-RTs were aligned using MAFFT (v7.525) [52] and provided to VeryFastTree (v4.0.5) [53] for phylogenetic analysis. The resulting newick trees were plotted using ggtree (v3.1.5) [54] with branches colored by their ancestral relationship.

## *Pucciniomycotina* Genomes and Annotations

The myrtle rust (*Austropuccinia psidii*) genome was downloaded from NCBI with accession ID ERS3953144 [3]. Genomes of the *Pucciniomycotina* fungus subdivision were obtained from the JGI Fungal Genomics Resource MycoCosm (https://mycocosm.jgi.doe.gov/pucciniomycotina/pucciniomycotina.info.html) (**Table S3**). Rust fungus genomes with contig N50 values higher than 100 kbp were used, for a total of 47 genomes (**Table S4**).

Benchmarking Universal Single-Copy Orthologs (BUSCOs) evaluations were performed on each genome by compleasm (v0.2.7) [40] with the odb12 database and the basidiomycota lineage that has 2409 universal single-copy orthologs. For genomes with duplicated BUSCO higher than 5%, the Redundans (v2020.01.28) [39] program was used to identify alternative haplotypes. Haplotype alignments with an identity higher than 90% and coverage of more than 80% were removed using "`--identity 0.9 --overlap 80`" parameters. compleasm was

again used to evaluate the purged genomes with the same parameter. The resulting duplicated BUSCO were considered biological duplications (**Table S4**). The haplotype genomes (de-duplicated) were used in the following analyses.

Genomes with single-copy BUSCOs higher than 50% were used to construct the phylogeny. Melame1 (BUSCO S:1134,D:1147,F:33,I:0,M:95) and MecolCla_1 (BUSCO S:826,D:1452,F:34,I:0,M:97) were not included in this analysis owing to the high proportion of duplicated BUSCO. The protein sequences of single-copy BUSCOs present in the remaining 45 genomes were extracted and further aligned using mafft (v7.525) [52] with a focus on local alignments using the "`--maxiterate 1000 --localpair`" parameters. The phylogeny was constructed using the aligned BUSCO and iqtree (v2.4.0) [41]. Each BUSCO alignment was treated as an individual partition [55], and ModelFinder [56] was used to identify the best model for each partition using the "`-m MFP`" parameter in iqtree. The Shimodaira–Hasegawa-like approximate likelihood ratio test (SH-aLRT) [57] and bootstrap tests were performed 1000 times each with parameters "-bb 1000 -alrt 1000" in iqtree. The bootstrap consensus tree was plotted by ggtree (v3.1.5) [58] in R.

Transposable elements in each fungal genome were annotated using EDTA (v2.2.2) [59] following the instructions of Benson et al. (2024) [60]. The "`--force 1`" parameter was used to overcome analysis in genomes with missing types of TEs (*i.e.*, LTR retrotransposons). The "`--anno 1`" parameter was used to obtain whole-genome TE annotations. The "`--cds`" parameter was used to provide coding sequences to EDTA to improve TE annotations. Primary coding sequences of each genome were obtained from JGI MycoCosm and Tobias et al. (2021) [3]. Solo and intact LTR retrotransposons were identified from whole-genome TE annotations using LTR_retriever (v3.0.1) [33]. The solo:intact ratio was calculated on the whole-genome level, which is the number of whole-genome solo LTR divided by the number of whole-genome intact LTR. "inf" values were given to genomes without intact LTR retrotransposons. The divergence of intact LTR retrotransposons was estimated by LTR_retriever (v3.0.1) [33] using the one-parameter Jukes-Cantor model [61].

## TIR Insertion Simulation with RandSeqInsert

To demonstrate RandSeqInsert's capabilities, three targeted simulations were conducted. Two simulations with 10 rice TIR transposons as donor sequences were inserted into the *Arabidopsis* Col-CEN mitochondrial chromosome (ChrM) [10] as the acceptor sequence. One simulation utilized the "`--visual`" parameter to generate DOT visualizations of the balanced sequence tree structure (**Fig. S1A**), while the second simulation's insertion events were recaptured using minigraph (v0.21) [29] with default parameters and visualized with Bandage [62]. A larger-scale simulation was then performed by inserting 30,000 rice TIR transposons into a 350 Mb random acceptor sequence under two conditions: with and without TSD (9 bp). The inserted elements were subsequently redetected using TIR-Learner (v3.0.7) (https://github.com/lutianyu2001/TIR-Learner) [63] with the "`--species others`" parameter and compared against RandSeqInsert's record to calculate the detection rate.

## Simulating Digital Replicas of *Arabidopsis thaliana* and Rust Fungus Genomes

The TAIR10 *A. thaliana* genome was downloaded from NCBI with accession ID GCF_000001735.4. The *Arabidopsis* TE library for simulation and TE removal was obtained from the EDTA GitHub repository (https://github.com/oushujun/EDTA/blob/master/database/) [59]. Three fungal genomes were chosen for simulations, including *Puccinia coronata* f. sp. *avenae* (PuccoNC29_1), *Cronartium ribicola* (Crori1), and *Austropuccinia psidii* (myrtle rust). TE libraries used for TE removal and simulating the fungal genomes were obtained via *de novo* TE annotation using EDTA (as described above). The Genome Approximation Mode of TEgenomeSimulator was used for simulations. The resulting digital replicas of the rust fungus genomes were further used to simulate LTR dynamics using PrinTE (see sections below).

## Forward Simulation of Rust Fungus Genomes

***Simulating LTR Dynamics in Rust Fungus Genomes***

***Cronartium ribicola***. We hypothesized that this genome is relatively stable, with high rates of LTR insertions and deletions. We aimed to evolve the genome with a relatively stable genome size. The digital replica of the *C. ribicola* genome was generated by TEGenomeSimulator with a size of 232.6 Mb, which was used as input for PrinTE using the '`--bed`' and '`--fasta`' parameters. Each experiment was simulated with 20,000 generations per step and a total of 200,000 generations (`-ge 200000 -st 20000`). A series of fixed insertion and deletion rates were used in different experiments, including the range of insertion rates from $1.0\times10^{-9}$ to $1.9\times10^{-8}$ per bp per generation and the range of deletion rates from $1.4\times10^{-9}$ to $4.0\times10^{-8}$ per bp per generation specified by the '`--fix`' parameter. Additionally, DNA substitution rates of $1.0\times10^{-8}$, $2.0\times10^{-9}$, and $5.0\times10^{-9}$ per bp per generation, k values of 0, 3, 6, and 10, and solo rates of 60, 70, 75, 80, and 95 were used in the parameter sweep analysis. The simulated genomes were summarized by PrinTE and LTR_retriever and compared with the real *C. ribicola* genome for the determination of the best parameters.

***Puccinia coronata* f. sp. *avenae.*** We hypothesized that this genome is relatively stable with low rates of LTR insertions and deletions. We aimed to evolve the genome with a relatively stable genome size. The digital replica of the *P. coronata* f. sp. *avenae* genome was generated by TEGenomeSimulator with a size of 112.7 Mb, which was used as input for PrinTE using the '`--bed`' and '`--fasta`' parameters. Each experiment was simulated with 1 million generations per step and a total of 10 million generations (`-ge 10000000 -st 1000000`). A series of fixed insertion and deletion rates were used in different experiments, including the range of insertion rates from $8.0\times10^{-12}$ to $5.0\times10^{-11}$ per bp per generation and the range of deletion rates from $9.6\times10^{-12}$ to $9.0\times10^{-11}$ per generation specified by the '`--fix`' parameter. Additionally, mutation rates of $2.0\times10^{-9}$ and $5.0\times10^{-9}$ per bp per generation, k values of 0, 3, and 6, and solo rates of 70 and 95 were used in the parameter sweep analysis. The simulated genomes were summarized by PrinTE and compared with the real *P. coronata* f. sp. *avenae* genome for the determination of the best parameters.

***Austropuccinia psidii***. We hypothesized that this was once an expanding genome but now a fast downsizing genome with a low rate of LTR insertion and a high rate of LTR deletion. Owing

to the relative phylogenetic closeness to *A. psidii*, we use the genome size and TE composition of *P. coronata* f. sp. *avenae* genome as an approximation of the ancestral state of *A. psidii* before genome expansion. We used PrinTE to simulate such an ancestral genome with parameters '`--cds_percent 20 --TE_num 6000 --chr_number 20 --size 113Mb --TE_mut_k 10 --TE_mut_Mmax 20`'. With a fixed deletion rate of $5.0\times10^{-13}$ per bp per generation, we explored insertion rates from $2.0\times10^{-12}$ to $9.8\times10^{-11}$ per bp per generation and identified that the insertion rate of $1.4\times10^{-11}$ per bp per generation could yield LTR divergence similar to the expansion phase of *A. psidii* (**Fig. S16A,B**). We thus used PrinTE to simulate a fast-expanding phase with parameters "`-k 0 --fix 1.4e-11,5e-13 --mutation_rate 2e-9 --solo_rate 95`" until it reached 1,178 Mb. For the contraction phase, we slowed down the insertion rate to $1.4\times10^{-13}$ per bp per generation and explored deletion rates from $3.0\times10^{-13}$ to $5.0\times10^{-11}$ per bp per generation. We identified that the deletion rate of $3.0\times10^{-11}$ per bp per generation could yield LTR divergence and genome size similar to those of the *A. psidii* genome (**Fig. S16A,B**). Finally, we used the outputs of the last iteration as inputs for PrinTE to simulate a fast-contracting phase with parameters "`-k 0 --fix 1.4e-13,3e-11 --mutation_rate 2e-9 --solo_rate 95`" until it reached 1000 Gb, approximately the size of the *A. psidii* genome.

### Use of AI in this study

ChatGPT 4o, o3-mini, o1, o3, o4-mini, and o4-high, Qwen3-235B-A22B, and DeepSeek R1 were used to help draft and improve scripts during the development of PrinTE and TEgenomeSimulator.

## Data Availability

All codes that produced the findings of the study, including all main and supplemental figures, are available at https://github.com/cwb14/TESS, under the GPL-3.0 license. TEgenomeSimulator is available at https://github.com/Plant-Food-Research-Open/TEgenomeSimulator under the GPL-3.0 license.

## Acknowledgments

We want to thank Drs. Jason Slot, James Pease, and H. Lisle Gibbs for helpful discussions of this project. We are also grateful to Drs. Chen Wu, Julie Blommaert, and David Chagné for reviewing this manuscript and providing insightful comments.

## Funding

This research was supported by the Ohio State University STEM Education Faculty Startup Award (S.O.), Undergraduate Research Access Innovation Seed Grant (S.O.), the United States Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA) Agriculture and Food Research Initiative (AFRI) Postdoctoral Fellowship (AWD-117797 to C.W.B.), and by

Plant & Food Research through their Kiwifruit Royalty Investment Programme (T.-H.C., O.A.B., C.H.D., and S.J.T.).

# Author information

Both Christopher W. Benson and Ting-Hsuan Chen contributed equally to this work and have the right to list their name first in their CV.

## Authors and Affiliations

Department of Molecular Genetics, Ohio State University, Columbus, OH, 43210, USA
Christopher W. Benson, Tianyu Lu & Shujun Ou

The New Zealand Institute for Plant and Food Research Limited, Lincoln 7608, New Zealand
Ting-Hsuan Chen & Susan J. Thomson

The New Zealand Institute for Plant and Food Research Limited, Palmerston North 4442, New Zealand
Olivia Angelin-Bonnet

The New Zealand Institute for Plant and Food Research Limited, Auckland 1025, New Zealand
Cecilia H. Deng

## Contributions

S.O., S.J.T., and C.H.D. conceived the project and supervised the work. T.L. developed RandSeqInsert, T.-H.C. developed TEgenomeSimulator, and C.W.B developed PrinTE and the TESS framework. C.W.B, T.-H.C., T.L., O.A.B., and S.O. performed the analysis. C.W.B, T.-H.C., S.J.T., C.H.D., and S.O. wrote the manuscript text. All authors reviewed the manuscript.

## Corresponding author

Correspondence to Ting-Hsuan Chen (Ting-Hsuan.Chen@plantandfood.co.nz) and Shujun Ou (ou.195@osu.edu).

# Ethics declarations

## Ethics approval and consent to participate

Not applicable.

## Competing interests

The authors declare that they have no competing interest.

1238 # References


1239 1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–
1240 921 (2001).
1241 2. Hufford, M. B. *et al.* De novo assembly, annotation, and comparative analysis of 26 diverse
1242 maize genomes. *Science* **373**, 655–662 (2021).
1243 3. Tobias, P. A. *et al.* Austropuccinia psidii, causing myrtle rust, has a gigabase-sized genome
1244 shaped by transposable elements. *G3 (Bethesda)* **11**, (2021).
1245 4. Serrato-Capuchina, A. & Matute, D. R. The Role of Transposable Elements in Speciation.
1246 *Genes* **9**, 254 (2018).
1247 5. Beric, A. *et al.* Comparative phylogenetics of repetitive elements in a diverse order of
1248 flowering plants (Brassicales). *G3 (Bethesda)* **11**, (2021).
1249 6. Qin, P. *et al.* Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden
1250 genomic variations. *Cell* **184**, 3542–3558.e16 (2021).
1251 7. Bayer, P. E. *et al.* Variation in abundance of predicted resistance genes in the Brassica
1252 oleracea pangenome. *Plant Biotechnol. J.* **17**, 789–800 (2019).
1253 8. Ou, S. *et al.* Differences in activity and stability drive transposable element variation in
1254 tropical and temperate maize. *Genome Res* **34**, 1140–1153 (2024).
1255 9. Hu, Y. *et al.* Gossypium barbadense and Gossypium hirsutum genomes provide insights
1256 into the origin and evolution of allotetraploid cotton. *Nat Genet* **51**, 739–748 (2019).
1257 10. Naish, M. *et al.* The genetic and epigenetic landscape of the Arabidopsis centromeres.
1258 *Science* **374**, eabi7489 (2021).
1259 11. Blommaert, J. Genome size evolution: towards new model systems for old questions. *Proc.*
1260 *Biol. Sci.* **287**, 20201441 (2020).
1261 12. Kapusta, A., Suh, A. & Feschotte, C. Dynamics of genome size evolution in birds and
1262 mammals. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E1460–E1469 (2017).

1263 13. Ricci, M., Peona, V., Guichard, E., Taccioli, C. & Boattini, A. Transposable elements activity
1264 is positively related to rate of speciation in mammals. *J. Mol. Evol.* **86**, 303–310 (2018).
1265 14. Pellicer, J., Hidalgo, O., Dodsworth, S. & Leitch, I. J. Genome size diversity and its impact
1266 on the evolution of land plants. *Genes (Basel)* **9**, (2018).
1267 15. Niu, S. *et al.* The Chinese pine genome and methylome unveil key features of conifer
1268 evolution. *Cell* **185**, 204–217.e14 (2022).
1269 16. Liang, Y. *et al.* The giant genome of lily provides insights into the hybridization of cultivated
1270 lilies. *Nat. Commun.* **16**, 45 (2025).
1271 17. Fernández, P. *et al.* A 160 Gbp fork fern genome shatters size record for eukaryotes.
1272 *iScience* **27**, 109889 (2024).
1273 18. Wright, S. I. & Agren, J. A. Sizing up Arabidopsis genome evolution. *Heredity (Edinb.)* **107**,
1274 509–510 (2011).
1275 19. Kovaka, S., Ou, S., Jenike, K. M. & Schatz, M. C. Approaching complete genomes,
1276 transcriptomes and epi-omes with accurate long-read sequencing. *Nat. Methods* **20**, 12–16
1277 (2023).
1278 20. Hoen, D. R. *et al.* A call for benchmarking transposable element annotation methods. *Mob.*
1279 *DNA* **6**, 13 (2015).
1280 21. Clough, J. E., Foster, J. A., Barnett, M. & Wichman, H. A. Computer simulation of
1281 transposable element evolution: random template and strict master models. *J Mol Evol* **42**,
1282 52–58 (1996).
1283 22. Kofler, R. SimulaTE: simulating complex landscapes of transposable elements of
1284 populations. *Bioinformatics* **34**, 1439 (2018).
1285 23. Bourgeois, Y., Ruggiero, R. P., Hariyani, I. & Boissinot, S. Disentangling the determinants
1286 of transposable elements dynamics in vertebrate genomes using empirical evidences and
1287 simulations. *PLoS Genet* **16**, e1009082 (2020).
1288 24. Deceliere, G., Letrillard, Y., Charles, S. & Biémont, C. TESD: a transposable element

1289 dynamics simulation environment. *Bioinformatics* **22**, 2702–2703 (2006).
1290 25. Verneret, M., Van Anthony, L., Faraut, T., Turpin, J. & Lerat, E. Particular sequence
1291 characteristics induce bias in the detection of polymorphic transposable element insertions.
1292 *bioRxiv* 2024.09.25.614865 (2024) doi:10.1101/2024.09.25.614865.
1293 26. Groza, C., Chen, X., Wheeler, T. J., Bourque, G. & Goubert, C. A unified framework to
1294 analyze transposable element insertion polymorphisms using graph genomes. *Nat*
1295 *Commun* **15**, 8915 (2024).
1296 27. Yue, J.-X. & Liti, G. simuG: a general-purpose genome simulator. *Bioinformatics* **35**, 4442–
1297 4444 (2019).
1298 28. Caballero, J., Smit, A. F. A., Hood, L. & Glusman, G. Realistic artificial DNA sequences as
1299 negative controls for computational genomics. *Nucleic Acids Res* **42**, e99 (2014).
1300 29. Li, H., Feng, X. & Chu, C. The design and construction of reference pangenome graphs
1301 with minigraph. *Genome Biol.* **21**, 265 (2020).
1302 30. Rodriguez, M. & Makałowski, W. Software evaluation for de novo detection of transposons.
1303 *Mobile DNA* **13**, 1–14 (2022).
1304 31. Tusso, S., Suo, F., Liang, Y., Du, L.-L. & Wolf, J. B. W. Reactivation of transposable
1305 elements following hybridization in fission yeast. *Genome Res.* **32**, 324–336 (2022).
1306 32. Nowoshilow, S. *et al.* The axolotl genome and the evolution of key tissue formation
1307 regulators. *Nature* **554**, 50–55 (2018).
1308 33. Ou, S. & Jiang, N. LTR_retriever: A Highly Accurate and Sensitive Program for
1309 Identification of Long Terminal Repeat Retrotransposons. *Plant Physiol.* **176**, 1410–1422
1310 (2018).
1311 34. Ma, J., Devos, K. M. & Bennetzen, J. L. Analyses of LTR-retrotransposon structures reveal
1312 recent and rapid genomic DNA loss in rice. *Genome Res.* **14**, 860–869 (2004).
1313 35. Swarbreck, D. *et al.* The Arabidopsis Information Resource (TAIR): gene structure and
1314 function annotation. *Nucleic Acids Res.* **36**, D1009–14 (2008).

36. Chen, T.-H. & Winefield, C. Comprehensive analysis of both long and short read transcriptomes of a clonal and a seed-propagated model species reveal the prerequisites for transcriptional activation of autonomous and non-autonomous transposons in plants. *Mob. DNA* **13**, 16 (2022).
37. Ahrendt, S. R., Mondo, S. J., Haridas, S. & Grigoriev, I. V. MycoCosm, the JGI's fungal genome portal for comparative genomic and multiomics data analyses. *Methods Mol. Biol.* **2605**, 271–291 (2023).
38. *Systematics and Evolution*. (Not Avail, 2014). doi:10.1007/978-3-642-55318-9.
39. Pryszcz, L. P. & Gabaldón, T. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res.* **44**, e113 (2016).
40. Huang, N. & Li, H. compleasm: a faster and more accurate reimplementation of BUSCO. *Bioinformatics* **39**, btad595 (2023).
41. Minh, B. Q. *et al.* IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
42. Devos, K. M., Brown, J. K. M. & Bennetzen, J. L. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome Res.* **12**, 1075–1079 (2002).
43. Edwards, H. M. & Rhodes, J. Accounting for the biological complexity of pathogenic fungi in phylogenetic dating. *J. Fungi (Basel)* **7**, 661 (2021).
44. Osmanski, A. B. *et al.* Insights into mammalian TE diversity through the curation of 248 genome assemblies. *Science* **380**, eabn1430 (2023).
45. Zoonomia Consortium. A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240–245 (2020).
46. G. M. Adel'son-Vel'skii, E. M. Landis, 'An algorithm for organization of information', Dokl. Akad. Nauk SSSR, 146:2 (1962), 263–266. https://www.mathnet.ru/php/archive.phtml?wshow=paper&jrnid=dan&paperid=26964&optio

1341 n_lang=eng.
1342 47. Gansner, E. R. & North, S. C. An open graph visualization system and its applications to
1343 software engineering. *Softw. Pract. Exp.* **30**, 1203–1233 (2000).
1344 48. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0. 2013--2015. Preprint at
1345 (2015).
1346 49. Zhao, D., Ferguson, A. A. & Jiang, N. What makes up plant genomes: The vanishing line
1347 between transposable elements and genes. *Biochim. Biophys. Acta* **1859**, 366–380 (2016).
1348 50. Tang, H. *et al.* JCVI: A versatile toolkit for comparative genomics analysis. *Imeta* **3**, e211
1349 (2024).
1350 51. Zhang, R.-G. *et al.* TEsorter: an accurate and fast method to classify LTR-retrotransposons
1351 in plant genomes. *Hortic. Res.* **9**, uhac017 (2022).
1352 52. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:
1353 improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
1354 53. Piñeiro, C., Abuín, J. M. & Pichel, J. C. Very Fast Tree: speeding up the estimation of
1355 phylogenies for large alignments through parallelization and vectorization strategies.
1356 *Bioinformatics* **36**, 4658–4659 (2020).
1357 54. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. Ggtree: An r package for
1358 visualization and annotation of phylogenetic trees with their covariates and other associated
1359 data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
1360 55. Chernomor, O., von Haeseler, A. & Minh, B. Q. Terrace aware data structure for
1361 phylogenomic inference from supermatrices. *Syst. Biol.* **65**, 997–1008 (2016).
1362 56. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S.
1363 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**,
1364 587–589 (2017).
1365 57. Anisimova, M. & Gascuel, O. Approximate likelihood-ratio test for branches: A fast,
1366 accurate, and powerful alternative. *Syst. Biol.* **55**, 539–552 (2006).

58. Xu, S. *et al.* Ggtree: A serialized data object for visualization of a phylogenetic tree and annotation data. *Imeta* **1**, e56 (2022).

59. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).

60. Benson, C. W., Heringer, P. & Ou, S. Four strategies for whole-genome annotation of transposable elements and repeats in maize. *Cold Spring Harb. Protoc.* (2024) doi:10.1101/pdb.prot108578.

61. Jukes, T. H. & Cantor, C. R. Evolution of Protein Molecules. in *Mammalian Protein Metabolism* 21–132 (Elsevier, 1969).

62. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350–3352 (2015).

63. Su, W., Gu, X. & Peterson, T. TIR-Learner, a New Ensemble Method for TIR Transposable Element Annotation, Provides Evidence for Abundant New Transposable Elements in the Maize Genome. *Mol. Plant* **12**, 447–460 (2019).