Reconstructing the Evolutionary History of Transposable Elements

Arnaud Le Rouzic[1,*], Thibaut Payen[1,3], and Aurélie Hua-Van[1,2]

[1]Laboratoire Évolution, Génomes, Spéciation, CNRS-LEGS-UPR9034, CNRS-IDEEV-FR3284, Gif sur Yvette, France

[2]Université Paris-Sud 11, Faculté des Sciences, Orsay, France

[3]Present address: UMR INRA/UHP, Interactions Arbres/Micro-Organismes, INRA-Nancy, Champenoux, France

*Corresponding author: E-mail: lerouzic@legs.cnrs-gif.fr.

Accepted: December 17, 2012

Abstract

The impact of transposable elements (TEs) on genome structure, plasticity, and evolution is still not well understood. The recent availability of complete genome sequences makes it possible to get new insights on the evolutionary dynamics of TEs from the phylogenetic analysis of their multiple copies in a wide range of species. However, this source of information is not always fully exploited. Here, we show how the history of transposition activity may be qualitatively and quantitatively reconstructed by considering the distribution of transposition events in the phylogenetic tree, along with the tree topology. Using statistical models developed to infer speciation and extinction rates in species phylogenies, we demonstrate that it is possible to estimate the past transposition rate of a TE family, as well as how this rate varies with time. This methodological framework may not only facilitate the interpretation of genomic data, but also serve as a basis to develop new theoretical and statistical models.

Key words: transposition activity, phylogeny, branching process, repeated sequences.

Introduction

As transposable elements (TEs) have no systematic role in genomes beyond their own perpetuation, they are generally considered as selfish DNA sequences (Doolittle and Sapienza 1980; Orgel and Crick 1980). Nevertheless, their activity consisting in self-promoting mobility and duplication has noticeable consequences on host genomes, including mutation, recombination, change in genome size, and modification of the regulation patterns (Hua-Van et al. 2011). They are virtually universal, and they probably have existed since the origin of life; describing the dynamical properties of TEs thus appears as a necessary step toward a better understanding of genome evolution (Lynch 2007).

The short- and long-term dynamics of TE families in their host genome has generated a significant amount of theoretical work in population and evolutionary genetics (Charlesworth B and Charlesworth D 1983; Charlesworth 1991; Charlesworth et al. 1994; Le Rouzic and Deceliere 2005). Population genetic models and simulations confirm that parasitic TEs could realistically invade and maintain for a long time in sexual populations. Theoretical approaches have also suggested that several long-term scenarios were possible, including the loss of all copies, or the persistence of TE activity, either as a transposition-selection equilibrium, or as a succession of burst and decay stages (Charlesworth B and Charlesworth D 1983; Le Rouzic and Capy 2006). Unfortunately, empirical insights remain scarce and information about TE dynamics in genomes, such as changes in the transposition rate or correlations between different TE families, do not cover enough species nor enough TE families to provide broad and general inference about genome evolution. The recent improvement in sequencing technology, as well as the availability of the corresponding data in public databases, makes it possible to anticipate significant progress on these issues. Yet, an important factor limiting the exploration of genome evolution remains the availability of efficient statistical and analytical tools able to extract meaningful and synthetic information from such a large amount of data.

As a consequence of their propensity to duplicate, TEs are present as multiple copies in genomes. The number of copies varies according to the TE family and the host species, from a very few insertion sequences in bacterial genomes (Chandler and Mahillon 2002) to hundreds of thousands of LINE and SINE elements in human (Lander et al. 2001). For RNA-intermediate elements (class I), duplication is directly induced by the "copy-and-paste" transposition mechanism, whereas

for DNA "cut and paste" transposons (class II elements), duplication arises indirectly via DNA replication and repair (Wicker et al. 2007). In any case, a transposition event may generate a duplicated copy, inserted into a new genomic site, with a sequence that is identical to the original element. From this point, copies accumulate mutations independently, and their divergence increases with time.

Reconstructing the phylogeny of TE copies from the genome sequence of an individual could thus be used as a basis to infer the evolutionary history of a TE family in the whole species, and represents a rich source of information about genome evolution (Kazazian 2004; Ray et al. 2009; Biémont 2010). With this article, we intend to describe a simple and satisfactory methodological framework to infer TE evolutionary history in genomes, based on the birth–death models that have been developed to infer speciation and extinction rates in phylogenies (Yule 1924; Kendall 1948; Nee et al. 1994). We then discuss how to interpret the distribution of TE activity in the context of existing theoretical models.

Materials and Methods

Transposition Model

Several evolutionary mechanisms are involved in the variation of the copy number in genomes. The number of elements increases by replicative transposition, which explains the maintenance of the genomic parasite. The transposition rate is not necessarily constant, it may be affected by various regulation mechanisms, or by the progressive loss of transposition activity by mutation accumulation on TE sequences. Meanwhile, copies can be lost by different processes, including transposition-related or spontaneous deletion. Natural selection may also affect TE copy number: by assuming a decrease in fitness associated to copy accumulation, individuals with less copies will reproduce more efficiently, thus reducing the average copy number at the next generation.

Formal population genetic models of TEs stem from the early 1980s (Hickey 1982; Charlesworth B and Charlesworth D 1983), see Charlesworth et al. (1994), Le Rouzic and Deceliere (2005), and Lynch (2007) for review. Even if more elaborated models (often not tractable analytically) have been developed since then (Quesneville and Anxolabéhère 1998; Le Rouzic and Capy 2005; Dolgin and Charlesworth 2006; Le Rouzic et al. 2007), we will stick here to the simpler framework described in Charlesworth B and Charlesworth D (1983), predicting the dynamics of the average number of copies per genome ($\bar{n}$) as:

$$\bar{n}_{t+1} \simeq \bar{n}_t \cdot (1 + u_t - v), \quad (1)$$

where u_t is the replicative transposition rate at time t, and v is the deletion rate. In this setting, all parameters are considered as constant, except the transposition rate u_t that can change with time. For simplicity, the impact of natural selection, which tends to decrease the probability of fixation of deleterious copies, is here considered together with transposition regulation, and thus included in u_t. In the simulations, all copies are able to transpose (which does not necessarily mean that they are all capable of producing the transposition machinery).

To use this setting in a phylogenetic context, two assumptions are necessary. First, in the original setting of Charlesworth B and Charlesworth D (1983), time steps were standing for generations. At an evolutionary scale, the transposition dynamics has to be assimilated to a continuous process, u and v becoming transposition and deletion rates per time unit. Second, the phylogenetic inference is generally drawn from a single sequenced genome, and the recent population process is ignored. The ancestral lineage of the sequenced individual is thus assumed to be representative of the whole species (i.e., recent transposition events could be different in another lineage, but their dynamics should be similar).

Birth–Death Models

A birth–death model describes a stochastic branching process in which branches can split or disappear in the course of time. In traditional phylogenetic analyses, branch splitting events correspond to speciations, and dead branches correspond to species extinctions. Here, we propose to use the same framework, with a different interpretation: splitting branches are duplication (transposition) events (followed by the fixation of the duplicated copy), and extinct branches feature deletion events (followed by the fixation of the deleted allele).

The simplest model involves only birth events with a constant rate (using the notation presented in the previous section, $u_t = u$ and $v = 0$), which describes a "pure birth" model or Yule process (after Yule 1924). Branch extinctions ($v > 0$) can be included in a more complex branching process as in Kendall (1948), but application to statistical inference must account for the fact that a splitting event can be noticed in a phylogeny only if both lineages maintain up to the present time. According to Nee et al. (1994), the waiting time t before the next observable splitting event is described by the following equation:

$$\text{Prob}(t \mid u, v) = P_{\text{split}} \times P_{\text{obs}}, \quad (2)$$

where P_{split} is the probability for a splitting event, which follows an exponential distribution, and P_{obs} the probability of observing this splitting event from survivor branches. The model is usually reparameterized with $r = u - v$, the net diversification rate, and $a = v/u$, the extinction fraction (Rabosky 2006). The expression of these probabilities, as well as the corresponding likelihood function, can be found in, for example, Nee et al. (1994). Maximizing this likelihood function numerically allows to get estimates for r and a (and thus for u and v).

Several extensions or alternatives to this model have been developed to account for smooth or rapid changes in diversification and/or extinction rates (Rabosky 2006; Stadler 2011). Here, we explored four models, available as contributed packages in R (version 2.14) (R Development Core Team 2011): the "pure birth" model, implemented in the function yule() from the package "ape" version 3.0–4 (Paradis et al. 2004), the "birth–death" model from the function bd(), package "laser" version 2.3 (Rabosky 2006), the exponential change in birth rate ($u_t = u_0 e^{-kt}$, k being the rate of the change) from a modified version of function fitSPVAR() in "laser," and the diversity-dependence model from function dd_ML() in package "DDD" version 1.2 (Etienne and Haegeman 2012), in which $u = u_0 - (u_0 - v)n/K$ (K being the diversity dependence parameter). Changes in fitSPVAR() include 1) the possibility to fit negative k values (increase in diversification rate with time) and 2) setting the extinction rate to 0. The corresponding code and scripts are available on demand. Support intervals of parameters were estimated from 100 bootstrapped trees (95% central values of the bootstrapped parameter distribution).

Tree Imbalance

Another meaningful piece of information that can be extracted from TE phylogenetic analysis is related to the balance (or imbalance) of the trees. In a perfectly balanced tree, all branches duplicate once, while the most unbalanced tree corresponds to the situation where all duplications happen in the same branch. In a TE-related context, balanced trees arise when all copies can duplicate at the same rate, while unbalanced trees correspond to "master copy" models when only one copy in the genome is able to transpose. Being able to quantify the balance of TE phylogenetic trees may thus lead to meaningful insights on transposition history.

The definition of mathematical and statistical tools to estimate phylogenetic tree imbalance has generated a significant amount of literature that cannot be explored here (see e.g., Kirkpatrick and Slatkin 1993; Aldous 2001; Blum and François 2006). We focused on a classical imbalance index, the β index. Index estimation by maximum likelihood (ML) and statistical analyses were performed with the package "apTreeshape" version 1.4–5 (Bortolussi et al. 2006) for R.

Interestingly, there is no general definition of balanced random trees. The literature reports two traditional models of random trees, the "Proportional to Distinguishable Arrangements" (PDA) model (assuming a uniform probability for all tree shapes), and the "Equal Rate Markov" (ERM) model, which corresponds to trees generated by a Yule process. Trees generated under the ERM model have a β index of 0, whereas PDA trees are characterized by $\beta = -1.5$. The β index can thus be interpreted along the following scale: imbalanced trees ($-2 < \beta < -1.5$), random trees ($-1.5 \leq \beta \leq 0$), and trees which are too perfectly balanced to be random ($0 < \beta < \infty$).

Simulations

Stochastic simulations were run to provide reference dynamics for interpretation. Simulations consider a unique genome reproducing clonally (the "average genome" of the species), and for simplicity, time steps are discrete. TE copies are followed individually and their pedigree is stored by the simulation program. The deletion rate v per time step is constant, and the transposition rate u_t can vary with time arbitrarily. The system evolves according to equation (1): every time step, $x_1 \sim \mathcal{P}(n_t \cdot u_t)$ new elements are created (all elements having equal probabilities of being the master copy; $\mathcal{P}(x)$ stands for the Poisson distribution of mean x), and $x_2 \sim \mathcal{B}(n_t, v)$ are randomly removed ($\mathcal{B}(N, p)$ stands for the Binomial distribution). Distance matrices and phylogenetic trees were reconstructed from the exact evolutionary relationships between elements (no further stochasticity is introduced to mimic the accumulation of mutations). Simulations were run for 30 time steps with four sets of parameters: 1) $u = 0.109$ and $v = 0$, 2) $u = 0.159$ and $v = 0.05$, 3) $u_0 = 0.1 \rightarrow u_{30} = 0.219$ and $v = 0.05$, and 4) $u_0 = 0.219 \rightarrow u_{30} = 0.1$ and $v = 0.05$ (the $\rightarrow$ symbol representing a linear change with time). These parameters were chosen so that the expected number of copies after 30 time steps should be 20. Simulations started with a unique copy, and 1,000 runs in which the final copy number was between 15 and 25 were kept for each parameter set.

The *Fot* Elements in *Fusarium*

We used real genomic data from a recent work by Dufresne et al. (2011) to illustrate this theoretical framework. *Fot* TEs are *Tc1-mariner-pogo* elements found in filamentous fungi. Four subfamilies extracted from the genome sequence of *Fusarium oxysporum* were selected for their average number of independent copies (a few dozen): *Fot2* (28 copies), *Fot3* (46 copies), *Fot5* (145 copies), and *Fot6* (38 copies). Duplicates with homologous flanking regions, corresponding to transposition-unrelated mechanisms (e.g., segmental duplication), have been removed from the data set (only one copy is randomly kept for each set of duplicates). Further details are provided in Dufresne et al. (2011).

The phylogenetic analysis was performed in R (version 2.14) (R Development Core Team 2011), using packages ape (Paradis et al. 2004) version 3.0–4 and phangorn (Schliep 2011) version 1.6–3. An ML phylogeny was derived for each *Fot* family, using a GTR + G (Gamma) model of substitutions. Trees were rooted with elements from other families. Ultrametric trees were calculated from the ML trees (without the outgroup) using the "pathd8" method (Britton et al. 2007), which happened to give visually more convincing results than

penalized likelihood (Sanderson 2002), or mean path length (Britton et al. 2002), perhaps because of the uneveness of the evolutionary rates across branches. Reproducing the analysis with mean path length ultrametric trees provide very similar results (not shown).

Results

Interpretation of Phylogenetic Patterns

In this article, we propose to quantify transposition activity over time from the distribution of transposition events. The steps required for such an analysis consist in 1) reconstructing the phylogeny of TE sequences from a clean and exhaustive sequence data set of the TE family in the studied genome, from which duplicates (copies gained by other mechanisms than transposition, e.g., polyploidization or segmental duplication) are removed, 2) estimating the age of the visible transposition events, corresponding to the nodes in the tree, and 3) inferring the past transposition dynamics from the branching pattern.

Simulation results illustrate how the divergence between homologous TE sequences reflects meaningful information about the transposition dynamics in this TE family. Transposition is an exponential process: if the transposition rate per copy is constant (fig. 1*A*), the number of new transpositions increases with the copy number (fig. 1*B*). As a result, a constant transposition rate mainly generates recent copies. One of the most convenient visualization tool is the "lineage through time" (LTT) plot, displaying the increase in the number of branches in the tree with time (figs. 1*C* and 2). An exponential increase of the number of lineages with time (linear trend on a logarithmic LTT plot) reflects a "pure birth" process with a constant transposition rate and no deletion. Departure from this linear pattern denotes deletions or changes in the transposition rate and can be used as a basis for parameter estimation.

Application to the Dynamics of *Fot* Elements in *F. oxysporum*

Four subfamilies of *Fot* elements, numbered *Fot2*, *Fot3*, *Fot5*, and *Fot6*, were retrieved from the genome of the filamentous fungus *F. oxysporum*, as described in Dufresne et al. (2011). All of these TE families are ancient families, elements displaying genetic distances up to 35%. In all four subfamilies, recent transposition events (identical or nearly identical sequences inserted in nonhomologous positions) were detected, suggesting that they are all still active. ML phylogenetic trees suggest important changes in the molecular evolutionary rates in some branches, most of them corresponding to repeat-induced-point mutations, a fungus-specific (but not very active in *F. oxysporum*) defense mechanism against selfish DNA (Cambareri et al. 1989; Galagan and Selker 2004). This may lead to poor temporal estimates for some nodes, but most copies remain unaffected, making further analysis on ultrametric trees (fig. 3) meaningful.

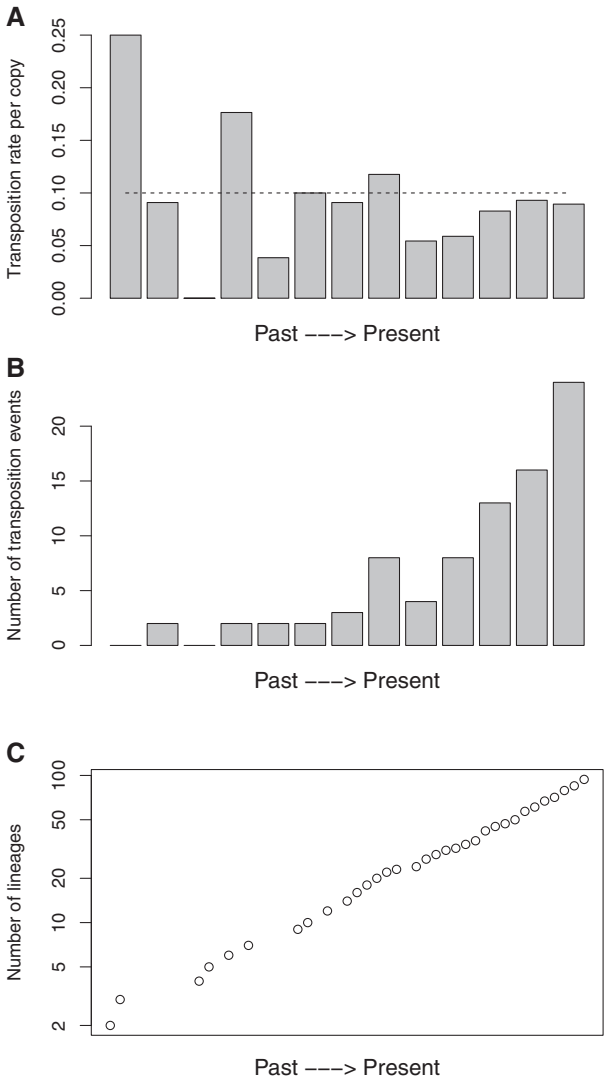

Fig. 1.—Single simulation of the temporal dynamics of a TE family with a constant transposition rate ($u = 0.1$ per copy and per time step), and no deletion ("pure birth" model). *X* axes are oriented from past to present in reconstructed dynamics (*A*, *B*, *C*) ($x = 0$ corresponds to the start of the transposition history, each bar stands for four successive generations). With a constant transposition rate per copy (dashed line on *A*), the number of copies increases exponentially. This increase is reflected by the log-linear pattern of the LTT plot (*C*), which can be used as a basis for reconstructing the dynamics of the TE family.

Branch lengths estimated by ML are corrected for multiple mutations, and are thus expected to be proportional to the evolutionary distance, assuming some approximative molecular clock. As all sequenced elements are present in the genomes of modern species, all the tips should be aligned when the tree scales with time: the corresponding ultrametric trees were obtained by the "pathd8" method, after removal of the outgroups (see Materials and Methods). We first applied a

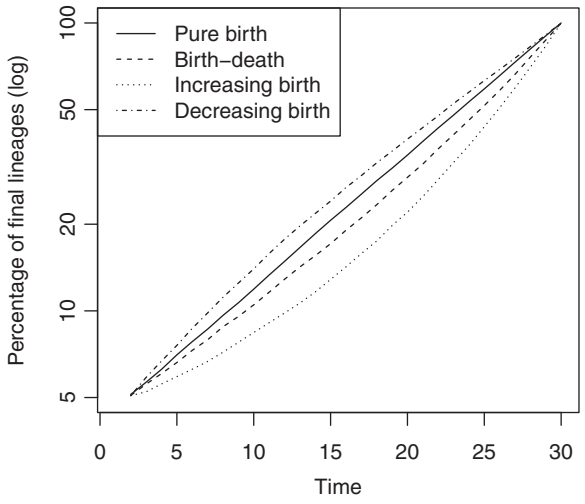

Fig. 2.—Simulated LTT plots in four scenarios. Each line is the average over 1,000 replicates. The pure birth model corresponds to a transposition-only model; the birth–death model features both transpositions and deletions; and the increasing and decreasing birth models represent linear changes in the transposition rate (see Materials and Methods for details). Different transposition dynamics generate different LTT profiles, illustrating how the branching pattern from phylogenetic trees can be used to estimate the transposition history.

"pure birth" model (constant transposition rate and no deletion) (table 1). The estimated transposition rates across the TE families are quite similar, between 0.09 and 0.16 per percentage of divergence. Nevertheless, the dynamics of these four families are not identical, since the birth–death model (allowing both transposition and deletion) could detect a non-null deletion rate for *Fot5*, whereas no significant deletions could be identified for the other families.

The resulting LTT (or more exactly, lineage-through-divergence) plots (fig. 4) suggest important departure from simple models. The curves for all *Fot* families are above the "pure birth" prediction, which suggests that the past rate of duplication per copy was higher than the current one. To check for changes in the transposition rate, we fit models in which transposition rates vary exponentially with time. Figure 5 illustrates the resulting dynamics, as well as the 95% support intervals calculated from bootstrapped phylogenies. At least two TE families show clear changes in their transposition dynamics: in *Fot2* and *Fot6*, the transposition rate tends to decrease with time. The slightly decreasing trends for *Fot3* and *Fot5* are not supported statistically.

Finally, we exploited an existing model for diversity-dependent speciation to test the hypothesis of transposition regulation. Transposition regulation assumes that the transposition rate decreases with the number of copies, which is necessary to avoid an exponential invasion of TEs in genomes. The model developed by Etienne et al. (2012) assumes that the "ecosystem" (in our case, the genome) has a carrying capacity K, so that the transposition rate varies with $1 - n/K$, where n is the number of TE copies of the family under consideration. For all four TE families, the diversity dependent model significantly outperforms the birth–death model, with Akaike Information Criterion (AIC) differences ranging from 15 units (*Fot2*) to 87 units (*Fot5*). However, estimated carrying capacities (the number above which transposition would stop completely) were well above the observed number of copies (*Fot2*, *Fot3*, *Fot5*, and *Fot6* occupy only 8%, 5%, 13%, and 4% of their theoretical niche, respectively). Although statistically significant, diversity-dependence remains moderate, and affects the transposition rate only marginally (the current transposition rate for all families is more than 85% of the estimated initial transposition rate when one copy only was present in the genome). This result supports the idea that transposition regulation by the number of copies is not strong enough to allow for a stable transposition–deletion equilibrium, although interpretation is obscured by the presence in the genome of TE copies caught in segmental duplications, which were not included in the phylogenetic analysis, but which could be involved in regulation.

Phylogenetic Tree Balance

The β index for tree imbalance was computed as detailed in the Materials and Methods section. ML estimates of β, as well as 95% support intervals calculated from 500 bootstraps, were as follows: $\hat{\beta}_{Fot2} = -1.02$ $(-1.75,\ 4.04)$, $\hat{\beta}_{Fot3} = -1.01$ $(-1.61,\ 0.35)$, $\hat{\beta}_{Fot5} = -1.03$ $(-1.30,\ -0.70)$, and $\hat{\beta}_{Fot6} = -1.16$ $(-1.78,\ -0.20)$. The estimates of tree imbalance are thus very similar across the four TE families, estimates being more precise in larger trees. All β estimates are consistent with random trees. Tree imbalance is intermediate between the two extreme models of random trees (the Yule process or ERM model, $\beta = 0$, and the uniform PDA model, $\beta = -1.5$). *Fot5* and *Fot6* trees exclude a Yule process as a generating mechanism ($\beta = 0$ being outside of the support interval), suggesting that the actual transposition rate differs across clades. However, the "master copy" hypothesis, which generates highly imbalanced trees ($\beta < -1.5$), can be statistically rejected for most families. Alternative indexes (Colless and Sackin indexes, as implemented in the package "apTreeshape," Bortolussi et al. 2006) provided identical results (tree imbalance intermediate between ERM and PDA models, not shown).

Discussion

Transposition Dynamics

With this article, our intention is to demonstrate how the phylogenetic pattern of repeated genomic sequences could be analyzed in terms of temporal dynamics. We showed that different transposition dynamics lead to different distributions of transposition events, and that it was possible to derive models to reconstruct transposition history from available

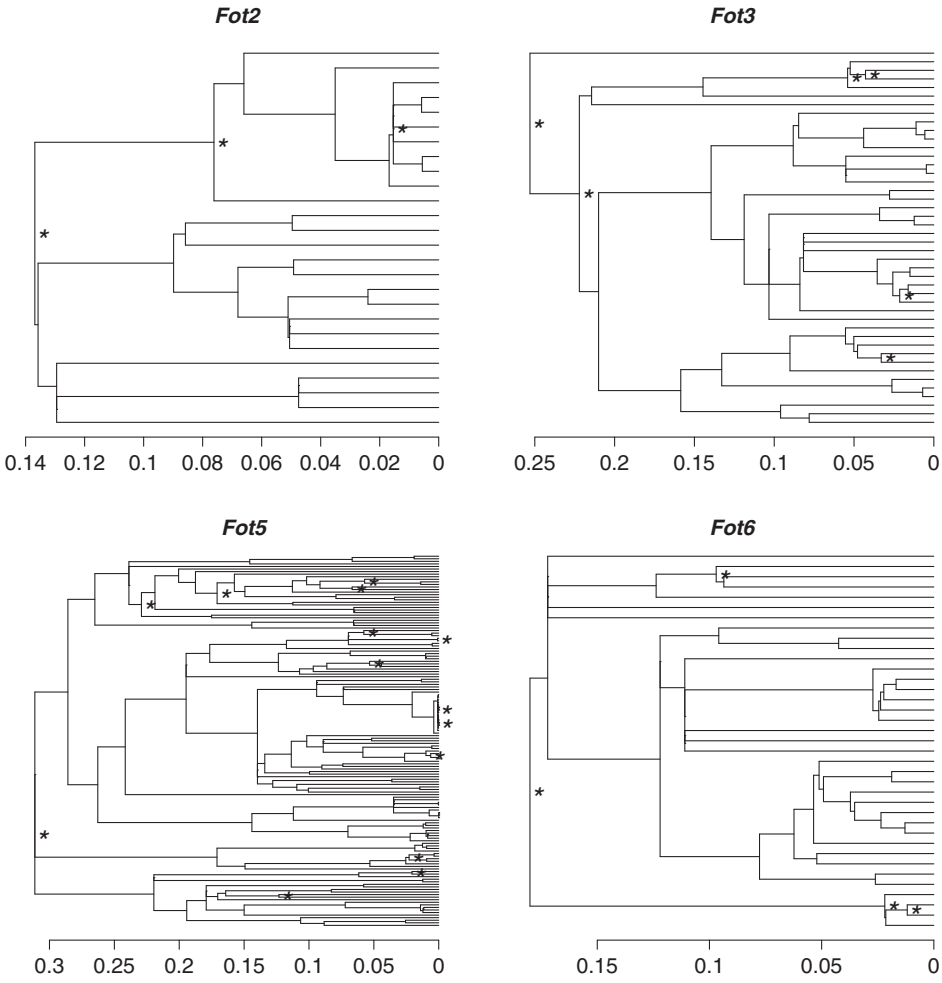

Fig. 3.—ML reconstructed phylogenies for the four *Fot* subfamilies. Trees were rooted with the other subfamilies. Ultrametric trees were obtained through the "pathd8" algorithm (see "Materials and Methods"). Asterisks (*) denote nodes that are supported by bootstrap scores ≥50.

sequence data, based on a quantitative statistical framework used for species phylogenies.

We believe that this strategy represents a significant improvement compared with the state of the art in genomics. The literature reports several ways to interpret phylogenetic and divergence data in similar contexts (Ray et al. 2008; Zerjal et al. 2009; Cordaux et al. 2010; Han et al. 2010; Dufresne et al. 2011). However, most of these methods are not devoid of limitations, biases, or caveats. Frequently, the age of a TE family is calculated as the average distance between copies and a consensus sequence (supposedly close to the ancestral sequence). Yet, this procedure does not allow the exploration of within-family dynamics. This issue is sometimes overcome by assuming several successive transposition bursts (Pace and Feschotte 2007), which is restricted to TE families with many copies. Visual comparison of tree topologies is qualitative only, and information about absolute branch lengths is disregarded. Alternatively, the distribution of pairwise distances between copies may provide quantitative results, but ancient transposition events (deep and bushy nodes in the tree) are counted several times, which severely hinders data interpretation. These approaches are difficult to apply to other species or TE families with smaller copy number or different transposition activity, and are probably not suitable for systematic exploration of available data. An exception lies in the ingenious method proposed by SanMiguel et al. (1998), which consists in estimating the insertion date of retro-elements based on the similarity between their two long-terminal repeats (LTRs), strictly identical after transposition. Unfortunately, this strategy can be applied only to complete LTR retro-elements, and remains associated with large sampling errors due to the small size of LTR sequences.

Model Limits

The dynamics of TE sequences in genomes remain quite a complex process, and a simple model necessarily relies on approximations. In particular, quantifying the statistical error

in phylogenetic analysis is known to be a complex issue (Felsenstein 1988; Wróbel 2008; Kumar et al. 2012), because errors are both quantitative (branch lengths) and qualitative (tree topology, selection of the evolutionary model). Here, we estimated errors using the same resampling strategy as for phylogeny: confidence intervals of, for example, transposition rates were derived from the distribution of estimated rates obtained by running the model on a large number of bootstrapped trees. This time-consuming resampling strategy has the advantage to be applicable to any phylogenetic reconstruction method.

However, estimating the sampling noise associated to parameter estimates does not inform about potential biases. Estimates of transposition dynamics are reliable only if the models on which they are based are good approximations of the real processes, including sequence alignment, phylogenetic reconstruction, tree datation, and transposition model. A critical step here is the estimation of an ultrametric tree (in which all tips are aligned and distances scale with time) from an ML tree with different branch lengths. The evolutionary rate of TE sequences is not very well understood, and is known to vary dramatically between TE clades, due to, for example, sequence inactivation (equivalent to pseudogenization), or more specifically in our example, repeat-induced point mutations, a fungus-specific regulation mechanism (Cambareri et al. 1989; Galagan and Selker 2004). Tree topology can also be affected by various biases; for instance, simulation studies show that poor data tend to generate imbalanced trees (see Mooers and Heard 1997 for review). The estimated branching dynamics (branch length and topology) thus rely on the robustness of a series of biological assumptions; improving the phylogenetic reconstruction (e.g., by implementing TE-specific features) may thus improve significantly the reliability of the inferred transposition history.

Table 1

Estimates of the Diversification Rate $r = u - v$ in the "Pure Birth" Model and in the "Birth–Death" Model (for Which the Extinction Fraction $a = v/u$ Is Also Provided)

	Pure Birth	Birth–Death
Fot 2		
r	0.155 (0.145, 0.168)	0.161 (0.144, 0.175)
a		0.000 (0.000, 0.000)
u	0.155 (0.145, 0.168)	0.161 (0.144, 0.175)
v		0.000 (0.000, 0.000)
Fot 3		
r	0.118 (0.111, 0.124)	0.121 (0.091, 0.126)
a		0.000 (0.000, 0.004)
u	0.118 (0.111, 0.124)	0.121 (0.112, 0.148)
v		0.000 (0.000, 0.051)
Fot 5		
r	0.118 (0.111, 0.125)	0.092 (0.067, 0.094)
a		0.004 (0.004, 0.006)
u	0.118 (0.111, 0.125)	0.157 (0.155, 0.197)
v		0.065 (0.063, 0.126)
Fot 6		
r	0.122 (0.114, 0.130)	0.126 (0.109, 0.134)
a		0.000 (0.000, 0.000)
u	0.122 (0.114, 0.130)	0.126 (0.109, 0.134)
v		0.000 (0.000, 0.000)

NOTE.—95% support intervals, calculated from 100 bootstrapped trees, are indicated between parentheses. Estimates of *u* and *v* calculated from *r* and *a* are also provided. *r*, *u*, and *v* are expressed in "events per percentage of divergence," whereas *a* is unitless.

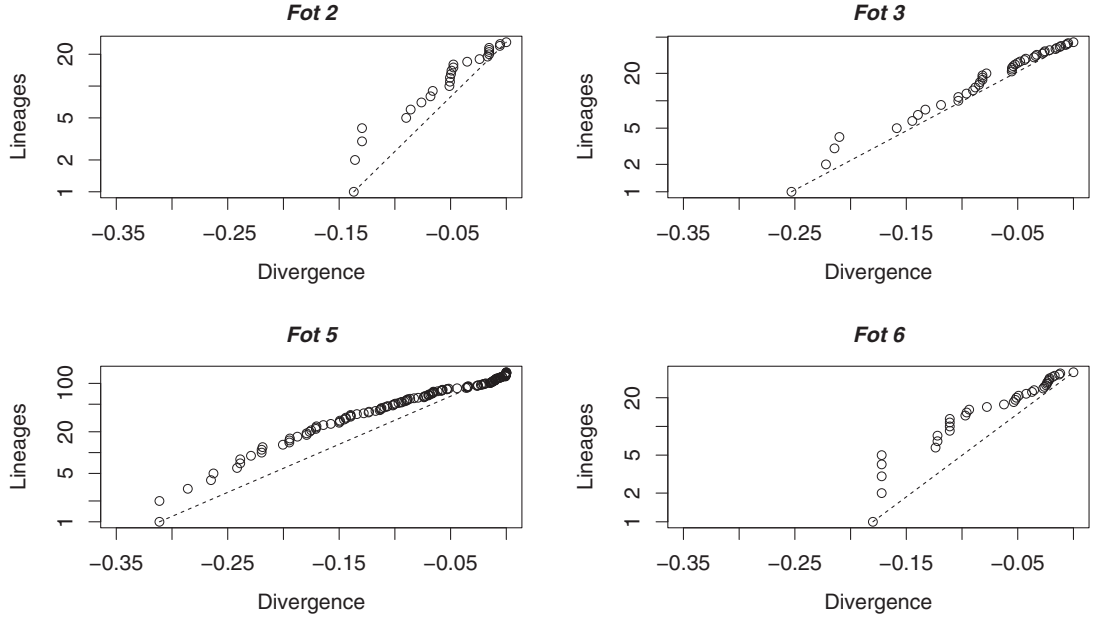

FIG. 4.—Lineage-through-divergence plots for the four *Fot* subfamilies. The dashed line illustrates the expectation for a "pure birth" model (constant transposition, no deletions).

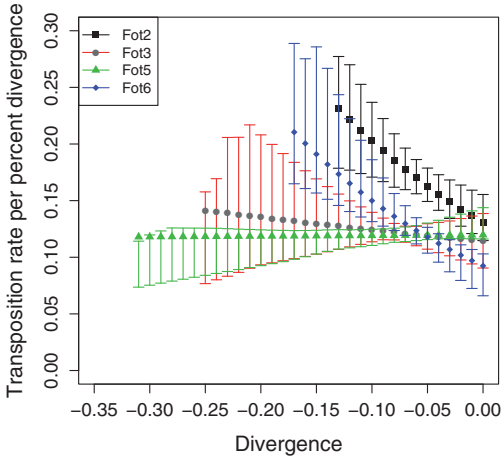

Fig. 5.—Illustration of the estimated ML exponential dynamics (dots), and the corresponding 95% support intervals from 100 bootstrapped trees.

Although powerful and widely used in phylogenetics, branching models should be interpreted carefully. One of the most problematic issues is the lack of power to compute the extinction rate (v in our case) compared with the net diversification rate ($r = u - v$), up to the point that some authors consider that extinction rates should not be estimated at all from phylogenies (Rabosky 2010). In our examples, a significant (but relatively small) deletion rate could be detected for one out of four *Fot* families. The estimated value of v is realistic, but alternative interpretation could be proposed, such as a recent increase in the transposition rate. More robust estimates of transposition rates could be obtained from more extensive data, for example, by comparing orthologous insertion sites between close species.

Interpreting variation of the transposition rate may also depend on the detailed nature of TEs. Here, we present an example based on cut-and-paste, class II TEs. In *Fot* elements, tree topologies appear to be roughly balanced, and most copies are able to transpose and to generate new branches in the tree, supporting (at least partially) the exponential Yule model. This pattern appears to be widespread for TE phylogenies (Cordaux et al. 2004). However, other TEs (such as class I elements) are known to generate a high proportion of "dead on arrival" copies after transposition (i.e., most transposition events are asymmetric and generate a nonfunctional copy), resulting in an extremely imbalanced tree. Therefore, in the latter case, known as the "master copy" model (Clough et al. 1996; Brookfield and Johnson 2006; Johnson and Brookfield 2006), the evolutionary dynamics should not be necessarily interpreted as a drop in transposition activity as long as the transposition rate per genome remains constant, even if the transposition rate per copy mechanically decreases with time. Both tree topology and branching dynamics, although almost independent statistically, thus provide complementary information to reconstruct the evolutionary history of repeated sequences.

Perspectives

A natural (yet, not trivial) evolution of the model should account for the activity of TE sequences. In general, genome scans reveal at least three functional categories: active copies (canonical elements), relic copies (equivalent to pseudogenes), and nonautonomous copies (unable to code for the transposition machinery, but mobile when *trans*-mobilized). Simulation models have shown that the relative proportion of each kind of copies may affect significantly the dynamics of the whole TE family (Le Rouzic et al. 2007; Boutin et al. 2012). Ideally, such a TE-specific evolutionary model should be taken into account in the phylogenetic reconstruction, including, for example, different mutation rates depending on the status of the copy, as well as the location of pseudogenization events in the tree based on the observed status of the sequences and the tree topology. Yet, implementing such a model may require deep changes in the phylogenetic algorithm.

Another issue with the most recent duplication events is that the branching model ignores recent population genetics mechanisms (such as natural selection against slightly deleterious TE copies), and that the phylogeny reconstructed from a single individual genome might provide a biased view of the recent transposition history. There is little doubt that, along with progress in sequencing, the genome of several individuals per species will be available soon as it is already the case with model species, which is likely to help fixing this issue (provided a suitable theoretical framework).

In any case, the nature of the genomic data makes it possible to obtain independent estimates of parameters of interest, which could validate phylogenetic models, or be used as fixed parameters to derive more complex models. For instance, deletion rates can be independently estimated by identifying and dating deletion events from TEs inserted in duplicated parts of the genome, which were not included in the phylogeny. The robustness of the procedure could also be improved by dating some of the tree nodes, by comparing insertions shared by close species, and inferring transposition timing based on estimates of speciation events from fossil data or phylogenies of conserved genes.

Reconstructing the activity dynamics of TEs from genome sequences thus requires to combine tools from bioinformatics, phylogenetic analysis, and population genetics. Here, we provide a methodological framework to estimate and interpret the pattern of transposition activity, using the statistical framework developed to infer speciation and extinction dynamics in species phylogenies. This framework can be complexified, and makes it possible to derive more efficient procedures and more realistic models. Given the rapid accumulation of new genome sequences, the development of a new set of tools

devoted to the study of repeated sequences appears as one of the keys for improving the efficiency of the analysis of such massive, costly, and informative data.

Acknowledgments

The authors thank P. Capy for useful discussion. This work was partly supported by the European Commission (Marie Curie-ERG 256507).

Literature Cited

Aldous D. 2001. Stochastic models and descriptive statistics of phylogenetic trees, from Yule to today. Stat Sci. 16:23–34.
Biémont C. 2010. A brief history of the status of transposable elements: from junk DNA to major players in evolution. Genetics 186(4): 1085–1093.
Blum MGB, François O. 2006. Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance. Syst Biol. 55(4):685–691.
Bortolussi N, Durand E, Blum M, François O. 2006. apTreeshape: statistical analysis of phylogenetic tree shape. Bioinformatics 22(3): 363–364.
Boutin TS, Le Rouzic A, Capy P. 2012. How does selfing affect the dynamics of selfish transposable elements? Mob DNA 3(1):5.
Britton T, Anderson CL, Jacquet D, Lundqvist S, Bremer K. 2007. Estimating divergence times in large phylogenetic trees. Syst Biol. 56(5):741–752.
Britton T, Oxelman B, Vinnersten A, Bremer K. 2002. Phylogenetic dating with confidence intervals using mean path lengths. Mol Phylogenet Evol. 24(1):58–65.
Brookfield JFY, Johnson LJ. 2006. The evolution of mobile DNAs: when will transposons create phylogenies that look as if there is a master gene? Genetics 173(2):1115–1123.
Cambareri EB, Jensen BC, Schabtach E, Selker EU. 1989. Repeat-induced G-C to A-T mutations in *Neurospora crassa*. Science 244:1571–1575.
Chandler M, Mahillon J. 2002. Mobile DNA II. Chapter: Insertion sequences revisited. Washington (DC): American Society for Microbiology Press. p. 305–366.
Charlesworth B. 1991. Transposable elements in natural populations with a mixture of selected and neutral insertion sites. Genet Res Camb. 57: 127–134.
Charlesworth B, Charlesworth D. 1983. The population dynamics of transposable elements. Genet Res Camb. 42:1–27.
Charlesworth B, Sniegowski P, Stephan W. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371:215–220.
Clough J, Foster J, Barnett M, Wichman H. 1996. Computer simulation of transposable element evolution: random template and strict master models. J Mol Evol. 42(1):52–58.
Cordaux R, Hedges DJ, Batzer MA. 2004. Retrotransposition of *Alu* elements: how many sources? Trends Genet. 20(10):464–467.
Cordaux R, Sen SK, Konkel MK, Batzer MA. 2010. Computational methods for the analysis of primate mobile elements. Methods Mol Biol. 628:137–151.
Dolgin ES, Charlesworth B. 2006. The fate of transposable elements in asexual populations. Genetics 174:817–827.
Doolittle W, Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature 284(5757):601–603.
Dufresne M, Lespinet O, Daboussi M, Hua-Van A. 2011. Genome-wide comparative analysis of *pogo*-like transposable elements in different *Fusarium* species. J Mol Evol. 73:230–243.
Etienne RS, Haegeman B. 2012. DDD: Diversity-dependent diversification. R package version 1.2. Available from: http://cran.r-project.org/web/packages/DDD (last accessed January 7, 2013).
Etienne RS, et al. 2012. Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. Proc Biol Sci. 279(1732):1300–1309.
Felsenstein J. 1988. Phylogenies from molecular sequences: inference and reliability. Annu Rev Genet. 22:521–565.
Galagan JE, Selker EU. 2004. RIP: the evolutionary cost of genome defense. Trends Genet. 20(9):417–423.
Han MJ, et al. 2010. Burst expansion, distribution and diversification of mites in the silkworm genome. BMC Genomics 11:520.
Hickey DA. 1982. Selfish DNA: a sexually-transmitted nuclear parasite. Genetics 101:519–531.
Hua-Van A, Le Rouzic A, Boutin TS, Filée J, Capy P. 2011. The struggle for life of the genome's selfish architects. Biol Direct. 6:19.
Johnson LJ, Brookfield JF. 2006. A test of the master gene hypothesis for interspersed repetitive DNA sequences. Mol Biol Evol. 23(2): 235–239.
Kazazian HH Jr. 2004. Mobile elements: drivers of genome evolution. Science 303(5664):1626–1632.
Kendall DG. 1948. On the generalized "birth-and-death" process. Ann Math Stat. 19:1–15.
Kirkpatrick M, Slatkin M. 1993. Searching for evolutionary patterns in the shape of a phylogenetic tree. Evolution 47(4):1171–1181.
Kumar S, Filipski AJ, Battistuzzi FU, Pond SLK, Tamura K. 2012. Statistics and truth in phylogenomics. Mol Biol Evol. 29(2):457–472.
Lander E, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409(6822):860–921.
Le Rouzic A, Boutin TS, Capy P. 2007. Long-term evolution of transposable elements. Proc Natl Acad Sci U S A. 104(49):19375–19380.
Le Rouzic A, Capy P. 2005. The first steps of transposable elements invasion: parasitic strategy vs. genetic drift. Genetics 169: 1033–1043.
Le Rouzic A, Capy P. 2006. Population genetics models of competition between transposable element subfamilies. Genetics 174(2): 785–793.
Le Rouzic A, Deceliere G. 2005. Models of the population genetics of transposable elements. Genet Res Camb. 85:171–181.
Lynch M. 2007. The origins of genome architecture. Sunderland (MA): Sinauer Associates.
Mooers AØ, Heard SB. 1997. Inferring evolutionary process from phylogenetic tree shape. Quart Rev Biol. 72(1):31–53.
Nee S, May RM, Harvey PH. 1994. The reconstructed evolutionary process. Philos Trans R Soc Lond B Biol Sci. 344(1309):305–311.
Orgel LE, Crick FHC. 1980. Selfish DNA: the ultimate parasite. Nature 284: 604–607.
Pace JK, Feschotte C. 2007. The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. Genome Res. 17(4):422–432.
Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.
Quesneville H, Anxolabéhère D. 1998. Dynamics of transposable elements in metapopulations: a model of *P* elements invasion in *Drosophila*. Theor Popul Biol. 54:175–193.
R Development Core Team. 2011. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.
Rabosky DL. 2006. Laser: a maximum likelihood toolkit for detecting temporal shifts in diversification rates from molecular phylogenies. Evol Bioinform. 2:273–276.
Rabosky DL. 2010. Extinction rates should not be estimated from molecular phylogenies. Evolution 64(6):1816–1824.
Ray DA, et al. 2008. Multiple waves of recent DNA transposon activity in the bat, *Myotis lucifugus*. Genome Res. 18(5):717–728.
Ray DA, Platt RN, Batzer MA. 2009. Reading between the lines to see into the past. Trends Genet. 25(11):475–479.

Sanderson MJ. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol Biol Evol. 19(1):101–109.
SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. 1998. The paleontology of intergene retrotransposons of maize. Nat Genet. 20(1):43–45.
Schliep K. 2011. Phangorn: phylogenetic analysis in R. Bioinformatics 27(4):592–593.
Stadler T. 2011. Mammalian phylogeny reveals recent diversification rate shifts. Proc Natl Acad Sci U S A. 108(15):6187–6192.
Wicker T, et al. 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 8(12):973–982.
Wróbel B. 2008. Statistical measures of uncertainty for branches in phylogenetic trees inferred from molecular sequences by using model-based methods. J Appl Genet. 49(1):49–67.
Yule GU. 1924. A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis. Philos Trans R Soc Lond B. 213:21–87.
Zerjal T, Joets J, Alix K, Grandbastien MA, Tenaillon MI. 2009. Contrasting evolutionary patterns and target specificities among three *Tourist*-like MITE families in the maize genome. Plant Mol Biol. 71(1–2):99–114.

Associate editor: Richard Cordaux