bioRxiv preprint doi: https://doi.org/10.1101/513572; this version posted January 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Ecological modeling of Long Interspersed elements reveals footprints of evolution and a role of chromatin in shaping their genome landscape in mammals

Silvia Vitali, Steven Criscione, Claudia Sala, Italo do Valle, Enrico Giampieri, Nicola Neretti, Gastone Castellani.

May 22, 2018

Abstract

Models derived from ecological theories have been applied to describe the dynamics of genomic transposable elements. Long Interspersed Elements (LINEs) are the most abundant class of transposable elements in mammals, still active in humans. A dynamical model is built here and applied to test if LINEs population can be modeled according to the neutral theory of biodiversity. Thereafter, by the introduction of a simple but realistic mechanism of competition for the internal promoters, the model undergoes a spontaneous breaking of the neutral assumption. Despite the apparent simplicity, the proposed model permits to cluster different Species by their Taxonomic Orders; to reveal several footprints of the evolutionary process in mammals, as the radiation of murine subfamily and Primates evolution; and, combined to chromatin state characterization of the LINE copies, to identify host-elements interaction evidences.

1 Introduction

Transposable elements (TEs) are DNA sequences that can move from one genomic location to another and, over evolutionary times increase their copy number within their host genome by several replication mechanisms. They constitute a large portion of most species' genomes, covering roughly 45% of the human genome [7]. Initially considered junk DNA, they are now known to play fundamental roles in the maintenance of genomic diversity and in the reshaping of gene regulatory networks [23], [14], [24], [2], [39], [9]. Currently, TEs activity in humans has been correlated with several genetic diseases and cancer [28], [17], [8].

The effect over the *host organism fitness* of the TEs copy number has been included in many models to study how TEs copies reach fixation in a population and what kind of mechanisms shape TEs abundances [5], [38], [11], [1], [32], [33], [31]. In these approaches the equilibrium between the selection pressure and the birth rate of TEs as well as self-regulation mechanisms are generally considered to limit the number of new insertions. Thus, the copies that reach fixation should have a neutral or eventually a positive effect on the genome [3]. The phenomenon of domestication, i.e., the recruitment by the cell of some TEs to carry out tasks related to their activity as the encoded transposition proteins or regions regulating their expression in the genome, is in fact widely observed in nature [42] [30].

bioRxiv preprint doi: https://doi.org/10.1101/513572; this version posted January 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A large diversity of TE sequences exists, classified in *classes*, *families*, *subfamilies* and *species* on the bases of their transposition mechanism (DNA or RNA-mediated), the proteins they encode, and finally the architecture and similarity of their DNA sequence. The relative abundance of different TEs classes, and, at the same *trophic level*, of TEs species, varies significantly between different host genomes. A recurrent concern is whether such heterogeneity and variability in the amount and diversity of TEs may reflect some organization due to host specific selection pressure, both at genomic and population level, or may be the result of stochastic forces at the level of the individual copies [41], [32]. The interdependence between TEs and their host genome, together with the replication mechanisms of the elements, suggests a strong parallelism between TEs dynamics in their genome and species community dynamics in their ecosystem [41], [35].

Both the niche theory, based on the partitioning of resources, space and time of action between competing species [6], and the neutral theory, in which stochastic mechanisms as demographic stochasticity, migration, and speciation are the most important forces shaping the community [19], have features suitable to describe the TEs ecosystem. Furthermore, TEs ecosystem contains some peculiarities that differentiate it from standard ecosystems [41]. First, TEs create and continuously reshape their own environment because the death copies, i.e., TE copies that lose any transposition ability, are the major part of the genomic landscape in which new copies may insert without deleterious effect on the cell functionalities. Second, the natural selection acts on two levels, the genome level and the host one. The selection at the level of the host induces TEs to evolve traits that constitute a selective disadvantage at the individual level, as for example a lower transposition rate [18] [31]. Phenomena related to the molecular nature of TEs may also occur, for example mutations, insertions and rearrangements, which may lead to functional variations.

Here we focus on the study of a particular family of TEs, the Long Interspersed elements (LINEs), modeling their copy number distribution inside a court of mammalian reference genomes under the hypothesis of *competitive neutrality [*26], i.e., absence of competitive differences among different types of entity. It means that all the copies of all elements in the community (the host genome) are characterized by the same transposition activity, sequence divergence, death rate [41].

LINEs are the most abundant class of TEs in mammals. They belong to the retroelements class, i.e., their replication is RNA mediated (figure 1). Full-length elements contain a promoter region (5'UTR), two protein coding regions (ORF1, ORF2) and a poly-A tail (3'UTR). The internal promoter directs transcription initiation permitting autonomous transposition independently of the specific location in the genome of the copy. However new insertions often result in low fidelity copies of the parent LINE element, with frequent truncations of the promoter region and protein coding defects, with complete or partial loss of the replication ability [8].

For their peculiar replication mechanism and evolutionary history, LINEs ecosystem in mammals results particularly suitable to be described by a birth death process under the neutral hypothesis. LINEs evolved often on a single lineage, in particular in Primates [22], with a subsequent appearance of active elements, making competition between different elements negligible. Coexistence of multiple L1 families or lineages is documented for ancient LINEs [36] and currently in mouse [27], where L1 frequently recruited novel 5'UTR sequences [37], suggesting that simultaneous activity of non-homologous promoters does not introduce a competition between

bioRxiv preprint doi: https://doi.org/10.1101/513572; this version posted January 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the elements. Finally, the genomic environment is unique to each of the TEs copy and full-length L1 copies may differ randomly in their level of transposition activity [4], [34]. Therefore, the stochasticity at the individual level strongly impact the structure of the entire community, supporting the neutral approach to describe the community dynamics.

2 Model

The model implemented to study the distribution of LINEs copy number is based on a master equation approach, in which several degrees of complexity can be included. We start our analysis by developing the most parsimonious model of neutrality. Then, we introduce a competition mechanism between the elements, that temporary breaks the neutral assumption. The model prediction has been tested on transposon abundances data in genomes of 46 mammalian species.

We modeled the way LINEs populated mammalian genomes over the course of evolution as a birth-death process of two interacting species: full-length (autonomous or active) and incomplete (non-autonomous or inactive) LINE copies. The number of active and inactive copies of one element in the host genome are denoted by the by n_A and n_I variables respectively. Active copies can generate new active LINEs by retro-transposition at a rate equal to $b_A n_A$ (birth process of active copies). Over the course of time, mutations and the host selection pressure inactivate active copies at a rate equal to $d_A n_A$ (death process of active copies). Some transposition events are incomplete such that the inserted copy is incapable of autonomous retro-transposition; for example, L1 insertions are often 5'-truncated (e.g. Figure 6B of [10]). We refer to the rate at which this process occurs as $b_I n_A$. Furthermore, a transcribed incomplete copy can hijack the retro-transposition machinery of autonomous copies to duplicate in a new location, a process called trans-complementation that can occur at a rate equal to $b_{AI} n_A n_I$. This phenomenon has been observed, for example, in LINE-1 retroelements, although it should happen at a much smaller rate than retro-transposition in *cis* [44]. Then, the birth rate of incomplete copies is given by $(b_{AI} + b_I/n_I) n_A n_I$ (birth process of inactive copies). The host selection may also prevent the fixation of copies that negatively affect its fitness by reducing this rate. Finally, when many mutations accumulate, inactivated copies cannot be distinguished from the random background of genomic sequences and they essentially disappear from the genome. This can also occur through excision of large genomic regions as well as by structural rearrangements. We refer to this process as the death rate of the incomplete copies d_I. This birth-death stochastic process can be described via the following two-dimensional master equation

$$\frac{dP}{dt} = (E_{nA}^{-} - 1) b_A n_A P + (E_{nA}^{+} E_{nI}^{-} - 1) d_A n_A P + (E_{nI}^{-} - 1) b_I n_A P + (E_{nI}^{-} - 1) b_{AI} n_A n_I P + (E_{nI}^{+} - 1) d_I n_I P \qquad (1)$$

where $P \equiv P(n_A, n_I, t)$, $E^{\pm} f(n) = f(n \pm 1)$ are called step operators [40] and their lower index indicates the variable on which the operator acts. Each term in the sum on the right-hand side of equation (1) represents one of the biological mechanisms described above (birth and death for active and inactive copies and trans-complementation).

The stationary distribution of (1) has not a closed form, however, it is possible to compute the marginal distributions, corresponding to the n_A and n_I species. Because, according to the available

bioRxiv preprint doi: https://doi.org/10.1101/513572; this version posted January 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

data, we can observe only n_I, we are interested in the characterization of the stationary distribution of the defective copies: $P(n_I)$.

The distribution $P(n_I)$ corresponds, indeed, to the relative species abundance (RSA) of the inactive species, i.e., the probability to observe a specie with a certain number of individuals in a community, to which it will be referred for sake of clarity as P_{nI}.

When equilibrium is reached for both active and inactive copies, the stationary solution (the RSA) is a negative binomial distribution. In fact, taking n_A as a constant the equation (1) for n_I reduces to a well-known ecological neutral model, described in [43].

By neglecting excision and trans-complementation processes a further and very interesting special case of the model is obtained. In the case that equilibrium for active species does not hold ($b_A << d_A$) the stationary solution P_{nI} is well approximated by a negative binomial distribution when the "absorbing state" for active copies is reached, with the notable difference that the expected value of the parameters is different from the previous case.

If trans-complementation does not produce a relevant contribution and the system is out of equilibrium, we expect to observe an RSA following a negative binomial distribution with a "number of failures" parameter close to one ($\Upsilon \sim 1$). Instead, if the trans-complementation process is relevant, the number of expected failures is much greater than one ($\Upsilon \sim (b_I + d_A)/b_{AI} >> 1$), due to the experimental observation that trans-complementation events are rarer if compared to retro-transposition in *cis*.

The distribution obtained must be normalized to one after removing the probability to observe zero individuals (the unobserved species). Mathematical details of the derivation and numerical simulations can be found in the Supplementary Material.

2.1 Competition mechanism

We propose a competition between two elements activated by the same promoter region as a stochastic process able to determine a deviation from the expected distribution, in particular generating a bimodal behavior described by a mixture of negative binomials.

When two elements (or more) characterized by the same promoter are contemporaneously active with n_1 and n_2 number of full-length copies respectively, the probability P_{poly} of a single polymerase to switch on one copy is rescaled by the fraction of active copies belonging to that specie: $P_{poly,1} = P_{poly} \cdot n_1/(n_1+n_2)$ and vice-versa. Thence the transposition rates b_A, b_I and b_{AI} are reduced by the same factor. A smaller birth rate will result in a lower abundance for the competing LINE species, in comparison to the elements not affected by the competition mechanism.

The introduction of the competition mechanism spontaneously breaks the hypothesis of neutrality, because in principle all LINEs species are equivalent in the model. The simultaneous activation of elements sharing the same promoter introduces a disadvantage for the species that compete for the molecular machinery, breaking the neutral assumption until the extinction of one of the competitors. We surmise that the distribution arising from competition is a mixture of two negative binomials, ad this is confirmed by numerical simulation for a range of parameters compatible with real data. (Supplementary Material),

bioRxiv preprint doi: https://doi.org/10.1101/513572; this version posted January 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$P_{RSA} = \alpha \cdot P_{rare} + (1 - \alpha) \cdot P_{abund} \qquad (2)$$

where α is the mixture coefficient, related to the probability that two elements compete, and with P_{rare} and P_{abund} representing respectively the population of rare elements, defined in our model by the losers of the competition, and the population of the abundant elements.

3 Results

The RSA of LINEs in 46 mammals' genomes have been fit by a negative binomial and a mixture of negative binomial through the approximate Bayesian computation method (ABC) to test neutral and competition models respectively. From the same prior distribution, we described 42 datasets, with the notable exception of Wallaby, Tasmania devil, Opossum, and Platypus, which are in fact the most isolated species of the group from an evolutionary point of view and are characterized by a small LINE's ecosystem. For most of the organisms under study, the ABC model selection score was comparable for the two models tested. Results of parameters posteriors distribution and comparison with RSA for each dataset are shown in Supplementary Materials.

ABC method already penalizes the model with the higher number of parameters because the phase-space is larger, and the two model results equivalently suitable from a purely statistical perspective. The law of parsimony supports the choice of the simplest theory when two alternatives possess equivalent power to describe the data. However, the mixture model allows a better characterization of the biological phenomena, suggesting that the introduction of the second component in the PDF enhance the ability to extract useful information from the data. Parameters associated to the mixture of two negative binomials (competition model) permit to separate the Host Species at the level of their taxonomic Order as can be seen in figure 2, while the neutral model does not produce such a good separation. We show in figure 2 only the most populated Orders ($n > 2$) for statistical reasons. In particular, the couples of parameters Υ_1, Υ_2 and x_1, x_2 seem a good representation to discriminate the host organisms in different taxonomic Orders. Within our description, such couples of parameters are related by the value of the disadvantage due to competition, on average. The values of the Υ_1, Υ_2 parameters, all close to one, indicates that a pure accumulation process is more convincing. Despite trans-complementation may take place, up to this description it is not a very relevant process in shaping the RSA of the community. Furthermore, the observation of negative binomial distribution within such model supports the idea that equilibrium between host and Elements do not hold, but a competition between host and Elements takes place.

Since transposons activity deeply contributes to shaping genomes and LINE's horizontal transfer is uncommon, we expect that LINEs abundance result a good indicator for phylogeny in mammals, in which this retro-transposon family is particularly abundant. This can be noticed by hierarchical clustering of the TEs abundances in different organisms (supplementary figure). Host species closer related possess more similar LINE's abundances because they have been inherited more recently from a common ancestor. For this reason, it cannot be excluded that the better performance of the mixture model is due to a better characterization of the statistical fluctuations in the RSA that have been inherited by common ancestors.

The ecosystem under study is composed, with a certain approximation, by the frozen populations of copies that reached fixation in the genomes after they lost transposition ability. LINE's species

bioRxiv preprint doi: https://doi.org/10.1101/513572; this version posted January 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

evolve inside and beside the genome with a turnover of active LINEs. By the knowledge on the time ranges of activity of different LINEs, we studied the evolution of such ecosystem as a function of time, in order to further investigate and identify time windows of competition. Data available from [16] have been used to order LINEs by their age of activity in human, chimp, rhesus macaque, mouse and rat datasets.

The list has been subdivided into time intervals (windows) containing a fixed number of elements. The number of elements in the sliding window can be arbitrarily chosen. We tested several windows lengths, from N = 15 to 40. The results of the fit were compatible for all the lengths, with smoother trends in time for larger windows. Here we show the results for N = 15, which is a compromise between a zoom in the action period and numerousness of the ecosystem for analysis purposes. Each window represents a picture of the RSA in a different evolution stage. This time-dependent ecosystem has been tested respect to the neutral model and the mixture model with the same ABC method and same prior distributions.

It results that in the primates under study neutrality is violated between the 40-65 interval of the rank, where the mixing coefficient of the mixture model is higher, and ABC model selection score suggests that a more complex description is suitable (figure 3). Similar results can be found for mouse and rat, where a preference for the mixture model is maintained in all ancient LINEs. The trend for the birth-death ratio and the influx parameters are reported in the supplementary materials. We clustered hierarchically the abundances of the elements belonging to competition window in primates (supplementary figure), and we observed that the bimodal behavior of rare and abundant elements is approximately maintained across all mammals included in this study, with the exception of the White Rhinoceros, which shows the opposite trend. Furthermore, species results mostly clustered by their taxonomic orders.

To test the hypothesis that competition is associated to similarity in the promoter region, we aligned pairwise the available consensus sequences of the 5'UTR from RepBase (http://www.girinst.org) [21] using ClustalW2 [25] and for each couple we calculated the distance between the promoter sequences. Significant similarity between 5'UTRs is observed for the following high and low copy number pairs (or group): L1M2-L1M2c and L1MA9-L1M3a-L1M3b-L1M3c-L1M3d (figure 4).

The expectation value of the copy number for the LINEs RSA is defined by the parameters of the negative binomial distribution by the relation:

$$\langle n_I \rangle = \frac{x}{1-x} \cdot Y \qquad (3)$$

In the case of the mixture model the parameters of the distribution of the abundant species can be employed. A transition in time to different average abundances can be identified for both the group of primates and rodents, as shown in figure 6, where the parameters of the negative binomial distributions describing the sliding windows RSA are plotted.

In the primates under study, we observe a transition to a lower average copy number, while in rat and mouse to a larger copy number. This transition can be observed in the neutral model as well as in the mixture model, in the component describing the elements with high copy number. The

bioRxiv preprint doi: https://doi.org/10.1101/513572; this version posted January 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

group associated with rare elements results much noisier, with parameters less correlated. The transition is more evident in the mixture description, for this reason, again, the mixture model looks more suitable to describe the system.

The impact of LINE retro-transposons on the whole structure of the genome can be partially quantified by studying the chromatin state of the genome where the insertions lie. Using chromatin state assignments in human [12] and mouse [45] genomes, and the coordinates of the respective TEs insertions from RepBase , we assigned to each LINEs copy a chromatin state, distinguishing between insertions in open and closed chromatin states, currently known as *euchromatin* and *heterochromatin* respectively. Since new insertions fixate in the germ line only, we refer to the state assignment in the embryonic stem cell. Multiple assignments have been treated classifying the combination of states into open, weakly open and closed chromatin, depending if the states identified belong mainly to one of these groups. Weakly open chromatin population has been added one time to open chromatin and after to closed chromatin and we found that this choice did not change significantly our results. The unknown state has been included in the closed chromatin group, which encloses the majority of the copies.

The average percentage of LINE copies inserted in euchromatic regions in the sliding window displays a decreasing trend with time ordered age in human and mouse (figure 5). However, in humans, it also shows a clear peak within the neutral time interval. The average percentage of copies in euchromatic regions is bigger for the windows with higher average copy number respect to the one with low copy number in human and in ancient elements in mouse. The presence of a higher fraction of rare species within the non-neutral time interval results then in agreement with the lower average percentage of insertions in euchromatin observed.

This relation is clarified considering the correlation between the number of insertions in euchromatin and the number of insertions in heterochromatin (figure 7). The linear correlation between the logarithm of the counts corresponds to a correlation of power-law type between the raw counts (any age assignment is considered):

$$N_{eu} = 2^{c_0 \pm \epsilon} N_{het}^{c} \qquad (4)$$

For human we have c = 1.18, c_0 = -4.58 and ϵ = 0.035, which correspond to the standard error in the estimate. The correlation coefficient is r = 0.96 and the p-value $p \sim 10^{-55}$. The superlinear correlation between the two quantity leads to the interesting result that a higher abundance, i.e., the sum of euchromatin and heterochromatin contributions, is related to a higher percentage of insertion in euchromatin states. The average trend of elements abundance in time result in fact slightly decreasing as well in figure 3. We hypothesize that the decreasing trend can be caused by the host selection pressure which on average select less invasive transposons.

In figure 5, referred to mouse, we observe a plateau in the percentage of euchromatin LINEs insertions followed by a decrease. However, the average abundance at a certain point drastically rears up. This corresponds in figure 7 to a transition to a different value of the coefficient c_0 in the correlation plot for mouse. For ancient LINE species the correlation between the number of insertions in euchromatin and the number of insertions in heterochromatin is closed to the one

bioRxiv preprint doi: https://doi.org/10.1101/513572; this version posted January 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

observed in human. Instead, more recent elements fall in a well separated cluster, associated to lower c_0 value, as it is highlighted by principal component analysis (PCA) too.

Given the same number of insertions in euchromatin states, a lower value for c_0 corresponds to a larger abundance, and, consequently, to a lower percentage of insertions in euchromatin.

The beginning of the transition in figure 7 for mouse is contemporary to the transition to higher average LINEs abundance shown in figure 5 and figure 6, and corresponds to the appearance of the LINE family Lx. Where the amplification of the LINE family Lx is associated with the murine subfamily radiation ~ 12 Myr according to [29] and [15]. In fact, the other elements characterizing this group are mainly murine specific.

In figure 7 referred to human, there is not a sharp transition between two different chromatin state distributions as observed for mouse. This is reasonable if we look at the problem from the perspective of the host organism fitness. A transition to a higher average copy number surely have a bad impact on the host fitness, because the probability of deleterious insertions in the genome increases, if it is not compensated by some self-regulatory strategy or host silencing mechanism. The combined sharp transitions observed in mouse agrees with this idea, and perhaps are the result of the host-elements interaction. Instead, in human we observe a transition to a lower average copy number (figure 5 and figure 6). This could be the reason why chromatin states distribution in human is not affected significantly, since further changes where not necessary to preserve the host fitness. The most ancient elements involved in such transition are indeed related to the evolutive differentiation of Primates, associated with the amplification of LIMA/LPB subfamilies ~ 70 - 100Myr (Khan, 2006).

4 Concluding Remarks

We analyzed and tested two variants of the neutral model proposed, with and without competition, to describe LINEs communities over 46 mammalian genomes, focusing on two particular but realistic regimes leading to negative binomial distribution, or a mixture of two.

We showed that the neutral model proposed can highlight some fundamental features of the dynamics of LINEs in mammalian genomes. Furthermore, the introduction of stochastic competition between elements reduces the level of noise permitting to distinguish different taxonomic orders on the bases of our model description.

The study of the evolution of the ecosystem by stratifying LINE subfamilies by age groups suggests that, at specific times during the evolution in the mammalian genome, multiple concurrently active LINE subfamilies might have been in direct competition for the promoter region. This approach also permits to identify evolutive transitions both in the three primates that in mouse and rat, supported by the characterization of the chromatin landscape of the elements and associated to the amplification of specific LINE's families.

bioRxiv preprint doi: https://doi.org/10.1101/513572; this version posted January 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The hypothesis that competition could have been shaped by the LINE 5'UTR structure is supported by the similarity measures of the 5'UTR sequences in concurrently active LINEs.

We want to stress at this point that the mechanism of competition proposed between LINE species is independent of the chromatin state distribution of the element copies, but act at the level of the elements affecting their abundances. Instead, chromatin state of the insertions should reflect the interaction of the element with the host, by mechanisms of silencing and self-regulation, affecting, as an example, the average abundance of elements in a genome and their chromatin states distribution.

Furthermore, the concept of neutrality we propose, that is that all elements have the same birth death rates, can be relaxed if elements are active mainly one by one, rescaling the time by a constant so that the rates can be properly adjusted.

Moreover, up to our model description, it appears that the equilibrium between host and elements does not hold. Thence LINEs possibly contribute as one of the driving forces of the evolutionary process, and as a source of innovation and variation in the genome equilibria, with deep impact on the fitness of the hosts. In addition, competition with a stochastic origin introduces deviations from the neutrality proposed between the elements supporting the introduction of variation in the spreading element sequences.

Finally, our approach also reveals several interesting footprints of evolution, and despite its simplicity, it results a powerful method for the characterization of genome landscapes, with promising applications in future studies.

5 Materials and Methods

5.1 Data sources

LINE abundances were calculated using RepeatMasker annotation (http://www.repeatmasker.org) for human genome build hg19 and 45 other mammalian species. LINE consensus sequences were downloaded from RepBase [20, 21] (http://www.girinst.org). Only a subset of the LINE consensus sequences contains the 5' UTR annotated in the RepBase associated report, which was selected for the analysis of LINE 5' UTR sequence. Chronological ordering of LINEs in human, chimp, rhesus macaque, mouse and rat was derived from Giordano et al. [16]. Chromatin structure data available for mouse [45] and human [12] were used to assign genomic copies of LINEs to open and closed chromatin states by knowledge of their coordinates in the reference genome. In the cited references chromatin structure assignment was conducted using ENCODE chromatin models using the ChromHMM method [13].

5.2 Statistical methods

The two models proposed have been tested on data by the implementation of an approximate Bayesian computation (ABC) method to fit the RSAs included in this study. The probability of the model given the data is $P(M|D) = P(D|M)\ P(M)$. In ABC $P(D|M)$ can be approximated by the ratio of the number of successes, correspondent to the number of accepted set of parameters, over the number of attempts for each model. The a priori probability for the two models is defined equal, then we can approximate the ABC model selection score to be the ratio of the fraction of successes for the two models.

bioRxiv preprint doi: https://doi.org/10.1101/513572; this version posted January 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

5.3 Numerical simulations

To test if the dynamical model can generate a negative binomial distribution beyond the given assumptions, we performed numerical simulations. We used Gillespie algorithm for the active copies' dynamics, for the inactive copies dynamics we used the tau-leap algorithm in the case of a pure accumulation process and a hybrid algorithm to simulate the dynamics of inactive copies with trans-complementation. The hybrid algorithm instead of a Gillespie was chosen to reduce the time of computation. It consists in the estimation of the expected number of inactive copies by ODE numerical integration to estimate the expected birth and death rates in that time interval and use a tau-leap algorithm to generate the stochastic increment associated to each time interval. Oracle comparison to the theoretically correct Gillespie algorithm was performed to test the accuracy of the hybrid simulation method. Simulations of the competition mechanism in both the regimes were performed to check if the solution was compatible with a mixture of negative binomials.

6 Supplementary Material

Supplementary tables and figures are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

7 Acknowledgments

This work was supported by the Italian Ministry of Education at University of Bologna (Alma Mater Studiorum), Department of Physics and Astronomy (DIFA), and in part by the following NIH grants: R56 AG050582-01 to N.N. and F31AG050365 to S.W.C., S.W.C. was also supported by the NIH Institutional Research Training Grant T32 GM007601. We aknowledge IMforFuture EU project and HARMONY EU project

References

[1] G. Abrusán and H. J. Krambeck. Competition may determine the diversity of transposable elements. *Theoretical Population Biology*, 70(3):364–375, 2006.

[2] C. Biémont. A brief history of the status of transposable elements: From junk dna to major players in evolution. *Genetics*, 186(4):1085–1093, 2010.

[3] S. Boissinot and A. V. Furano. Adaptive evolution in LINE-1 retrotransposons. *Journal of Molecular Biology and Evolution*, 18:2186–2194, 2001.

[4] B. Brouha et al. Hot l1s account for the bulk of retrotransposition in the human population. In *Proceedings of the National Academy of Sciences USA*, pages 5280–5285, 100, 2003.

[5] B. Charlesworth and C. H. Langley. The population genetics of drosophila transposable elements. *Annual review of genetics*, 23:251–287, 1989.

[6] J. M. Chase and M. A. Leibold. *Ecological Niches: Linking Classical and Contemporary Approaches*. University Press, Chicago, 2003.

bioRxiv preprint doi: https://doi.org/10.1101/513572; this version posted January 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[7] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[8] R. Cordaux and M.A. Batzer. The impact of retrotransposons on human genome evolution. *Nature Review Genetics*, 10(10):691–703, 2009.

[9] M. Cowley and R. J. Oakey. Transposable elements re-wire and fine-tune the transcriptome. *PLoS Genetics*, 9:1, 2013.

[10] S. W. Criscione, Y. Zhang, W. Thompson, J. M. Sedivy, and N. Neretti. Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC genomics*, 15:583, 2014.

[11] G. Deceliere, S. Charles, and C. Biémont. The dynamics of transposable elements in structured populations. *Genetics*, 169(1):467–474, 2005.

[12] J. Ernst and M. Kellis. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature biotechnology*, 28(8):817–825, 2010.

[13] J. Ernst and M. Kellis. Chromhmm: automating chromatin-state discovery and characterization. *Nature Methods*, 9:215–216, 2012.

[14] C. Feschotte. The contribution of transposable elements ot the evolution of regulatory networks. *Nature Reviews Genetics*, 9(5):397–405, 2008.

[15] A. V. Furano, B. E. Hayward, P. Chevret, F. Catzeflis, and K. Usdin. Amplification of the Ancient Murine Lx Family of Long Interspersed Repeated DNA Occurred During the Murine Radiation. *Journal of Molecular Evolution*, 38:18–27, 1994.

[16] J. Giordano, Y. Ge, Y. Gelfand, G. Abrusan, G. Benson, and P. E. Warburton. Evolutionary history of mammalian transposons determined by genome-wide defragmentation. *PLoS Computational Biology*, 3(7):1321–1334, 2007.

[17] J. Gonzalez and D. A. Petrov. Evolution of genome content: Population dynamics of transposable elements in flies and humans. *Methods in Molecular Biology*, pages 361–383, 2012.

[18] J. S. Han and J. D. Boeke. A highly active synthetic mammalian retrotransposon. *Nature*, 429:314–318, 2004.

[19] S. P. Hubbell and L. B. de Água. The unified neutral theory of biodiversity and biogeography: reply. *Ecology*, 85(11):3175–3178, 2004.

[20] J. Jurka. Repbase update: A database and an electronic journal of repetitive elements. *Trends in Genetics*, 16(9):418–420, 2000.

[21] J. Jurka, V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. Repbase update, a database of eukaryotic repetitive elements. *Cytogentic and Genome Research*, 110:462–467, 2005.

[22] H. Khan, A. Smit, and S. Boissinot. Molecular evolution and tempo of amplification of human line-1 retrotransposons since the origin of primates. *Genome Research*, 16(1):78–87, 2006.

bioRxiv preprint doi: https://doi.org/10.1101/513572; this version posted January 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[23] M. G. Kidwell and D. Lisch. Transposable elements as sources of variation in animals and plants. In *Proceedings of the National Academy of Sciences USA*, pages 7704–7711, 94(15), 1997.

[24] G. Kunarso, N. y. Chia, J. Jeyakani, C. Hwang, X. Lu, Y. s. Chan, H. h. Ng, and G. Bourque. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature Genetics*, 42(7):631–634, 2010.

[25] M.A. Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, and D.G. Higgins. Clustal w and clustal x version 2.0. *Bioinformatics*, 23(21):2947–8, 2007.

[26] S. Linquist, K. Cottenie, T. A. Elliott, B. Saylor, S. C. Kremer, and T. R. Gregory. Applying ecological models to communities of genetic elements: The case of neutral theory. *Molecular Ecology*, 24(13):3232–3242, 2015.

[27] M. L. Mears and C. A. Hutchinson. The evolution of modern lineages of mouse L1 elements. *Journal of Molecular Evolution*, 52:51–62, 2001.

[28] M. Munoz-López and J. L. García-Pérez. DNA transposons: nature and applications in genomics. *Current genomics*, 11(2):115–28, 2010.

[29] E. Pascale, E. Valle, and A. V. Furano. Amplification of an ancestral mammalian LI family of long interspersed repeated DNA occurred just before the murine radiation. *Proceedings of the National Academy of Sciences USA*, 87:9481–9485, 1990.

[30] R. Rebollo, M. T. Romanish, and D. L. Mager. Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes. *Annual Review of Genetics*, 46:1, 2011.

[31] A. Le Rouzic, T. S. Boutin, and P. Capy. Long-term evolution of transposable elements. *Proceedings of the National Academy of Sciences USA*, 104(49):19375–19380, 2007.

[32] A. Le Rouzic and P. Capy. The first steps of transposable elements invasion: Parasitic strategy vs. genetic drift. *Genetics*, 169(2):1033–1043, 2005.

[33] A. Le Rouzic and P. Capy. Population genetics models of competition between transposable element subfamilies. *Genetics*, 174(2):785–793, 2006.

[34] M. C. Seleme et al. Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity. *Proceedings of the National Academy of Sciences USA*, 103:6611–6616, 2006.

[35] F. Serra, V. Becher, and H. Dopazo. Neutral Theory Predicts the Relative Abundance and Diversity of Genetic Elements in a Broad Array of Eukaryotic Genomes. *PLoS ONE*, 8:6, 2013.

[36] A. F. Smit, G. Tóth, A. D. Riggs, and J. Jurka. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *Journal of molecular biology*, 246(3):401–417, 1995.

[37] A. Sookdeo, C. M. Hepp, M. aMcClure, and S. Boissinot. Revisiting the evolution of mouse line-1 in the genomic era. *Mobile DNA*, 4(1), 2013.

bioRxiv preprint doi: https://doi.org/10.1101/513572; this version posted January 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[38] C.J. Struchiner, M.G. Kidwell, and J.M.C. Ribeiro. Population dynamics of transposable elements: copy number regulation and species invasion requirements. *Journal of Biological Systems*, 13(4):455–475, 2005.

[39] A. Testori, L. Caizzi, S. Cutrupi, and O. Friard. M. De Bortoli, dcora', and m. caselle. *The role of Transposable Elements in shaping the combinatorial interaction of Transcription Factors BMC genomics*, 13(1), 2012.

[40] N. G. van Kampen. *Stochastic Processes in Physics and Chemistry*. North-Holland, Amsterdam, 1981.

[41] S. Venner, C. Feschotte, and C. Biémont. Dynamics of transposable elements: towards a community ecology of the genome. *Trends in Genetics*, 25(7):317–323, 2009.

[42] J. N. Volff. Turning junk into gold: Domestication of transposable elements and the creation of new genes in eukaryotes. *BioEssays*, 28(9):913–922, 2006.

[43] I. Volkov, J. R. Banavar, S. P. Hubbell, and A. Maritan. Neutral theory and relative species abundance in ecology. *Nature*, 424(6952):1035–1037, 2003.

[44] W. E. I. Wei, N. Gilbert, S. L. Ooi, J. F. Lawler, E. M. Ostertag, H. H. Kazazian, J. E. F. D. Boeke, and J. V. Moran. Human L1 Retrotransposition:. *Society*, 21(4):1429–1439, 2001.

[45] F. Yue et al. A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, 515(7527):355–64, 2014.

8 Figures

bioRxiv preprint doi: https://doi.org/10.1101/513572; this version posted January 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

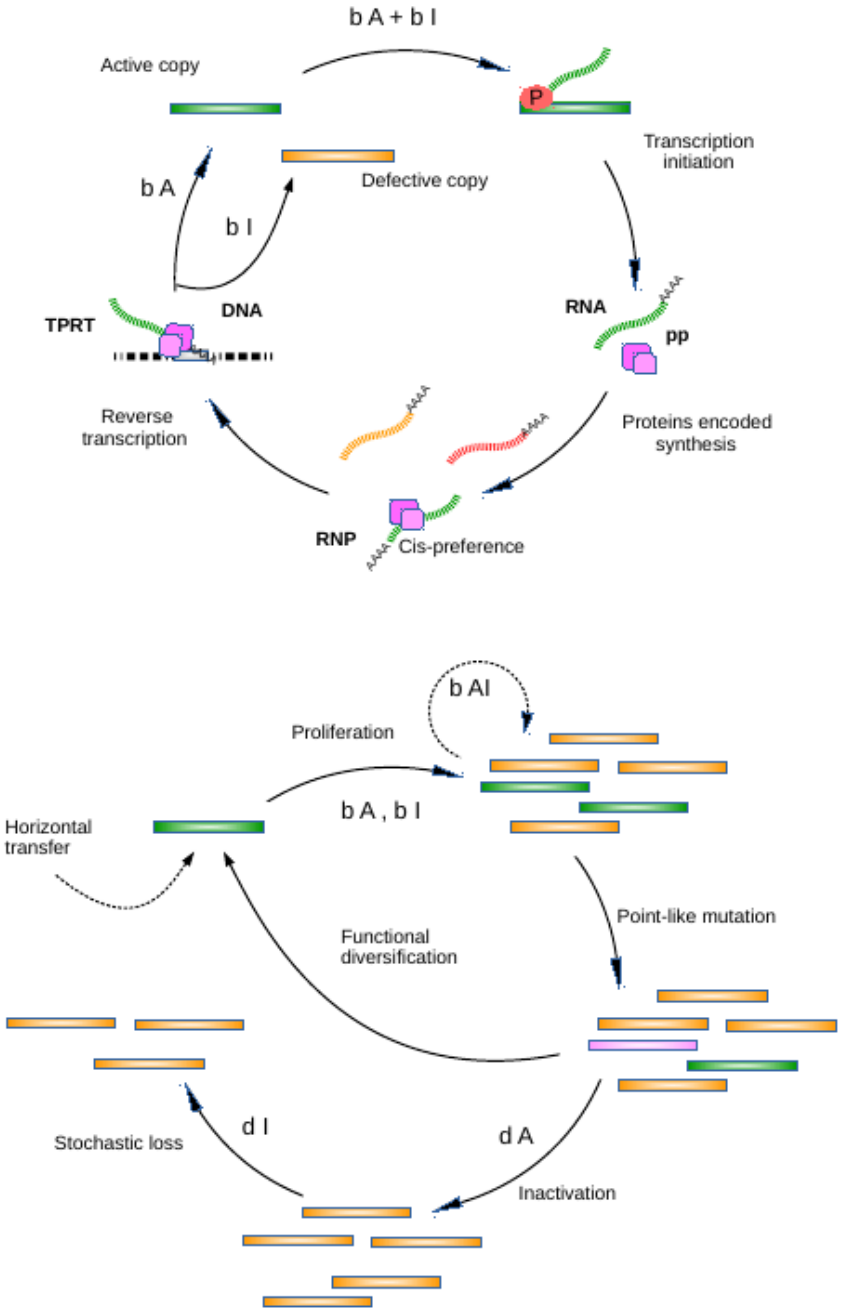

Figure 1: **LINEs transposition and activity cycle diagram.** The diagram of the birth process of defective and full-length copies from a full-length master copy by retro-transposition and element life cycle is shown. a) When the transcribed RNA reaches the cytoplasm, the protein encoding regions ORF1 and ORF2 are translated to an RNA-binding protein and a protein with endonuclease and reverse-transcriptase activities, respectively. Both proteins show a strong *cis*-preference; consequently, they preferentially associate with the RNA transcript that encoded them to produce what is called a ribonucleoprotein (RNP) particle. After coming back into the nucleus, the proteins

bioRxiv preprint doi: https://doi.org/10.1101/513572; this version posted January 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

on RNA can open a nick in DNA and produce a DNA copy of the template through a process termed target-primed reverse transcription (TPRT). The resulting new insertion is a low fidelity copy of the parent LINE element with frequent 5' truncations and protein coding defects, and often losing replication capacity [8].

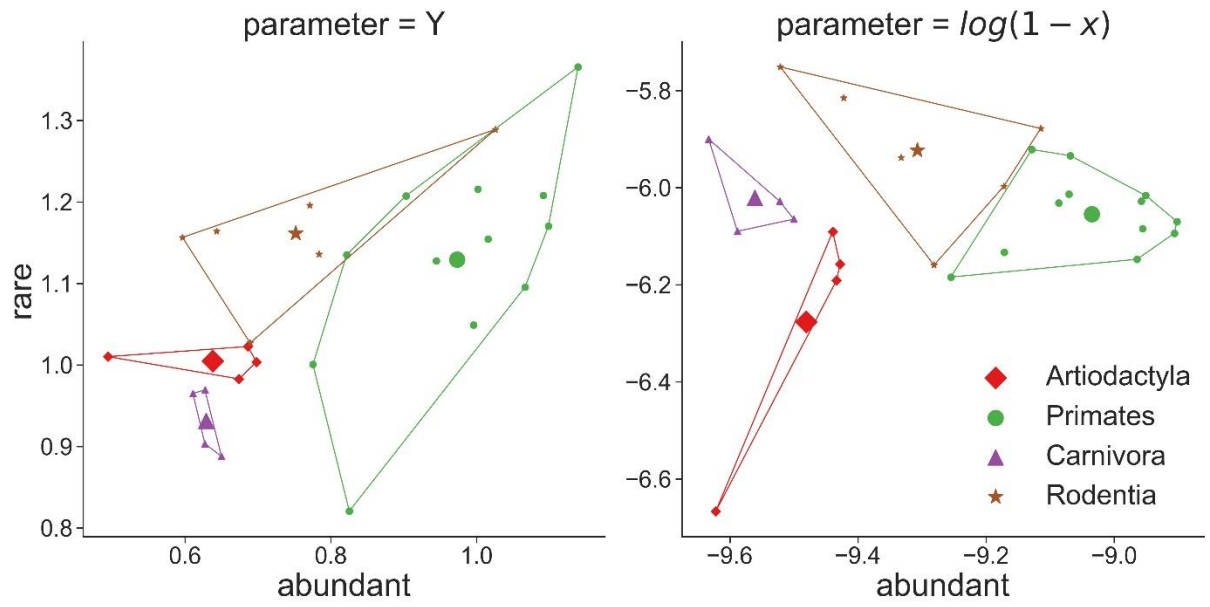

Figure 2: **Fit of the RSA of LINEs with a mixture model clusters different mammalian Orders.** The set of optimized parameters obtained by fitting a mixture model of two negative binomials on mammalian LINE RSAs are able to separate the most represented Taxonomic Orders: Υ abundant respect to Υ rare and $(1 - x)$ abundant respect to $(1 - x)$ rare respectively.

bioRxiv preprint doi: https://doi.org/10.1101/513572; this version posted January 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

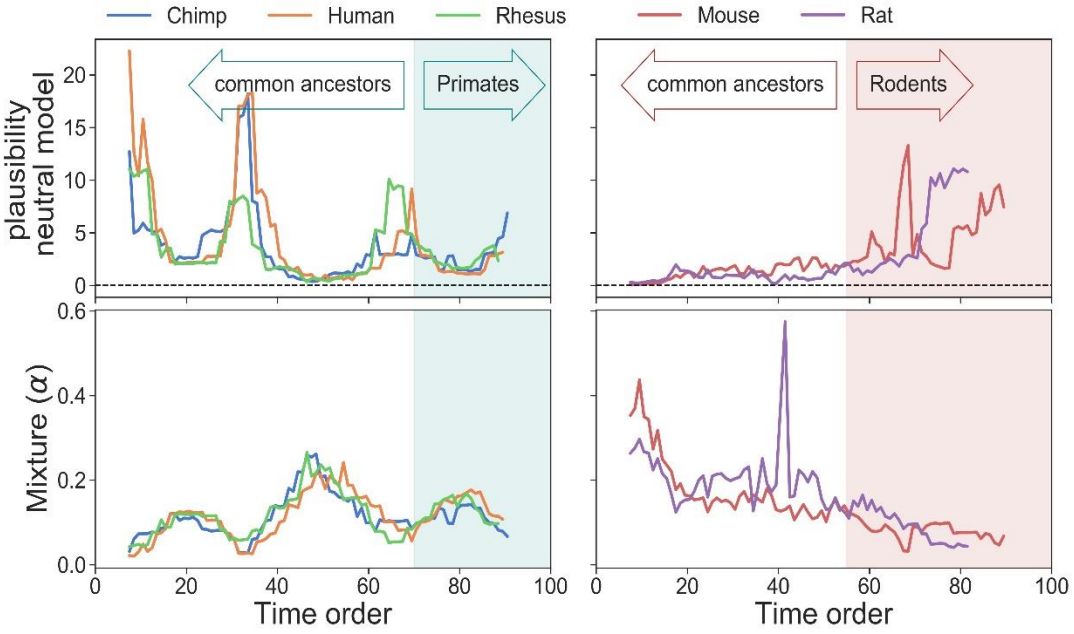

Figure 3: **Comparison of the ABC model selection score with the mixture coefficient during the evolution of the LINE's ecosystem.** Data available from [16] have been used to rank order LINE elements by their age of activity. The rank has been subdivided into intervals containing a fixed number of contiguous elements (N = 15), each interval has been used as a sample ecosystem to fit both the neutral model and the mixture model. Upper panels: The ABC model selection approach was used to compare the goodness of fit between the two models. The ABC model selection approach shows that a mixture model provides a better fit between rank positions 40 and 65 of the time ordered age of LINEs (" non-neutral time interval"). Lower panels: Estimation of the mixture coefficient *a* during evolution. This coefficient represents the proportion of species associated with the regime in the mixture model, described by the negative binomial with a lower mean (i.e., LINE subfamilies with fewer elements in the genome). When *a* is higher, it indicates a significant presence of rare LINE's species.

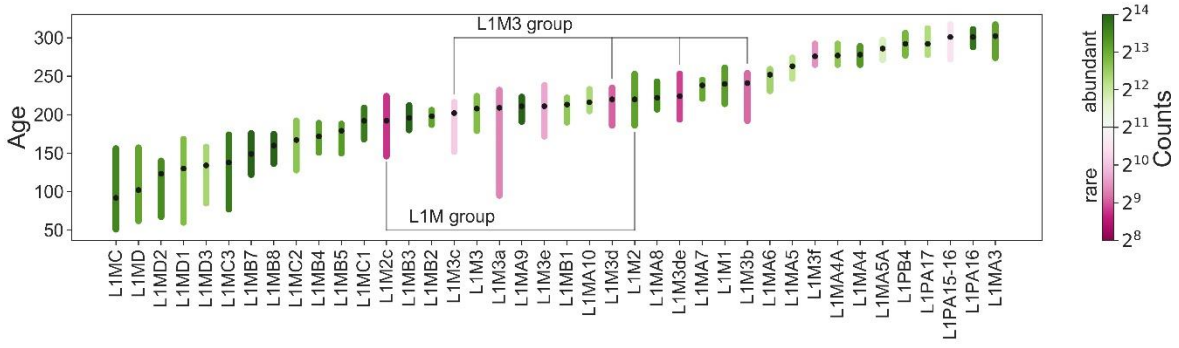

Figure 4: **5'UTR similarity between competing LINE retro-transposons in human.** In [22] is suggested that different L1 elements may coexist without competing if the 5'UTR is different,

bioRxiv preprint doi: https://doi.org/10.1101/513572; this version posted January 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

while a species will overcome the others if the 5'UTR are very similar and might compete for the same factors. The available consensus sequences of the 5'UTR of LINEs in the human genome have been aligned pairwise, with ClustalW2. The minimum distance is achieved between couples (or groups) of elements with similar ages and having high and low copy number respectively. Range of activity shown by the green bars. Abundance shown by the color legend. Significant similarity between 5'UTRs is observed for the following high and low copy numbers groups: L1M2-L1M2c (black stars), L1M3a-L1M3b-L1M3c-L1M3d (red dots).

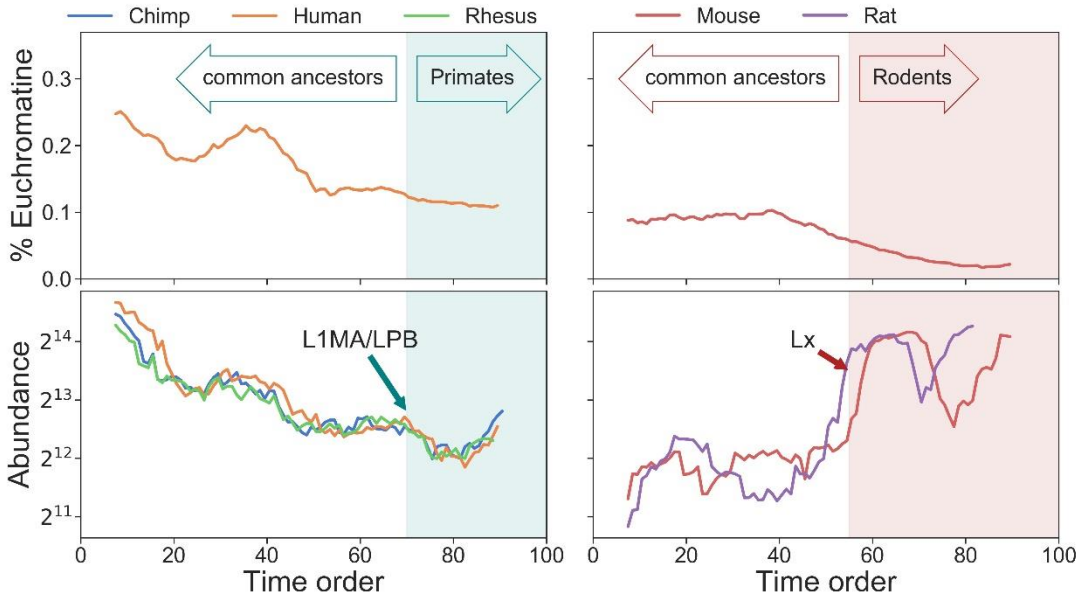

Figure 5: **Comparison of insertion's percentage inside euchromatin and expected abundance during the evolution of the LINE's ecosystem.** Data available from [16] have been used to rank order LINE elements by their age of activity. The rank has been subdivided into intervals containing a fixed number of contiguous elements ($N = 15$), each interval has been used as a sample ecosystem to fit both the neutral model and the mixture model. Upper panels: The percentage of LINE copies inserted in euchromatic regions displays a decreasing trend with time ordered age in human. Lower panels: The copy number of LINE (euchromatin and heterochromatin insertions) displays a decreasing trend with time order as well.

bioRxiv preprint doi: https://doi.org/10.1101/513572; this version posted January 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

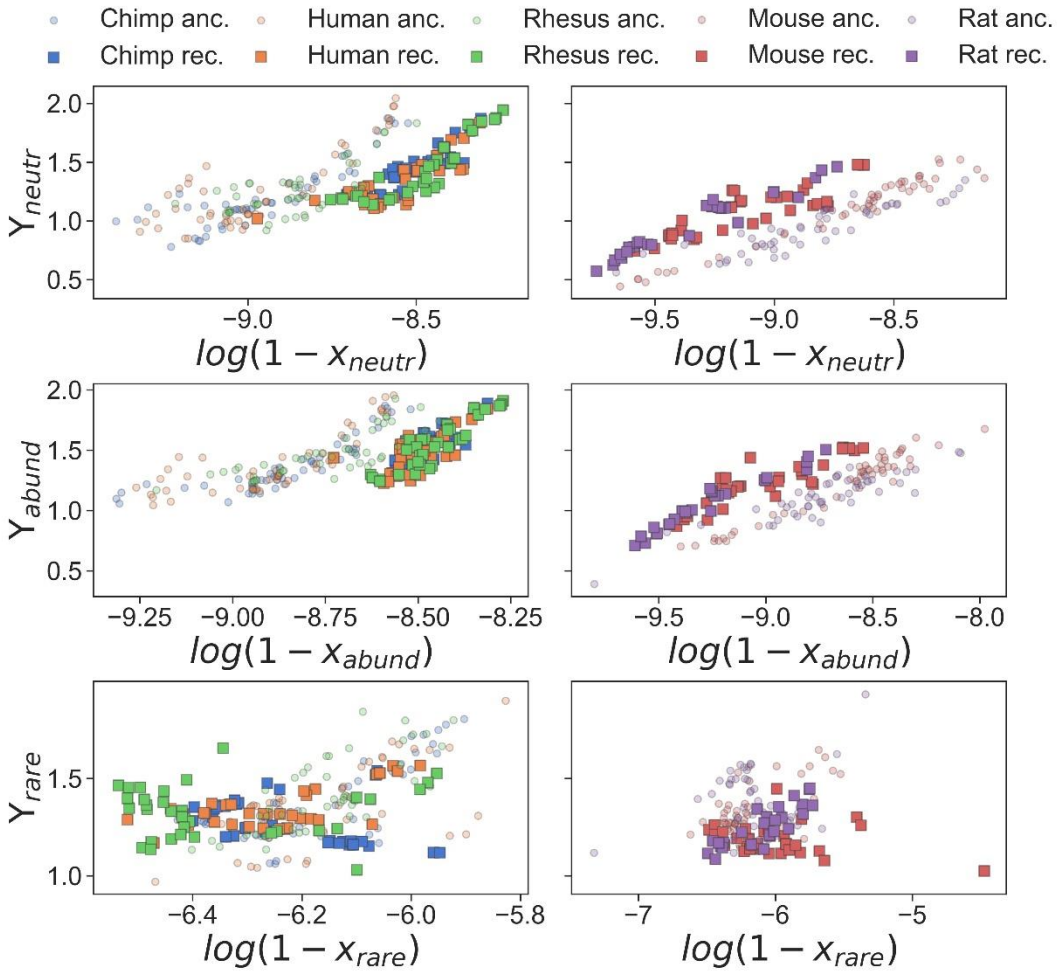

Figure 6: **Space of parameters of LINEs evolution in three primates and two murine genomes shows evidences of radiation events.** The space of parameters describing sliding window ecosystem of LINEs in human, chimpanzee, rhesus macaque (left panels) and mouse, rat (right panels) is shown. x and Υ parameters are correlated by the expected value (mean) of the distribution. Upper panels refer to the neutral model, middle panels refer to the group of the mixture model with highest copy number, lower panels refer to the group of rare elements of the mixture model. Black circles indicated the most recent elements, associated with Primates evolution and Murine radiation. Two groups can be distinguished in upper and middle panels associated to a transition in time for the most recent elements. They lead to a lower average copy number for the primates and to higher average copy number for the murine genomes.

bioRxiv preprint doi: https://doi.org/10.1101/513572; this version posted January 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

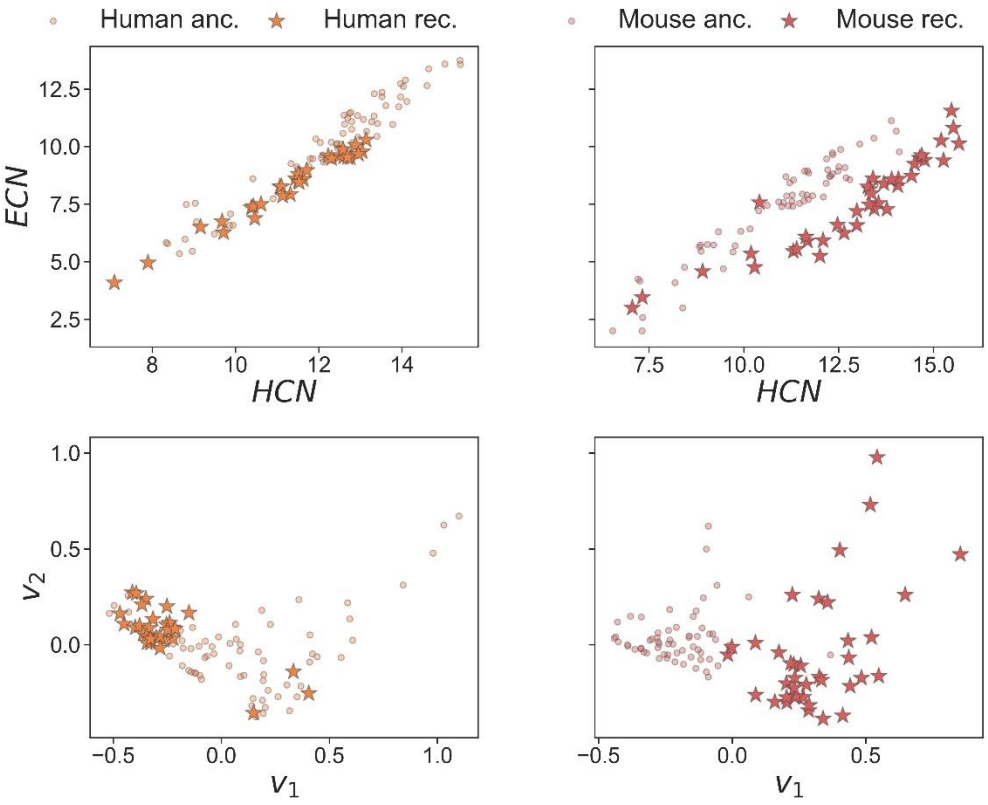

Figure 7: **Nonlinear correlation between number of insertions in euchromatin and heterochromatin states shows host genome adaptation mechanism.** Upper panels: Scatter plot in *log*2 scale of the number of insertions in euchromatin respect that in heterochromatin for each LINE specie. The number of insertions in euchromatin and heterochromatin states results correlated by a power law with exponent ~ 1.2. Lower panels: The group of most recent elements is well separated by PCA (age variable included).