Integration of abundances and chromatin state data of Long INterspersed Elements reveals dynamics transitions during evolution in mammalian genomes

Silvia Vitali[1,3,*,+], Enrico Giampieri[1,+], Steven Criscione[2], Claudia Sala[1], Italo do Valle[1,4], Nicola Neretti[2], and Gastone Castellani[1]

[1] *University of Bologna, Department of Physics and Astronomy, 40127 Bologna, Italy.*
[2] *Brown University, Department of Molecular Biology, Cell Biology and Biochemistry,Providence, RI 02906 USA*
[3] *Current affiliation: Basque Center for Applied Mathematics, 48009 Bilbao, Spain*
[4] *Current affiliation: Northeastern University, 177 Huntington Ave. Boston, MA 02115, USA*
[*] *svitali@bcamath.org*
[+] *these authors contributed equally to this work*

May 6, 2020

Abstract

Genome ecology and evolutionary biology have being increasingly investigated by interdisciplinary approaches, complementing experimental techniques with advanced modeling and statistical methods. Both disciplines, with distinct perspectives, have been successful in giving theoretical insights of the processes that happen inside and shape the genomes. Distinguishing between evolutionary and ecological origin of genomes patterns is not easy, and often the two approaches dedicate to well separated topics. Here, we integrate data of Long-INterspersed Elements (LINEs) abundances in 46 mammalian genomes with the insertions chromatin configuration, and their estimated age of amplification, to study the evolution of LINEs ecosystem inside and together with the genome landscape. We describe LINEs amplification dynamics by a birth-death process with assumption of competitive neutrality. Then, a competition mechanism for the internal promoter is introduced, spontaneously breaking the neutral assumption. We show that LINEs abundances, as well as the inherent model rates, cluster according to the host taxonomic order. The temporal variation of these rates combined with the average abundances and chromatin state of LINEs copies highlights host-elements interaction and taxa-specific element appearance, such as Lx, associated to the radiation

of the murine subfamily, and LIMA/LPB sub-families, related to primates evolution.

Keywords— Long Interspersed Elements, evolution, transposons ecology, mammals, Approximate Bayesian Computation method

Introduction

In the last decades an increasingly large number of genomes has been annotated [28] thanks to the availability of high-throughput sequencing methods and the development of algorithms and bioinformatics tools to process such data [27]. Additionally, extensive studies for chromatin state characterization in different cell lines of several organisms have been published [12, 11, 46], as well as results concerning genomes structure [32] and genomes content statistics [15] and evolution [14]. Therefore, a huge amount of data is available for many organisms nowadays.

Transposable elements (TEs) constitute a large portion of most species' genomes, covering roughly 45% of the human genome [7]. Also known as *selfish genes*, they are protein coding DNA sequences that can move from one genomics location to another and/or increase their copy number within the host-genome. Horizontal transfer of TEs between different organisms is also common for several sub-types of TEs.

Host organism and TEs interaction has been often described as a host-parasite relationship. The impact over the host organism fitness of TEs appearance and amplification has been extensively investigated to study how TEs copies reach fixation in a population and how TEs abundances are shaped [5, 41, 10, 1, 34, 35, 33]. The loss of fitness due to TEs could be responsible for the appearance of mechanisms of regulation and self-regulation, to limit the number and the impact over the genome functionality of new TE insertions [3, 18, 42]. The appearance of these regulation mechanisms is especially significant for TE sub-types which are unable to transfer from one host to another.

The co-evolution of TEs with the genomes and the interaction between these two parties make TEs a fundamental player [25, 31], heavily involved in both evolutionary biology and genome ecology.

Interdisciplinary approaches should provide advanced modeling techniques to genome ecology and evolutionary biology to extract most information as possible from the available data. Advanced Bayesian modeling approaches have been already applied successfully for dating the appearance of TEs in genomes, overcoming some limitations of standard techniques such as consensus sequence divergence or phylogenetic analysis [14].

TE sequences are classified into two classes on the basis of the transposition mechanism: DNA-transposons and retro-transposons. Their transposition mechanism is respectively DNA- or RNA-mediated. Each class is further partitioned into subclasses, families, subfamilies and elements, on the basis of: their structural organization, the proteins they encode, the sharing of specific insertions, deletions or substitutions. A TE family is composed by the ensemble of copies that share $> 80\%$ of their DNA sequence [43]. From an ecological perspective a single copy could be considered as an individual, and all the individuals belonging to the same family or subfamily as constituting a species, meaning that they occupy the same ecological niche in the genome [43]. Thus, all the elements belonging to the same family or subfamily can be located at the same trophic level. In the present work we reserve the word *species* to the element level for modeling purposes. Therefore, different elements belonging to the same family or sub-family are treated as different species, each one with its own number of copies. The copy number of an element constitutes the abundance, or the number of individuals, of that species. Then, according to the previous classification, all the TE

species belonging to the same family or subfamily can be located at the same trophic level.

A large diversity of TE sequences exists in the genomes in terms of biodiversity, biomass, and abundances. The Relative Species Abundance (RSA) of TEs varies significantly in different organisms and great variability in the amount of TE species and abundances can be observed also at the same trophic level. There is currently no agreement about the relative importance of the possible sources of this variability: such as the host specific selection pressure, both at the genome and host population levels, and the stochasticity of the forces acting on the individual copies [43, 34, 24].

The interdependence between the TE community and the host genome, together with the replication mechanisms of the elements, suggests a strong parallelism between TEs dynamics in the genome and species community dynamics in the ecosystem [43, 37]. Both the niche and the neutral theory have features convenient to describe TEs ecosystem. Niche theory is based on the partitioning of resources between competing species [6]. In the neutral theory, the stochastic mechanisms as demographic stochasticity, migration, and speciation are the most important forces shaping the community [17]. However, TEs ecosystem contains some peculiarities that differentiate it from standard ecosystems [43]. TEs create and continuously reshape their own environment, because the copies that lose any transposition ability generate a large part of the genomics landscape in which new copies may insert without deleterious effect on the cell functionalities. Furthermore, the natural selection acts on two levels, the genome level and the host one. In this regard, we distinguish *transposon ecology*, describing the interaction of TEs with the genome and cellular environment only, from *genome ecology*, which includes the interaction with the external environment mediated by the host fitness [24].

The selection at the level of the host could eventually induce TEs to evolve traits that constitute a selective disadvantage at the individual level, as for example a lower transposition rate [16, 33]. Phenomena related to the molecular nature of TEs may also occur, for example mutations, insertions and sequence rearrangements, which may lead to functional variations of the elements.

Here we focus on the study of a particular family of *non-long terminal repeats*: the Long Interspersed Elements (LINEs). We restrict the model to LINE family only, discarding the possible interaction with different TEs entities. Future studies may include information about inter family interactions, for example Short INterspersed Elements (SINE) -LINE parasitism [29].

We model LINEs copy number distribution inside a cohort of mammalian reference genomes under the hypothesis of competitive neutrality [23]. Competitive neutrality represents the absence of competitive differences among different LINE species. Thus, all the copies of all elements in the community could be characterized by the same transposition activity, sequence divergence, and death rate [43]. The variability of the *abiotic* component (in this context genes, repetitive sequences, and intracellular components) of the ecosystem is further reduced by including only mammalian genomes in the study.

LINEs are the most abundant family of TEs in mammals, in terms of biomass. They belong to the retro-elements class, which means that their replication is RNA mediated (figure 1). They are also incapable of horizontal transfer [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, and permits autonomous transposition. 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 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 new insertions often result in low fidelity copies of the parent LINE [8]. 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 [9]). Furthermore, a transcribed incomplete copy can hijack the retro-transposition machinery of autonomous copies to duplicate into a new location: a process called trans-complementation. The phenomenon of trans-complementation has been observed, for example, in LINE-1 retroelements, although it should happen at a much smaller rate than retro-transposition in *cis* [**Wei2001**].

Despite occasional re-activation of inactive elements has been observed in certain diseases, LINE community in mammalian genomes is mainly composed by the collection of defective and/or silenced copies of inactive elements that reached fixation in the genome. The stratification of such elements in the genome, in the absence of horizontal transfer, can be used to infer the changes in time that may have occurred in the dynamics of LINEs.

Some peculiarities of LINEs replication mechanism and their evolutionary history inside mammalian genomes support our hypothesis that LINEs community in mammals could be successfully described by a birth-death process under competitive neutrality 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 lineages is documented for ancient LINEs [38] and currently in mouse [26], where L1 frequently recruited novel 5'UTR sequences [40], suggesting that simultaneous activity of non-homologous promoters does not introduce a competition between the elements. Finally, the genome environment is unique to each of the TE copies and full-length L1 copies may differ randomly in their level of transposition activity [4, 36]. Therefore, the stochasticity at the individual level could have a significant impact on the structure of the entire community, supporting the neutral approach to describe the community dynamics.

Here we model the distribution of LINEs copy number through a Master Equation approach [2, 21], under the hypothesis of competitive neutrality and an alternative hypothesis of competition for the promoter region. The competition is occasional between two species and induces a reduction of the birth rates of the competing species, breaking spontaneously the hypothesis of competitive neutrality. The two models are nested. They are tested by fitting the RSA of LINEs communities through a hierarchical Approximate Bayesian Computation (ABC) method. A sliding window analysis is applied to study the evolution of the RSA in time by the same ABC method together with the chromatin state characterization of the individual copies.

Results and methods are summarized in the respective sections. The parameters expectation associated to several regimes of the model and the results are extensively analysed in the discussion section.

Results

The RSA of LINEs in 46 mammalian genomes have been fit by a negative binomial and by a mixture of negative binomial through hierarchical ABC method (see, section methods for details). From the same prior distribution, we obtained different posteriors of the model parameters for the RSA of 42 datasets, with the exception of Wallaby, Tasmania devil, Opossum, and Platypus, for which the rate of success of ABC method was extremely small in comparison to the others 42. Posteriors distribution and model comparison with the LINEs RSA are shown in Supplementary Materials (supplementary figures 6-15). The goodness-of-fit of the two models, assessed through the likelihood-ratio test, shows that the models are comparable for the majority of the

data set. In figure 2 the expected value of the parameters are shown, color labeled by the taxonomic order of the corresponding host genome.

LINEs community in human, chimp, rhesus macaque, mouse and rat genomes have been sorted by their relative time of appearance and amplification according to published genome-wide defragmentation results [14]. For each of these collections we performed a sliding window analysis (see section methods for details) of RSA distribution patterns, average percentage of insertions in open chromatin regions, and average abundance. Here we show the results for the windows length $N = 15$. Windows lengths of $N = 30, 40$ produce consistent results, with smoother temporal trends.

The sliding windows of RSA distribution patterns have been tested with the same method applied previously to the entire timeline. In figure 3 the likelihood-ratio test of the two models and the mixture coefficient α of the competition model are shown. The results for the other parameters are reported in the supplementary materials.

In figure 4 the sliding window analysis of the average abundance and the average percentage of insertions in open chromatin regions [46, 12] is shown. The average percentage of LINE copies belonging to open chromatin regions displays a decreasing temporal trend for all the organisms for which such information was available (human and mouse). The average abundance is decreasing for the primates cohort and display a fast transition to larger abundances in the rodents cohort.

In figure 5 the estimated time range of activity of LINE elements inside the human genome[14] belonging to the windows range most significant for competition (largest values of the mixture coefficient in figure 3) is shown. Elements are color labeled by their abundance in the human genome. The presence of significant similarity between 5'UTRs pairs (see section for details) is highlighted for the following high and low copy number pairs (or group): L1M2-L1M2c and L1MA9-L1M3a-L1M3b-L1M3c-L1M3d.

Hierarchical clustering of the LINEs abundances inside the 46 genomes, for the elements shown in figure 5, is reported in the supplementary material (supplementary figure 19). Elements are color labeled according to their abundance. Rare or abundant classification of the elements across all the mammals included in this study is consistent, except for the White Rhinoceros, which shows the opposite trend. The corresponding clustering for the host species is also mostly in agreement with the taxonomic classification.

The study of the correlation between the negative binomial parameters x, Υ obtained with the sliding window analysis of the RSA patterns is shown in figure 6.

The negative binomial parameters x, Υ of both cohorts can be clearly separated in two clusters, divided by the time of appearance of specific elements (LIMA/LPB for primates and Lx for rodents). One cluster contains all the samples before the appearance, the other all the sample after. Thus, a transition in time of the correlation of the parameters can be identified for both the group of primates and rodents. The same transition can be observed for the neutral model description as well as for the competition model (mixture of two negative binomials), by looking at the component describing the elements with large copy number.

The panel in figure 6 containing the parameters of the negative binomial component describing the less abundant LINEs do not display significant correlation. The other parameters describing the sliding windows RSA are shown in supplementary figures 17-18.

By using chromatin state assignments in human [12] and mouse [46] genomes, and the coordinates of the respective LINEs insertions from RepBase, we assigned to each LINE copy a chromatin configuration, distinguishing between open and closed states (*euchromatin* and *heterochromatin*). In figure 7 the correlation between the number of insertions in euchromatin (ECN) and the number of insertions in heterochromatin regions (HCN) and the corresponding 2D principal component analysis (PCA) is shown. Two clusters are clearly visible for mouse data set. The elements can be classified in the two clusters according to the same time threshold identified in figure 6 for mouse.

Information about the time of appearance is included in the PCA, to highlight the separation between the two clusters. The logarithms of ECN and HCN in figure 7 are strongly correlated. To estimate the correlation, a linear regression between the logarithm of the counts has been performed. This correlation corresponds to a power-law relationship between the raw counts:

$$N_{Eu} = 2^{c_0 \pm \epsilon} (N_{Het})^c \,. \tag{1}$$

We obtained for human: $c = 1.18$, $c_0 = -4.58$ and $\epsilon = 0.035$, which correspond to the standard error in the estimate. The correlation coefficient is $r = 0.96$ with p-value $p \sim 10^{-55}$. For the most ancient group of LINE in mouse we obtained $c = 1.12$, $c_0 = -4.84$, $\epsilon = 0.043$, $r = 0.96$, $p \sim 10^{-33}$, for the most recent we obtained $c = 1.03$, $c_0 = -5.84$, $\epsilon = 0.0.084$, $r = 0.90$, $p \sim 10^{-13}$.

The beginning of the transition in figure 7 for the mouse data set is contemporary to the transition of the parameters in figure 4 and figure 6 for the rodents cohort. Regarding the primates cohort, in 7 we highlighted the same group of elements defined by the transition in 6 for sake of clarity, but a sharp transition as the one observed for the rodents cohort is absent.

Discussion

We modeled the way LINEs populated different mammalian genomes as a birth-death process of two interacting sub-types: full-length (autonomous or active) copies and incomplete (non-autonomous or inactive) copies. The biological processes that characterize LINEs retro-transposition activity (replication, mutation, disappearance, extinction, etc.) have been described in terms of transition rates by a Master Equation (ME) approach (equation 2). The stochastic processes described by the ME, with the corresponding rates, are depicted in figure 1.

We analyzed and tested the two variants of the model proposed, with and without competition, to describe LINEs communities over 46 mammalian genomes. We focused on the two realistic dynamics regimes characterized by a specific asymptotic stationary solution to describe LINEs RSAs: a negative binomial distribution in case of competitive neutrality, and a mixture of two negative binomial distributions in case of direct competition between elements.

The negative binomial distribution depends on two parameters: the "probability of success", x, and the "number of failures", Υ. 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 $\Upsilon \sim 1$. Instead, if the trans-complementation process is relevant we expect $\Upsilon \sim \frac{(b_I + d_A)}{b_{AI}} >> 1$, due to the experimental observation that trans-complementation events are rarer if compared to retro-transposition in cis.

We considered that simultaneous activation of elements sharing the same promoter may introduce a disadvantage for the species that compete for the molecular machinery. According to our model a competitor could appear, with a certain probability, every time a new active copy is created in the system. Thereafter, the extinction of one of the competitors restores the neutrality.

Both the variants of the model reproduce the fundamental features of LINEs RSA patterns inside 42 of 46 mammalian genomes. The four exceptions are Wallaby, Tasmania devil, Opossum, and Platypus, which are characterized by a smaller number of resident LINE elements. From an evolutionary perspective they also correspond to a well defined subgroup of host genomes (marsupials and monotremes). Instead, the 42 genomes successfully described by the model belong to the Eutheria clade.

The expectation of the parameters of the competition model (mixture of two negative binomials) permits to separate the host species at the level of taxonomic order

(figure 2 and supplementary figure 5). Instead, the neutral model produces a less pronounced separation between taxa (figure 2). The couples of parameters Υ_1, Υ_2 and x_1, x_2 seem the best representation to discriminate the host organisms in different taxonomic orders. Within our description, such couples of parameters are related by the average value of the disadvantage due to competition ($\sim \frac{\langle b_{A,1} \rangle}{\langle b_{A,2} \rangle}$). The value of the failure parameter remain always closed to one ($\Upsilon \sim 1$), suggesting that a pure accumulation process is more convincing ($b_{AI} \approx 0, d_I \approx 0$). Hence, despite trans-complementation may take place, our results suggest it is not a very relevant process in shaping the RSA of the community. The shape of RSA, together with the rareness of trans-complementation events, supports the idea that equilibrium between host and LINE population does not hold in general ($b_A << d_A$), but a competition between the host and LINE species takes place. In fact, in the case of equilibrium for the active sub-type dynamics ($b_A \approx d_A$), numerical simulations display heavier tails in the generated RSA (supplementary figure X). Such excess of abundant elements could be compensated by a higher rate of excision from the genome (d_I) (supplementary figure X). However, this parameter configuration results less convincing, at least in primates, because more ancient LINE elements are more abundant on average than the recent one (figure 4).

For the majority of the organisms under study, the likelihood-ratio test for the two model examined was of order one, suggesting no clear preference between the assumptions of competitive neutrality and competition. The ABC method applied already penalizes the model with the higher number of parameters because the phase-space is larger, hence the two model results equivalently acceptable 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 introduction of competition between LINE species permits to better distinguish different taxonomic orders based on our model description, suggesting that the introduction of the second component in the PDF enhance the ability to extract useful information from the data and a better characterization of the biological phenomena (figure 2).

In fact, hierarchical clustering of the LINEs abundances in different organisms (supplementary figure 16) displays the same tendency to aggregate host organisms belonging to the same taxa. Horizontal transfer is very uncommon for the LINE family and LINEs mutually evolve with the genome with a turnover of active species. Hence, closely related host species could possess more similar LINEs 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 (supplementary figure X).

The evolution of the LINEs community has been investigated by sorting LINE elements according to their time of appearance and amplification in the host genomes [14] and by applying a sliding window analysis. The resulting time-dependent host-specific RSA have been tested for the neutral and competition models, by applying the same ABC method and prior distributions employed previously for the entire timeline. Where the mixing coefficient of the mixture model is higher, ABC model selection score supports the presence of competition (figure 3). The sliding window analysis 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 (figures 3 and supplementary figures X). Despite the time-dependent RSA is host-specific, some patterns are recurrent between the majority of the host organisms under study supporting the hypothesis of inheritance of the abundances (supplementary figures X). The hypothesis that competition could have been shaped by the LINE 5'UTR structure is supported by the similarity (measures of distance between pairwise aligned sequences) of the 5'UTR sequences in concurrently active LINEs (figure 5).

The transition in the correlation of the parameters x, Υ in figure 6 is thus associated to a transition to lower average abundance for primates and higher average abundance for the two rodents. These transitions are in agreement with the trend in time of the abundance (figure 4) and further supported by the transition in the chromatin landscape of the LINE copies for mouse (7) and by the concurrent amplification of host-specific LINE elements.

The average percentage of LINE copies inserted in euchromatin regions in the sliding window displays a decreasing trend with time ordered age in human and mouse (figure 4). However, in humans, it also shows a clear peak within the neutral time interval. The average percentage of copies in euchromatin regions is bigger for the windows with higher average copy number respect to the one with low copy number in human and in ancient LINE species 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.

The super linear correlation observed between the number of copies belonging to open and closed chromatin regions (figure 7) lead to the interesting result that a higher copy number, i.e., the sum of euchromatin and heterochromatin contributions, is associated to a higher percentage of insertions in euchromatin states. A reduction of the average percentage of insertions in euchromatin regions thus corresponds to a reduction in the average abundance. Moreover, the presence of the same type of correlation in human and mouse genomes, shared by all the most ancient elements, indicates the existence of a common pattern in the chromatin landscape of LINEs in mammals.

In agreement with such prediction, in primates, the decreasing trend of the insertion percentage in open chromatin regions is accompanied by a decreasing trend of LINE species abundance in time (figure 4). On the contrary, referred to mouse, the average abundance at a certain point drastically rears up. The time point at which the abundance rears up corresponds, in figure 7, to a transition to a smaller value of the coefficient c_0 (see section Results), while for ancient LINE species the correlation trend in mouse genome is very close to the one observed in human. Given the same number of insertions in open chromatin regions, 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 is defined as the time of appearance of the most ancient element belonging to the cluster of recent elements. In mouse data set, such transition is contemporary to the transition to higher average LINEs abundance shown in figure 4 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 ago according to [30, 13]. 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 (figure 4)surely have a bad impact on the host fitness, if it is not compensated by some host defence mechanisms, because the probability of deleterious insertions in the genome increases. The combined sharp transitions in the chromatin landscape and abundances, further associated to the evolutive transition of the host genome (murine subfamily differentiation), observed in mouse support this idea, and are perhaps the result of the competition between the host and LINEs. In human instead, a transition to a lower average copy number is observed (figure 6 and 4). 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. Indeed, the most ancient LINE species involved in the transition depicted in figure 6 are related to the evolutive differentiation of Primates, associated with the amplification of LIMA/LPB subfamilies $\sim 70 - 100$ Myr ago [22].

Conclusion

The present research is a first attempt to answer some of the fundamental questions concerning LINEs dynamics and co-evolution with the genome by combining different data types through an interdisciplinary approach. Data of LINEs abundance in 46 mammalian genomes [39] have been integrated with data about the relative time of appearance and amplification of the elements inside a sub-group of host genomes [14], and with data about the current chromatin state of the LINEs insertions inside human and mouse genomes [46, 12].

The mechanism of competition proposed between LINE species is independent of the chromatin state distribution of the copies, but acts at the level of the LINE species affecting their abundances. Instead, the chromatin state of LINE copies and average abundance should reflect the interaction of LINE species with the host, by mechanisms of silencing (for example methylation) and self-regulation (for example selection of elements with lower birth rates or with specific genomics region preference of the new insertions).

The analysis shows that LINEs abundances, as well as the inherent average birth-death rate obtained through the model, cluster according to the host organism taxonomic order. Model selection and promoter similarity analysis support the idea that ancient sub-groups of LINEs could have been in direct competition within mammalian genomes. Furthermore, the variation in time of the model parameters, combined with the average abundances and chromatin state of LINEs copies, displays evidences of host-elements interaction and features highlighting taxa-specific element appearance, such as Lx, associated to the radiation of the murine subfamily, and LIMA/LPB subfamilies, associated to primates evolution. The sliding window analysis shows that the decreasing trend of euchromatin percentage of insertions is shared by primates and rodents cohort. Chromatin information highlights also a super linear correlation between the insertions in open chromatin regions and insertions in closed chromatin regions. This type of correlation indicates that a increasingly small percentage of copies inserted in open chromatin regions will be associated to an increasingly small abundance of the element (sum of the number of element insertions in open chromatin and in closed chromatin regions). In fact, Lx appearance, associated to an increase in the average LINEs abundance in rodents, is also characterized by a transition to a different correlation regime of the number of insertions in different chromatin configurations.

We believe that interdisciplinary research could positively contribute to improve field specific research methodologies, and possibly complete and enrich the perspective of the specialists. The possibility to map in different systems (hosts) the evolution of genomics elements, with careful considerations about the nature of such elements, could become a powerful tool to understand present and past dynamics of the entities that contribute in shaping the genome, and to bridge evolutionary biology and genome ecology investigations. Despite the relative simplicity of the data types included in the analysis, the results obtained are in agreement with several independent studies [22, 38, 26, 30] and encourage the application of similar approaches by integrating more refined information at the copy level, such as DNA sequence, chromosome location and genome patterns (CG/AT content, chromatin configuration, etc.), to investigate specific questions at the edge of evolutionary biology, genome ecology, and transposon ecology fields.

Methods

Data sources

LINE abundances were calculated using RepeatMasker annotation (http://www.repeatmasker.org) [39] for human genome build hg19 and 45 other mammalian species. LINE consensus

sequences were downloaded from RepBase [19, 20] (http://www.girinst.org). Chronological ordering of LINEs in human, chimp, rhesus macaque, mouse and rat genome was derived from genome wide defragmentation results [14]. Chromatin structure data are available for mouse [46] and human [12]. The employed chromatin state assignment was conducted by using ENCODE chromatin models from the ChromHMM method [11].

Chromatine state assignment

Chromatin structure data were used to assign each LINE copy to open or closed chromatin state by the knowledge of their coordinates in the reference genome. Open and closed chromatin states were defined and located according to the available classification for mouse [46] and human [12] for the germ line. Multiple assignments of the same genome region have been treated by classifying the combination of states into open, weakly open and closed chromatin, depending if the assigned configurations belong mainly to one of these groups. We tested grouping weakly-open chromatin population with both open chromatin and closed chromatin to assess if this choice affected our results, but did not observed any significant difference. Thus, the weakly-open and the unknown state have been included in the closed chromatin group, which encloses most of the LINE copies.

Sliding window analysis

LINEs species are sorted by their relative time of appearance and amplification in the genome according to specific host genome analysis [14] (different genomes may show variations of this time series). By dividing the time series into intervals (windows), each of them containing a fixed number of elements, a different realization of the LINEs ecosystem is obtained. By sliding the window, a different sub-sample of elements active in a distinct evolution stage of the genome can be selected, representing a picture of the LINE community in a different evolutionary stage of the genome. The time series of the windows depicts several variations of the LINEs community in time in terms of: RSA patterns; average abundance; and average percentage of insertions in different chromatin configurations. The RSA patterns are described by the same ABC method applied to the entire timeline.

Stochastic model

The LINEs dynamics inside the genome is described via the following two-dimensional Master Equation:

$$\begin{aligned}\frac{dP}{dt} &= (E^{-}_{n_A} - 1)b_A n_A P + (E^{+}_{n_A}E^{-}_{n_I} - 1)d_A n_A P \\ &+ (E^{-}_{n_I} - 1)b_I n_A P + (E^{-}_{n_I} - 1)b_{AI} n_A n_I P \\ &+ (E^{+}_{n_I} - 1)d_I n_I P.\end{aligned} \tag{2}$$

where we consider $P \equiv P(n_A, n_I, t)$ and the Van Kampen step operators [21] $E^{\pm}_n f(n, m) = f(n \pm 1, m)$, the lower index n indicates the variable on which the operator acts, the upper index $+$ or $-$ determines the direction of the unitary change of the value of the variable n. n_A and n_I represent respectively the number of full-length copies (autonomous in self replication) and the number of defectives copies (non-autonomous or inactive) of a specific element in a genome. Each term in on the right-hand side of equation 2 represents one of the stochastic process examined to define LINEs dynamics: b_A, d_A, rate of birth and death of full-length copies; b_I, d_I, rate of birth and death of defective copies; b_{AI}, rate of birth of defective copies by trans-complementation. The stationary distribution of the bidimensional model (equation

2) doesn't have a closed form, however, it is possible to compute the marginal distributions, corresponding to the n_A and n_I species. The distribution $P(n_I)$, to which it will be referred as P_{n_I}, corresponds to the *relative species abundance* (RSA) of the LINE species, which represents the probability to observe a species with a certain number of *individuals* (the copy number) inside a community (the genome).

When equilibrium is reached for both active and inactive copies, the RSA corresponds to a negative binomial distribution. In fact, taking n_A as a constant, the equation 2 for n_I describes a well-known ecological neutral model, successfully applied to coral reefs and rain forests [45]. If equilibrium for active species does not hold ($bA << dA$) and excision and trans-complementation processes are neglected, the stationary solution P_{n_I} is still approximated by a negative binomial distribution, obtained when the "absorbing state" is reached ($n_A = 0$). The two regimes can be distinguished because the expected value of the parameters are different for biological considerations (see section discussion).

A spontaneous breaking of the competitive neutrality assumption is introduced by assuming that contemporary activation by the same promoter region of two LINEs reduces the birth rates b_A, b_I and b_{AI} in equation 2 by a factor $n_1/(n_1 + n_2)$ where n_1 and n_2 are the full-length copies of the two competing elements. 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, and will induce deviations from the expected distribution by generating a bimodal behavior. We surmise that the distribution arising from this type of competition is a mixture of two negative binomials (equation 3) for a range of parameters compatible with real data (confirmed by numerical simulation):

$$P_{RSA} = \alpha \cdot P_{rare} + (1 - \alpha) \cdot P_{abund}\,, \tag{3}$$

where α is the mixture coefficient, related to the probability that two elements compete, and with P_{rare} and P_{abund} representing respectively the RSA of rare elements.

The negative binomial distribution depends on two parameters: the "probability of success", x, and the "number of failures", Υ. The expectation value of the copy number for the LINEs RSA is defined by the parameters of the negative binomial distribution by the relation:

$$\langle n_I \rangle = \frac{x}{1 - x} \cdot \Upsilon \tag{4}$$

where, in the case of competition, the parameters of the distribution of the abundant species, which contains the majority of the LINE species, can be employed.

Hierarchical approximate Bayesian computation

A hierarchical approximate Bayesian computation (ABC) method has been implemented in Python to fit the RSAs included in this study. The method can be schematized by the following procedure: (i) a set of parameters is taken from uninformative prior distributions to built a negative binomial or a mixture of two negative binomials distribution representing the RSA distribution; (ii) a sample of abundances is generated according to such distribution; (iii) the set of parameters taken from the priors is accepted if the distance between the generated sample and the empirical LINEs abundance of at least one of the 46 genomes is under a certain threshold, (iv) the procedure (i-iii) is repeated $\sim 10^6$ times to built a posterior of the parameters; (v) the posterior is used to replace the uninformative prior with a more informative one, to increase the statistics and reduce computation efforts. The procedure (i-iv) is then repeated for each of the 46 data set to build a posterior distribution of the parameters specific for each of the data set. The posterior distribution of the parameters represents the probability that a certain set of parameters describes the data according to the model. The expectation value of the parameters with respect to the posterior

distribution is thereafter employed to describe the data. Model selection is performed according to the likelihood-ratio test information criterion, approximated to be the ratio of successes (number of set of parameters accepted over the total number of set tested) for the two models.

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 for the case $b_{AI} = 0$ and a hybrid algorithm when trans-complementation is considered. The hybrid algorithm was chosen instead of the Gillespie one to reduce the time of computation. It consists in the estimation of the expected number of inactive copies by ordinary differential equation (ODE) numerical integration at the beginning of each time interval $\bar{n}_{I,t}$. Then, tau-leap algorithm is applied to generate a stochastic increment associated to the time interval $\Delta n_{I,t}$. The number of inactive copies at the end of the time interval is thus determined by the sum $n_{I,t+1} = \bar{n}_{I,t} + \Delta n_{I,t}$. 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 ($b_{AI} = 0$ or $b_{AI} > 0$) were performed to check if the solution was compatible with a mixture of negative binomials.
More details about these methods can be found in the Supplementary Materials (supplementary figures 1-4).

Acknowledgements

The results presented in this work are part of the PhD Thesis of SV [44]. This work has been supported by the Italian Ministry of Education at University of Bologna (Alma Mater Studiorum), Department of Physics and Astronomy (DIFA), by the Basque Center for Applied Mathematics (Bilbao, Spain) 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 also aknowledge IMforFuture EU project and HARMONY EU project.

Author contributions statement

N.N. and G.C. conceived the study, S.W.C. and I.F.V. prepared the data sets, E.G., S.V. and C.S. implemented the ABC pipeline and simulations, S.V. performed and developed the analysis.

Author competing interest statement

All authors reviewed the manuscript and declare no conflict of interest.

References

[1] G. Abrusán and H. J. Krambeck. "Competition may determine the diversity of transposable elements". In: *Theoretical Population Biology* 70.3 (2006), pp. 364–375.

[2] Animesh Agarwal et al. "On the precision of quasi steady state assumptions in stochastic dynamics". In: *The Journal of Chemical Physics* 137.4 (2012), p. 044105. DOI: 10.1063/1.4731754.

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

[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*. 100, 2003, pp. 5280–5285.

[5] B. Charlesworth and C. H. Langley. "The population genetics of Drosophila transposable elements". In: *Annual review of genetics* 23 (1989), pp. 251–287.

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

[7] International Human Genome Sequencing Consortium. "Initial sequencing and analysis of the human genome". In: *Nature* 409.6822 (2001), pp. 860–921.

[8] R. Cordaux and M.A. Batzer. "The impact of retrotransposons on human genome evolution". In: *Nature Review Genetics* 10.10 (2009), pp. 691–703.

[9] S. W. Criscione et al. "Transcriptional landscape of repetitive elements in normal and cancer human cells". In: *BMC genomics* 15 (2014), p. 583.

[10] G. Deceliere, S. Charles, and C. Biémont. "The dynamics of transposable elements in structured populations". In: *Genetics* 169.1 (2005), pp. 467–474.

[11] J. Ernst and M. Kellis. "ChromHMM: automating chromatin-state discovery and characterization". In: *Nature Methods* 9 (2012), pp. 215–216.

[12] J. Ernst and M. Kellis. "Discovery and characterization of chromatin states for systematic annotation of the human genome." In: *Nature biotechnology* 28.8 (2010), pp. 817–825.

[13] A. V. Furano et al. "Amplification of the Ancient Murine Lx Family of Long Interspersed Repeated DNA Occurred During the Murine Radiation". In: *Journal of Molecular Evolution* 38 (1994), pp. 18–27.

[14] J. Giordano et al. "Evolutionary history of mammalian transposons determined by genome-wide defragmentation". In: *PLoS Computational Biology* 3.7 (2007), pp. 1321–1334.

[15] Paci Giulia et al. "Characterization of DNA methylation as a function of biological complexity via dinucleotide inter-distances". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* (2016). DOI: http://doi.org/10.1098/rsta.2015.0227.

[16] J. S. Han and J. D. Boeke. "A highly active synthetic mammalian retrotransposon". In: *Nature* 429 (2004), pp. 314–318.

[17] S. P. Hubbell and L. B. de Água. "The unified neutral theory of biodiversity and biogeography: reply." In: *Ecology* 85.11 (2004), pp. 3175–3178.

[18] M. Imbeault, P. Helleboid, and D. Trono. "KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks." In: *Nature* 543 (2017), pp. 550–554. DOI: https://doi.org/10.1038/nature21683.

[19] J. Jurka. "Repbase Update: A database and an electronic journal of repetitive elements". In: *Trends in Genetics* 16.9 (2000), pp. 418–420.

[20] J. Jurka et al. "Repbase Update, a database of eukaryotic repetitive elements". In: *Cytogentic and Genome Research* 110 (2005), pp. 462–467.

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

[22] H. Khan, A. Smit, and S. Boissinot. "Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates". In: *Genome Research* 16.1 (2006), pp. 78–87.

[23] S. Linquist et al. "Applying ecological models to communities of genetic elements: The case of neutral theory". In: *Molecular Ecology* 24.13 (2015), pp. 3232–3242.

[24] S. Linquist et al. "Distinguishing ecological from evolutionary approaches to transposable elements". In: *Biological Reviews* 88.3 (2013), pp. 573–584. ISSN: 14647931.

[25] Z. Lippman et al. "Role of transposable elements in heterochromatin and epigenetic control". In: *Nature* 430 (2004), pp. 471–476. DOI: https://doi.org/10.1038/nature02651.

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

[27] L. Milanesi et al. "Trends in modeling biomedical complex systems". In: *BMC bioinformatics* 10 (2009). DOI: doi:10.1186/1471-2105-10-S12-I1..

[28] "National Center for Biotechnology Information (NCBI)". In: *Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information* (1988). URL: Available%20from:%20https://www.ncbi.nlm.nih.gov/.

[29] I Ogiwara et al. "Retropositional parasitism of SINEs on LINEs: identification of SINEs and LINEs in elasmobranchs." In: *Molecular Biology and Evolution* 16.9 (Sept. 1999), pp. 1238–1250. ISSN: 0737-4038. DOI: 10.1093/oxfordjournals.molbev.a026214. eprint: https://academic.oup.com/mbe/article-pdf/16/9/1238/9593471/mbe1238.pdf. URL: https://doi.org/10.1093/oxfordjournals.molbev.a026214.

[30] 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." In: *Proceedings of the National Academy of Sciences USA* 87 (1990), pp. 9481–9485.

[31] L.M. Payer and K.H. Burns. "Transposable elements in human genetic disease." In: *Nat Rev Genet* 20 (2019), pp. 760–772. DOI: https://doi.org/10.1038/s41576-019-0165-8.

[32] Michael Rosenthal et al. "Bayesian Estimation of 3D Chromosomal Structure from Single Cell Hi-C Data". In: *bioRxiv* (2018). DOI: 10.1101/316265. eprint: https://www.biorxiv.org/content/early/2018/05/07/316265.full.pdf. URL: https://www.biorxiv.org/content/early/2018/05/07/316265.

[33] A. Le Rouzic, T. S. Boutin, and P. Capy. "Long-term evolution of transposable elements". In: *Proceedings of the National Academy of Sciences USA* 104.49 (2007), pp. 19375–19380.

[34] A. Le Rouzic and P. Capy. "The first steps of transposable elements invasion: Parasitic strategy vs. genetic drift". In: *Genetics* 169.2 (2005), pp. 1033–1043.

[35] A. Le Rouzic and P. Capy. "Population genetics models of competition between transposable element subfamilies". In: *Genetics* 174.2 (2006), pp. 785–793.

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

[37] 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". In: *PLoS ONE* 8 (2013), p. 6.

[38] A. F. Smit et al. "Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences." In: *Journal of molecular biology* 246.3 (1995), pp. 401–417.

[39] AFA Smit, R. Hubley, and P. Green. "RepeatMasker Open-4.0". In: *http://www.repeatmasker.org* (2013).

[40] A. Sookdeo et al. "Revisiting the evolution of mouse LINE-1 in the genomic era." In: *Mobile DNA* 4.1 (2013).

[41] C.J. Struchiner, M.G. Kidwell, and J.M.C. Ribeiro. "Population dynamics of transposable elements: copy number regulation and species invasion requirements". In: *Journal of Biological Systems* 13.4 (2005), pp. 455–475.

[42] W. Sun et al. "Pathogenic tau-induced piRNA depletion promotes neuronal death through transposable element dysregulation in neurodegenerative tauopathies." In: *Nat Neurosci* 21 (2018), pp. 1038–1048. DOI: https://doi.org/10.1038/s41593-018-0194-1.

[43] S. Venner, C. Feschotte, and C. Biémont. "Dynamics of transposable elements: towards a community ecology of the genome". In: *Trends in Genetics* 25.7 (2009), pp. 317–323.

[44] S. Vitali. "Modeling of Birth-Death and Diffusion Processes in Biological Complex Environments." In: *Ph.D. Thesis, Department of Physics and Astronomy, University of Bologna, Bologna, Italy* (2018).

[45] I. Volkov et al. "Neutral theory and relative species abundance in ecology." In: *Nature* 424.6952 (2003), pp. 1035–1037.

[46] F. Yue et al. "A comparative encyclopedia of DNA elements in the mouse genome." In: *Nature* 515.7527 (2014), pp. 355–64.

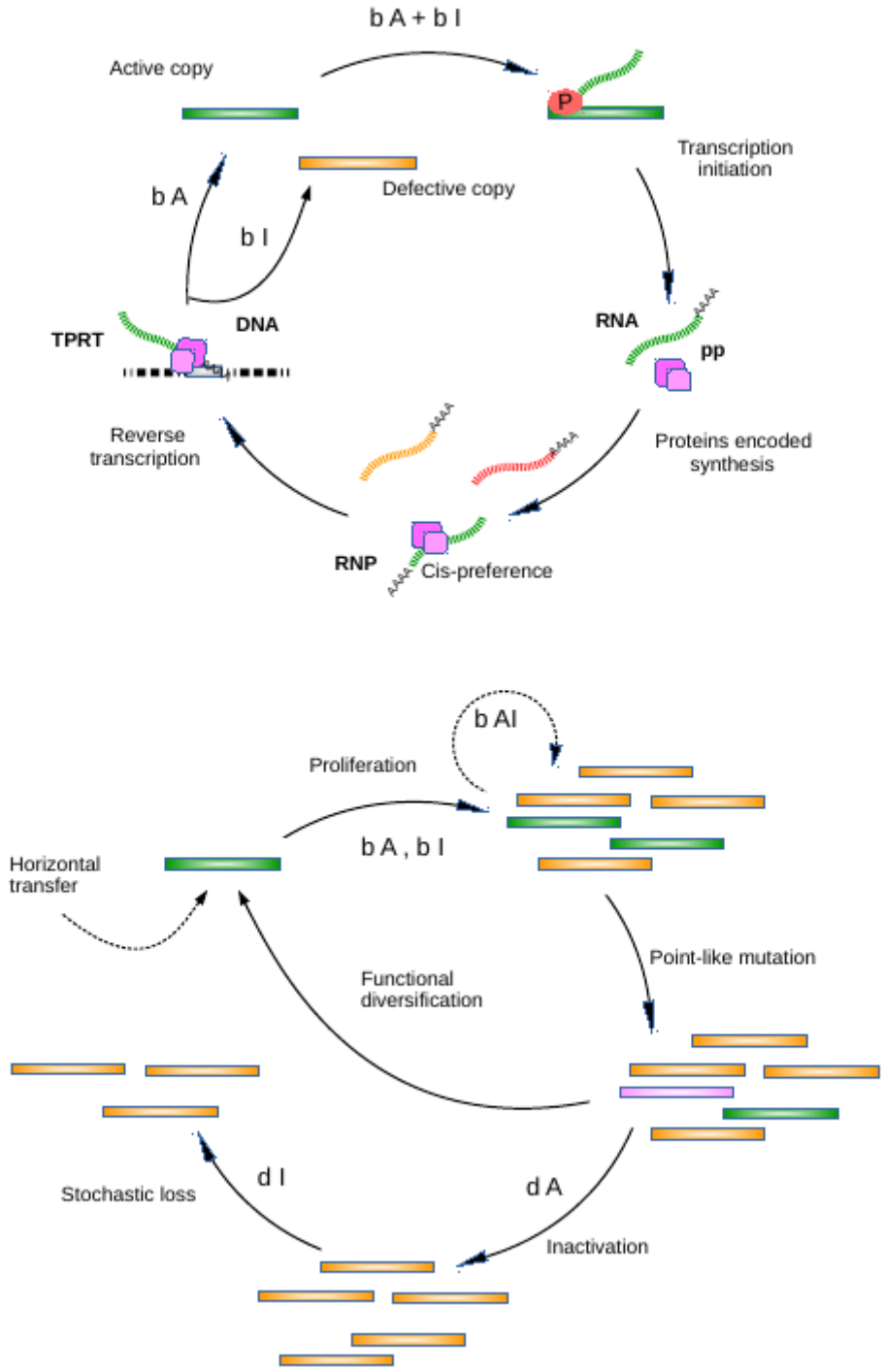


Figure 1: **LINEs transposition and activity cycle diagram.** Diagram of the birth-death process of defective (orange) and full-length copies (green) from a full-length master copy by retro-transposition and cycle of element amplification in the genome. Occasional point-like mutation (or other) generate a new element species (pink) which start the cycle again and eventually become a competitor.

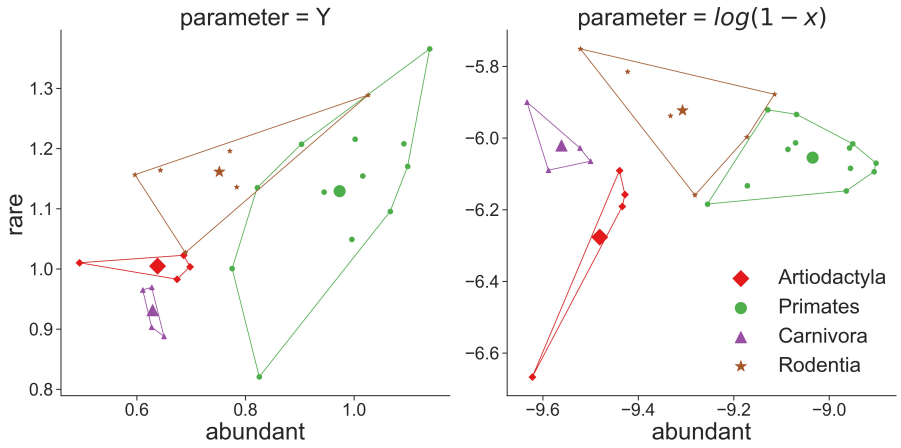


Figure 2: **Competition model description clusters different mammalian orders.** Set of optimized parameters (posteriors expectation value) obtained by ABC method for the competition model (mixture of two negative binomials). The couples $\Upsilon_{rare}, \Upsilon_{abund}$ and $log(1 - x_{rare}), log(1 - x_{abund})$ are shown for different hosts. Convex hull of same host order parameters are highlighted by straight lines. The average of the parameters per host order are shown by the larger markers.

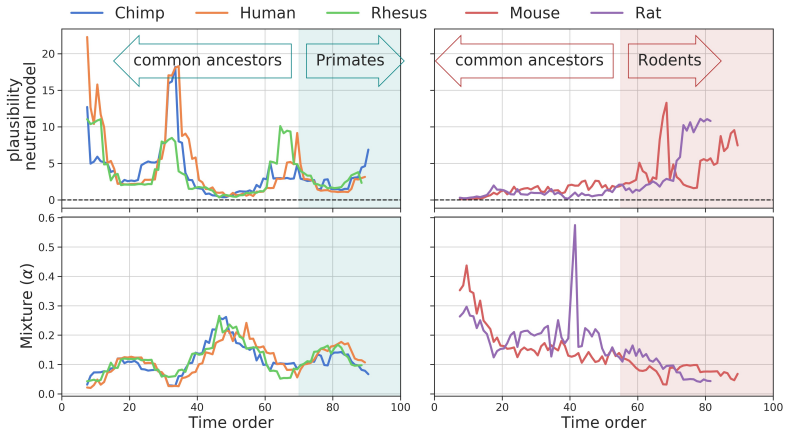


Figure 3: **Comparison of the two model performance during the evolution of the LINEs ecosystem in primates and rodents.** Upper panels: plausibility of the neutral model calculated in term of log-likelihood ratio test of the neutral model over competition model) by the sliding window approach. Lower panels: estimation of the mixture coefficient by the sliding window approach. Larger values for the mixture coefficient α are associated to lower plausibility of the neutral model. The portion of elements highlighted in different colors belong to different clusters in figure 6.

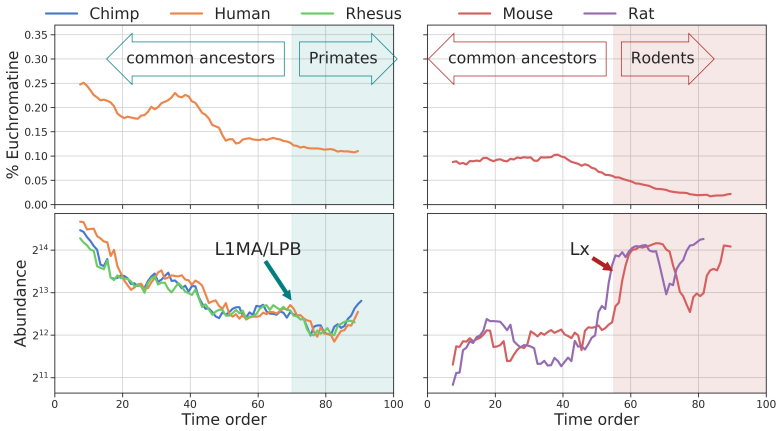


Figure 4: **Sliding window of LINE insertions percentage inside euchromatin and expected LINEs abundance in primates and rodents.** Upper panels: percentage of LINE copies inserted in euchromatin regions calculated by the sliding window approach. Lower panels: LINEs average copy number (sum of euchromatin and heterochromatin insertions) calculated by the sliding window approach. The portion of elements highlighted in different colors belong to different clusters in figure 6.

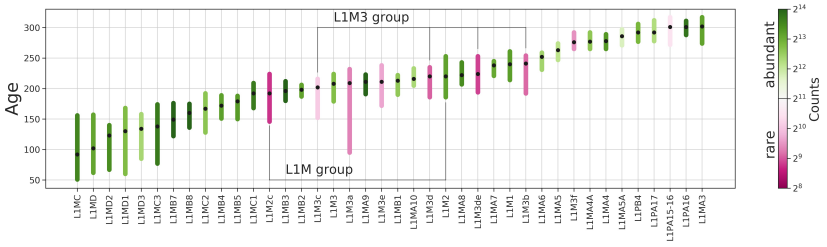


Figure 5: **5'UTR similarity between competing LINE retrotransposons in human.** The available consensus sequences of the 5'UTR of LINEs in the human genome have been aligned pairwise, with ClustalW2. The minimum distance is achieved between couples (or groups) of elements with similar ages and having high and low copy number respectively. Range of activity shown by the bar length. Abundance shown by the color legend. Significant similarity between 5'UTRs is observed for the following high and low copy numbers groups: L1M2-L1M2c , L1M3a-L1M3b-L1M3c-L1M3d.

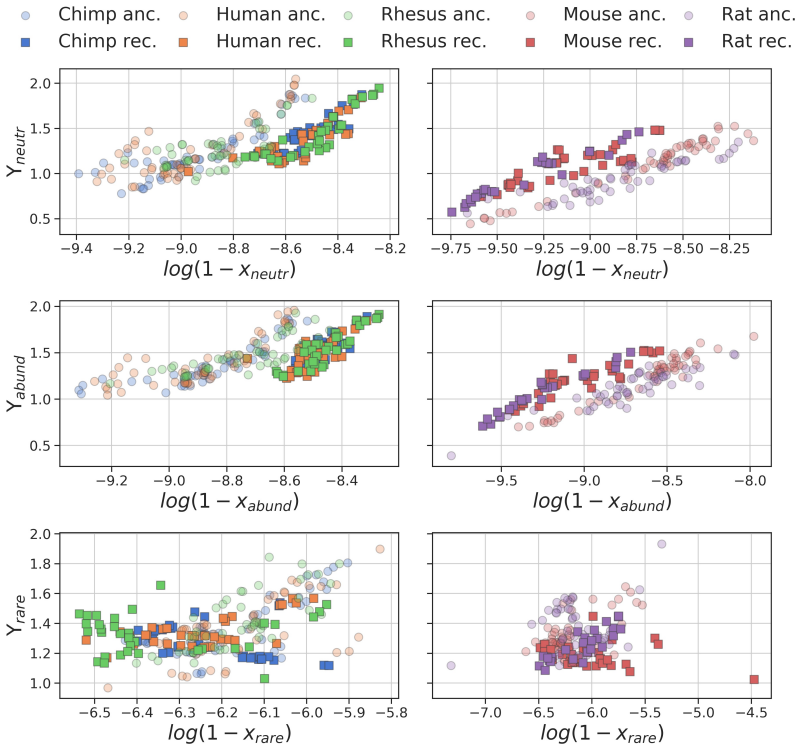


Figure 6: **Space of parameters of the two model tested by the sliding window approach in primates and rodents.** The space of parameters describing sliding window ecosystem of LINEs in human, chimpanzee, rhesus macaque (left panels) and mouse, rat (right panels) is shown. x and Υ parameters are correlated 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 largest copy number, lower panels refer to the group of less abundant elements. Circles indicated the most ancient elements. Transition between the two cluster are associated with specific LINE species appearance.

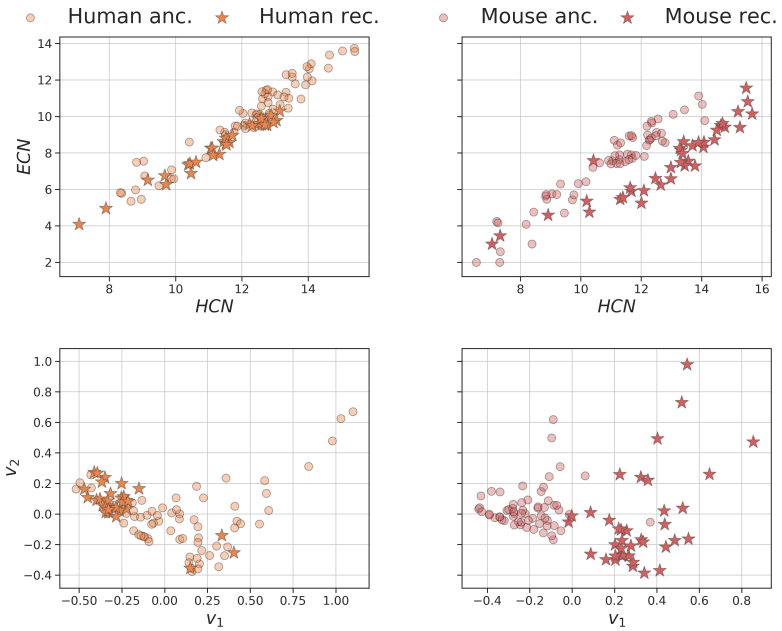
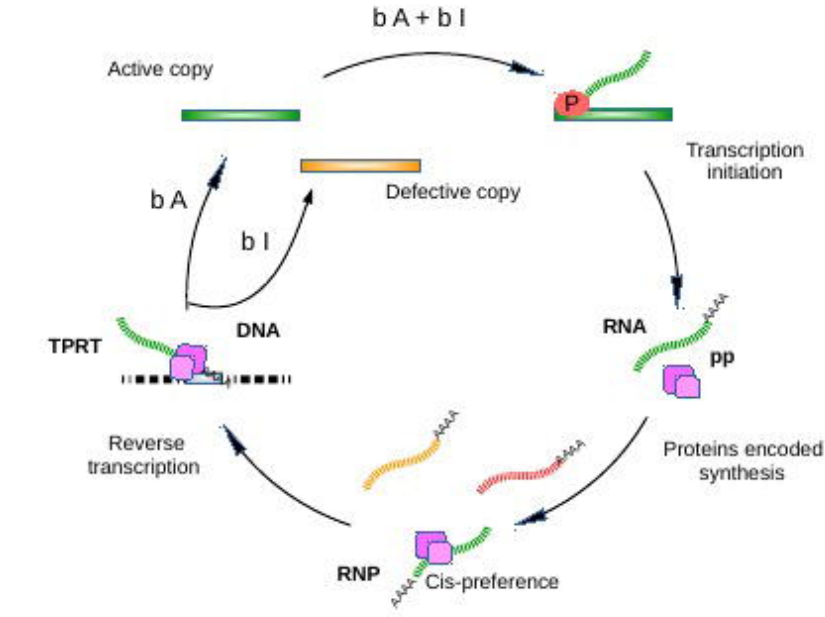
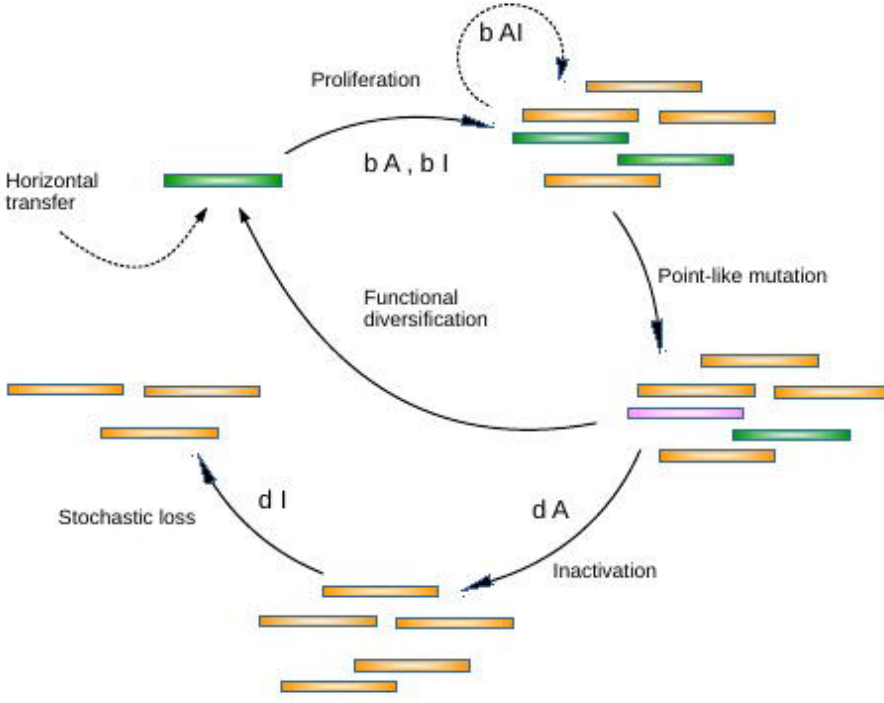
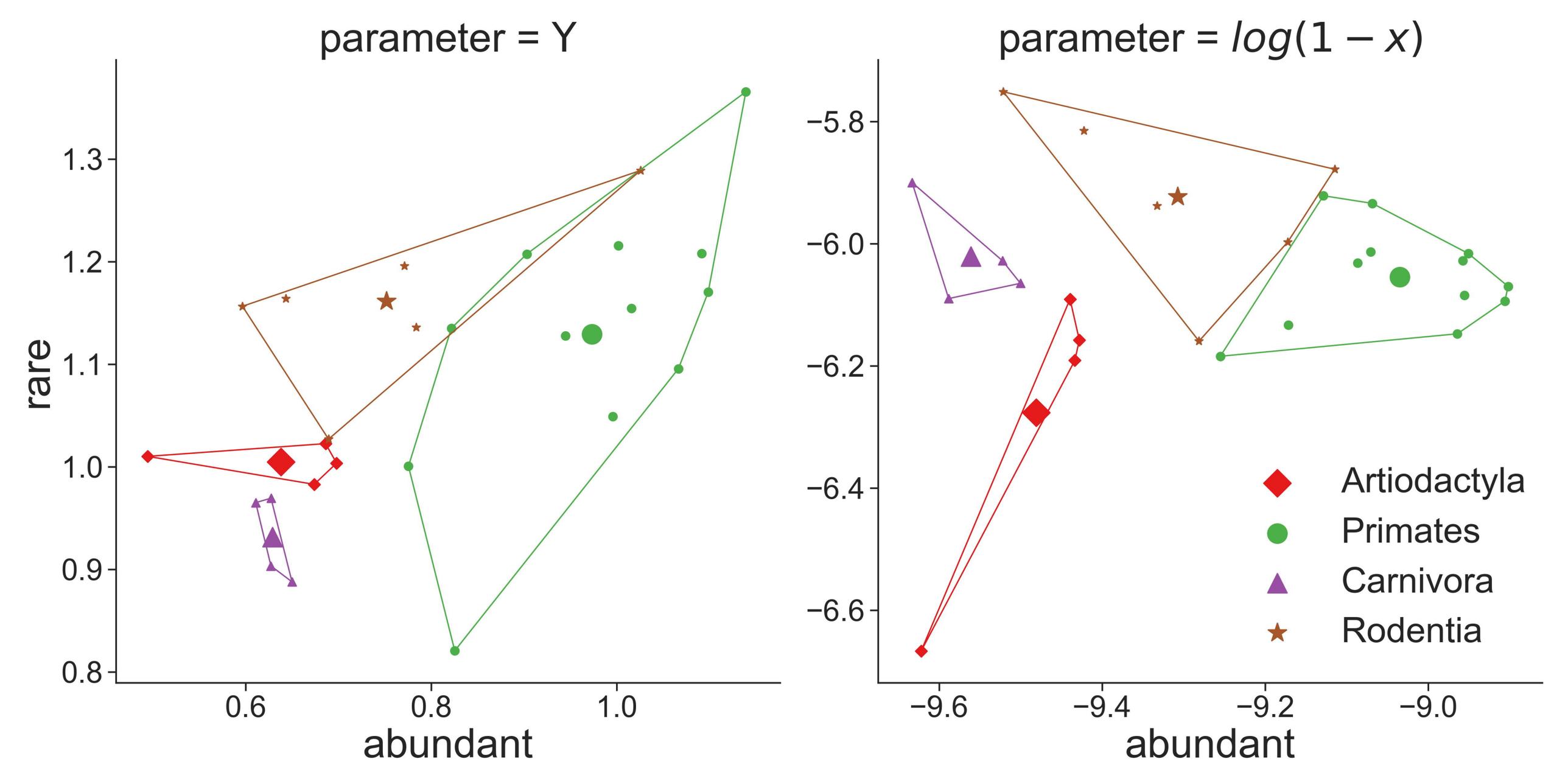
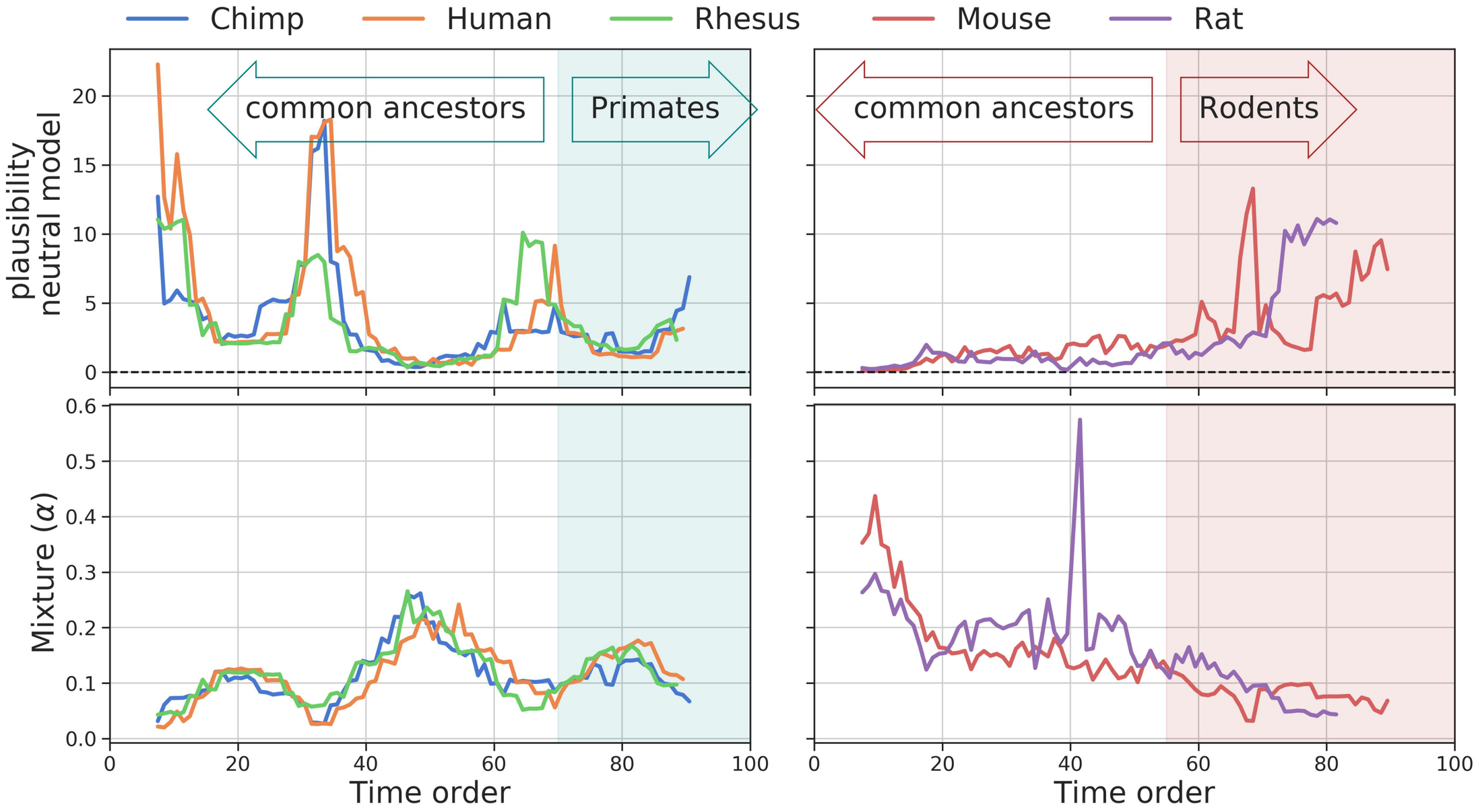


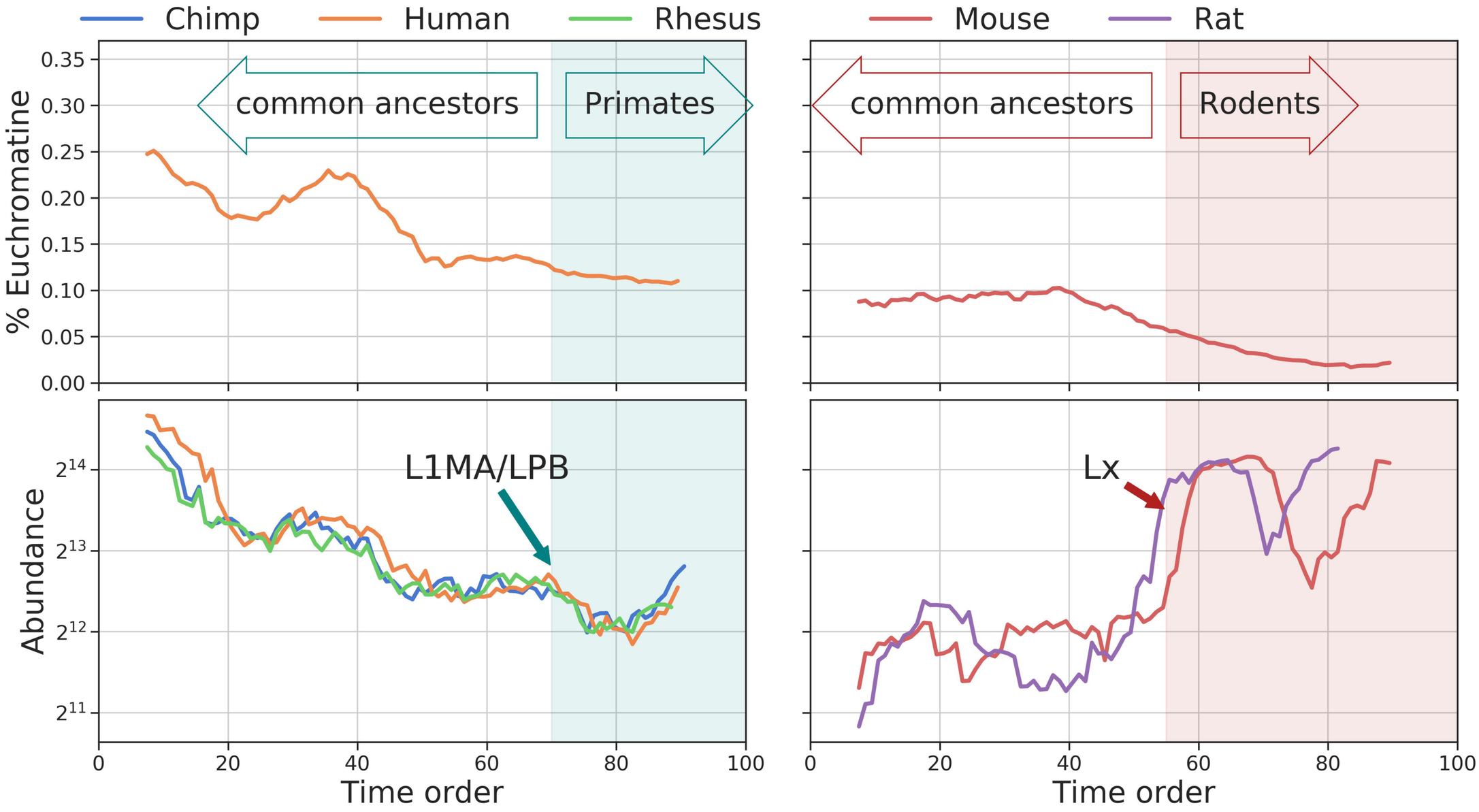
Figure 7: **Number of insertions in euchromatin and heterochromatin states in human and mouse.** Upper panels: Scatter plot in log_2 scale of the number of insertions in euchromatin respect that in heterochromatin for each LINE specie. The number of insertions in euchromatin and heterochromatin states results correlated by a power law. Lower panels: PCA of the number of insertions in euchromatin, in heterochromatin, and relative time of appearance. The portion of elements highlighted in different colors belong to the two different clusters in figure 6: ancient elements (circles), recent elements (stars).

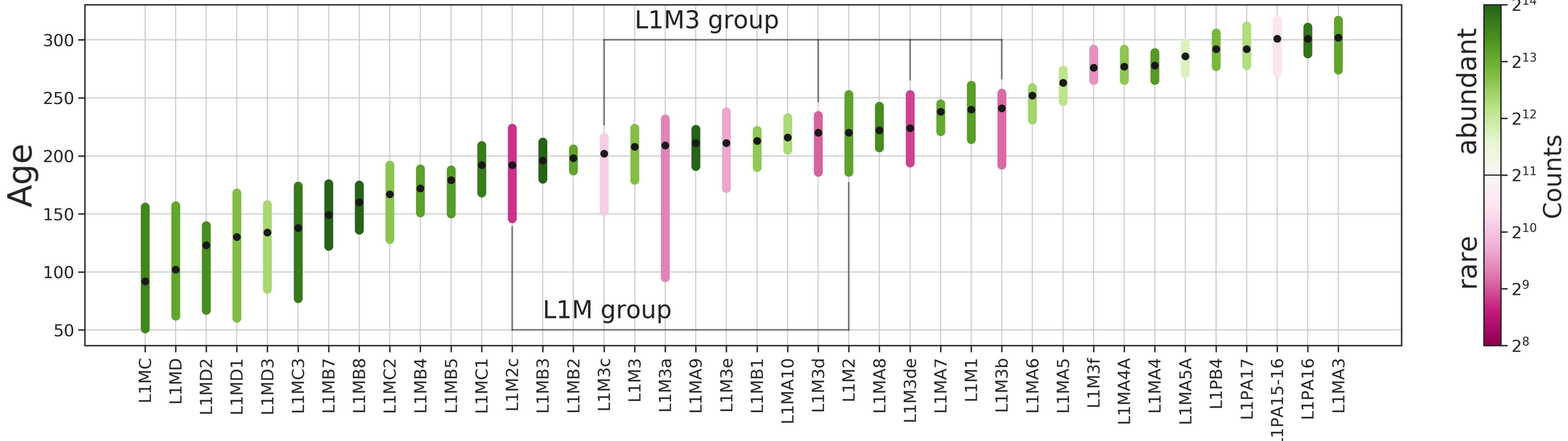

b A + b I
Active copy
P
Transcription initiation
Defective copy
b A
b I
TPRT
DNA
RNA
AAAA
pp
Reverse transcription
Proteins encoded synthesis
RNP
Cis-preference


b AI
Proliferation
b A , b I
Horizontal transfer
Point-like mutation
Functional diversification
Stochastic loss
d I
d A
Inactivation


parameter = Y
parameter = log(1 − x)
rare
abundant
abundant
1.3
1.2
1.1
1.0
0.9
0.8
0.6
0.8
1.0
−5.8
−6.0
−6.2
−6.4
−6.6
−9.6
−9.4
−9.2
−9.0
Artiodactyla
Primates
Carnivora
Rodentia


Chimp
Human
Rhesus
Mouse
Rat
common ancestors
Primates
common ancestors
Rodents
plausibility
neutral model
Mixture (α)
Time order
Time order


Chimp
Human
Rhesus
Mouse
Rat
% Euchromatine
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
common ancestors
Primates
common ancestors
Rodents
Abundance
2^{14}
2^{13}
2^{12}
2^{11}
L1MA/LPB
Lx
0
20
40
60
80
100
Time order
Time order


L1M3 group
L1M group
Age
300
250
200
150
100
50
L1MC
L1MD
L1MD2
L1MD1
L1MD3
L1MC3
L1MB7
L1MB8
L1MC2
L1MB4
L1MB5
L1MC1
L1M2c
L1MB3
L1MB2
L1M3c
L1M3
L1M3a
L1MA9
L1M3e
L1MB1
L1MA10
L1M3d
L1M2
L1MA8
L1M3de
L1MA7
L1M1
L1M3b
L1MA6
L1MA5
L1M3f
L1MA4A
L1MA4
L1MA5A
L1PB4
L1PA17
L1PA15-16
L1PA16
L1MA3
abundant
rare
Counts
2^{14}
2^{13}
2^{12}
2^{11}
2^{10}
2^{9}
2^{8}