Transposable elements activity reveals punctuated patterns of speciation in Mammals

Ricci Marco[1]* & Peona Valentina[1], Guichard Etienne[1], Boattini Alessio[1]
[1]University of Bologna, Department of Biological, Geological and Environmental Sciences, Italy

Abstract

Transposable elements (TEs) play an essential role in shaping eukaryotic genomes and in organismal diversification. It has been hypothesized that bursts of TEs may correspond to punctuated events of speciation (CArrier SubPopulation, Epi-Transposon, TE-Thrust hypotheses), thus it is expected that highly differentiated taxa bear highly active TEs in their genomes. To test this hypothesis, we designed two new parameters: the Relative Rate of Speciation (RRS) and the Density of Insertions (DI). These parameters measure, respectively, how much the taxa are undergoing an adaptive radiation and the magnitude of TE activity in their genomes. We call “hot” and “cold” those genomes with high and low DI, respectively. In this study, we test the association between RRS and DI (“Cold Genome Hypothesis”) in Mammalian families and superorders. Furthermore, since the age of TEs can be inferred by calculating the distance from their respective consensus sequences, we subset TEs in different age classes in order to study the evolution of genomes at different time scales. Here, we consider “recent” TEs (divergence <1%) and “less recent” TEs (divergence <5%). Comparing the TE activity in 19 pairs of species belonging to different mammalian families and the average TE activity of sampled species from the four superorders of Placentalia, we show that taxa with high RRS correlate with “hot genomes” and taxa with low RRS correlate with “cold genomes”. Specifically, the density of recent insertions correspond to recent macroevolutionary events while the density of less recent insertions to older events. These results are fully coherent with the “Cold Genome Hypothesis”. In addition, our study supports the Punctuated Equilibria theory in both the phases of radiation and stasis, corroborating the hypothesis that Mammals evolved through punctuated

mechanisms rather than through gradualistic ones.

Introduction

1.1 Gradualism and Punctuated Equilibria

The debate between phyletic gradualism and punctuated equilibria (PE) in evolutionary biology is still intense[1-3]. Phyletic gradualism was embraced by the Modern Synthesis as the evolutionary dynamics for speciation implying a gradual accumulation of mutation throughout time until genetic incompatibilities break up the gene flow. The theory of punctuated equilibria proposes that "[...] evolution is concentrated in very rapid events of speciation (geologically instantaneous, even if tolerably continuous in ecological time). Most species, during their geological history, either do not change in any appreciable way, or else they fluctuate mildly in morphology, with no apparent direction"[4]. Evidences in favor of punctuated equilibria come from diverse fields such as paleontology[4-6], phylogenesis[7,8], experimental evolution[9] and cancer research[10,11]. At the same time, also gradualism gained its evidences[12,13] in various living models.

1.2 The evolutive power of TEs

Transposable elements are far from being junk DNA[14]. In the last years they have been continuously linked to essential cellular activities such as the telomeres repairing[14,16], rewiring of transcriptional networks[17,18], regulation of gene expression[19], ectopic recombination alongside other chromosomal rearrangements[20]. TEs are key contributors of evolution and gave rise to evolutionary features[14,21-27] of utmost importance like the V(D)J system of acquired immunity[22,28,29] and the placenta[30] and they are also involved in embryogenesis[31,32] and neurogenesis[33-35]. Mobile elements make genomes fluid, dynamic and organisms evolvable[35,36].

Given their huge impact on shaping genomes, transposable elements are thought to contribute to the formation of reproductive barriers facilitating speciation[32,38-41]. Some authors

proposed to associate models between high TE activity and organismal differentiation[42-44]. The environment and its influence on the epigenetic structure of the cell modulates the mobilization of TEs. The disruption of the epigenome potentially leads to bursts of activity, thus to a rapid accumulation of genomic variability necessary for phenotypic innovation and speciation[42,43]. Moreover, the diversification of TE families is likely to coincide with events of genetic drift (CArrier SubPopulation hypothesis[44]). The same authors also observe that taxa poor in species seem to be poor in TE content[41,42,44]. Interestingly, these hypotheses[42,43,44] can explain both gradualistic and punctuated evolution on the base of the genome content of TEs.

The debate between gradualism and punctuated equilibria is still open but TEs can shed some light on it.

1.3 Hot and cold genomes

Organisms owe their ability to diversificate to the plasticity of their genomes and the activity of TEs substantially contributes to the genomic plasticity. Hence, we expect that a positive relationship exists between the TEs activity and the extant biodiversity. For example, within Mammals the order Monotremata is the most ancient and the poorest in living species (Figure 1); accordingly, the platypus genome harbors the lowest number of recently mobilized elements[31]. Is it possible that taxa with low rates of speciation are associated with genomes with inactive TEs? Building from the the platypus case, we aim to widen the perspective to a general evolutionary framework that we call the “Cold Genome Hypothesis”. According to our hypothesis, genomes with highly active TEs (“hot genomes”) belong to taxa with high rates of speciation whereas genomes with inactive TEs (“cold genomes”) belong to taxa with low rates of speciation (Figure 2A). We investigate the TE activity using the Density of Insertion (DI) of elements and (as previously proposed in literature[44]) the number of TE families at different time scales. Furthermore, we introduce the concept of

Relative Rate of Speciation (RRS) in order to establish the magnitude of rates of speciation within a given group of taxa (Figure 2B).

Viable elements act like an evolutionary driving force (leading to punctuated events/bursts of insertions/hot genomes) but cells have a plethora of molecular mechanisms aimed to modulate and eventually to fully repress the expression of TEs. Molecular mechanisms seems to intervene in temporally discrete periods to modulate the activity of TEs[45] – for example young LINE-1 elements are repressed via methylation while old ones are repressed by the KRAB/KAP1 system[45] – potentially leading to a paucity of innovative elements hence to both a genomic and macroevolutionary stasis (cold genomes).

In this paper we investigate the mechanisms of speciation in Mammals showing how the gradualistic model is not able to describe the evolution of the extant biodiversity whereas our new evolutionary framework, that includes PE theory and the genomic impact of transposable elements (TEs), is a more powerful and fitting model.

2 Materials and Methods

2.1 Datasets

We retrieved the number of species for all the 152 mammalian families listed in the last mammalian phylogeny[46] from Catalogue of Life[47]. The crown age of the mammalian families was estimated from their timed phylogenetic tree[46].

The profile of TE activity in mammals was retrieved from a previous study[44].

This paper shows the TEs content of 31 mammalian genomes; among them 28 species belong to Placentalia and represent the four superorders Laurasiatheria (L, 9 species), Euarchontoglires (E, 14 species), Afrotheria (A, 3 species) and Xenarthra (X, 2 species).

2.2 Rate of Speciation and Relative Rate of Speciation

We calculated the Rate of Speciation (RS) with the formula: RS = NS / CA, were NS is the

total number of species for the analyzed taxonomical family and CA is the Crown Age of the same taxon.

The Relative Rate of Speciation (RRS) is a binomial parameter, based on the age of the clade of interest and its species richness, that, given a pair of taxa, identifies which one has the highest speciation activity. Briefly, the taxon showing a higher rate of speciation (RS) and, at the same time, a lower age and a greater number of extant species (NS), has a “high” (+) Relative Rate of Speciation and a putative “hot genome”; consequently the other taxon has a “low” RRS and a putative “cold genome” (for details see Figure 2B). This parameter can not be applied on every pair of taxa, for example in this study we couldn’t compare the family Hominidae with Cercopithecidae because the first is younger but still poorer in species than the latter. In this case it is not possibile to distinguish which one has the higher rate of speciation.

RRS is grounded on one assumption: the different number of species found in different taxa is influenced by diverse frequencies of speciation and not only by the random effect of extinctions.

2.3 TE families count and Density of Insertion

For every species’ genome considered, TE families were analyzed by Jurka and colleagues[44]. For each TE family a consensus sequence was produced, representing the reference with which all TEs are compared; the higher the divergence of a sequence from its consensus, the older the sequence is. The mobile elements diverging less than 1% from their consensus sequences and their respective TE families were pooled in the “1% dataset”, which represent the most recently mobilized elements, whereas the ones that diverge less than 5% formed the “5% dataset”, i.e. the list of both recent and ancient insertions.

In order to better evaluate the activity of mobile elements in the considered genomes/species, we propose a new parameter herein called Density of Insertion (DI). DI is calculated

according to the formula: DI = NI/GS, where NI is the total Number of Insertions (of elements contained in the 1% or 5% datasets) and GS is the Genome Size in gigabases. Accordingly, DI is measured in insertions for gigabase (ins/Gb). We called “1% DI” the parameter calculated using the 1% dataset and “5% DI” the one that used the 5% dataset.

All the parameters measuring TEs activity were also averaged between species belonging to the same taxonomical family, in order to perform analyses on a larger scale. Using the same method, 1% DI and 5% DI were also averaged in the four Placetalia superorders and standard errors were calculated.

2.4 Pairwise taxonomical families comparison

The relationship between number of species and clade age was tested using a linear regression model (*lm* function in R, stats package). Moreover, a Spearman non-parametric correlation test was carried out on the same data with the *cor.test* function in R.

In this study, we chose to compare taxonomical families within the same taxonomical order, whose species shared most of their evolutionary history. That way we minimize the background noise caused by evolutionary factors other than transposable elements. Single species were chosen to represent their taxonomical families; for all the chosen pairs, RRS was calculated in order to define which one has, relatively to the other, a putative hot genome and which has a putative cold one. The family pairs involved must present different RRS; on the whole we tested our hypothesis on 19 families of six orders of mammals (Table 1), which were compared through their 1% DI and 5% DI and the count of TE families of the 1% and 5% datasets.

The correspondence between putative hot/cold genomes (based on the RRS) and actual hot/cold genomes (using the four TEs activity parameters) was tested with a paired Wilcoxon Signed Rank Test, using the *wilcox.test* function in R.

2.5 Superorders comparison

The same comparison was performed on the four Placentalia superorders. Laurasiatheria (L) and Euarchontoglires (E) have differentiated more recently and have a higher number of species than Afrotheria (A) and Xenarthra (X); we could thus perform four comparisons: A - E, A - L, X - E, X - L (Table 2), in this case using only the 1% DI and 5% DI averaged parameters.

We also chose to merge the analogous datasets - e.g. pooling the respective 1% datasets together - of Xenarthra and Afrotheria (cold groups and sister clades) and the ones of Euarchontoglires and Laurasiatheria (hot groups and sister clades), in order to obtain datasets large enough to run the Wilcoxon Test on them (*wilcox.test* function in R).

2.6 Correlation between RS and TE activity

Furthermore, we tested the relationship between rate of speciation and the averaged parameters describing the activity of TEs between families across the whole Mammalia class. This was done by checking the existence of a non-parametric statistical correlation (using R function *cor.test* with the Spearman method) and by generating a linear regression model (*lm* function in R, stats package) (Table S2).

3 Results

3.1 Clade age is not related to species richness in Mammalia

From a neodarwinian point of view, species continuously accumulate mutations that would eventually lead to the differentiation and speciation; therefore, older taxonomical groups should have had more time to accumulate biodiversity, leading to an overabundance of species in comparison to younger groups[12]. We tested the phyletic gradualism model in Mammalia by investigating the relationship between clade age and species richness of the 152 mammalian families, using both a linear regression model and a non-parametric correlation test.The calculated regression coefficient is slightly negative (-0.1104) and the R^2 is

very low (0.00041, P-value 0.8039), hence there is no statistically significant association between the two variables (Figure S1). There is no significant correlation between clade age and species richness either (rho 0.01815343, P-value 0.8243). This model does not seem to describe mammalian evolution accurately, as their differentiation pattern seems to behave in a more complex way.

3.2 TEs activity and speciation correlate in the whole Mammalia class

In Figure 3 we show the TEs activity measured for all the families, which are arranged in increasing order of RS (calculated as described in 2.2). There is an increasing trend in all the parameters from left (low Rate of Speciation) to right (high Rate of Speciation).

All of them also show significant correlation with the Rate of Speciation (Supplementary Material Table 2A). In order of significance, the parameters that better relate to the rate of speciation in the whole Mammalia class are: 5% DI (P-value < 0.005), 1% dataset families (P-value < 0.005), 5% dataset families (P-value < 0.05) and 1% DI (P-value < 0.05).

Linear regression models for all the parameters in function of the Rate of speciation were estimated (Table S2). All models show positive angular coefficients and have significant P-values.

3.3 Families with higher Relative Rate of Speciation show higher TE activity

Mammalian phylogeny seems to support an evolutionary framework in which short bursts of diversification are alternated with longer periods of relative stasis, similarly to what is stated in the punctuated equilibria theory[4-6].

In order to validate this observation we tested two key factors, the activity of Transposable Elements and the newly introduced Relative Rate of Speciation, relating them in pairwise

comparisons of species representing taxonomical families (belonging to the same order). In every pair, RRS was evaluated and the two species were sorted: the one with higher/lower RRS was labeled as the species with a putative hot/cold genome respectively. It was not always possible to determine RRS for all the potential pairs of families/species. For instance, we cannot compare the families Hominidae and Cercopithecidae: Hominidae seems to have a lower rate of speciation, with only 7 living species against the Cercopithecidae's 159, but the family Cercopithecidae is 8 Mya older, which means that this taxon could have accumulated more species in a larger amount of time.

The four parameters of TEs activity, used as proxies for evolvability and differentiability as aforementioned, were measured and the levels of 1% DI in all the observed pairs are shown in Figure 4A. Further tests were performed in order to assess the relationship between hot/cold genomes (as assessed by the RRS) and the higher/lower ability to speciate of the clades (measured through the four TEs activity parameters). The content of TEs among the species is not normally distributed so we statistically validated the relationship using the Wilcoxon test: all the W scores and P-values are highly significant (Supplementary Material Table 1). In order of significance (and descriptive efficiency), the parameters that better explain the relationship between hot/cold genomes and TEs content are (P-values in parentheses): 1% DI (0.0009651), 1% dataset families (0.003745), 5% dataset families (0.005329) and 5% DI (0.04937). The descriptive efficiency of the chosen parameters is also presented in Figure 4B. When there is no association between the better descriptive parameter (1% DI) and RRS, the other parameters show no association either.

For 1% DI, 16 out of 19 pairs follow the expected trend of association between DI values and RRS. Among those, 13 pairs show a difference in DI of at least one order of magnitude, up to 300-fold higher in the pair *Homo sapiens* - *Tarsius syrichta*.

Among the three exceptions, the pair belonging to the order Insectivora shows the biggest difference in Density of Insertion, which is three-fold higher in *Erinaceus europaeus* than in

Sorex araneus; in the other two pairs the DI difference is quite lower (*Microcebus murinus - Callithrix jacchus* and *Otolemur garnettii - Callithrix jacchus*). Despite these exceptions, the statistical support clearly suggests that, in Mammals, TEs are associated with adaptive radiation.

It is worth noticing that *Canis lupus* is the only species that, despite the lower total number of mobile elements insertions (lower DI: 194 ins/GB), harbors one more specific TE family (4) than its paired species *Felis catus* (higher DI: 1446 ins/GB, 3 TE families). The other case in which the 1% DI and the number of 1% dataset TE families show discordance is the pair *Tarsius syrichta - Otolemur garnettii*. *Otolemur garnettii* has in fact more new integrants (DI: 51 Ins/Gb) than *Tarsius syrichta* (DI: 6 Ins/Gb) but the same number of TEs families (2 each). In both these cases, the greater diversity in TE families (or absence thereof) does not reflect the relative hotness of the species genome, which is better described by the impact of the elements on their genome. From the example presented and the statistical tests, Density of Insertion is a more sensible parameter than the abundance of TEs families, since the diversity in active families is not always a proxy of the activity of mobile elements in the genome.

The parameters measuring TEs activity confirmed to be accurate even at an intra-family level (as shown in the case of Muridae and Hominidae). In fact, keeping as constant the taxa with lower RRS and interchanging different genera of the same family as a comparison point, we observed variations in the parameters values, but a statistically similar relevance of their association with the RRS (Figure 4B).

The results obtained point to state that the activity of TEs does not vary randomly within the mammalian phylogeny and that the RRS shows strong association with this activity. Therefore, RRS allows prediction of the relative level of TE mobility between two taxa, which in turn is highly related to their ability to differentiate and speciate.

3.4 Ancient TE bursts correlate with the ancient history of Pla-

centalia

Once proved that DI is the most descriptive parameter for intra order activity of TEs, we tested our hypothesis at a higher taxonomic level.

We pooled together superorders of placentals with Higher RRS, Laurasiatheria and Euarchontoglires, and compared them with the merged superorders sharing lower RRS, Xenarthra and Afrotheria. The combined groups consist of 22 species belonging to hot superorders and 5 belonging to cold superorders. For both 1% and 5% datasets, Xenarthra and Afrotheria have an average DI more than tree fold lower than Laurasiatheria and Euarchontoglires. Specifically for the 5% dataset, the pairs A - E and X - E show a ratio of about 1:5; for the pairs A - L and X - L the ratio is about of 1:8. Minor differences can be observed for the average density of insertions from the 1% dataset. In increasing order we have X - E (1:3), E - L (1:4), A - E (1:7), A - L (1:10). The comparison between (X + A) and (E + L) at 1% of divergence from the consensus is non-significant with a P-value of 0.09734.

Barplots in Figure 4C show that, in both comparisons, standard errors of average DI do not overlap. The Wilcoxon test, performed for the average 5% DI parameter, showed significant difference between the two groups (P-value 0.03939). Although the same tendency is clearly visible for the average 1% DI parameter, the Wilcoxon did not give statistically significant results (P-value 0.097). While this is definitely influenced by the small number of species compared (the lower RRS group listed only 5 species, the other one 22), the difference in statistical significance of the data may also have a biological explanation. The activity of TEs reflects the history of genomes and their evolution: the divergence of the elements from their consensus is proportional to their age and thus the different datasets describe different periods of time and related taxonation events. For instance, the density of insertion at 5% divergence is the least accurate parameter to study recent speciation (Fi-

gure 4B) but it works efficiently in the study of older macroevolutive events such as the origin of the four superorders of Placentalia. Since some species of the same pool were separated for a long time, the heterogeneity of values for 1%DI gives a particularly high variance.

Even though TEs seem to be an important factor for the adaptive radiations, the genomic signal they left behind in the extant biodiversity is exposed, as many biological phenomena, to a background noise.

4 Discussion

According to the neodarwinian point of view, species continuously accumulate mutations that would eventually lead to the differentiation and speciation. Therefore, older taxonomical groups have had more time to accumulate species, taking to an overabundance of species compared to younger groups[12].

This relationship seems to be inexistent in Mammals (Figure S1). On the other hand, transposable elements are powerful facilitators of evolution[20-23,40,41] and they are tightly associated to the evolutionary history of mammals[30,31]. Is it possibile that the pattern of speciation in mammals is reflected by the activity of transposable elements?

A first evidence is given by the correlation tests between the rate of speciation (RS, calculated as the number of species divided by the age of the clade) and the activity of TEs (density of insertion and number of TE families) that are all significant indicating that a positive relationship potentially exists.

In order to further test this hypothesis, we made up two parameters that accordingly measure the presence of adaptive radiations and the magnitude and impact of TEs activity on genomes: the Relative Rate of Speciation (RRS) and the Density of Insertion (DI).

The Relative Rate of Speciation is a binomial parameter (higher/lower) designed to establish, within a pair of taxa, which one has the higher rate of speciation based on age and

species richness. Specifically, a taxon has a higher RRS than another one if it is younger and richer in species.

In order to measure the magnitude and impact of transposable elements on genomes, we introduced the Density of Insertion, i.e. the number of insertions per gigabase. Specifically, we distinguish the DI calculated on recent TEs (<1% divergence from consensus) and on less recent elements (<5% divergence from consensus). We also considered the number of TE families (1% and 5% of divergence from consensus), as proposed by Jurka[44], to assess the activity of TEs. We call “hot” and “cold” those genomes with high and low DI/ number of TE families, respectively. It is worth noting, the two biological parameters (phylogeny and mobilome activity) are mutually independent since the first is based on genetic distances and the latter on the non-coding part of genomes.

Once the species with higher and lower RRS are determined, their (1% and 5%) DI are compared pairwise. Given a pair of taxa, we hypothesize that the one with the lower RRS has a putative cold genome, while the other one is defined as the putative hot genome. Here we test if the RRS and the state of the genomes (high/low RRS and hot/cold genome accordingly) show a predictable pattern.

From the available[44] dataset, we selected 19 pairs of species that follow the RRS rules (Table 1). First, we analyze 19 pairs of mammalian families (belonging to the same order) using the RRS associated with 1% and 5% DI (based on 1% and 5% TE families). Second, we shift to the four superorders of Placentalia using the RRS and the mean 1% and 5% DI (Table 2).

Our results (Figure 4B and Table 3) show that the activity of transposable elements is highly associated with the phylogenetic position of the species, revealing that the hot genomes are preferentially found in species belonging to clades with higher RRS. Given the effectiveness of the four parameters taken into consideration (Figure 4B, Table 3), the 1% DI is the most accurate one and the most related to the rate of speciation (Table S2).

However, three pairs of species do not fit the model with 1% DI and, at the same time, they show concordant outcomes for all the other three parameters (Figure 4B). Among these three pairs of species, the pair *Erinaceus europeaus* - *Sorex araneus* belong to the family Insectivora whose phylogenetic positions are not firmly established[48], meaning that this particular pair of species might not be compatible with the evolutionary frame of the RRS. In addition, in the order Carnivora *Felis catus* shows a value of DI 1%, that is five fold higher than *Canis lupus*, but the dog genome harbors one exclusive TE family. Consequently, we can suppose that the density of insertions is a better proxy of genomic diversification than the number of TEs families. Despite these exceptions, TEs in Mammals are significantly associated with their adaptive radiation.

We also tested the Cold Genome Hypothesis at an higher taxonomic level than comparing the four superorders of Placentalia using the parameter of the average DI. As for the 1% dataset (Figure 4C), the putative colder taxa are well differentiated in absolute values and the standard errors of the pairs X - A and X - E overlap one another only slightly. On the other hand the comparison at 5% of divergence is significant (0.03939) (Figure 4C). It must be noticed that given the biological paucity of species in the groups Xenarthra and Afrotheria, their 1% and 5% datasets are very small and heterogeneous (i.e. low statistical power). For this reason it is hard to statistically compare them to the large datasets of Euarchontoglires and Laurasiatheria.

The discrepancy between the results of the two datasets may be interpreted from an evolutionary point of view. The density of insertion at 5% is the least accurate parameter to study recent events (Figure 4B and Table 3) but it works efficiently in the study of older macroevolutive events such as the origin of the four superorders of Placentalia (Figure 4C). The divergence of the elements from their consensus tend to reflect their age and thus the different datasets describe, accordingly to age, different periods of time and related taxonation events. Even though TEs seem to be an important factor for the adaptive

radiations, the genomic signal they left in the extant biodiversity is exposed, as many biological phenomena, to a background noise.

The cold genome hypothesis supports punctuated equilibria in both the punctuated events and stasis. The punctuated events are limited periods in time in which occurs a great mole of changes and they may correspond with the state of hot genome. In these periods the creative contribution of TEs would guide the adaptive radiation. The stasis coincides with the state of cold genome, whereas through the molecular mechanisms above-mentioned, the genomes are less “fluid” and the evolutive thrust is limited.

According to the Modern Synthesis, macroevolution is nothing else than the long-term result of microevolution. This vision should reconcile itself with the periods of acceleration of the evolutive rates and with the periods of stasis. The stasis is usually explained with the lack of genetic variability or with the stabilizing selection[48], but both the explanations are problematic. Species evolve rapidly in the short-term but not in the long-term even when they possess large genetic variability as in the Grants’ 30 year long experiment on *Geospiza fortis*[50,51]. Lieberman and Dudgeon[52] cast doubts on the stabilizing selection observing that the species of interest showed more morphological changes during periods of environmental stability than across periods of abiotic and biotic changes. Our study on TEs seems to give a more consistent evolutionary framework. TEs can explain[43] how a species shifts from its adaptive peak to another one and they can exert a huge influence on the insurgence of reproductive barriers, whether these are caused by genetic remodeling or regulatory innovation[44,53].

Finally, the state of hot and cold can be considered as a proxy of the genomic biodiversity and, consequently, of the potential production of new phenotypes or of the tendency to stasis (and possibly to extinction).

5 Conclusions

In this study, we put side by side the activity of TEs and the rate of speciation in Mammals. Despite the background genomic noise, our results indicate firmly that the TEs activity does not vary randomly within the phylogeny of Mammals: the more rich (in TEs) a clade is, the more its mobilome is active. We call this evolutionary pattern “cold genome hypothesis”. Moreover, our results suggest that TEs activity coincides with the evolutionary pattern predicted by punctuated equilibria.

To establish this relationship, we defined two new parameters: 1) the Relative Rate of Speciation (RRS), which is aimed to detect potential punctuated equilibria events, and 2) the Density of Insertions (DI), that evaluates the degree of TEs activity.

Together, these two parameters revealed a pattern of TEs activity that coincides with the evolutionary pattern predicted by punctuated equilibria. Thus, we show how punctuated equilibria describe the macroevolution in Mammals and their signature can be detected from the activity of transposable elements.

Furthermore, we evidence that TEs insertions describe the occurrence of clades accordingly to their age: old insertions are a proxy for older events of taxonation (mammalian superorders), recent for late events (mammalian families).

Previous studies, empirical or theoretical[38-45], linked the TE proliferation to the emergence of new species, but here the first time we depict an evolutionary framework and a computational approach in order to validate or invalidate the hypothesis.

The next steps of our work are: 1) to test the Cold Genome hypothesis in clades other than Mammals; 2) to understand whether bursts of TEs occur before (as one of the causes) or after (as effect) the speciation events.

References

1. Pennell, M., Harmon, L. & Uyeda, J. Is there room for punctuated equilibrium in macroevolution?. *Trends in Ecology & Evolution* 29, 23-32 (2014).

2. Lieberman, B. & Eldredge, N. What is punctuated equilibrium? What is macroevolution? A response to Pennell et al. *Trends in Ecology & Evolution* 29, 185-186 (2014).

3. Casane, D. & Laurenti, P. Why coelacanths are not 'living fossils'. *Bioessays* 35, 332-338 (2013).

4. Gould, S. & Eldredge, N. Punctuated equilibria: the tempo and mode of evolution reconsidered. *Paleobiology* 3, 115-151 (1977).

5. Eldredge, N. & Gould, S. Punctuated equilibrium: an alternative to phyletic gradualism. *Models in Paleobiology* 82-115 (1972).

6. Gould, S. *The structure of evolutionary theory*. (Belknap Press of Harvard University Press, 2002).

7. Pagel, M., Venditti, C. & Meade, A. Large Punctuational Contribution of Speciation to Evolutionary Divergence at the Molecular Level. *Science* 314, 119-121 (2006).

8. Mattila, T. & Bokma, F. Extant mammal body masses suggest punctuated equilibrium. *Proceedings of the Royal Society B: Biological Sciences* 275, 2195-2199 (2008).

9. Lenski, R. & Travisano, M. Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proceedings of the National Academy of Sciences* 91, 6808-6814 (1994).

10. Baca, S. et al. Punctuated Evolution of Prostate Cancer Genomes. *Cell* 153, 666-677 (2013).

11. Cross, W., Graham, T. & Wright, N. New paradigms in clonal evolution: punctuated equilibrium in cancer. *J. Pathol.* (2016). doi:10.1002/path.4757

12. McPeek, M. & Brown, J. Clade Age and Not Diversification Rate Explains Species Richness among Animal Taxa. *The American Naturalist* 169, E97-E106 (2007).

13. Stadler, T., Rabosky, D., Ricklefs, R. & Bokma, F. On Age and Species Richness of Higher Taxa. *The American Naturalist* 184, 447-455 (2014).

14. Biemont, C. A Brief History of the Status of Transposable Elements: From Junk DNA to Major Players in Evolution. *Genetics* 186, 1085-1093 (2010).

15. Farkash, E. & Prak, E. DNA Damage and L1 Retrotransposition. *Journal of Biomedicine and Biotechnology* 2006, 1-8 (2006).

16. Muotri, A., Marchetto, M., Coufal, N. & Gage, F. The necessary junk: new functions for transposable elements. *Human Molecular Genetics* 16, R159-R167 (2007).

17. Lynch, V., Leclerc, R., May, G. & Wagner, G. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nature Genetics* 43, 1154-1159 (2011).

18. Kunarso, G. et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature Genetics* 42, 631-634 (2010).

19. Faulkner, G. et al. The regulated retrotransposon transcriptome of mammalian cells. *Nature Genetics* 41, 563-571 (2009).

20. Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics* 9, 397-405 (2008).

21. Fedoroff, N. Transposable Elements, Epigenetics, and Genome Evolution. *Science* 338, 758-767 (2012).

22. Kapitonov, V. & Jurka, J. RAG1 Core and V(D)J Recombination Signal Sequences Were Derived from Transib Transposons. *PLoS Biology* 3, e181 (2005).

23. Feschotte, C. & Pritham, E. DNA Transposons and the Evolution of Eukaryotic Genomes. *Annu. Rev. Genet.* 41, 331-368 (2007).

24. Emera, D. & Wagner, G. Transposable element recruitments in the mammalian placenta: impacts and mechanisms. *Briefings in Functional Genomics* 11, 267-276 (2012).

25. Piskurek, O. & Jackson, D. Transposable Elements: From DNA Parasites to Architects of Metazoan Evolution. *Genes* 3, 409-422 (2012).

26. Richardson, S. et al. The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. *Microbiology Spectrum* 3, (2015).

27. Carmi, S., Church, G. & Levanon, E. Large-scale DNA editing of retrotransposons accelerates mammalian genome evolution. *Nature Communications* 2, 519 (2011).

28. Koonin, E. & Krupovic, M. Evolution of adaptive immunity from transposable elements combined with innate immune systems. *Nat Rev Genet* 16, 184-192 (2014).

29. Huang, S. et al. Discovery of an Active RAG Transposon Illuminates the Origins of V(D)J Recombination. *Cell* 166, 102-114 (2016).

30. Lavialle, C. et al. Paleovirology of 'syncytins', retroviral env genes exapted for a role in placentation. *Philosophical Transactions of the Royal Society B: Biological Sciences* 368, 20120507-20120507 (2013).

31. Gerdes, P., Richardson, S., Mager, D. & Faulkner, G. Transposable elements in the mammalian embryo: pioneers surviving through stealth and service. *Genome Biol* 17, (2016).

32. Friedli, M. & Trono, D. The Developmental Control of Transposable Elements and the Evolution of Higher Species. *Annual Review of Cell and Developmental Biology* 31, 429-451 (2015).

33. Notwell, J., Chung, T., Heavner, W. & Bejerano, G. A family of transposable elements co-opted into developmental enhancers in the mouse neocortex. *Nature Communications* 6, 6644 (2015).

34. Levin, H. & Moran, J. Dynamic interactions between transposable elements and their hosts. *Nat Rev Genet* 12, 615-627 (2011).

35. Singer, T., McConnell, M., Marchetto, M., Coufal, N. & Gage, F. LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes?. *Trends in Neurosciences* 33, 345-354 (2010).

36. Schrader, L. et al. Transposable element islands facilitate adaptation to novel environments in an invasive species. *Nature Communications* 5, 5495 (2014).

37. Hof, A. et al. The industrial melanism mutation in British peppered moths is a transposable element. *Nature* 534, 102-105 (2016).

38. Belyayev, A. Bursts of transposable elements as an evolutionary driving force. *Journal of Evolutionary Biology* 27, 2573-2584 (2014).

39. Senerchia, N. et al. Differential introgression and reorganization of retrotransposons in hybrid zones between wild wheats. *Molecular Ecology* 25, 2518-2528 (2016).

40. Oliver, K., McComb, J. & Greene, W. Transposable Elements: Powerful Contributors to Angiosperm Evolution and Diversity. *Genome Biology and Evolution* 5, 1886-1901 (2013).

41. Feiner, N. Accumulation of transposable elements in Hoxgene clusters during adaptive radiation of *Anolis lizards*. *Proceedings of the Royal Society B: Biological Sciences* 283, 20161555 (2016).

42. Oliver, K. & Greene, W. Transposable elements and viruses as factors in adaptation and evolution: an expansion and strengthening of the TE-Thrust hypothesis. *Ecol Evol* 2, 2912-2933 (2012).

43. Zeh, D., Zeh, J. & Ishida, Y. Transposable elements and an epigenetic basis for punctuated equilibria. *BioEssays* 31, 715-726 (2009).

44. Jurka, J., Bao, W. & Kojima, K. Families of transposable elements, population structure and the origin of species. *Biology Direct* 6, 44 (2011).

45. Castro-Diaz, N., Friedli, M. & Trono, D. Drawing a fine line on endogenous retroelement activity. *Mobile Genetic Elements* 5, 1-6 (2015).

46. Meredith, R. et al. Impacts of the Cretaceous Terrestrial Revolution and KPg Extinction on Mammal Diversification. *Science* 334, 521-524 (2011).

47. www.catalogueoflife.com

48. Symonds, M. Phylogeny and life histories of the ‘Insectivora’: controversies and consequences. *Biological Reviews* 80, 93-128 (2005).

49. Bokma, F. Evolution as a Largely Autonomous Process. *Interdisciplinary Evolution Research* 87-112 (2015). doi:10.1007/978-3-319-15045-1_3

50. Grant, P. Unpredictable Evolution in a 30-Year Study of Darwin's Finches. *Science* 296, 707-711 (2002).

51. Grant, B. & Grant, P. Fission and fusion of Darwin's finches populations. *Philosophical Transactions of the Royal Society B: Biological Sciences* 363, 2821-2829 (2008).

52. Lieberman, B. & Dudgeon, S. An evaluation of stabilizing selection as a mechanism for stasis. *Palaeogeography, Palaeoclimatology, Palaeoecology* 127, 229-238 (1996).

53. Jurka, J., Bao, W., Kojima, K., Kohany, O. & Yurka, M. Distinct groups of repetitive families preserved in mammals correspond to different periods of regulatory innovations in vertebrates. *Biology Direct* 7, 36 (2012).

Figures and Tables

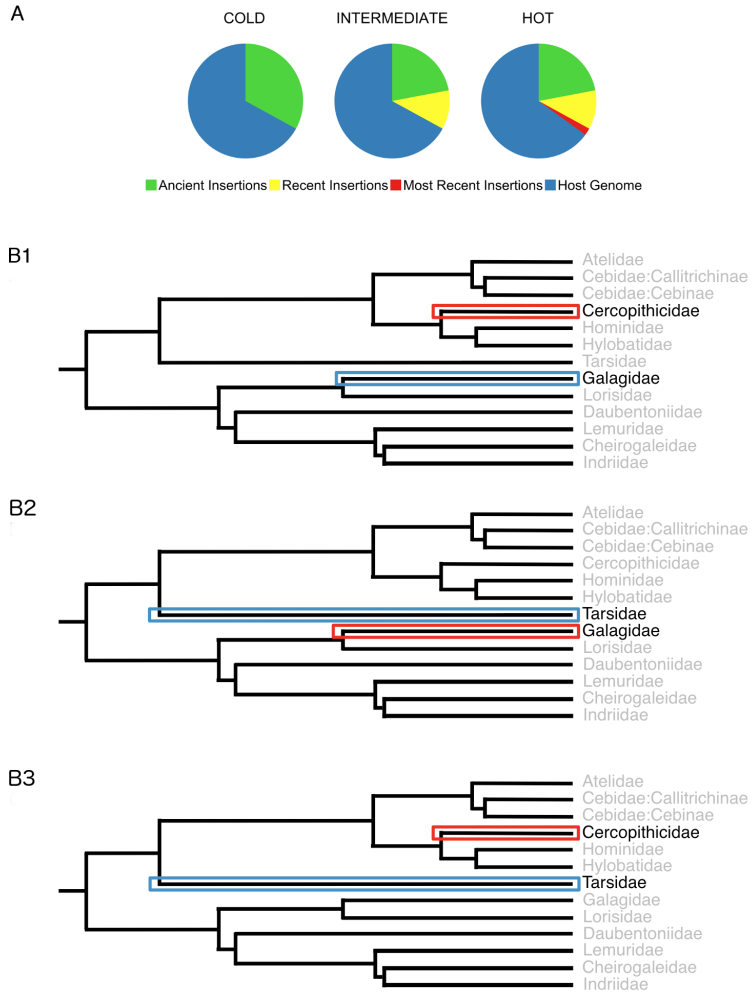


Figure 2. A) Modelization of the Cold Genome hypothesis. The cold genome presents only non-viable TEs. The intermediate genome model contains a fraction of TEs recently mobilized (divergent less than 5% from their consensus sequence). The hot genome has also a fraction of TEs more recently mobilized (divergent less than 1% from their consensus sequence).
B) Exemplified Relative Rate of Speciation in families of the order Primates. B1) Galagidae is older and poorer in species than Cercopithecidae so it has a lower RRS. Since our parameter is relative, in B2) Galagidae is younger and richer in species than Tarsidae, so in this case Galagidae has a higher RRS than Tarsidae. B3) Cercophitecidae is younger and richer in species than Tarsidae so it has a higher RRS than Tarsidae.

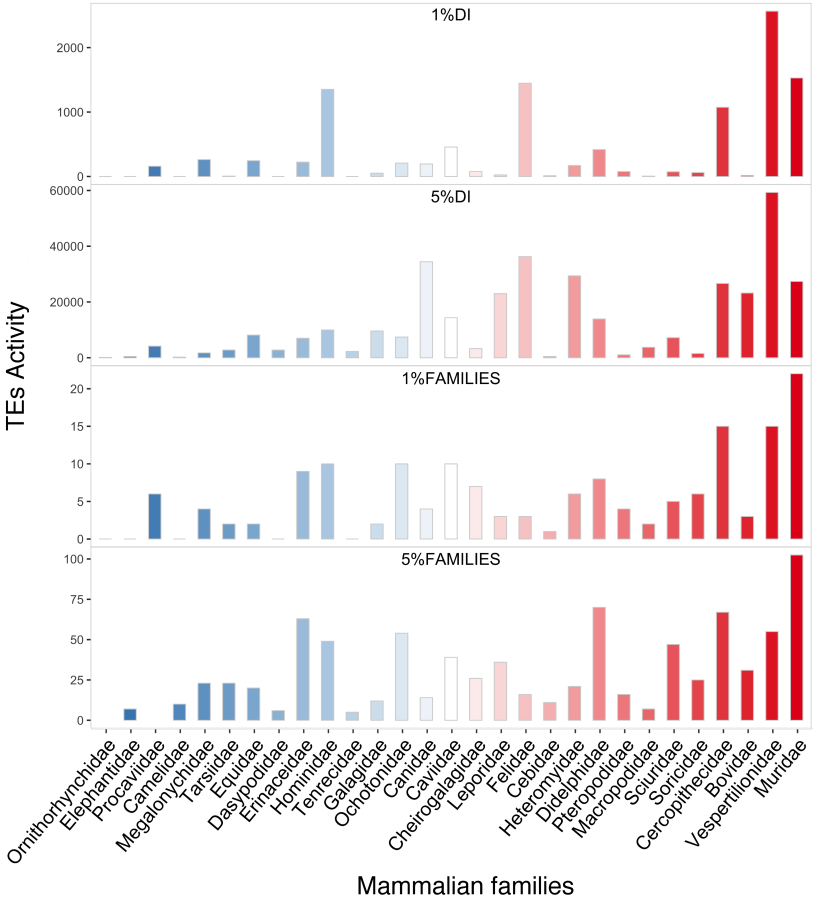
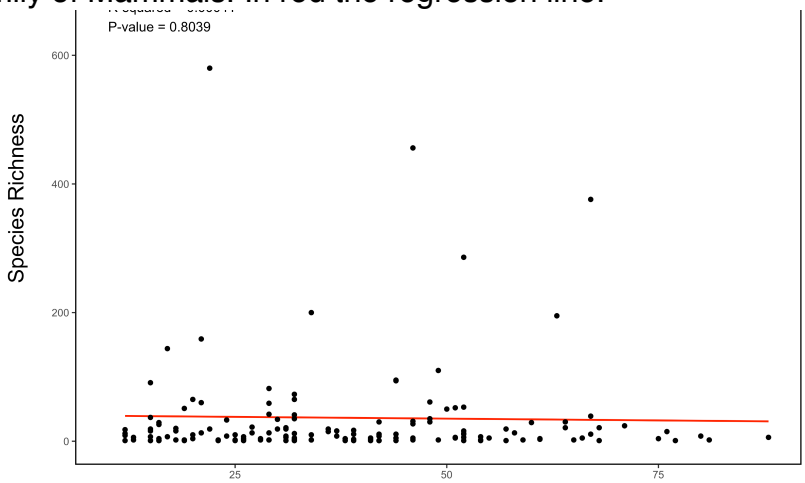


Figure 3. Relationship between the rate of speciation (RS) and TE activity in 29 mammalian families. The families are arranged in crescent order of RS.

Figure S1. Relationship between number of species and crown clade age. Each point represents a family of Mammals. In red the regression line.



Microcebus murinus	*Macaca mulatta*	✓	✓	✓	✓
Microcebus murinus	*Callithrix jacchus*	✗	✗	✗	✗
Otolemur garnetti	*Macaca mulatta*	✓	✓	✓	✓
Otolemur garnettii	*Microcebus murinus*	✓	✓	✓	✗
Otolemur garnettii	*Callithrix jacchus*	✗	✗	✗	✗
Dipodomys ordii	*Mus musculus*	✓	✓	✓	✗
Dipodomys ordii	*Rattus norvegicus*	✓	✓	✓	✗
Spermophilus tridecemlineatus	*Mus musculus*	✓	✓	✓	✓
Spermophilus tridecemlineatus	*Rattus norvegicus*	✓	✓	✓	✓

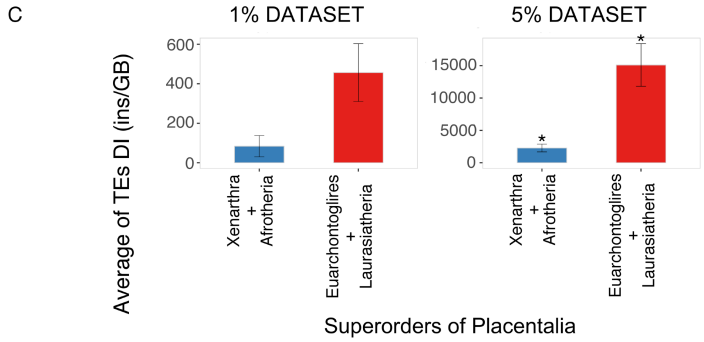
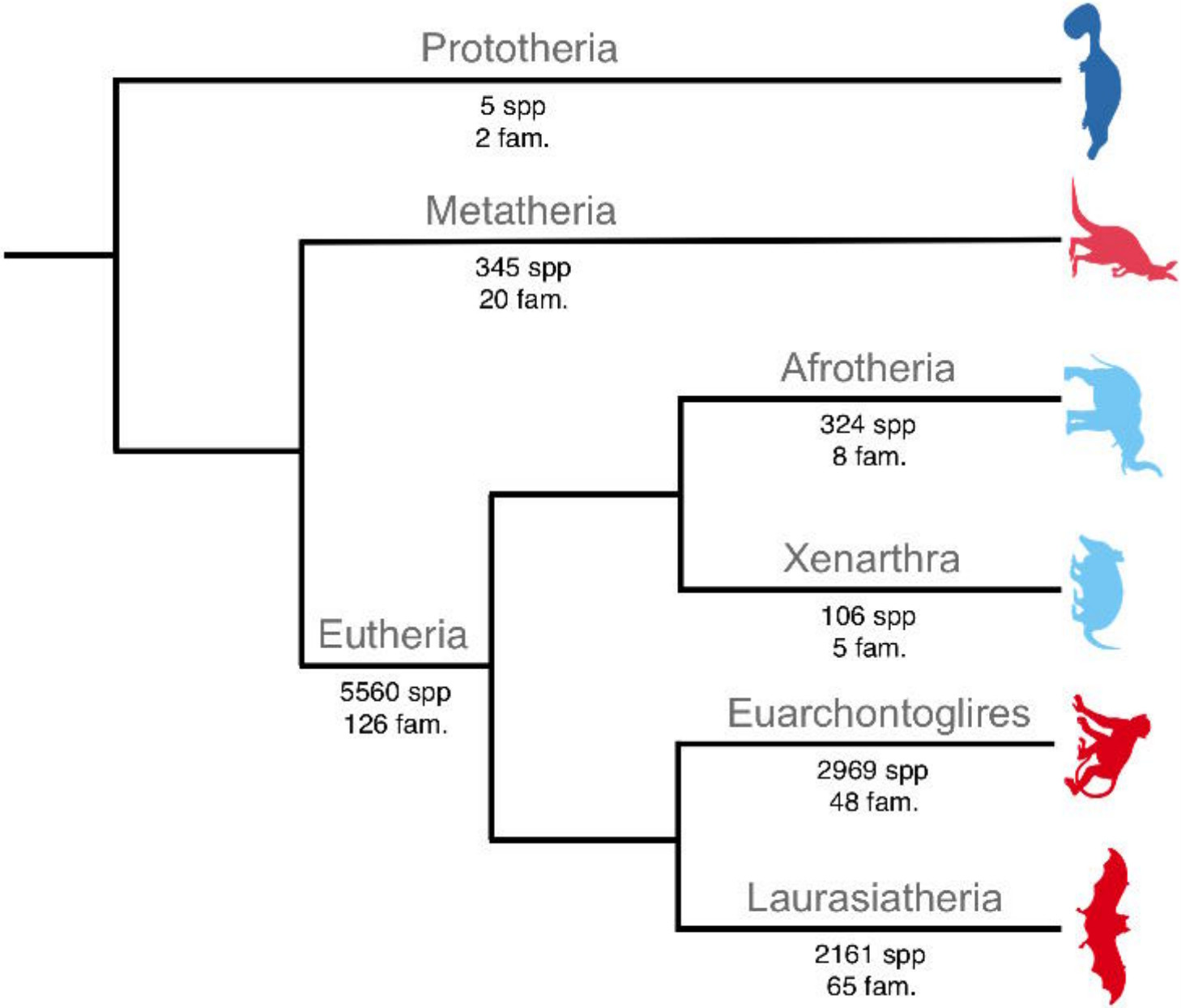
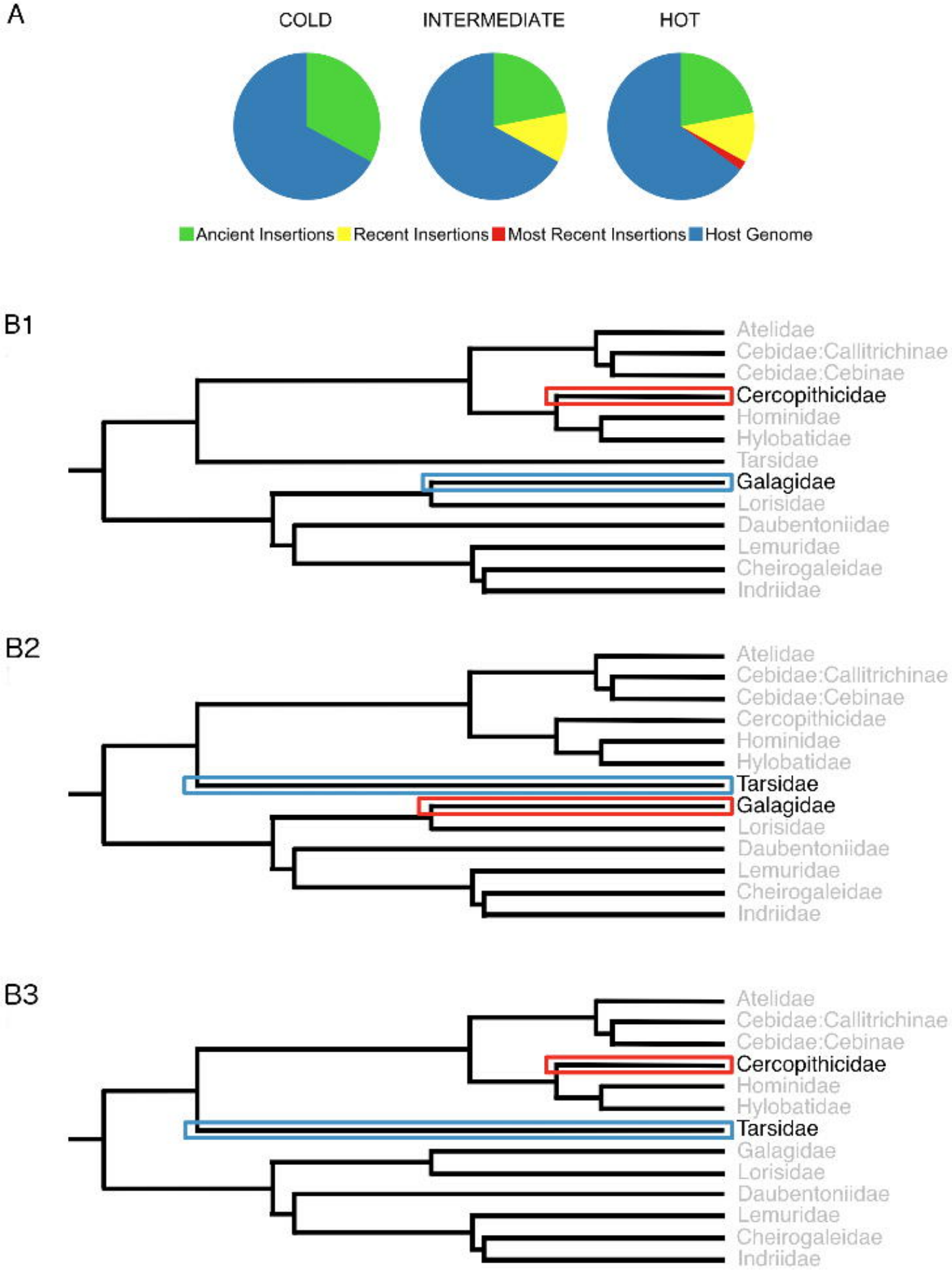
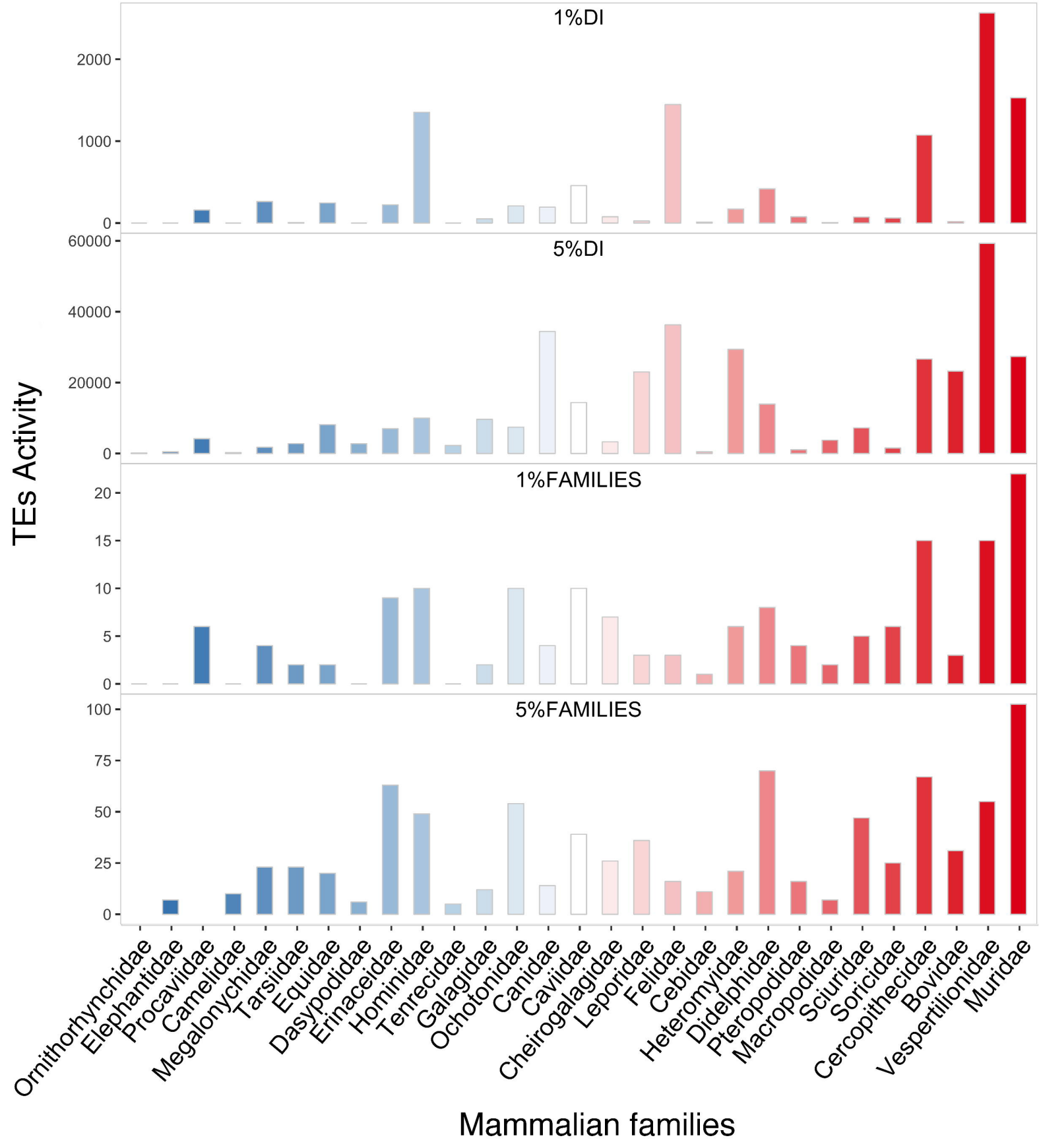


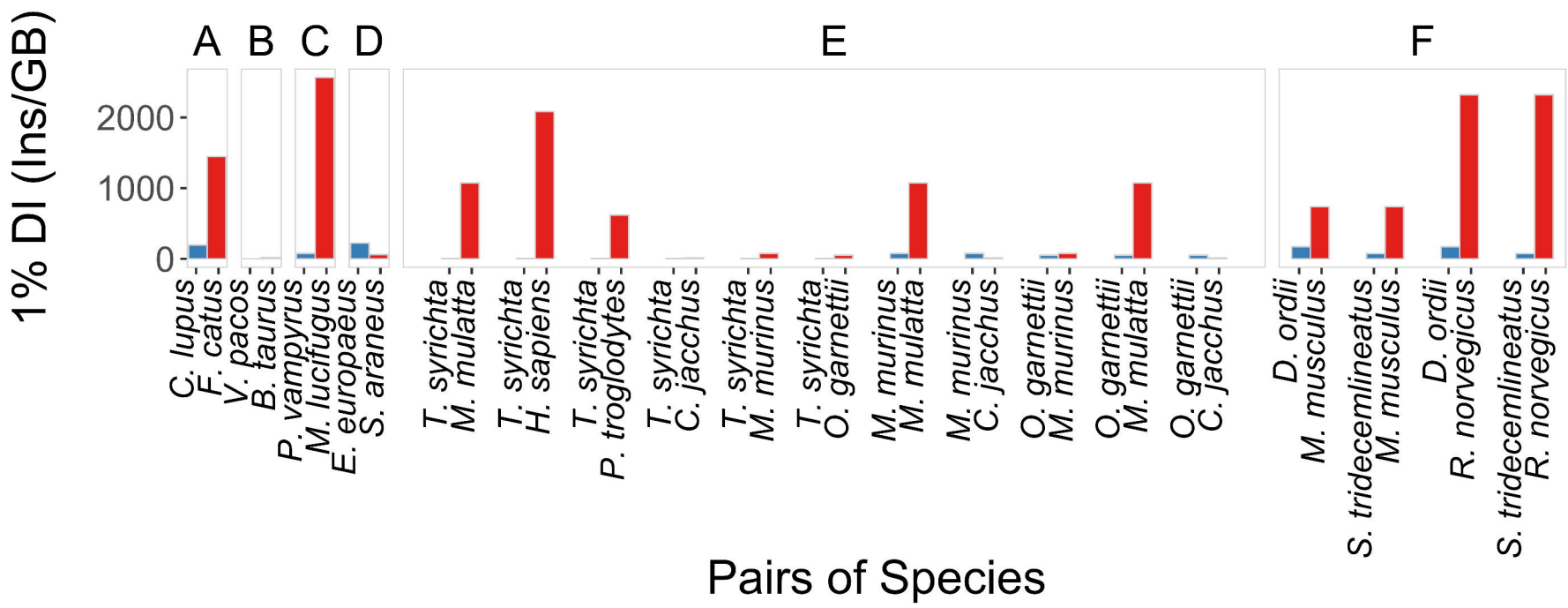
Figure 4. A) Comparison of the DI 1% in the 19 pairs of Mammals. Blue bars: putative cold genomes; Red Bars: putative hot genomes. The barplot is divided per order: A) Carnivora, B) Cetartiodactyla, C) Chiroptera, D) Insectivora, E) Primates, F) Rodentia.
B) Efficiency of the four parameters used. Green ticks and red crosses indicate, respectively, the pairs of species where the model fits or not.
C) Comparison of the DI 1% and DI 5% in the 4 superorders of Placentalia. Blue bars: putative cold genomes; Red Bars: putative hot genomes. * pvalue < 0.05


Prototheria
5 spp
2 fam.
Metatheria
345 spp
20 fam.
Afrotheria
324 spp
8 fam.
Xenarthra
106 spp
5 fam.
Eutheria
5560 spp
126 fam.
Euarchontoglires
2969 spp
48 fam.
Laurasiatheria
2161 spp
65 fam.

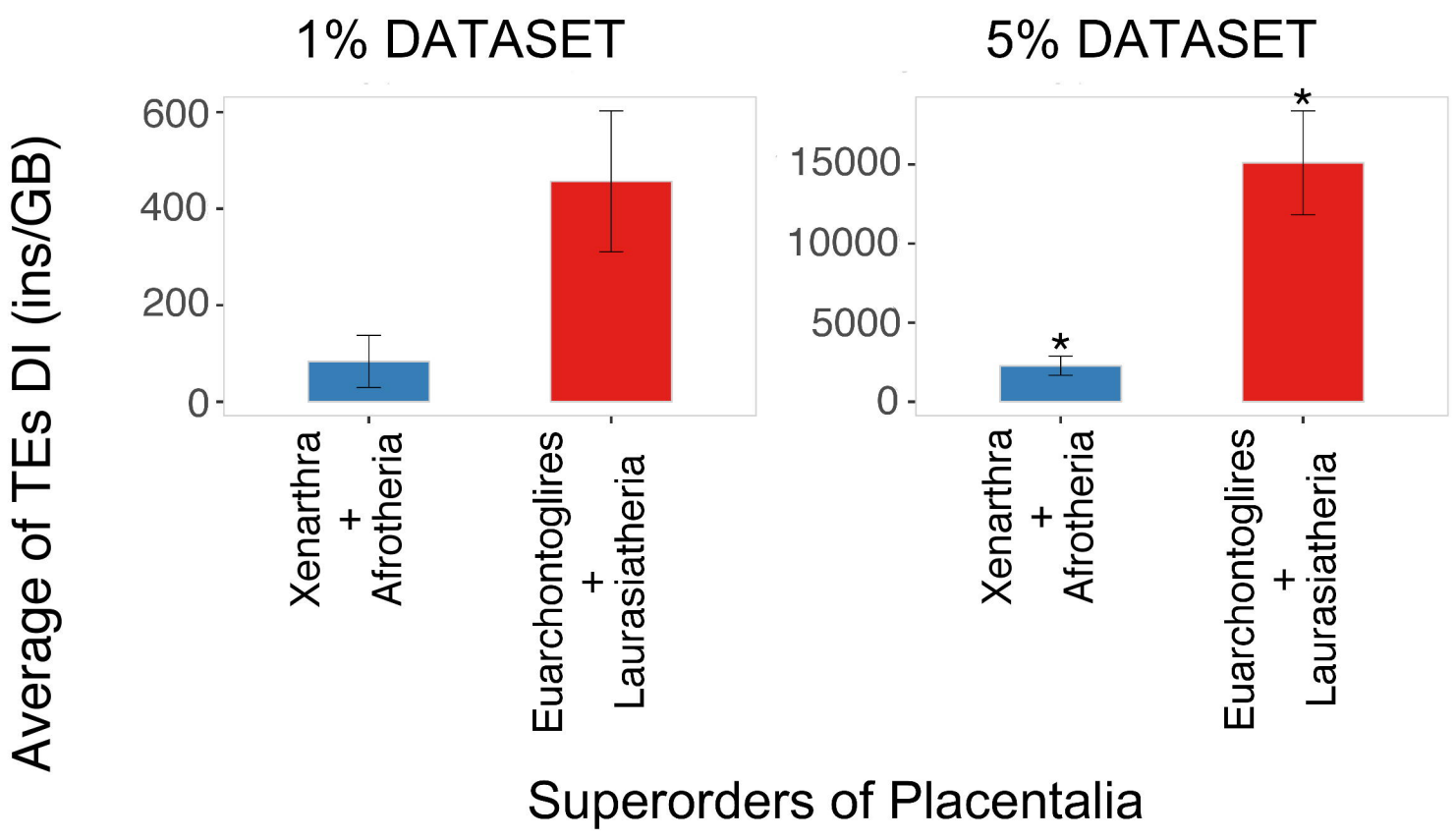

A
COLD
INTERMEDIATE
HOT
Ancient Insertions
Recent Insertions
Most Recent Insertions
Host Genome
B1
Atelidae
Cebidae:Callitrichinae
Cebidae:Cebinae
Cercopithicidae
Hominidae
Hylobatidae
Tarsidae
Galagidae
Lorisidae
Daubentoniidae
Lemuridae
Cheirogaleidae
Indriidae
B2
Atelidae
Cebidae:Callitrichinae
Cebidae:Cebinae
Cercopithicidae
Hominidae
Hylobatidae
Tarsidae
Galagidae
Lorisidae
Daubentoniidae
Lemuridae
Cheirogaleidae
Indriidae
B3
Atelidae
Cebidae:Callitrichinae
Cebidae:Cebinae
Cercopithicidae
Hominidae
Hylobatidae
Tarsidae
Galagidae
Lorisidae
Daubentoniidae
Lemuridae
Cheirogaleidae
Indriidae


1%DI
5%DI
1%FAMILIES
5%FAMILIES
TEs Activity
Mammalian families
Ornithorhynchidae
Elephantidae
Procaviidae
Camelidae
Megalonychidae
Tarsiidae
Equidae
Dasypodidae
Erinaceidae
Hominidae
Tenrecidae
Galagidae
Ochotonidae
Canidae
Caviidae
Cheirogalagidae
Leporidae
Felidae
Cebidae
Heteromyidae
Didelphidae
Pteropodidae
Macropodidae
Sciuridae
Soricidae
Cercopithecidae
Bovidae
Vespertilionidae
Muridae

A



B

Low RRS	High RRS	DI 1%	Families 5%	Families 1%	DI 5%
Canis lupus	*Felis catus*	✓	✓	✗	✓
Lama pacos	*Bos taurus*	✓	✓	✓	✓
Pteropus vampyrus	*Myotis lucifugus*	✓	✓	✓	✓
Erinaceus europaeus	*Sorex araneus*	✗	✗	✗	✗
Tarsius syrichta	*Macaca mulatta*	✓	✓	✓	✓
Tarsius syrichta	*Homo sapiens*	✓	✓	✓	✓
Tarsius syrichta	*Pan troglodytes*	✓	✓	✓	✓
Tarsius syrichta	*Callithrix jacchus*	✓	✗	✗	✗
Tarsius syrichta	*Microcebus murinus*	✓	✓	✓	✓
Tarsius syrichta	*Otolemur garnettii*	✓	✗	✗	✓
Microcebus murinus	*Macaca mulatta*	✓	✓	✓	✓
Microcebus murinus	*Callithrix jacchus*	✗	✗	✗	✗
Otolemur garnetti	*Macaca mulatta*	✓	✓	✓	✓
Otolemur garnettii	*Microcebus murinus*	✓	✓	✓	✗
Otolemur garnettii	*Callithrix jacchus*	✗	✗	✗	✗
Dipodomys ordii	*Mus musculus*	✓	✓	✓	✗
Dipodomys ordii	*Rattus norvegicus*	✓	✓	✓	✗
Spermophilus tridecemlineatus	*Mus musculus*	✓	✓	✓	✓
Spermophilus tridecemlineatus	*Rattus norvegicus*	✓	✓	✓	✓

C