Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

BIODIVERSITÀ ED EVOLUZIONE

Ciclo XXV

Settore scientifico-disciplinare di afferenza: BIO-05 Zoologia

Settore concorsuale di afferenza: 05/B1 Zoologia e Antropologia

TRANSPOSABLE ELEMENTS DYNAMICS IN TAXA WITH DIFFERENT REPRODUCTIVE STRATEGIES OR SPECIATION RATE

Presentata da: Dr. Marco Ricci

Coordinatore Dottorato

Prof. Barbara Mantovani

Relatore

Prof. Barbara Mantovani

Esame finale anno 2014

INDEX

Chapter1

INTRODUCTION

1.1 Transposable Elements

Most part of eukaryotic genomes is composed by repetitive sequences: these can be divided in tandem and interspersed repeats. The latter are in most part Transposable Elements (TEs). These genetic elements have the skill to move along the genome (McClintock 1950). Over the past 60 years, research on TEs showed their tremendous biodiversity, mainly due to the different transposition mechanisms, leading to TEs ordering into families, superfamilies and classes. Class I is formed by retrotransposons; these are transposable elements moving through the element transcription, reverse transcription and insertion of the cDNA in the genome. Such replicative cycle is carried out by recruiting RNA polymerase of the host, while the reverse transcription is performed by enzymes encoded by the retrotransposons. Retrotransposons are divided into two superfamilies, LTR and non-LTR elements. LTR retrotransposons are strongly related to retroviruses, the main difference between them being the presence of a functioning ENV protein that enable the formation of the envelope in the cell membrane of the host and the resulting spread of the virus in the external environment. LTR retrotransposons comprise two regions called Long Terminal Repeats, such sequences regulating the activation and termination of transcription and the integration of the cDNA in the host genome. Among the two LTRs they contain ORFs for GAG proteins involved in the formation of retroviral particles and for POL which encodes an aspartic protease, reverse transcriptase proteinase (AP), reverse transcriptase, RNase H and DDE integrase.

The LTR superfamily comprises two families of non-autonomous retrotransposons The LArge Retrotransposon Derivatives (LARDs), Terminal Repeat retrotransposons In Miniature (TRIMs).

(Reviewed by Wicker et al. 2007). The non-LTR superfamily is divided between autonomous and non-autonomous elements. Autonomous non-LTR retrotransposons are present in all eukaryotes and are probably the oldest retrotransposable elements present in animal genomes. Phylogenetic analysis reveal that all non-LTR families date to the Cambrian period (Harmit et al. 1999). All families are provided with a POL-like region coding for reverse transcriptase and endonuclease. Unlike LTR retroptransposons, non-LTR elements can use the host DNA repair enzymes to complete their replication cycle. Non autonomous non-LTR retrotransposons are segments provided with a transcriptional promoter at the 5' end and a sequence for recruitment of reverse transcriptase. The enzymes for reverse transcription are provided by a non-LTR autonomous retrotransposon which in practice are parasitized. Two families of non autonomous non-LTR retrotransposon are currently known: Short INterspersed Elements (SINEs) and Sine Vntr Alu (SVA) (reviewed by Rebollo et al. 2010). Class 2 includes all the transposable elements that don't perform the reverse transcription in their replicative cycle. This class consists mainly of transposons that move through a "cut and paste" mechanism and are characterized by Terminal Inverted Repeats (TIRs) of 10-200 bps flanking the sequence encoding transposase. This enzyme cuts the element at the level of TIRs, the subsequent insertion occurring in a new target site of the host genome. Miniature inverted repeated elements (MITEs) are a class I of non-autonomous elements composed of two TIRs, flanking a short sequence devoid of an ORF. These TEs are trans-activated by transposases belonging to autonomous elements which recognize TIRs as triggers of transposition. There are two TE families of Class II that use an alternative apparatus for replication: the Helitron family, coding for the enzyme helicase which replicates DNA by a rolling circle mechanism and the Polinton elements, the TEs of greatest length yet discovered. Their size can be up to 20 kb. They are the only transposons encoding a DNA polymerase (Pritham et al. 2007). The replicative mechanism suggested provides the excision of a single strand followed by extra chromosomal

replication and subsequent integration in a new site. Forty years ago Susumu Ohno coined the term "Junk DNA" to define the non-coding part of eukaryotic genomes. Among the various types of Junk DNA, transposons were defined as "selfish DNA" meaning that they were not specifically contributing to the phenotype (Orgel and Crick 1980). Currently available data indicate that TEs are involved in a big amount of aspects of eukaryotic organisms. The amount of TE is directly correlated with the genome size (Kidwell 2002). They are involved in the network of gene regulation, as the SINE's derived micro RNAs (Ponicsan et al. 2010). Surprising discoveries have recently been made about the involvement of retrotransposons in developmental biology, as the mobilization of retrotransposons L1 during development of the human brain to generate somatic mosaicism in the brain and increase intra individual neuronal variability (Thomas and Muotri 2012; Singer et al. 2010). The most important aspect of the study of TEs, is their relationship with the evolution of genomes genomes and genes. RAG1/RAG2 recombinase genes are responsible for the hipermutation of V(D)J loci. This is the central mechanism for the specific acquired immunity in vertebrates. The catalytic process of cut and paste of RAG1/RAG2 complex is probably derived from a transposase belonging to an ancient class II TE (Fugmann 2010). Many other TE mediated mechanisms are known, such as exon shuffling (Moran et al. 1999), unequal crossover, gene duplications (Muotri et al. 2007), TE exonizations (Schmitz and Brosius 2011), inversions (Lee et al. 2008). These data allow us to look at TEs like the main cause of mutations and increase of complexity of eukaryotic genomes. The relationship between TE and host genome is therefore dual. Even if nowadays the role of simple genomic parasite is ruled out, a deleterious effect on the genome is widely documented. In *Homo sapiens* the disruption of gene function by TE in the germ line results in genetic diseases as Neurofibromatosis, Choroideremia, Cholinesterase deficiency, Apert syndrome, Dent's disease, β-thalassemia, and Walker-Warburg syndrome. In addition, TE insertions in somatic tissues have been correlated with various type of tumors as the disruption of BRCA1 and BRCA2 genes in breast cancer and APC in colon cancer (Belancio et al. 2009.).

1.2 Relationship among TE dynamics and reproductive strategies

The main theory tells that in bisexual taxa TEs proliferation is tolerated because of the possibility to counteract their accumulation through recombination, while unisexuals must repress TEs activity to avoid their indiscriminate accumulation leading towards extinction. Computer simulations demonstrated that small unisexual taxa populations may go extinct by a Muller's ratchet-like mechanisms (Dolgin and Charlesworth 2006). To test what happens in nature, evolutionary biology needs suitable biological models. In rotifers, the species of the class Bdelloidea are known to be ancient asexuals. The comparisons with their sister clade Monogononta showed a reduced activity of class I TEs (Arkipova and Meselson 2005). The same trend was observed in the crustacean *Daphnia pulex* (Branchiopoda Cladocera), the species comprising both cyclic and obligatory parthenogenetic populations. Transposon display assays showed that cyclically parthenogenetic populations have more class II TEs than obligate parthenogens (Shaak et al. 2010). *Leptopilina clavipes* (Insecta Hymenoptera) is a parasitoid wasp that present gonochoric reproduction and a lineage of the same species that was rendered unisexual by *Wolbachia*-induced parthenogenesis. The analyses performed showed higher amount of class II and of some families of LTR and non-LTR retrotransposons in the unisexual lineage differently from what expected and observed in other taxa. (Kraaijeveld et al. 2012). To clarify this aspect further animal models are needed.

1.3 A new model for the study of TE dynamics: the genus *Bacillus*

The animal kingdom comprises 1.263.186 species so far described, the 75% of which belongs to the Class Insecta. Actually, insects are the animal taxon under the biggest radiation (http://www.catalogueoflife.org/col/). Phasmida is one of the 11 orders belonging Polyneoptera. It is composed by 3000 species diffused in all continents. Phasmida are characterized by cryptic mimetism, leading to they being generally known as leaf and stick insects. Among stick insects, the genus *Bacillus* is distributed in the Mediterranean area and shows one of the best known examples of reticulated evolution (see Fig 1.1). It comprises three parental species (*Bacillus grandii* – gonochoric; *Bacillus rossius* - facultative parthenogen; *Bacillus atticus* - obligate parthenogen). These three species gave origin to a variegated group of hybrids: the diploid parthenogen *Bacillus whitei (Bacillus rossius* /*Bacillus grandii grandii*), the triploid parthenogen *Bacillus lynceorum* (*Bacillus rossius* / *Bacillus grandii grandii* / *Bacillus atticus*). The Sicilian species of this taxon have allowed the discovery of two additional reproductive strategies: hybridogenesis and androgenesis. In the first instace an hybrid *Bacillus rossius* /*Bacillus grandii* female transmits to the progeny exclusively the *Bacillus rossius* haploset through the elimination of paternal chromosomes during oogenesis and spermatogenesis. The hybrid condition is restored trough fertilization by the syntopic *Bacillus grandii* male. The second one consists in the fertilization of a *Bacillus rossius* egg devoid of the maternal haploset by two sperms of a *Bacillus grandii* male. The progeny is therefore given by a nuclear *Bacillus grandii* and a mitochondrial *Bacillus rossius* genome. (Mantovani and Scali 1992; Mantovani et al. 1999; Scali et al. 2003).

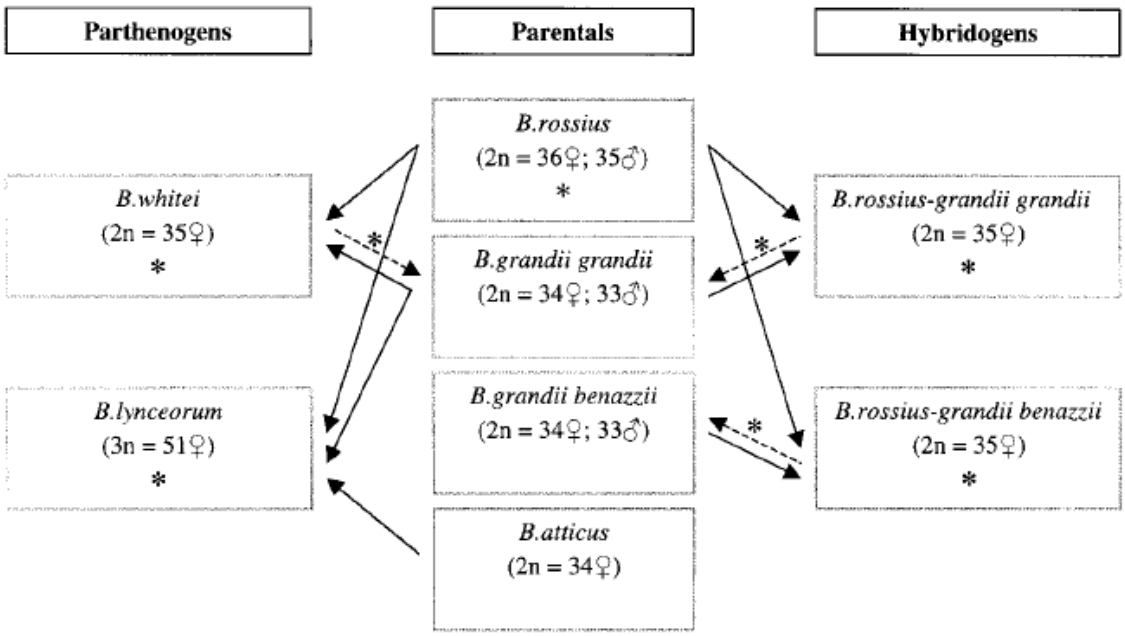


Figure 1.1 Scheme of reticulated evolution in *Bacillus* taxa. Asterisks indicate *B. rossius* mtDNA haplotype, sketched arrows indicate movement of *Bacillus rossius* mtDNA during hybridogenesis and androgenesis (from Scali et al. 2003).

1.4 Relationship between TE dynamics and speciation

The possibility that TE activity is a facilitating factor for speciation is an issue that has developed considerably in recent years. To face this question it is important to make a big assumption: the number of living species in a taxon must reflect the effective rate of speciation. So in their recent history, ancient taxa with fewer species should have experienced a reduced speciation rate with respect to the rate of extinction. Starting from this assumption, some authors have developed different models of TE-related speciation. In recent years, many authors are trying to formulate a model in this regard. The TE-Trust hypothesis suggests that TEs promote the origin of new lineages and drive lineage divergence through the engineering of specific traits (Oliver and Greene 2011). The Carrier Subpopulation Hypothesis proposes that the origin and fixation of new TE families are facilitated by genetic drift derived from the formation of small subpopulations. The higher rate of transposition of new families increases the probability of reproductive isolation due to genetic

diversification (Jurka et al. 2011). One of the most intriguing aspects of this research is the possibility that the short periods of TE-proliferation coincide with short periods of morphological change and speciation predicted by the theory of punctuated equilibria (Eldredge and Gould 1977). Up to now we know that physiological stresses can cause epigenetic reprogramming. This implicates that the usual epigenetic silencing of TEs can be disrupted increasing TEs expression. The higher mutation rate caused by a TE-burst could lead a population to fast speciation events and to the reaching of new adaptive peaks: this theory is called Epi-transposon hypothesis (Zeh et al. 2009). In this kind of investigations, one of the main problems to solve is to determine the exact timing of TE burst and species diversification (Rebollo et al. 2010). The most used approach to solve this problem is to relate a certain taxon specific TE with the age of the taxon itself and then for each family the age is refined calculating the substitution rate based on the average percentage divergence of individual copies to their consensus. (Bailey et al.2003, Pace and Fechotte 2007, Ray et al. 2008). The consequence hypothesized for the lacking of a suitable repertoire of TEs is the stasis of the taxon with its total extinction or the survival of living fossils (Oliver and Greene 2011).

1.5 Thesis aims

In the light of the above reported literature, my main interests during the PhD research period were:

1) to verify if there are differences in the dynamics of transposable elements and repated sequences in general in the unisexual taxon *Bacillus atticus* respect to the gonochoric *Bacillus grandii* and gonochoric population of *Bacillus rossius* through the analyses of genomic random libraries.

These aspects led to the publication “Random DNA libraries from three species of the stick-insects *Bacillus* genus (Insecta: Phasmida): repetitive DNA characterization and first observation of polyneopteran MITEs” and it represents Chapter 2.

2) To evaluate the possibility that Transposable Elements dynamics affects the rate of speciation I designed an evolutionary framework that allows comparisons between taxa with different rates of speciation. These aspects will be dealt in Chapter 3. I developed this part of my PhD thesis through the collaboration of the Equipe Eléments transposables, Evolution, Populations Leaded by Cristina Vieirà at the LBBE Universitè Lyon-1 France.

Chapter 2

RANDOM DNA LIBRARIES FROM THREE SPECIES OF THE GENUS *BACILLUS* (INSECTA: PHASMIDA): REPETITIVE DNA CHARACTERIZATION AND FIRST OBSERVATION OF POLYNEOPTERAN MITEs

NRC Research Press

ARTICLE

Random DNA libraries from three species of the stick insect genus *Bacillus* (Insecta: Phasmida): repetitive DNA characterization and first observation of polyneopteran MITEs

Marco Ricci, Andrea Luchetti, Livia Bonandin, and Barbara Mantovani

Abstract: The repetitive DNA content of the stick insect species *Bacillus rossius* (facultative parthenogenetic), *Bacillus grandii* (gonochoric), and *Bacillus atticus* (obligate parthenogenetic) was analyzed through the survey of random genomic libraries roughly corresponding to 0.006% of the genome. By repeat masking, 19 families of transposable elements were identified (two LTR and six non-LTR retrotransposons; 11 DNA transposons). Moreover, a de novo analysis revealed, among the three libraries, the first MITE family observed in polyneopteran genomes. On the whole, transposable element abundance represented 23.3% of the genome in *B. rossius*, 22.9% in *B. atticus*, and 18% in *B. grandii*. Tandem repeat content in the three libraries is much lower: 1.32%, 0.64%, and 1.86% in *B. rossius*, *B. grandii*, and *B. atticus*, respectively. Microsatellites are the most abundant in all species. Minisatellites were only found in *B. rossius* and *B. atticus*, and five monomers belonging to the *Bag320* satellite family were detected in *B. atticus*. Assuming the survey provides adequate representation of the relative genome, the obligate parthenogenetic species (*B. atticus*), compared with the other two species analyzed, does not show a lower transposable element content, as expected from some theoretical and empirical studies.

Key words: genomic sequence survey, miniature inverted repeats (MITEs), stick insects, tandem repeats, transposable elements.

Résumé : Les auteurs ont analysé le contenu en ADN répété chez les phasmes *Bacillus rossius* (à parthénogenèse facultative), *Bacillus grandii* (gonochorique) et *Bacillus atticus* (à parthénogenèse obligatoire) en explorant aléatoirement des banques génomiques correspondant à environ 0,006 % du génome. En masquant les séquences répétées, 19 familles d'éléments transposables ont été identifiées (deux familles de rétrotransposons à LTR et six familles sans LTR; 11 familles de transposons à ADN). De plus, une analyse de novo a permis de découvrir, au sein des trois banques, la première famille d'éléments MITE au sein des polynéoptères. Globalement, l'abondance des éléments transposables représentait 23,3 % du génome chez le *B. rossius*, 22,9 % chez le *B. atticus* et 18 % chez le *B. grandii*. Les séquences répétées en tandem étaient beaucoup moins abondantes au sein des trois banques : 1,32 %, 0,64 % et 1,86 % respectivement chez le *B. rossius*, le *B. grandii* et le *B. atticus*. Les microsatellites étaient les plus abondants chez les trois espèces. Des minisatellites n'ont été observés que chez le *B. rossius* et le *B. atticus*, tandis que cinq monomères de la famille de satellites *Bag320* ont été détectés de manière unique chez le *B. atticus*. En supposant que cet échantillonnage du génome soit suffisamment représentatif de chaque génome, il s'avère que le génome de l'espèce à parthénogenèse obligatoire (*B. atticus*) ne présente pas un contenu en éléments transposables inférieur à celui des deux autres espèces, tel que le prédisaient certaines études théoriques et empiriques. [Traduit par la Rédaction]

Mots-clés : relevé de séquences génomiques, séquences répétées inversées miniatures (MITE), phasmes, répétitions en tandem, éléments transposables.

Introduction

A significant fraction of eukaryotic genomes harbours DNA sequences repeated either in tandem (head-to-tail arranged) or interspersed (Richard et al. 2008).

Tandem repeats are made by monomeric units, whose length is comprised between two and hundreds of base pairs (bp), organized in arrays ranging from few to thousands of units. They can be categorized into the following three main classes: microsatellites, minisatellites, and satellite DNAs (satDNAs). Although the three classes cannot be discriminated on the sole basis of unit length (Charlesworth et al. 1994; Richard et al. 2008), their approximate monomer length ranges can be considered as 2–10, 11–100, and >100 bp, respectively.

Interspersed repeats are, mainly, transposable elements (TEs), i.e., sequences able to move from one genomic location to another (Makałowski et al. 2012). There are two main classes of TEs: class I elements, moving via an RNA intermediate (retrotransposons), and class II elements, moving via a DNA intermediate (transposons). Within the two TE classes, autonomous elements, which are able to encode the proteins necessary for their transposition, and nonautonomous elements, which parasitize the transposition machinery of an autonomous partner, can be further distinguished (Makałowski et al. 2012). Among the latter, short interspersed elements (SINEs) are the most diverse and represented retrotransposons; their sequence is composed of (*i*) an RNA-related head, (*ii*) an anonymous body, and (*iii*) a simple sequence repeat tail. To date, the absence of SINEs has been reported only in species of *Drosophila* (Kramerov and Vassetzky 2011). Miniature inverted repeats (MITEs) are nonautonomous DNA transposons; they usually derive from autonomous elements through deletion of the internal protein-coding sequence and are characterized

Received 25 June 2013. Accepted 29 October 2013.

Corresponding Editor: T.E. Bureau.

M. Ricci, A. Luchetti, L. Bonandin, and B. Mantovani. Dipartimento di Scienze Biologiche, Geologiche e Ambientali (BiGeA), Università di Bologna, via Selmi 3, 40126 Bologna, Italy.

Corresponding author: Marco Ricci (e-mail: marco.ricci19@unibo.it).

Genome **56**: 729–735 (2013) dx.doi.org/10.1139/gen-2013-0107

Published at www.nrcresearchpress.com/gen on 7 November 2013.

by terminal or subterminal inverted repeats. MITEs are mainly found in plants, but they are also well represented in metazoan genomes (Feschotte et al. 2002).

Far from being just "junk DNA", repetitive DNA sequences are known to be biologically relevant for the host genome. SatDNAs, for example, are involved in structural functions, being the major component of centromeres in almost all eukaryotes (Plohl et al. 2008). On the other hand, TE genomic dynamics impact the host in different ways. For example, in many instances, TE insertions have been found to modify gene structures, gene expression profiles, and to promote recombination, thus introducing genetic diversity and even adaptive changes (Kazazian 2004). Besides positive interactions, repeated DNAs may also impact negatively on the host genome: for example, contraction or expansion of mini- and microsatellite arrays are directly involved in cancer development and other human pathologies (Galindo et al. 2011; El-Murr et al. 2012), while TE insertions can be deleterious for their involvement in gene disruption, negative alteration of gene expression, and ectopic recombination (Kazazian 2004).

The very existence of repeated DNAs has been long considered paradoxical because their accumulation should not be easily tolerated by the host genome (Charlesworth et al. 1994). In principle, TE accumulation or large expansions of tandem repeat arrays are contrasted by recombination that helps in eliminating deleterious alleles, for example, a TE insertion or a too large tandem array. Organisms with high genetic diversity and bisexual reproduction, therefore, would eliminate more efficiently an overload of repeated sequences. This relationship has been well depicted in the evolutionary hypothesis known as Muller's ratchet (Brookfield and Badge 1997; Wright and Schoen 1999): nonrecombining genomes would accumulate deleterious mutations that will drive them to extinction. However, the observation of unisexual and asexual taxa persisting over evolutionary time requires that mechanisms exist that are able to avoid the deleterious mutation load: the absence of TEs (or a very low TE activity) can be one such mechanism (Arkhipova and Meselson 2000, 2005; Sullender and Crease 2001). It is worth noting, though, that exceptions occur in that unisexual taxa show the same TE load than bisexual taxa (Kraaijeveld et al. 2012). Moreover, generally speaking, less virulent or even favourable parasites (here including TEs) can be selected in an unisexual or asexual lineage as a result of the strictly vertical transmission under these reproductive circumstances: this would allow both host and parasite to survive (Fine 1975; Bull et al. 1991; Wright and Finnegan 2001).

The *Bacillus* stick insect species complex is restricted to the Mediterranean area and shows a number of different reproductive biology issues. The genus comprises three so-called parental species: *Bacillus rossius*, with bisexual and parthenogenetic populations; the strictly bisexual *Bacillus grandii*; and the obligate parthenogenetic *Bacillus atticus*. Interspecific hybridization between or among the parental species produced both unisexual taxa and hybridogenetic lineages (Scali et al. 2003).

To go through the evolution of repeated DNAs, we undertook a genomic sequence survey by randomly cloning genomic fragments of the three parental species *B. rossius*, *B. grandii*, and *B. atticus*, obtaining low coverage DNA libraries. Low coverage sequencing, albeit giving partial genomic information, may provide a quick snapshot of the genome content, especially regarding repetitive DNA. For example, to de novo isolate SINE elements, the random sequencing of a relatively small portion of the genome is a recommended strategy (Nishihara and Okada 2008). In other instances, low coverage genomic surveys (even <0.1×) yield enough data for a good picture of the repeat content (Rasmussen and Noor 2009; Leese et al. 2012).

Here, we present the first results based on repeat masking and de novo characterization of repeated DNAs in genomes of *Bacillus*.

Table 1. Transposable element (TE) families found in the three analyzed genomes.

TE class/ superfamily	TE family	*Bacillus rossius*	*Bacillus grandii*	*Bacillus atticus*
Class I				
LTR		11	5	9
	Bel	2	2	*
	Gypsy	9	3	9
non-LTR		2	2	2
	Jockey	*	1	*
	L2B	n.f.	n.f.	1
	Nimb	1	*	*
	Outcast	*	1	*
	Penelope	1	*	*
	RTE	n.f.	*	1
Class II		16	18	20
	Academ	1	*	*
	Chapaev	*	*	1
	Harbinger	2	*	*
	hAT	5	1	2
	Helitron	1	3	3
	Kolobok	*	2	*
	Mariner	5	6	3
	P	*	1	*
	PiggyBac	*	1	*
	Polinton	2	3	10
	Sola	n.f.	1	1
Total		29	25	31

Note: The number of clones showing significant homology with listed TE families is given. Asterisks indicate TE presence verified through Southern Blot analysis. n.f., not found.

Data presented here will constitute the starting point for further analysis, aiming to clarify the relationships between repetitive DNA sequences and the reproductive biology of the host species.

Materials and methods

Samples and genomic DNA isolation

A *B. rossius* female (Patti, Sicily, gonochoric population), a *B. grandii* male (Ponte Manghisi, Sicily), and a *B. atticus* female (Scoglitti, Sicily) were utilized for the analyses. Gut-deprived specimens were maintained at –80 °C until the DNA isolation was performed through a standard phenol–chloroform procedure.

DNA library construction

For each library, 2 μg of genomic DNA was partially digested with *Eco*RI restriction enzyme (Invitrogen, Carlsbad, Calif., USA) for 2 h at 37 °C. Fragments were ligated to *Eco*RI-adapters (5′-CTCGTAGACTGCGTACC-3′; 5′-AATTGGTACGCAGTCTAC-3′) and then amplified with adaptor-specific primers (5′-GACTGC GTACCAATTCN-3′). After a 1% agarose gel electrophoresis, fragments between 800–1200 bp were recovered by gel extraction and cloned into a pGem-T Easy Vector (Promega, Madison, Wis., USA) used to transform *E. coli* DH5α competent cells (Invitrogen, Carlsbad, Calif., USA). Recombinant colonies were screened by PCR amplification with T7/SP6 primers, under standard PCR conditions. In total, 196 clones per species were sequenced at Macrogen Inc. (Korea).

Southern blot analysis

For each of the 14 TE families missing in at least one of the three libraries, probes were obtained by PCR amplification using specifically designed primers (supplementary data, Table S1)[1]. The PCR program was as follow: initial denaturation at 95 °C for 2 min; 35 cycle of denaturation at 95 °C for 30 s, annealing at 48 °C for 30 s, elongation at 72 °C for 30 s; and a final elongation step at

[1]Supplementary data are available with the article through the journal Web site at http://nrcresearchpress.com/doi/suppl/10.1139/gen-2013-0107.

Table 2. Nucleotide identity of transposable element (TE) families.

TE class	TE family	Species	Avg. identity (%)	*N*	Total matches (bp)
Intraspecific comparisons					
Class I	*Bel*	*Bacillus grandii*	99	2	885
	Gypsy	*Bacillus rossius*	96	7	3340
		Bacillus grandii	97	2	902
		Bacillus atticus	97	3	3063
Class II	*Helitron*	*Bacillus grandii*	66	2	596
		Bacillus atticus	67	2	202
	Mariner	*Bacillus rossius*	99	2	821
		Bacillus grandii	92	5	3630
	Polinton	*Bacillus grandii*	98	2	735
		Bacillus atticus	85	8	3180
Interspecific comparisons					
Class I	*Gypsy*	*Bacillus rossius-grandii*	97	5	4374
		Bacillus rossius-atticus	96	4	2642
		Bacillus grandii-atticus	96	3	1812
Class II	*Helitron*	*Bacillus rossius-grandii*	69	2	326
	Mariner	*Bacillus rossius-atticus*	94	2	867
	Polinton	*Bacillus rossius-atticus*	68	4	797
		Bacillus grandii-atticus	91	6	3054

Note: *N*, number of clones compared.

72 °C for 4 min. PCR reactions were performed with the GoTaq amplification kit (Promega, Madison, Wis., USA), using 30 ng of genomic DNA. Twenty microlitres of each amplification product was separated on a 1.5% agarose gel and Southern blotted onto a positively charged nylon membrane. Hybridization was performed using the AlkPhos labelling and detection kit (GE Healthcare, Pittsburg, Pa., USA) following the manufacturer's protocol. Stringency washes allowed up to 10% of nucleotidic divergence between probes and target DNA.

Sequence analysis

TE identification was done by repeat masking on Repbase Update database with CENSOR web tool (Kohany et al. 2006). We took into account only the hits with nucleotidic score >500 or amino acidic score >300, or those presenting simultaneously amino acidic score >200 and positives >0.5. For family identification, we considered accurate only the alignments with amino acidic score >200 and positives >0.5. Sequence identity for each TE family has been calculated as Σ IA × (LA / LSA), where IA is the identity of the alignment (max identity = 1), LA is the length of the alignment (bp), and LSA is the length of the sum of the alignments (bp).

A de novo search of interspersed repeats was done by self-comparison of each library; an e-value $\leq 10^{-5}$ was set to define significant high scoring segment pairs. Copy number of de novo identified interspersed repeats was calculated following the formula: (No. of occurrences in the library × genome size) / library size.

Differences in relative abundance of scored TEs in the three libraries were tested with repeated measure analysis of variance (ANOVA), followed by post-hoc paired *t* test with Holm correction.

Neighbor-joining tree, using uncorrected *p*-distance, was calculated with MEGA v.5 (Tamura et al. 2011); nodal support was obtained after 500 bootstrap replicates.

Tandem repeat searches were performed by Phobos v.3.3.11 (Mayer 2010), allowing extend exact search, repeat unit size from 2 to 500 bp, and at least four consecutive units.

Sequences were deposited in GenBank under accession numbers KF256266–KF256815.

Results and discussion

For each species, 196 random genomic fragments ~1000 bp long were obtained; on the whole, 144 250, 144 896, and 173 946 bp have been sequenced for *B. rossius*, *B. grandii*, and *B. atticus*, respectively. The genome sizes of these three stick insect species are 2.12–1.90, 2.55–2.11, and 2.25 Gbp, respectively; therefore, the sequencing corresponds to less than 0.006%–0.007% genomic coverage. The average GC content calculated within sequenced libraries is 40.3%, ranging from 39.3% in *B. atticus* to 42.2% in *B. rossius*.

Transposable elements

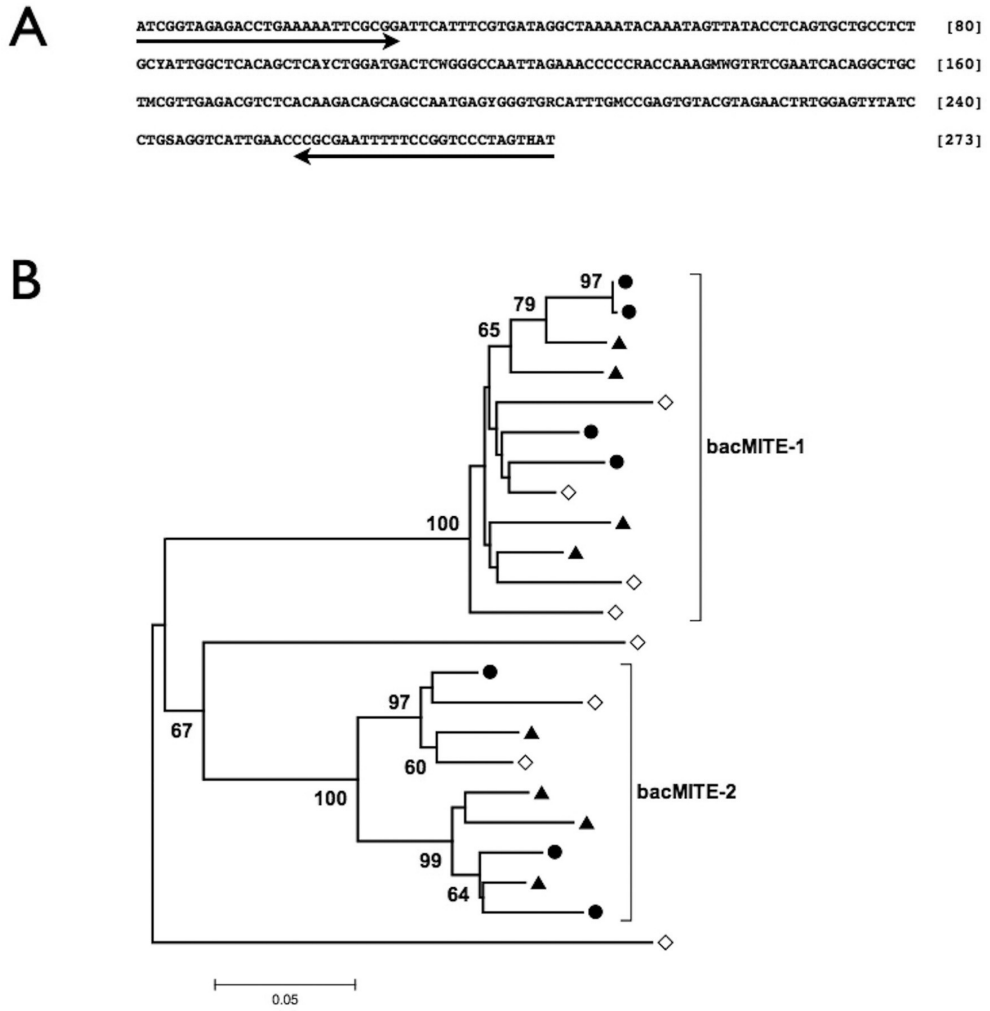
Repeat masking on the three libraries revealed 85 clones containing either class I or class II TEs.

Class I mobile elements belong to the following eight families: *Bel*, *Gypsy*, *Jokey*, *L2B*, *Nimb*, *Outcast*, *Penelope*, and *RTE*. *Bel* and *Gypsy* are LTR retroelements, while the other six are non-LTR retrotransposons (Table 1). Eleven families of class II elements have been identified as follows: *Academ*, *Chapaev*, *Harbinger*, *hAT*, *Helitron*, *Kolobok*, *Mariner*, *P*, *piggyBac*, *Polinton*, and *Sola* (Table 1).

All families have been identified in the three libraries either by clone sequencing or through Southern blot analysis, with the exceptions of *L2B*, *RTE*, and *Sola* in *B. rossius* and *L2B* in *B. grandii* (Table 1). On the whole, the parthenogenetic *B. atticus* shows the presence of all families and shares the majority of them with *B. grandii*, as it could be expected on the basis of phylogenetic relationships (Mantovani et al. 2001). The absence of some families in the gonochoric genomes of *B. rossius* and *B. grandii* may witness the greater ability of bisexuals in dealing with TEs; moreover, considering the time elapsed since the species splitting (23–17 Myr ago; Mantovani et al. 2001), these families could have had enough time for accumulating nucleotide divergence over 10%, and thus becoming undetectable under the Southern blot conditions used in this analysis.

For each library, TE families detected in at least two clones with overlapping regions were analyzed to evaluate the level of similarity (Table 2). These comparisons involved the LTR elements *BEL* and *Gypsy*, and the DNA families *Helitron*, *Mariner*, and *Polinton*. Identity values indicate a substantial intraspecific conservation of sequences, possibly being copies of the same element. The only exception is given by the *Helitron* elements in *B. grandii* and *B. atticus*, with identity values falling to 66% and 67%, respectively. Interspecific comparisons involve the LTR *Gypsy* and the class II elements *Helitron*, *Mariner*, and *Polinton* (Table 2). *Gypsy* and *Mariner* appear well conserved across species, as well as *Polinton* between *B. grandii* and *B. atticus*. On the other hand, a low degree of identity is found for *Helitron* in the *B. rossius* versus *B. grandii* comparison and *Polinton* in the *B. rossius* versus *B. atticus* comparison. On the whole, *Helitron* appears the less conserved element both within

Fig. 1. (A) The 50% majority rule consensus sequences of bacMITE elements. Arrows mark the terminal inverted repeats (TIRs). (B) Neighbor-joining tree based on uncorrected *p*-distance between bacMITE sequences; bootstrap values are calculated after 500 replicates. The two clusters corresponding to the two subfamilies have been indicated. Empty diamonds, *Bacillus rossius* clones; filled circles, *Bacillus grandii* clones; and filled triangles, *Bacillus atticus* clones.



and between genomes: further investigations will clarify if this could be due to higher element diversity.

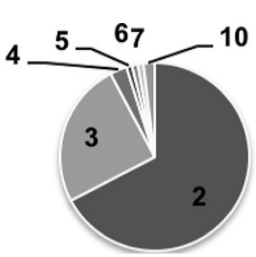
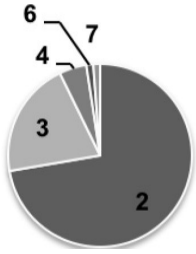
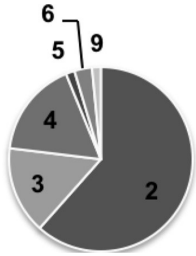
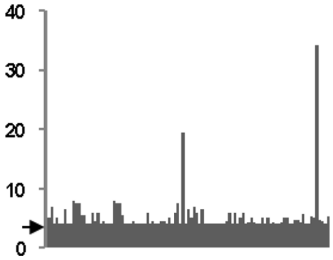
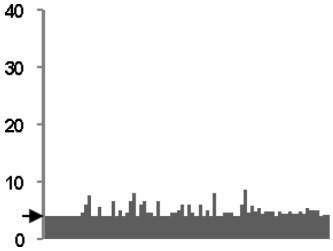
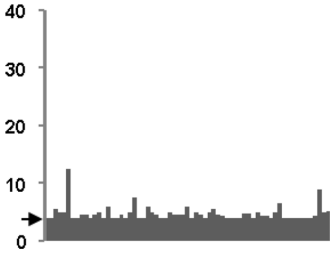
De novo identification of interspersed repeats led to the characterization of 23 homologous nucleotidic stretches 124–277 bp long, distributed in 21 clones from the three genomes, and sharing 77.3% sequence similarity. Aligned sequences gave a consensus length of 273 bp with terminal inverted repeats (TIRs; Fig. 1A). The short length and TIRs are common features of MITEs, small nonautonomous DNA transposon (Feschotte et al. 2002): we, therefore, named the retrieved sequences as bacMITEs. In the phylogenetic analysis, sequences are distributed in two main clusters having 100% nodal support; only two sequences fall outside the two clusters and may represent highly diverging or recombinant elements (Fig. 1B). Therefore, two possible MITE subfamilies can be identified (bacMITE-1 and bacMITE-2), showing within-subfamily sequence identity ranging from 81.5% to 91.2%. The relative copy numbers are 5.6×10^4, 4.8×10^4, 5.6×10^4, for bacMITE-1, and 5.6×10^4, 6.4×10^4, 5.6×10^4, for bacMITE–2, in *B. rossius*, *B. grandii*, and *B. atticus*, respectively. No flanking target site duplications (TSD) have been identified, but this could be due to the low number of identified full-length elements (three bacMITE-1 and one bacMITE-2). To our knowledge, bacMITEs are the first MITE elements ever found within a polyneopteran genome.

Table 3. Occurrences of transposable element (TE), relative percentages of TE classes among positive clones, and class II/class I ratio.

	Bacillus rossius	*Bacillus grandii*	*Bacillus atticus*
Total TEs/library	23.3%	18.0%	22.9%
Class I			
LTR	27.5%	11.8%	20.5%
non-LTR	10.0%	14.7%	6.8%
Class II	62.5%	73.5%	72.7%
Class II/class I	1.667	2.778	2.667

It is worth noting that no SINEs have been identified in this survey. Of course, this could be due to the limited sequencing

Table 4. Tandem repeat abundance (bp) detected in the three libraries.

	Bacillus rossius	*Bacillus grandii*	*Bacillus atticus*
Microsatellite	1.00% (103)	0.64% (83)	0.54% (65)
Unit length distribution (bp)			
Array length distribution (bp)			
Minisatellite	0.32% (2)	0.00% (0)	0.36% (3)
Unit length/copy number	15/11 45/7 —	n.f. — —	16/8 19/6 30/10
Satellite	0.00% (0)	0.00% (0)	0.96% (2)
Total	1.32% (105)	0.64% (23)	1.86% (70)

Note: The number of loci is given in parentheses. For microsatellite loci, array distributions are also shown (*x* axis, unit length; *y* axis, copy number; arrow indicates the four copy threshold used as the minimum tandem array length during Phobos v.3.3.11 search; each bin represents a single locus). n.f., not found.

relative to the genome size of *Bacillus* spp., but it should be noted that, in other instances, low coverage sequencing gave at least some SINE sequences. For example, in the termite *Reticulitermes lucifugus*, 25 SINE sequences belonging to four distinct SINE families have been found by random sequencing ~130 000 bp (approximately 0.012% of the termite genome; Luchetti and Mantovani 2011). In the hyrax genome, the random sequencing of 63 000 bp (~0.002% of the whole genome) revealed 26 AfroSINE elements (Nishihara and Okada, 2008). It is, thus, possible to hypothesize that the genome of *Bacillus* lacks SINEs, as observed in species of *Drosophila* (Kramerov and Vassetzky 2011), or that they are very poorly represented. Further sequencing or focused experiments on SINE search will allow more definitive conclusions.

As a general picture, LTR, non-LTR, and DNA transposons have quite different representativeness ($p = 0.002$), though this abundance distribution is not significantly different among the surveyed libraries ($p = 0.521$).

DNA transposons are generally prevailing in all libraries (class II versus LTR, $p = 0.00060$; class II versus non-LTR, $p = 0.00032$), their amount ranging from 1.666- to 2.778-fold higher than that of retrotransposons. This variability is not unexpected, taking into account that the class II/class I ratio may vary from ~100% class II to ~100% class I elements. For example, in two closely related insect species, the Culicidae mosquitoes *Anopheles gambiae* and *Aedes aegypti*, this ratio ranges from ~0.7 to ~2.3, respectively (reviewed in Feschotte and Pritham 2007). Within class I elements, LTR families outnumber non-LTR families (Table 3), but this difference is not significant ($p = 0.17433$).

Tandem repeats

On the whole, less than 2% of the sequenced genomic fragments in each species of *Bacillus* is constituted by tandem repeats.

As expected, microsatellite loci (2–10 bp) are more represented than minisatellite or satellite DNA. Di- and trinucleotide loci are the most abundant, with few instances of longer units (Table 4). However, most of the retrieved loci retains the minimum length imposed during the repeat search or remains below the 10 copies; only two loci in *B. rossius* and one locus in *B. atticus* showed longer arrays. In the former species, a dinucleotide array showed 19 repeats and a pentanucleotide locus has up to 34 repeat units. In the latter taxon, a dinucleotide array is made by 12 repeats (Table 4).

Minisatellites (11–100 bp) occurred only in *B. rossius* and *B. atticus* libraries, with repeat unit length ranging from 16 to 45 bp and copy number comprised between 6 and 10 (Table 4). Interestingly, a 45-mer array found in *B. rossius* has significant homologies with clones from both *B. grandii* and *B. atticus* libraries. Sequence analysis indicates that one clone from *B. grandii* (gra_af4) and five from *B. atticus* (att_ab7, att_ah5, att_ah7, att_ah11, and att_bc8) have from two to three tandemly arranged 45-mer repeat units, with an overall repeat units sequence identity of 62%. Therefore, the 5′ and the 3′ array flanking regions were compared to check if the same genomic locus harbours this minisatellite in all the three genomes: while the 3′ end flanking region did not show any homology among clones, the 5′ flanking region was significantly conserved even between species (74.1% pairwise identity; 84.4% of identical sites), with the exception of the clone att_ah11. The analysis of the consensus sequence generated from the alignment of 5′ flanking regions did not give any similarity with any known sequence in public databases. Tandem repeat arrays flanked by

homologous sequences are, actually, commonly found: in fact, they can originate within the TE sequence (Mogil et al. 2012; Sharma et al. 2013) or, in most cases, they are generated at retrotransposon tails when they reintegrate into new genomic locations (Lopez-Giraldez et al. 2006; Meglécz et al. 2007; Coates et al. 2009, 2011; Luchetti and Mantovani 2009, 2011). However, retrotransposons usually generate microsatellite loci, while in this case a 45-bp unit made the tandem array. At present, it is not possible to further explain such occurrence, especially because the homologous flanking sequence is not similar to any known TE sequence. However, it would be interesting to check whether this minisatellite locus has been generated upon the insertion of a retrotransposon or by some recombinative mechanism during the evolution of the genome in species of the genus *Bacillus*.

Satellite DNA sequences (>100 bp) have been retrieved only in the *B. atticus* library, where five *Bag320* monomers (Mantovani et al. 1997; Cesari et al. 2003; Luchetti et al. 2003) have been found in two clones (att_ac9 and att_ba1). With respect to previously isolated *Bag320* sequences, they show a sequence similarity with *B. atticus* specific monomers ranging from 94.1% to 96.0%. No *Bag320* sequences have been found in the genomes of *B. rossius* or *B. grandii*. For the former species, this was quite expected, as it is known that this satellite DNA occurs at a very low copy number in this genome and was only isolated by PCR amplification (Cesari et al. 2003). Its absence in the *B. grandii* library is, however, more surprising because this genome contains the highest copy number of the satellite family (15%–20% in *B. grandii* versus 2%–5% in *B. atticus*; Mantovani et al. 1997). As a general consideration, the use of *Eco*RI restriction enzyme for library production may have biased the genomic sampling of *Bag320* sequences, as they do not contain its cutting site in the considered species; therefore, its sampling from *B. atticus* genomes could be considered as mere chance.

On the whole, the sequenced libraries represent a small fraction of the whole genome of *B. rossius*, *B. grandii*, and *B. atticus*, with less than 1% genomic coverage. Yet, the survey allowed the retrieval of a number of repetitive DNAs, either interspersed or not. Most of the main TE families are represented, and a MITE family, the first ever discovered in polyneopteran insects, has been de novo characterized. Moreover, mini- and microsatellite loci were found even if characterized by short arrays. Unfortunately, it is not possible to make strong comparisons on the representativeness of repetitive DNAs within the sequenced libraries, as no polyneopteran genomes have been sequenced so far and the only polyneopteran DNA library available has been built from the termite species *R. lucifugus*. In this library (covering 0.012% of the genome), four SINE and one putative MITE families were de novo characterized (Luchetti and Mantovani 2011; A. Luchetti, unpublished data); moreover, 11 minisatellite and 298 microsatellite loci were isolated in a single termite species. Therefore, despite the smaller genome (~1 versus >2 Gbp), termites appear to have a more repetitive genomic landscape than stick insects. As a final remark, it is interesting to point out the relative TE content showed by unisexual (*B. atticus*) as opposed to gonochoric taxa (*B. rossius* and *B. grandii*). Both theoretical and empirical studies evidenced that parthenogenetic, i.e., low recombining, genomes should avoid TE accumulation to escape the effects of Muller's ratchet; otherwise, TE load would raise without the possibility of elimination eventually leading to the host species extinction (Arkhipova and Meselson 2000, 2005; Sullender and Crease 2001; Dolgin and Charlesworth 2006). On the other hand, in the parasitoid wasp *Leptopilina clavipes*, unisexual and bisexual lineages showed no difference in the overall TE genomic coverage (Kraaijeveld et al 2012). In line with this, *B. atticus* genome appears to have (at least) the same TE content of the bisexual species. As argued by Wright and Finnegan (2001), TE prevalence studies in obligate unisexual genomes are difficult to interpret mainly because they are derived from bisexual ancestors. Thus, at the moment, it is impossible to state whether *B. atticus* and *L. clavipes* TEs have been selected for less harmful elements or they are still in the "elimination phase", i.e., the two taxa are not unisexual since enough time to allow an efficient clearance of TE load. On the whole, although based on small datasets and considering the different reproductive strategies, the variation of ploidy, and the presence of hybrid taxa that characterize the *Bacillus* complex, the present survey provides interesting preliminary data to undertake further analysis in species of the genus *Bacillus*.

Acknowledgements

This work was funded by Canziani and RFO funds to B.M. M.R. produced the three libraries and carried out repeat masking; A.L. performed de novo TE searches and tandem repeat characterization; L.B. performed Southern blot analysis; B.M. supervised the project; and M.R., A.L., and B.M. wrote the article. Authors wish to thank the two reviewers whose criticisms and suggestions helped to substantially improve the paper.

References

Arkhipova, I., and Meselson, M. 2000. Transposable elements in sexual and ancient asexual taxa. Proc. Natl. Acad. Sci. U.S.A. **97**(26): 14473–14477. doi:10.1073/pnas.97.26.14473. PMID:11121049.

Arkhipova, I., and Meselson, M. 2005. Deleterious transposable elements and the extinction of asexuals. Bioessays, **27**(1): 76–85. doi:10.1002/bies.20159. PMID:15612027.

Brookfield, J.F.Y., and Badge, R.M. 1997. Population genetics models of transposable elements. Genetica, **100**(1–3): 281–294. doi:10.1023/A:1018310418744. PMID:9440281.

Bull, J.J., Molineux, I.J., and Rice, W.R. 1991. Selection of benevolence in a host-parasite system. Evolution, **45**(4): 875–882. doi:10.2307/2409695.

Cesari, M., Luchetti, A., Passamonti, M., Scali, V., and Mantovani, B. 2003. Polymerase chain reaction amplification of the *Bag320* satellite family reveals the ancestral library and past gene conversion events in *Bacillus rossius* (Insecta Phasmatodea). Gene, **312**: 289–295. doi:10.1016/S0378-1119(03)00625-5. PMID:12909366.

Charlesworth, B., Sniegowski, P., and Stephan, W. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature, **371**(6494): 215–220. doi:10.1038/371215a0. PMID:8078581.

Coates, B.S., Sumerford, D.V., Hellmich, R.L., and Lewis, L.C. 2009. Repetitive genome elements in a European corn borer, *Ostrinia nubilalis*, bacterial artificial chromosome library were indicated by bacterial artificial chromosome end sequencing and development of sequence tag site markers: implications for lepidopteran genomic research. Genome, **52**(1): 57–67. doi:10.1139/G08-104. PMID:19132072.

Coates, B.S., Kroemer, J.A., Sumerford, D.V., and Hellmich, R.L. 2011. A novel class of miniature inverted repeat transposable elements (MITEs) that contain hitchhiking $(GTCY)_n$ microsatellites. Insect. Mol. Biol. **20**(1): 15–27. doi:10.1111/j.1365-2583.2010.01046.x. PMID:20977507.

Dolgin, E.S., and Charlesworth, B. 2006. The fate of transposable elements in asexual populations. Genetics, **174**(2): 817–827. doi:10.1534/genetics.106.060434. PMID:16888330.

El-Murr, N., Abidi, Z., Wanherdrick, K., Svrcek, M., Gaub, M.P., Fléjou, J.F., et al. 2012. *MiRNA* genes constitute new targets for microsatellite instability in colorectal cancer. PLoS ONE, **7**(2): e31862. doi:10.1371/journal.pone.0031862. PMID:22348132.

Feschotte, C., and Pritham, E. 2007. DNA transposons and the evolution of eukaryotic genomes. Annu. Rev. Genet. **41**: 331–368. doi:10.1146/annurev.genet.40.110405.090448. PMID:18076328.

Feschotte, C., Zhang, X., and Wessler, S.R. 2002. Miniature inverted-repeat transposable elements (MITEs) and their relationship with established DNA transposons. *In* Mobile DNA II. *Edited by* N. Craig, R. Craigie, M. Gellert, and A. Lambowitz. A.S.M. Press, Washington, D.C. pp. 1147–1158.

Fine, P.E.M. 1975. Vectors and vertical transmission: an epidemiologic perspective. Ann. N.Y. Acad. Sci. **266**: 173–194. doi:10.1111/j.1749-6632.1975.tb35099.x. PMID:829470.

Galindo, C.L., McIver, L.J., Tae, H., McCormick, J.F., Skinner, M.A., Hoeschele, I., et al. 2011. Sporadic breast cancer patients' germline DNA exhibit an AT rich microsatellite signature genes chromosomes. Cancer, **50**(4): 275–283. PMID:21319262.

Kazazian, H.H. 2004. Mobile elements: drivers of genome evolution. Science, **303**(5664): 1626–1632. doi:10.1126/science.1089670. PMID:15016989.

Kohany, O., Gentles, A.J., Hankus, L., and Jurka, J. 2006. Annotation, submission and screening of repetitive elements in Repbase: Repbase Submitter and Censor. BMC Bioinformatics, **25**(7): 474. PMID:17064419.

Kraaijeveld, K., Zwanenburg, B., Hubert, B., Vieira, C., De Pater, S., Van Alphen, J.J.M., et al. 2012. Transposon proliferation in an asexual parasi-

toid. Mol. Ecol. **21**(16): 3898–3906. doi:10.1111/j.1365-294X.2012.5582.x. PMID: 22548357.

Kramerov, D.A., and Vassetzky, N.S. 2011. Origin and evolution of SINEs in eukaryotic genomes. Heredity, **107**(6): 487–495. doi:10.1038/hdy.2011.43. PMID: 21673742.

Leese, F., Brand, P., Rozenberg, A., Mayer, C., Agrawal, S., Dambach, J., et al. 2012. Exploring Pandora's box: potential and pitfalls of low coverage genome surveys for evolutionary biology. PLoS ONE, **7**(11): e49202. doi:10.1371/journal.pone.0049202. PMID:23185309.

López-Giráldez, F., Andrés, O., Domingo-Roura, X., and Bosch, M. 2006. Analyses of carnivore microsatellites and their intimate association with tRNA-derived SINEs. BMC Genomics, **7**: 269. doi:10.1186/1471-2164-7-269.

Luchetti, A., and Mantovani, B. 2009. *Talua* SINE biology in the genome of the *Reticulitermes* subterranean termites (Isoptera, Rhinotermitidae). J. Mol. Evol. **69**(6): 589–600. doi:10.1007/s00239-009-9285-7. PMID:19904483.

Luchetti, A., and Mantovani, B. 2011. Molecular characterization, genomic distribution and evolutionary dynamics of Short INterspersed Elements in the termite genome. Mol. Genet. Genomics, **285**(2): 175–184. doi:10.1007/s00438-010-0595-7.

Luchetti, A., Cesari, M., Carrara, G., Cavicchi, S., Passamonti, M., Scali, V., and Mantovani, B. 2003. Unisexuality and molecular drive: *Bag320* sequence diversity in *Bacillus* taxa (Insecta Phasmatodea). J. Mol. Evol. **56**(5): 587–596. doi:10.1007/s00239-002-2427-9. PMID:12698295.

Makałowski, W., Pande, A., Gotea, V., and Makałowska, I. 2012. Transposable elements and their identification. Methods Mol. Biol. **855**: 337–359. doi:10.1007/978-1-61779-582-4_12. PMID:22407715.

Mantovani, B., Tinti, F., Bachmann, L., and Scali, V. 1997. The *Bag320* satellite DNA family in *Bacillus* stick insects (Phasmatodea): different rates of molecular evolution of highly repetitive DNA in bisexual and parthenogenetictaxa.Mol.Biol.Evol.**14**(12):1197–1205.doi:10.1093/oxfordjournals.molbev.a025729. PMID:9402731.

Mantovani, B., Passamonti, M., and Scali, V. 2001. The mitochondrial cytochrome oxidase II gene in *Bacillus* stick insects: ancestry of hybrids, androgenesis, and phylogenetic relationships. Mol. Phylogenet. Evol. **19**(1): 157–163. doi:10.1006/mpev.2000.0850. PMID:11286500.

Mayer, C. 2010. Phobos 3.3.11. Available from http://worldwidewords.rub.de/spezoo/cm/cm_phobos.htm [accessed 22 April 2013].

Meglécz, E., Anderson, S.J., Bourguet, D., Butcher, R., Caldas, A., Cassel-Lundhagen, A., et al. 2007. Microsatellite flanking region similarities among different loci within insect species. Insect. Mol. Biol. **16**(2): 175–185. doi:10.1111/j.1365-2583.2006.00713.x. PMID:17298557.

Mogil, L.S., Slowikowski, K., and Laten, H.M. 2012. Computational and experimental analyses of retrotransposon-associated minisatellite DNAs in the soybean genome. BMC Bioinformatics, **13**(Suppl. 2): S13. doi:10.1186/1471-2105-13-S2-S13. PMID:22536864.

Nishihara, H., and Okada, N. 2008. Retroposons: genetic footprints on the evolutionary paths of life. Methods Mol. Biol. **422**: 201–225. doi:10.1007/978-1-59745-581-7_13. PMID:18629669.

Plohl, M., Luchetti, A., Meštrović, N., and Mantovani, B. 2008. Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. Gene, **409**(1–2): 72–82. doi:10.1016/j.gene.2007.11.013. PMID:18182173.

Rasmussen, D.A., and Noor, M.A.F. 2009. What can you do with 0.1× genome coverage? A case study based on a genome survey of the scuttle fly *Megaselia scalaris* (Phoridae). BMC Genomics, **10**: 382. doi:10.1186/1471-2164-10-382.

Richard, G.-F., Kerrest, A., and Dujon, B. 2008. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol. Mol. Biol. Rev. **72**(4): 686–727. doi:10.1128/MMBR.00011-08. PMID:19052325.

Scali, V., Passamonti, M., Marescalchi, O., and Mantovani, B. 2003. Linkage between sexual and asexual lineages: genome evolution in *Bacillus* stick insects. Biol. J. Linn. Soc. **79**: 137–150. doi:10.1046/j.1095-8312.2003.00172.x.

Sharma, A., Wolfgruber, T.K., and Presting, G.G. 2013. Tandem repeats derived from centromeric retrotransposons. BMC Genomics, **14**: 142. doi:10.1186/1471-2164-14-142. PMID:23452340.

Sullender, B.W., and Crease, T.J. 2001. The behavior of a *Daphnia pulex* transposable element in cyclically and obligately parthenogenetic populations. J. Mol. Evol. **53**(1): 63–69. PMID:11683324.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. 2011. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance and maximum parsimony methods. Mol. Biol. Evol. **28**: 2731–2739. doi:10.1093/molbev/msr121. PMID:21546353.

Wright, S., and Finnegan, D. 2001. Genome evolution: sex and the transposable element. Curr. Biol. **11**: R296–R299. doi:10.1016/S0960-9822(01)00168-3. PMID: 11369217.

Wright, S.I., and Schoen, D.J. 1999. Transposon dynamics and the breeding system. Genetica, **107**(1–3): 139–148. doi:10.1023/A:1003953126700. PMID: 10952207.

Chapter 3

TRANSPOSABLE ELMENTS ANALISYS IN TAXA WITH DIFFERENT SPECIATION RATE

3.1 Materials and methods

3.1.1 Taxa speciation rate evaluation

I assume that the species richness currently observable in a taxon should represent the effective rate of speciation. Using the phylogenetic tree of mammals produced by Meredith et al. (2011; Fig 3.1**)** all the families containing the species analyzed by Jerzy Jurka (Jurka et al. 2011) was taken in account. Then, the families with different rate of speciation belonging to the same order were chosen; in this way the distance between the species is reduced to a minimum and the data are collected for genomes with a similar history. If family one has a lower number of species and an older age with respect to family two, the rate of speciation of taxon one is lower than taxon two (Fig 3.2). Data on the number of known living species for each taxon were collected from Catalogue of Life (http://www.catalogueoflife.org/col/). And Taxonomy Browser (http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/).

3.1.2 Dataset construction

I exploited the data produced by Jurka et al. (2011). and is analyzed the content of transposable element of 31 mammal species for which whole genomes are sequenced. Species are listed in table 3.1. For each species Jurka showed both the total number of families and the total number of Transposable Elements with a divergence from their consensus sequences lower then 1%. Then He repeated the analysis with Transposable Elements with a divergence from their consensus sequences lower then 5%. From now on I will indicate the two groups of Trasposable elements as 1% Dataset and 5% Dataset. Analyzing the Mammalia phylogenetic tree of Meredith et al. 2011, using the evolutionary framework described in paragraph 3.1.1, I made 19 pairs of species using the dataset of Table 3.1. Each pair is composed of species that belong to different families of the same order. Species are listed indicating first the one with the lower rate of speciation followed by the species belonging to the family with the higher rate of speciation. From this dataset, 19 species pairs were considered for the intraorder comparisons; these are with the species with lower rate of speciation listed as first:

Order Insectivora: *Erinaceus europaeus-Sorex araneus*; Order Carnivora: *Canis lupus-Felis catus;* Order Cetartyodactyla: *Lama pacos-Bos taurus;* Order Rodentia: *Dipodomys ordii-Mus musculus; Spermophilus tridecemlineatus-Mus musculus; Dipodomys ordii-Rattus norvegicus; Spermophilus tridecemlineatus-Rattus norvegicus;* Order Primates*: Tarsius syrichta-Macaca mulatta; Tarsius syrichta-Homo sapiens; Tarsius syrichta-Pan troglodites; Tarsius syrichta-callithrix jacchus; Tarsius syrichta-Microcebus murinus; Tarsius syrichta-Otolemur garnettii; Microcebus murinus-*

Macaca mulatta; Microcebus murinus-Callithrix jacchus; Otolemur garnettii-Macaca mulatta; Otolemur garnettii-Callithrix jacchus.

Observing the phylogeny of the entire class Mammalia (Figure 3.1), the superorders Laurasiatheria and Euarchontoglires have been formed more recently and have a higher number of species than Xenartra and Afrotheria. I therefore performed a large scale analysis by splitting the 28 species of Placentalia available from the dataset analyzed by Jerzy Jurka in their four superorders: Laurasiantheria-*Bos taurus, Canis lupus familiaris, Equus caballus, Erinaceus europaeus, Felis catus, Lama pacos, Myotis lucifugus, Pteropus vampyrus, Sorex araneus;* Euarchontoglires-*Callithrix jacchus, Cavia porcellus, Dipodomys ordii, Homo sapiens, Macaca mulatta, Mus musculus, Otolemur garnettii, Ochotona princeps, Orictolagus cuniculus, Microcebus murius, Pan troglodites, Rattus norvegicus, Spermophilus tridecemlineatus, Tarsius syrichta;*

Xenarthra-*Choloepus ofmanni, Dasypus novemcinctus;* Afrotheria-*Echinops telfair, Loxodonta africana, Procavia capensis;* The comparisons are (with the superorder with lower rate of speciation listed as first):

Xenarthra-Laurasiatheria, Xenarthra-Euarchontoglires, Afrotheria-Laurasiateria, Afrotheria-Euarchontoglires. Among the 31 species of mammals availables from the dataset produced by Jurka et al. (2011) I selected seven species of which I have downloaded the genome and did the analysis of the total content of TE through Repeat Masker. This will allow us to repeat the comparison taking into account also the oldest insertions. The species selected are: the pair *Lama pacos-Bos taurus* for the superorder Laurasiatheria and the pair *Dipodomys ordii-Rattus norvegicus* for the superorder Euarchontoglires. I decided to repeat this analysis comparing three species belonging to

different superorders, in order to observe the dynamics among more distant genomes. The taxa are listed by increasing rate of speciation: *Ornitoryncus anatinus* (order Monotremata), *Dasypus novemcinctus* (order Cingulata), *Myotis lucifugus* (order Chiroptera). The analysis of total TE fraction was extended outside mammal class. Four species of the class Aves of which: two with lower rate of speciation-*Gallus gallus* and *Meleagris gallopavo* (order Galliformes) that list 219 species, two with a higher rate of speciation-*Geospiza fortis* and *Taeniopygia guttata* (order Passeriformes) that lists 5828 species, Two species of Osteichthyes among them the well known living fossil *Latimeria chalumnae* (Sarcopterygii) that list eight species and *Danio rerio* (Actinopterygii) that lists 31182. Sarcopterygii have a lower rate of speciation with respect to Actinopterygii. The evolutionary distance between the two lineages overcome the 420 Mya.

The analysis was finally extended to the class Insecta with the selection of a pair of species of the genus *Drosophila*: *Drosophila virilis* and the Hawaiian *Drosophila grimshawii*. I classified Hawaiian species as a group with a higher rate of speciation respect the north American group to which *Drosophila virilis* belongs considering that it had a radiation of about 1000 species becoming from a common ancestor that colonized the islands about 25 Mya ago (O'Grady et al. 2011).

3.1.3 Proxy evaluation for generation time

To verify the possibility that species with short generation time and consequent high population size are positive correlated with the rate of speciation, I suggest an indicative value . The parameter is considered as proxy for generation time. It is calculated suming up the months of gestation and the months needed to reach the sexual maturity from data available in The Animal Diversity Web database (http://www.animaldiversity.ummz.umich.edu/accounts/). This parameter was applied in the 19 intra-order comparisons of mammals listed in paragraph 3.1.2.

3.1.4 TE insertion density evaluation

Exploiting data produced by Jurka on the 31 mammal genomes (Jurka et al. 2011) the following formulawas applied: DI = NI/GS where: DI = TE insertion Density, NI = Number of insertions in the genome, GS = Genome Size in gigabases. This calculation was performed for the 1% and the 5% dataset. The insertions density values in one Gigabase will be defined with: ins/Gb.

3.1.5 Average density of TE insertions

The average density of TE insertions for 1% and 5% dataset from their consensus sequence was analyzed in Laurasiatheria, Euarchontoglires, Xenartra and Afrotheria comparing them in the light of their speciation rate with in the following four comparisons: Laurasiatheria-Xenartra, Laurasiatheria-Afrotheria, Euarchontoglires-Xenartra, Euarchontoglires-Afrotheria. Statistical validation was carried out calculating the 95% confidence interval.

3.1.6 Transposable Elements percentage evaluation

Computational searches were carried out from September 2012 through February 2013 against 16 animal genomes on data deposited at NCBI (http://www.ncbi.nlm.nih.gov/genome/) at the LBBE Universitè Lyon-1 in collaboration with the equipe Transposable Element Evolution Population leaded by Cristina Vieirà.

The GenBank accession numbers are: *Rattus norvegicus:* AABR06000000, *Dipodomis ordii:* ABRO01000000, *Ictidomys tridecemlineatus:AGTP01000000, Bos taurus:*AAFC03000000, *Vicugna pacos:*ABRR01000000, *Myotis lucifugus:*AAPE02000000, *Dasipus novemcinctus:*AAGV03000000, *Ornithorhynchus anatinus:*AAPN01000000, *Geospiza fortis:*AKZB01000000, *Taeniopygia guttata:*ABQF01000000, *Meleagris gallopavo:*ADDD01000000, *Gallus gallus:* AADN03000000, *Latimeria chalumnae:* AFYH01000000, *Danio rerio:*CABZ01000000, *Drosophila virilis:*AANI01000000, *Drosophila grimshawii:*AAPT01000000.

Whole genome were analyzed through Repeat masker (Smit et al. 1998) using the library of the subphylum Vertebrata for mammals, birds and fishes, and of the phylum Arthropoda for *Drosophila virilis* and *Drosophila grimshawii.* The percentage of total Transposable elements in bps was evaluated summing up the percentages of single superfamilies: autonomous non-LTR retrotransposons, SINEs, LTR retrotransposon and DNA transposable elements.

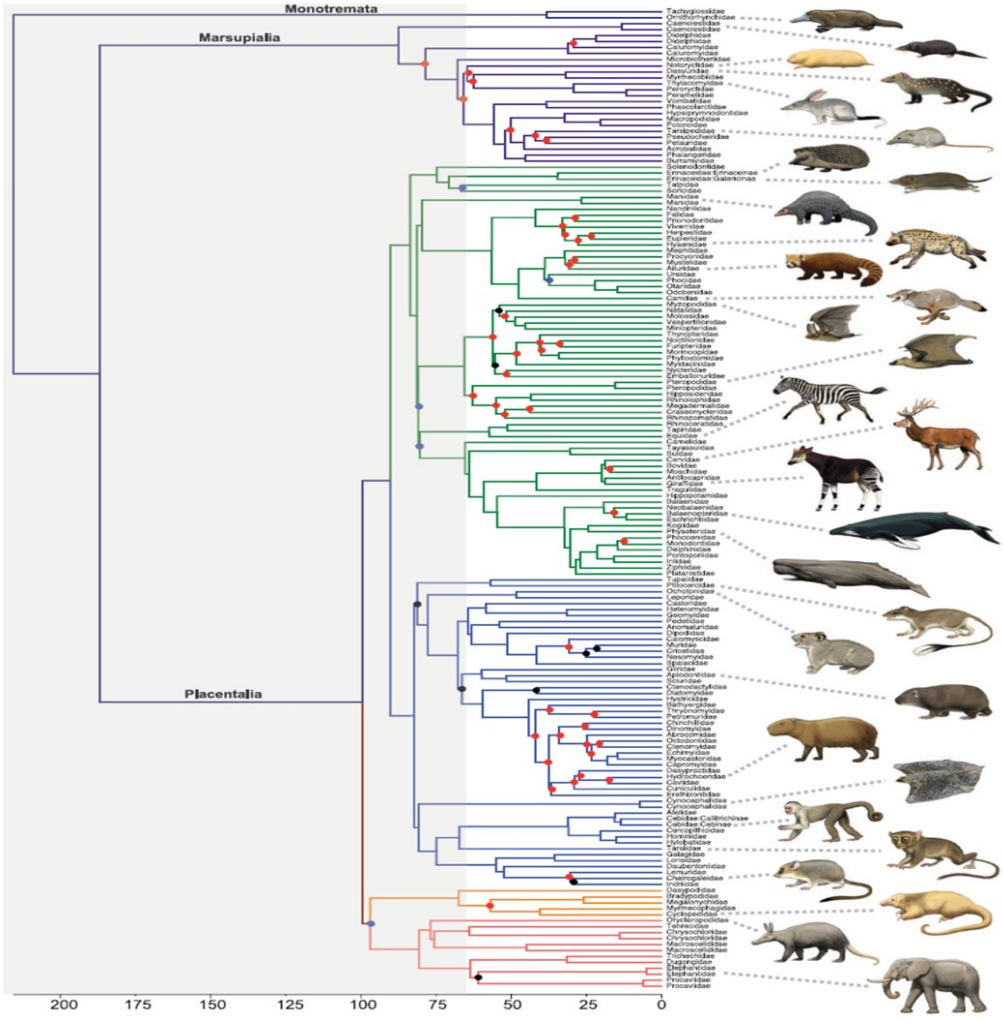

Monotremata
Marsupialia
Placentalia
200
175
150
125
100
75
50
25
0

Figure 3.1.

Phylogeny of the class Mammalia from Meredith et al. (2011)

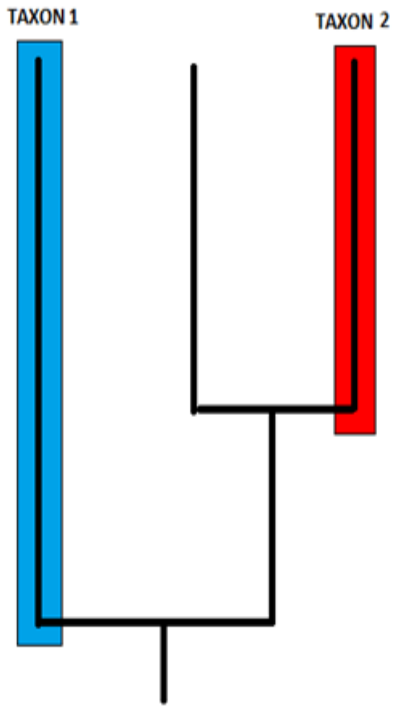


figure 3.2.

Exemplification of taxa with different rate of speciation: age of taxon 1 > age of taxon 2; number of species of taxon 1 < number of species of taxon 2.

SPECIES	FAMILY	FAMILY n° species	proxy for generation time	TE Families from 5% dataset	TE Families from 1% dataset	DI from 5% dataset	DI from 1% dataset
Bos taurus	Bovidae	156	21	3	31	23232	17
Canis lupus familiaris	Canidae	38	11	4	14	34441	194
Callithrix jacchus	Cebidae	62	20	11	1	471	12
Cavia porcellus	Cavidae	14	4,4	10	39	14351	457
Choloepus ofmanni	Bradipodidae	4	45	4	23	1750	262
Dasypus novemcinctus	Dasipododae	21	19	0	6	2767	0
Dipodomys ordii	Heteromidae	59	3	6	21	29383	171
Echinops telfair	Tenrecidae	30	2,88	0	5	2267	0

Equus caballus	Equidae	9	42,7	2	20	8133	245
Erinaceus europaeus	Erinaceidae	24	9,63	9	63	6991	223
Felis catus	Felidae	40	9	3	16	36290	1446
Homo sapiens	Hominidae	5	177	12	48	11712	2084
Lama pacos	Camelidae	6	29,85	0	10	219	0
Loxodonta africana	Elefantidae	3	154	0	7	457	0
Macaca mulatta	Cercopitecidae	82	54	15	67	26636	1073
Macropus eugenii	Macropododae	65	16,88	2	7	3745	6
Otolemur garnettii	Galagidae	15	10,13	2	12	9629	51
Monodelphis domestica	Didelphidae	87	4,92	70	8	13921	417
Mus musculus	Muridae	1330	2,2	23	113	27571	737
Myotis lucifugus	Vespertilionidae	320	8,83	15	55	59290	2565
Ochotona princeps	Ochotonidae	30	11,87	10	54	7427	209
Orictolagus cuniculus	Leporidae	61	17,17	3	36	23007	25
Ornitoryncus anatinus	Ornitorincidae	1	21,57	0	10	116	0

Microcebus murius	Cheirogaleidae	34	24,33	7	26	3279	77
Pan troglodites	Hominidae	5	163,7	8	50	8218	620
Procavia capensis	Procavidae	4	23	6	40	4127	159
Pteropus vampyrus	Pteropodidae	166	29,53	4	16	1051	77
Rattus norvegicus	Muridae	1330	4,27	21	92	27160	2321
Sorex araneus	Soricidae	376	10,17	6	25	1532	61
Spermophilus tridecemlineatus	Sciuridae	273	12	5	47	7201	73
Tarsius syrichta	Tarsiidae	5	23	2	23	2782	6

Table 3.1. Total genomic data of mammal species analyzed: TE Families from 5% dataset and 1% dataset = total number of families of transposable elements deviating <5% and 1% from their consensus sequences. DI from 5% dataset and 1% dataset = density of insertion of transposable elements deviating <5% and 1% from their consensus sequences.

3.2 Results

The first parameter considered is the proxy for generation time. Ten out of 19 comparisons do not follow the expected trend (see fig 3.3); then in our species sample the generation time should not be strictly correlated with the rate of speciation.

The 19 comparisons among the mammalian species selected were repeated for the number of families of TE belonging to 5% dataset. In this way we tested the species with the parameter suggested by Jerzy Jurka in the Carrier Subpopulation Hypothesis. Five comparisons out of 19 don't confirm the expected trend: the Insectivora *Erinaceus europaeus-Sorex araneus* and four pairs within the order Primates *Tarsius syrichta-Callithrix jacchus, Tarsius syrichta-Otolemur garnettii, Microcebus murinus-Callithrix jacchus* and *Otolemur garnettii-Callithrix jacchus* (Figure 3.4).

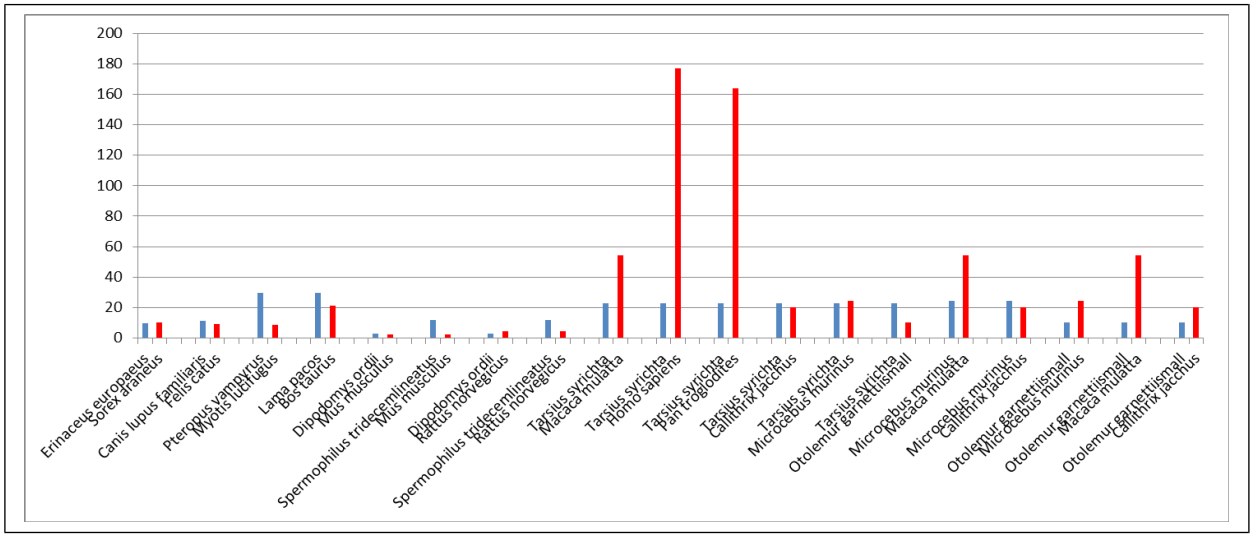


Figure 3.3. Comparison of proxy for generation time 19 mammal's pairs selected. X axis: blue bars; species of the family with lower rate of speciation, red bars species of the family with higher rate of speciation. Y axis proxy for generation time in months.

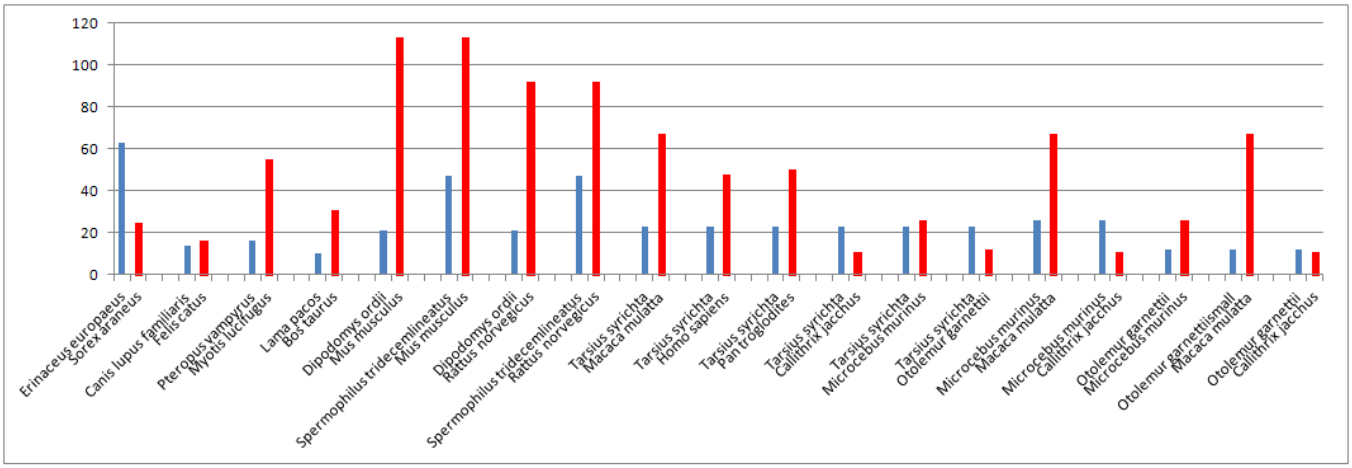


Figure 3.4. Comparison of number of families of TE of 19 mammal's pairs selected. X axis: blue bars; species from taxa with lower rate of speciation, red bars; species from taxa with higher rate of speciation. Y axis: number of families of TE belonging to 5% dataset.

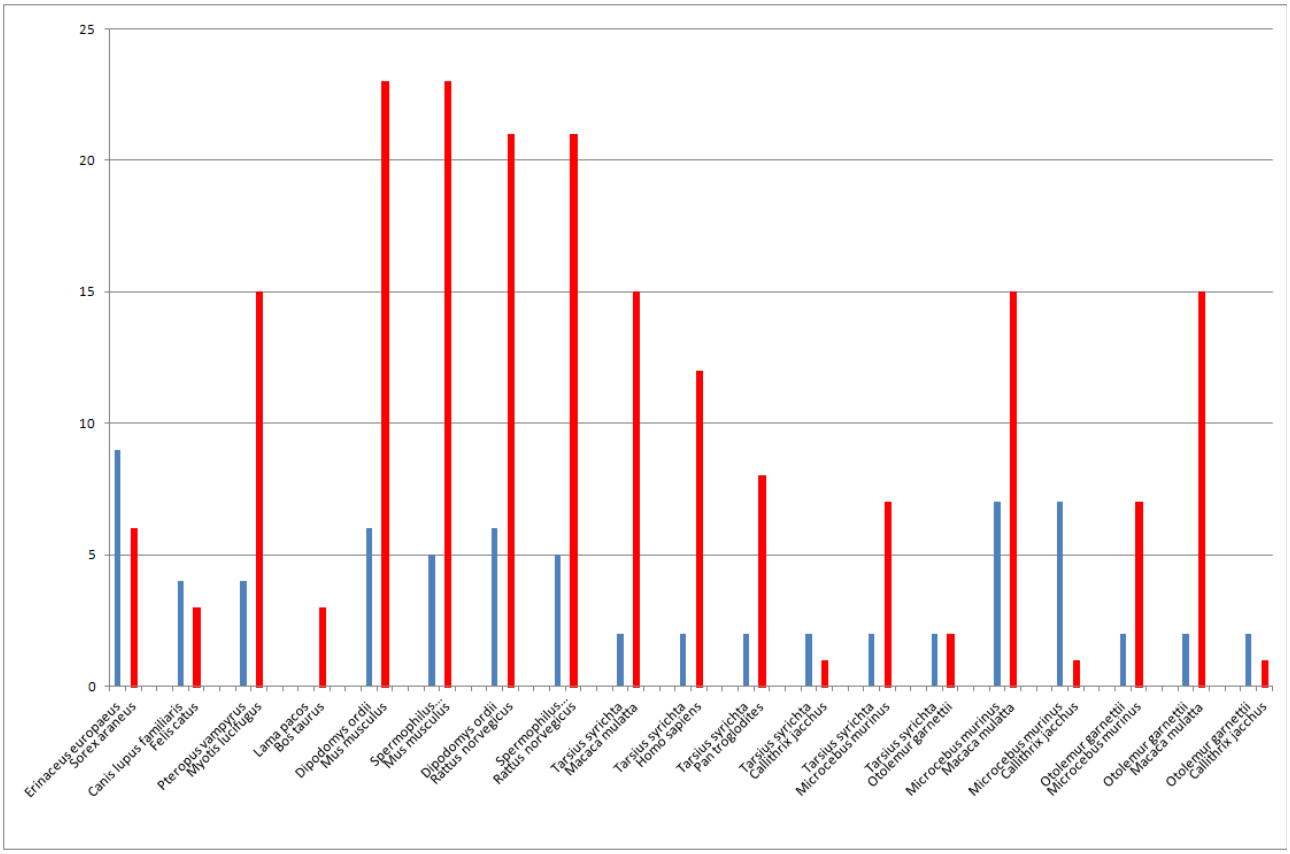


Figure 3.5. Comparison of number of families of TE of 19 mammal's pairs selected. X axis: blue bars; species from taxa with lower rate of speciation, red bars; species from taxa with higher rate of speciation. Y axis: number of families of TE belonging to the 1% dataset.

The comparison was repeated for families belonging to the 1% dataset (See Figure 3.5). It is possible to observe that five out of 19 comparisons show more TE families in the taxon with lower rate of speciation; in the comparison *Tarsius syrichta-Otolemur garnettii* both species contain two families. On the whole this parameter appears to be less related with the rate of speciation considering the "addition" of Carnivora. Indeed *Canis lupus familiaris* contains one more family than *Felis catus.*

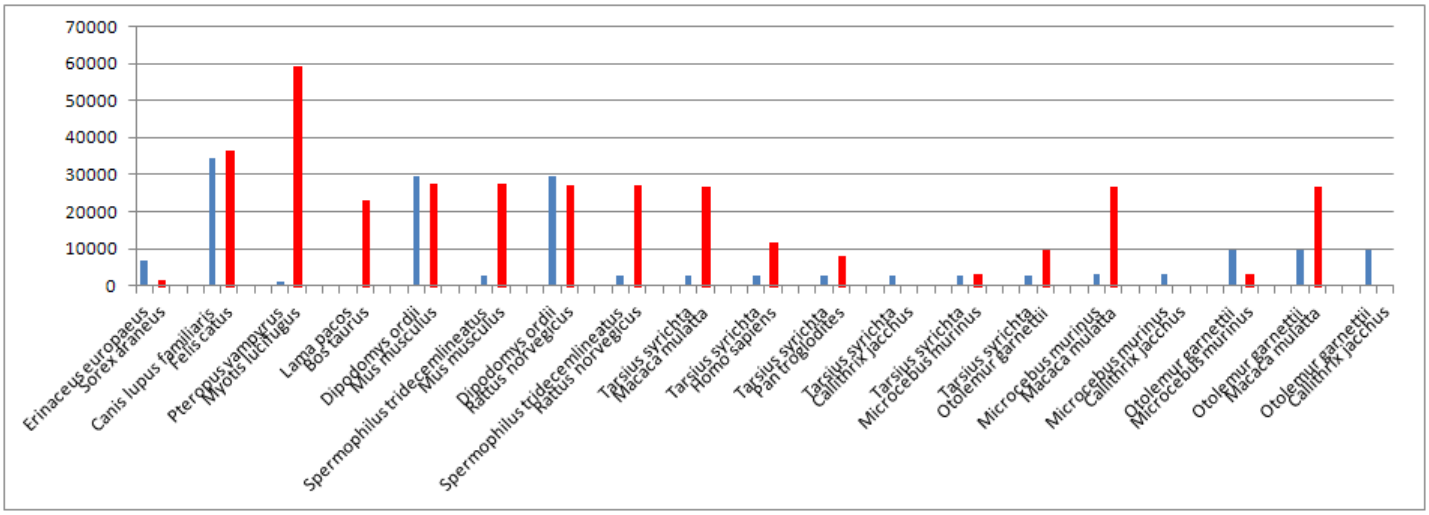


Figure 3.6. Comparison of density of insertion of TE of 19 mammal's pairs selected. X axis: blue bars species from taxa with minor rate of speciation, red bars; species from taxa with higher rate of speciation. Y axis: inserctions for Gigabase of TE belonging to the 5% dataset.

Looking at the comparisons among TE insertion densities belonging to the 5% dataset it appears that the pair belonging to the order Insectivora, four pairs of the order Primates and two in the order Rodentia do not follow the expected trend (see Fig. 3.6). The last parameter considered is the density of insertions of elements belonging to the 1% dataset. This parameter showed the most statistically significant relation between TE content and rate of speciation. The only data that do not follow the expected trend are represented by the pair of Insectivora species *Erinaceus europaeus-Sorex araneus* and by the Primates *Microcebus murinus-Callithrix jacchus* and *Otolomer garnettii-Callithrix jacchus* (see figure 3.7). These three exceptions appear in all the previous analyses (see Fig.3.6,3.5,3.4).

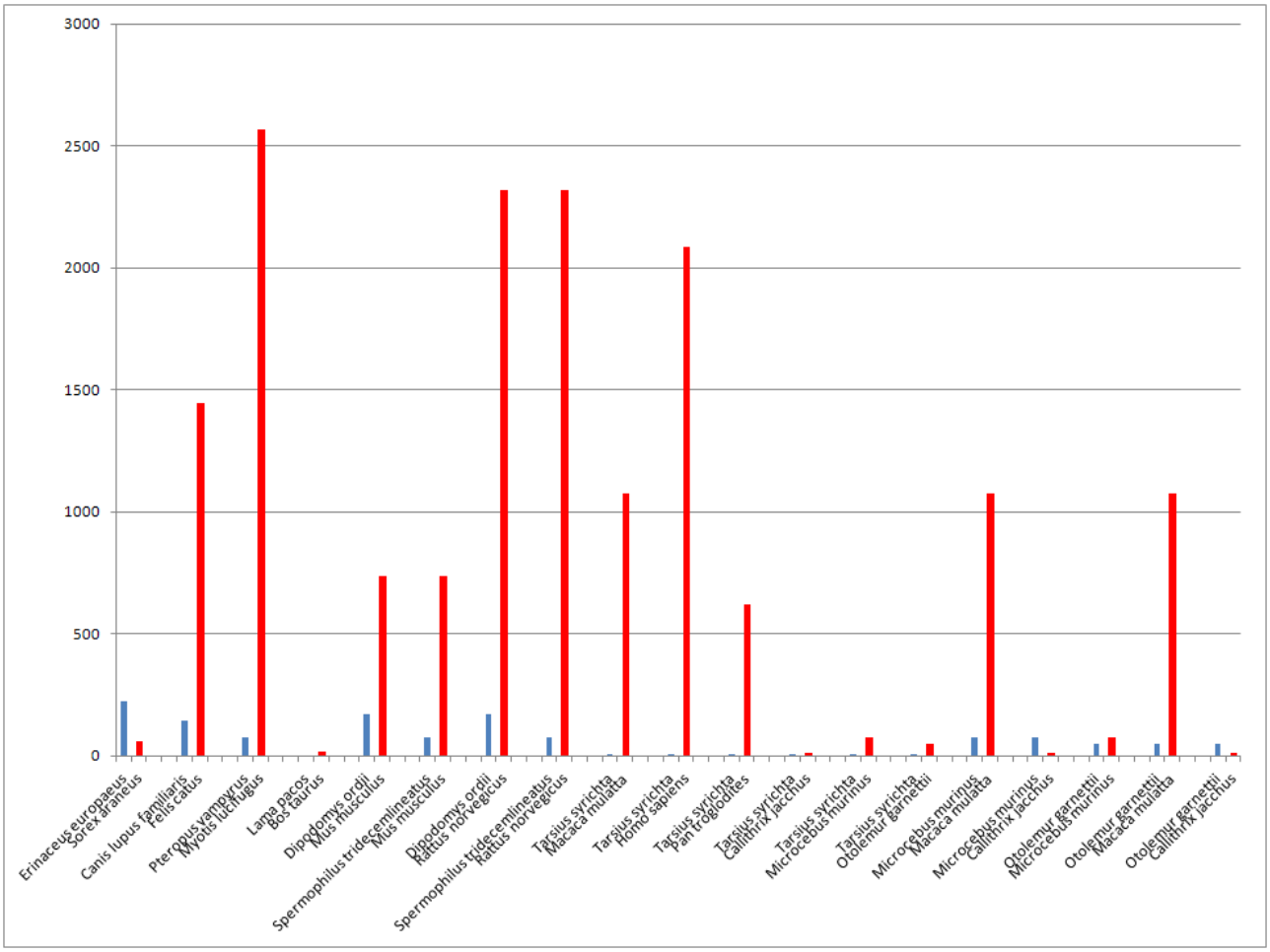


Figure 3.7. Comparison of density of insertion of TE of 19 mammal's pairs selected. X axis: blue bars; species from taxa with minor rate of speciation, red bars; species from taxa with higher rate of speciation. Y axis: inserctions for Gigabase of TE belonging to the 1% dataset.

PARAMETER	W	P
Proxy for generation time	73	0,4
TE_families_5%	26	0,01
TE_families_1%	19	0,01
DI_5%	47	0,05
DI_1%	18	0,01

Table 3.3. Wilcoxon test for not normally distributed five parameters. W value = calculated values for the each one of the five analyzed parameters. The lower P levels to find statistically significant differences within each parameter are displayed. P values corresponds with two tailed significance level.

Values calculated for all parameters analyzed are not normally distributed then the validation was carried out with the non parametric test of Wilcoxon. Considering the 19 pairs of species analyzed, I can state that the proxy for generation time does not have any relation with the rate of speciation while all parameters connected to Transposable Elements display significant differences between the number of pairs that follows the expected trend and the others. The density of insertions of Tranposable Elements belonging to the 5% dataset is related to the rate of speciation with a significance of 95%, for the other three parameters the significance is of 99%. The parameters considered above relate only to recently mobilized TE. Two pairs of species were selected among the 19 previously analyzed and their TE total content in bps was calculated. Fig 3.8 shows an example of how the evolutionary framework is applied on the pairs of Ceratiodactyla species. The family Camelidae (with eight living species) is more than 40 mya older with respect to the Bovidae one (that counts 156 extant species; see Fig 3.1). These features make this comparison particularly suitable in the light of the rate of speciation; in fact looking to the density of insertions we have strong evidences of a link between TE and rate of speciation (see Fig 3.9 a,b.). The fraction of total TE shows the same trend with less than 30% for *Lama pacos* and 45% for *Bos taurus* (Fig. 3.9c). Observing the four TE superfamilies separately it is possible to notice that they follow different trend. Class II TE and LTR elements density is higher in *Lama pacos* than in *Bos taurus* (Fig 3.9c). The same analyses were carried out in the order Rodentia. The ratio among the number of species ascribed to Heteromidae and Muridae is 1/20 as in Camelidae and Bovidae but the difference in the taxa age is only about 10 mya (Fig 3.1). It is noteworthy that, while the density of total insertions

belonging to the 1% dataset is significantly higher in *Rattus norvegicus* with respect to D*ipodomys ordii* (2321 *vs* 171), the opposite trend is observed for density of insertions of elements belonging to the 5% dataset (27160 in *Rattus norvegicus* against 29383 in *Dipodomis ordii*). On the other hand, if we observe the percentage of total TE, the hypothesis of a content in TE positively correlated with the rate of speciation fits since *Rattus norvegicus* has a TE content higher (almost 40%) than *Dipodomys ordii* (nearly 25%). Considering the TE superfamilies separately the trend is the same for all of them except for SINEs that are more than double in *Dipodomys ordii* than in *Rattus norvegicus* (see Fig. 3.11).

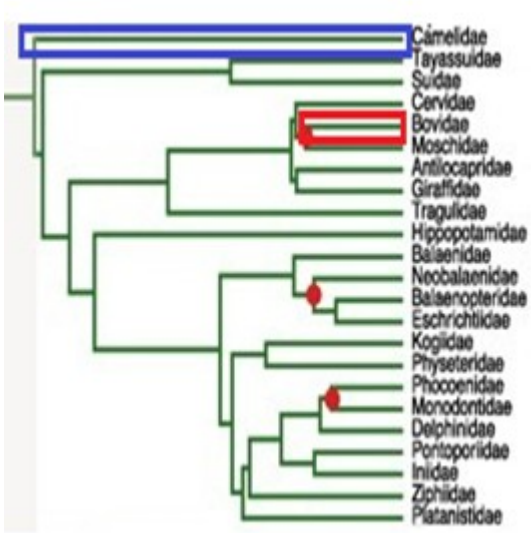


Figure 3.8. Species analyzed from Cetartyodactyla order; blue rectangle: family selected for lower rate of speciation; red rectangle: family selected for higher rate of speciation.

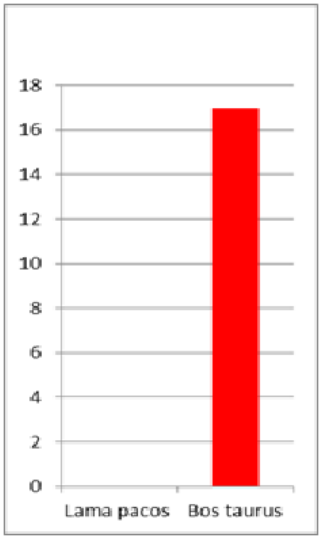


Figure 3.9.a. X axis; pair of species analyzed for Cetartyodactyla order. Y axis; density of insertion of transposable elements belonging to the 1% dataset.

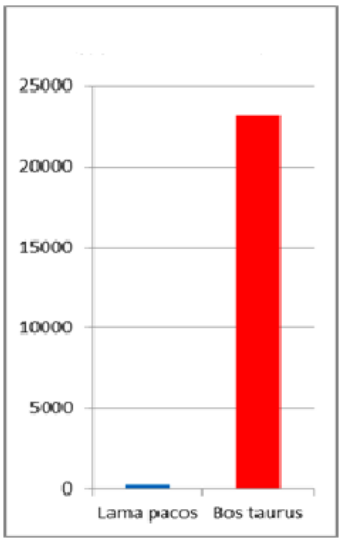


Figure 3.9.b. X axis: pair of species analyzed for Cetartyodactyla order. Y axis: density of insertions of transposable elements belonging to the 5% dataset.

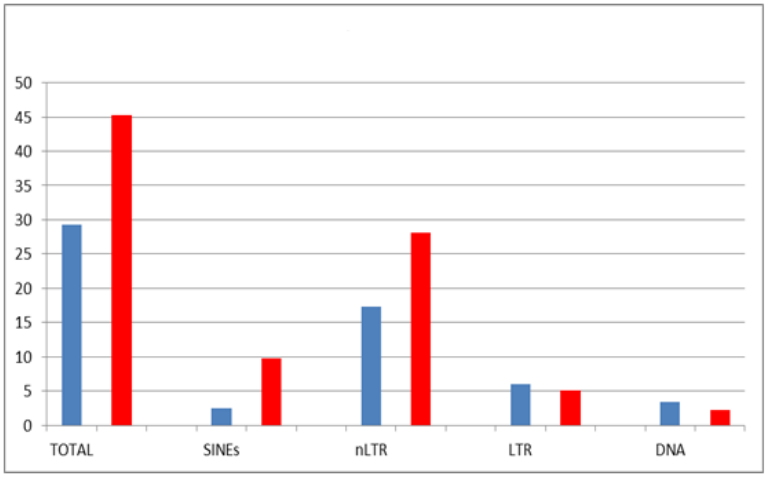


Figure 3.9.c. X axis: blue bars; *Lama pacos*, red bars; *bos taurus*. Y axis: percentage of genome occupied by transposable elements.

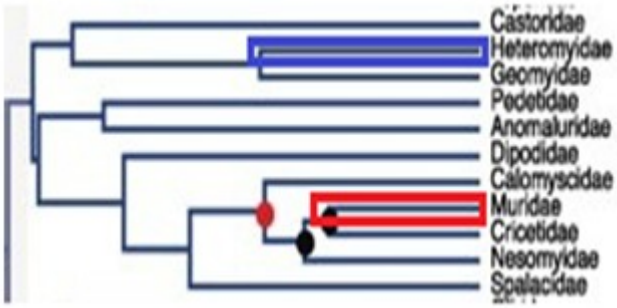


Figure 3.10. Species analyzed from Rodentia order; blue rectangle: family selcted for lower rate of speciation; red rectangle: family selcted for higher rate of speciation.

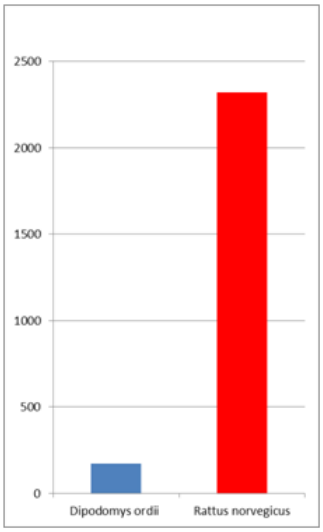


Figure 3.11.a. X axis: pair of species analyzed for Cetartyodactyla order. Y axis: density of insertions of transposable elements belonging to the 1% dataset.

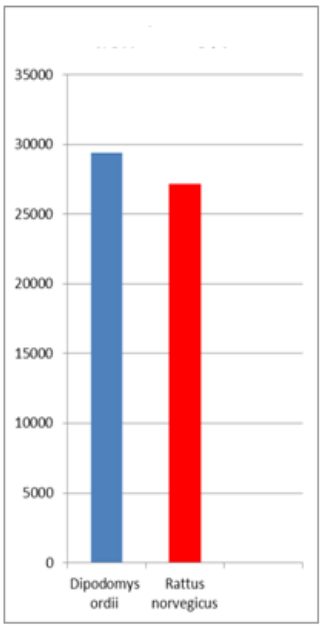


Figure 3.11.b. X axis: pair of specie analyzed for Rodentia order. Y axis: density of insertions of transposable elements belonging to the 5% dataset.

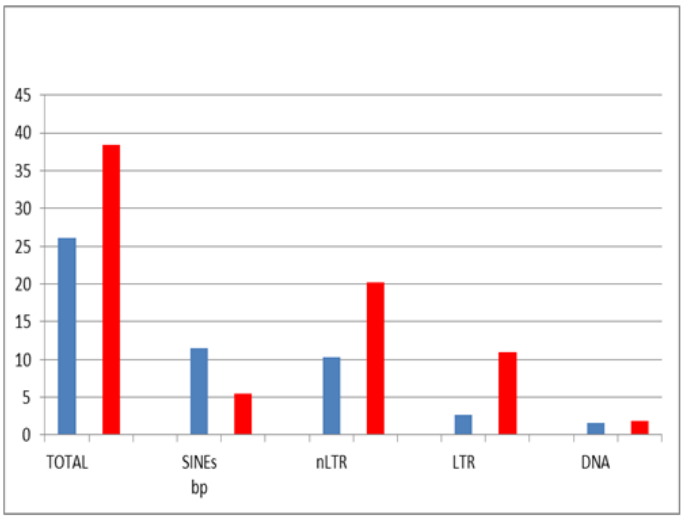


Figure 3.11.c. X axis: blue bars*; Dipodomys ordii*, red bars *Rattus norvegicus*. Y axis: percentage of genome occupied by transposable elements.

The calculation of average density of Transposable Elements insertions (average ins/Gb) for each of the four superorders belonging to subclass Placentalia displayed that Afrotheria and Xenarthra have an average ins/Gb lower than Laurasiantheria and Euarchontoglires for both 5% and 1% datasets (see Table 3.3). The pairs Euarchontoglires-Xenarthra and Euarchontoglires-Afrotheria show a ratio about 5/1. The Laurasiatheria-Xenarthra and Laurasiatheria-Afrotheria ratios are about 10/1 for the average insertions belonging to the 5% dataset (see figure 3.12). Minor differences can be observed for average insertions from the 1% dataset. Laurasiatheria-Xenarthra comparison has a ratio of 3/1, while Laurasiatheria- Afrotheria one is of 2/1. The paucity of available species in the taxa with low rate of speciation (two for Xenartra and three for Afrotheria) makes the sample analyzed

insufficient to have a statistical support for TE insertions belonging to the 1% dataset (Fig. 3.13), while for the density of insertions from the 5% dataset we are 95% confident that values calculated reflect the real trend of the species belonging to the four taxa (Fig. 3.12).

	Euarchontoglires	Laurasiatheria	Xenarthra	Afrotheria
Av-DI TEs 5%	12393	19019	2258,5	2284
Conf-int 5%	5005	11912	2270	2526
Av-TEs 1%	401	536	131	52
Conf-int 1%	256	509	478	124

Table 3.3. Av-DI TEs 5%,1% = average of densities of insertions of transposable elements belonging to the 5% and 1% dataset. Conf-int 5%,1% = 95% confidence interval for densities of insertions of transposable elements belonging to 5% and 1% dataset.

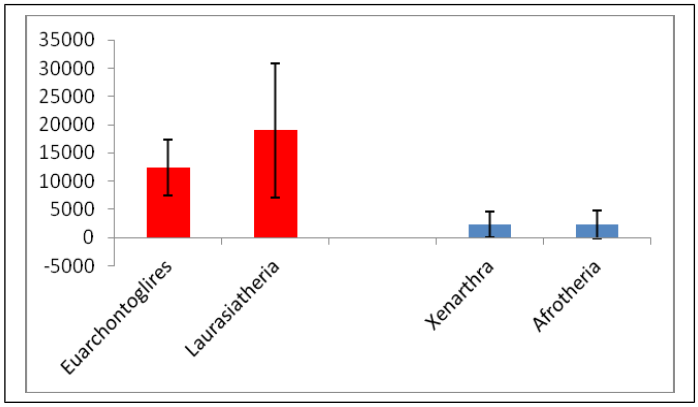


Figure 3.12.

X axis: red bars taxa with higher rate of speciation, blue bars taxa with lower rate of speciation. Y axis: average density of insertions of transposable elements belonging to the 5% dataset.

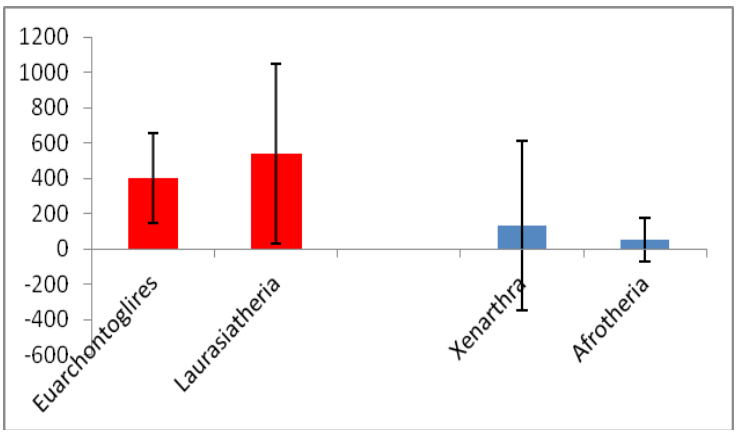


Figure 3.13. X axis: red bars taxa with higher rate of speciation, blue bars taxa with lower rate of speciation. Y axis: average density of insertions of transposable elements belonging to the 1% dataset.

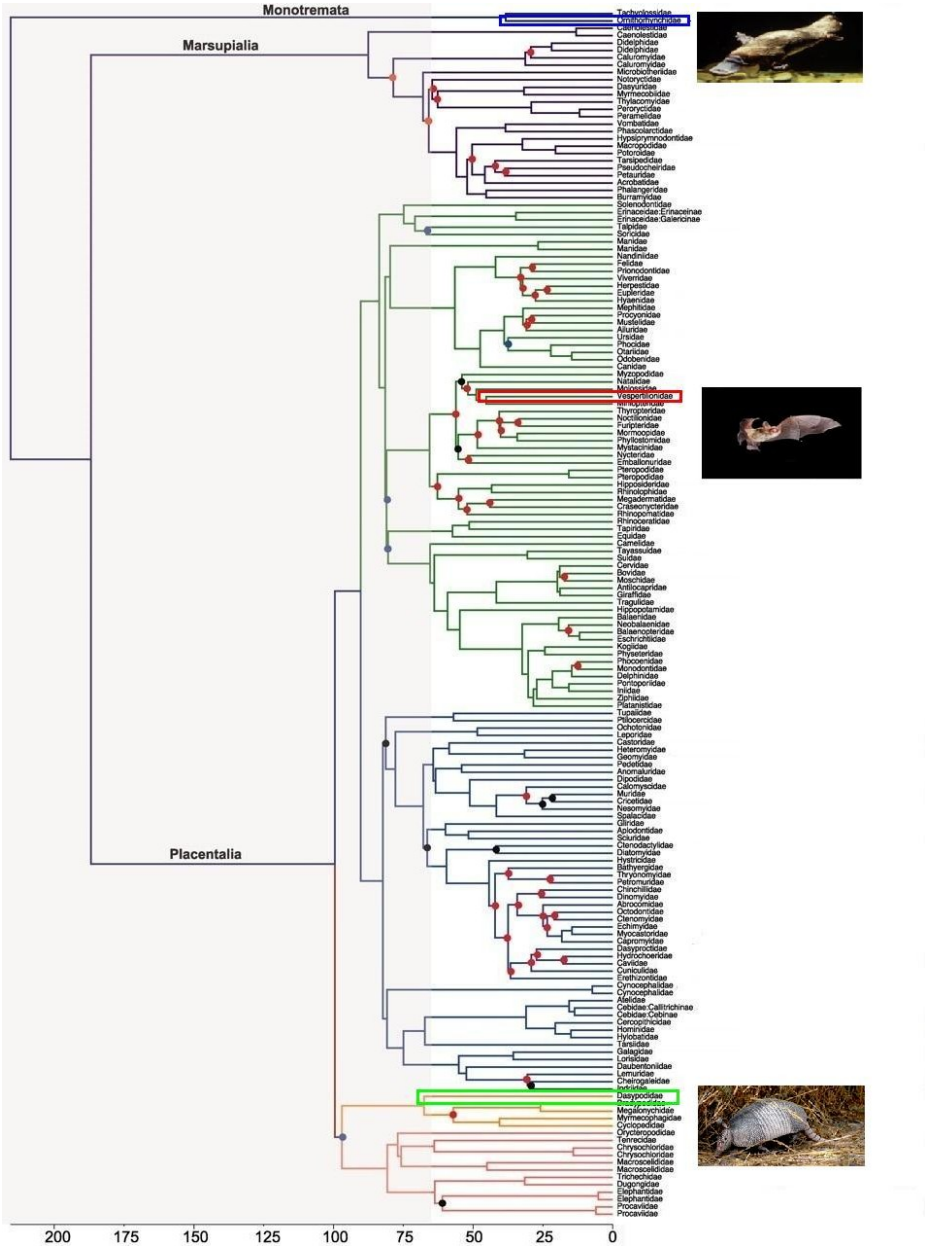


Figure 3.14. Selection of three species belonging to Mammalia: blue rectangle; lowest rate of speciation, *Ornithorhynchus anatinus* order Monotremata. Green rectangle; intermediate rate of speciation, *Dasypus novemcinctus* Order Cingulata. Red rectangle; highest rate of speciation, *Myotis lucifugus* Order Chiroptera.

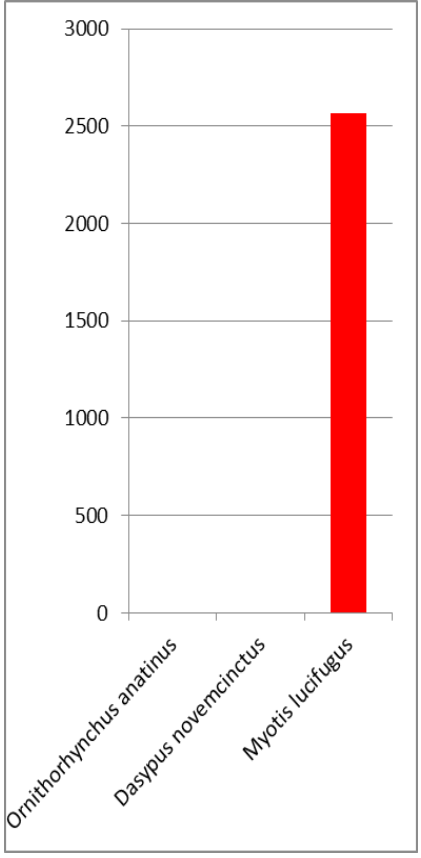


Figure 3.15.a X axis: Species listed from left side by growing Reltive Rate of Speciation . Y axis: density of insertions of transposable elements belonging to 1% dataset.

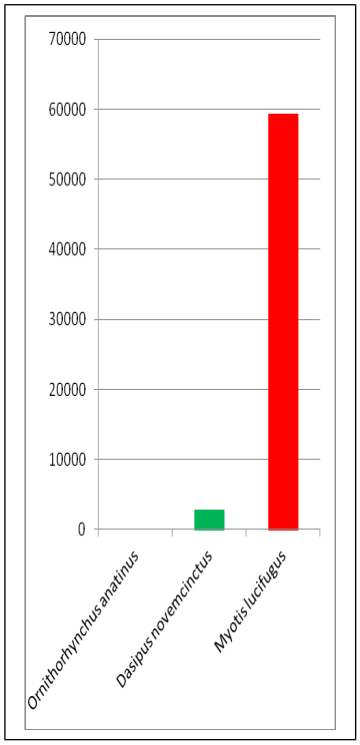


Figure 3.15.b. X axis: Species listed from left side by growing Reltive Rate of Speciation. Y axis: density of insertions of transposable elements belonging to the 5% dataset.

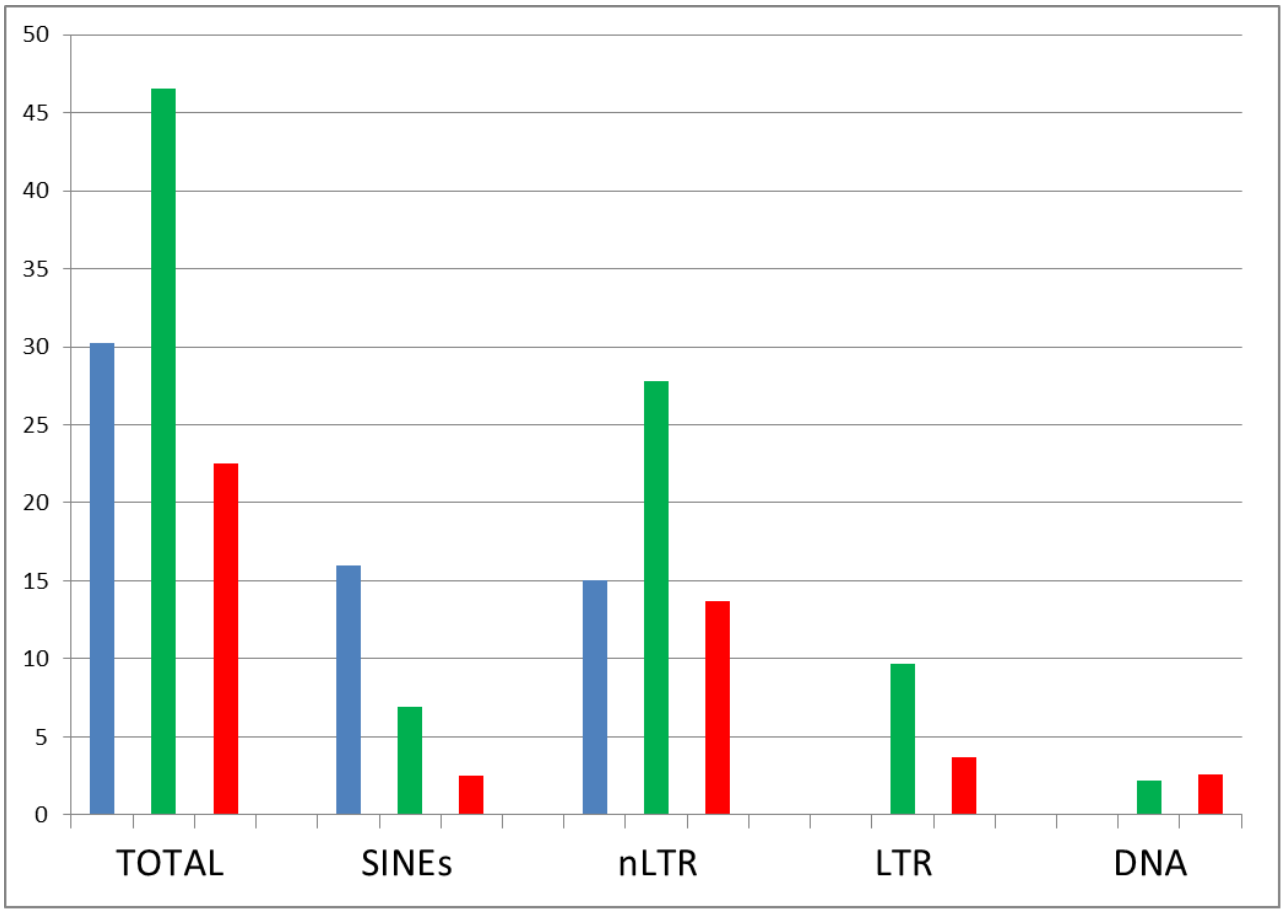


Figure 3.15.c. X axis: blue bars; *Ornithorhynchus anatinus*, green bars; *Dasypus novemcinctus*, red bars; *Myotis lucifugus*. Y axis: percentage of genome occupied by transposable elements.

Three mammal species have been selected for a comparison between orders with increasing rate of speciation (see fig. 3.14). The density of insertions of TE from 1% dataset reveals that only *Myotis*

lucifugus presents a recent activity of TEs with more than 2565 insertions in one GB (Fig. 3.15.a). The density of insertions of TE belonging to the 5% dataset reveals a very low value for *Ornithorhynchus anatinus* (116 ins/Gb), an higher density of insertions for *Dasypus novemcinctus* (2.767 ins/Gb) and the highest value in *Myotis lucifugus* (59290 ins/Gb Fig. 3.15.b). Insertions from the 5% dataset reveal a slight trace of activity in *Ornithorhynchus anatinus*. *Dasypus novemcinctus* has a twenty fold higher density of insertions while *Myotis lucifugus* shows the highest value with a four hundred-fold higher density than *Ornithorhynchus anatinus*. Both parameters appear directly correlated with the rate of speciation. The situation changes with data of total content of TEs. The analysis shows that the total TE content is not correlated with the rate of speciation. *Myotis lucifugus* contains the lowest amount with 22% followed by *Ornithorhynchus anatinus* with 30% and *Dasypus novemcinctus* with 47% (see Fig 3.15.c).

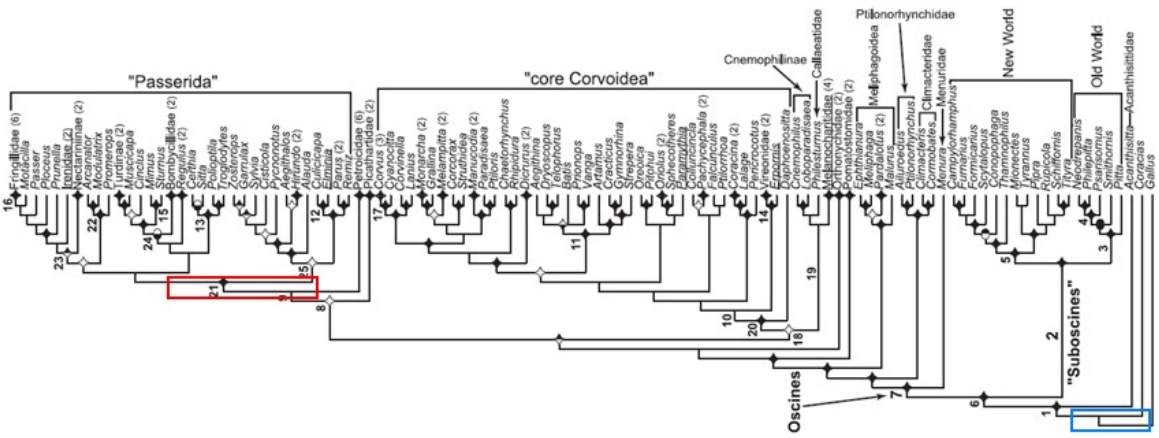


Figure 3.16. Neoaves phylogenetic tree from Barker et al. 2004. Red rectangle Order Passeriformes, blue rectangle order Galliformes.

To verify if the trend observed in the class Mammalia is generalizable, the analysis of percentage of genome containing TE was repeated in the class Aves. The phylogeny of Neoaves places the order Galliformes in a basal position, while the order Passeriformes appears to be the most recent (see fig. 3.16). The differences in their number of living species is considerable: Galliformes counts 219 living species and Passeriformes 5828. For both taxa the genomes of two species were analyzed (*Gallus gallus* and *Meleagris gallopavo* in Galliformes. *Geospiza fortis* and *Taeniopygia guttata* in Passeriformes). We analyzed the total TE coverage with *Gallus gallus* and *Meleagris gallopavo*

against *Geospiza fortis* and *Taeniopygia guttata* genomes. The total TE content does not appear related to the rate of speciation; *Meleagris gallopavo* and *Gallus gallus* contains about 6,5% and 5%, respectively, while *Taeniopygia guttata* and *Geospiza fortis* show about 7% and 6,5%, respectively. In all analyzed species only two superfamilies of non-LTR retroposon and LTR retroposons are present. It can be noted a greater amount of LTR in Passeriformes (3,5% for *Taeniopygia guttata* and 2,8% for *Geospiza fortis)* with respect to Galliformes (*Meleagris gallopavo* and *Gallus gallus* have a content of 0,9 and 0,1 %, respectively).

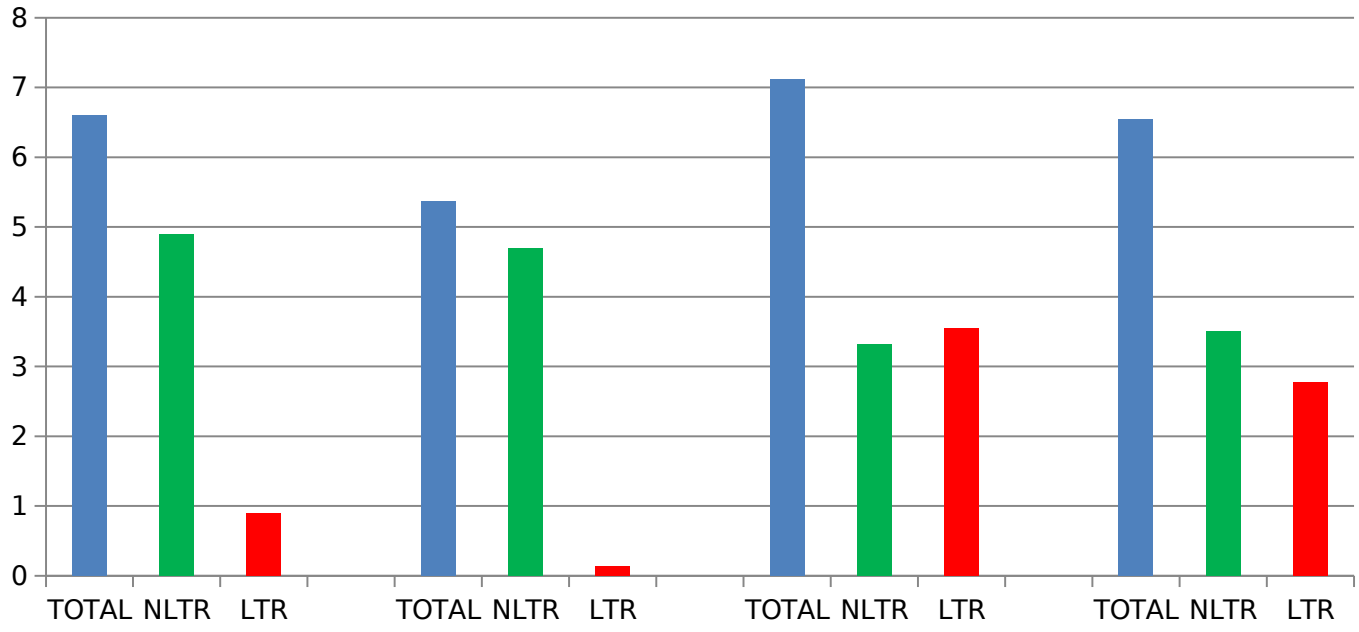


Figure 3.17. X axis: A; *Meleagris gallopavo*, B; *Gallus gallus*, C; *Taeniopygia guttata* D; *Geospiza fortis.* Y axis: percentage of genome occupied by transposable elements.

The last analysis on vertebrates taxa was carried out through the comparison between the well known living fossil *Latimeria chalumnae* and *Danio rerio*, the evolutionary distance between them being more than 400 Mya. The taxon Sarcopterygii ("lobe finned fish") contains eight species, while Actinopterygii ("ray finned fish") lists 31182 species. The total TE content follows the expected trend: *Latimeria chalumnae* shows a 14% of total TE content, while for *Danio rerio* the value raises to 48%. Despite this dramatic difference, not all TE superfamilies follow the same trend. SINEs are more represented in *Latimeria chalumnae* with 7%, while *Danio rerio* contains only the 3%. Non-LTR retroposons are 5% in *Latimeria chalumnae* and 3% in *Danio rerio.* LTR retroposons follow the expected trend with the 1% in *Latimeria chalumnae* and the 5% in *Danio*

rerio. DNA transposons show the biggest difference with 0,7% in *Latimeria chalumnae* and 36% in *Danio rerio* (See fig. 3.18).

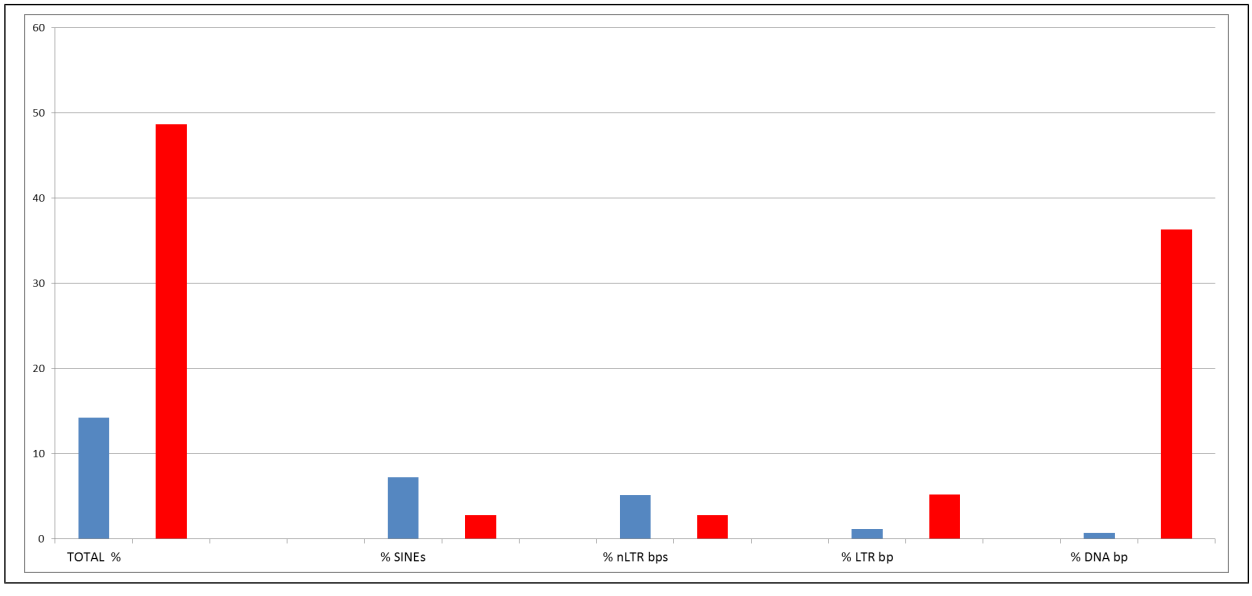


Figure 3.18. X axis: blue bars; *Latimeria chalumnae*, red bars; *Danio rerio*. Y axis: percentage of genome occupied by transposable elements.

Finally I compared the TE content of two insects, *Drosophila grimshawii* belonging to the group under fast speciation of Hawaiian species and *Drosophila virilis*. *D. grimshawii* presents the highest fraction of TEs, 14% while *D. virilis* shows the 10%. Non-LTR elements are the only TEs more represented in *D. virilis* with the 3% *vs* 2% of *D. grimshawii* (See fig. 3.19).

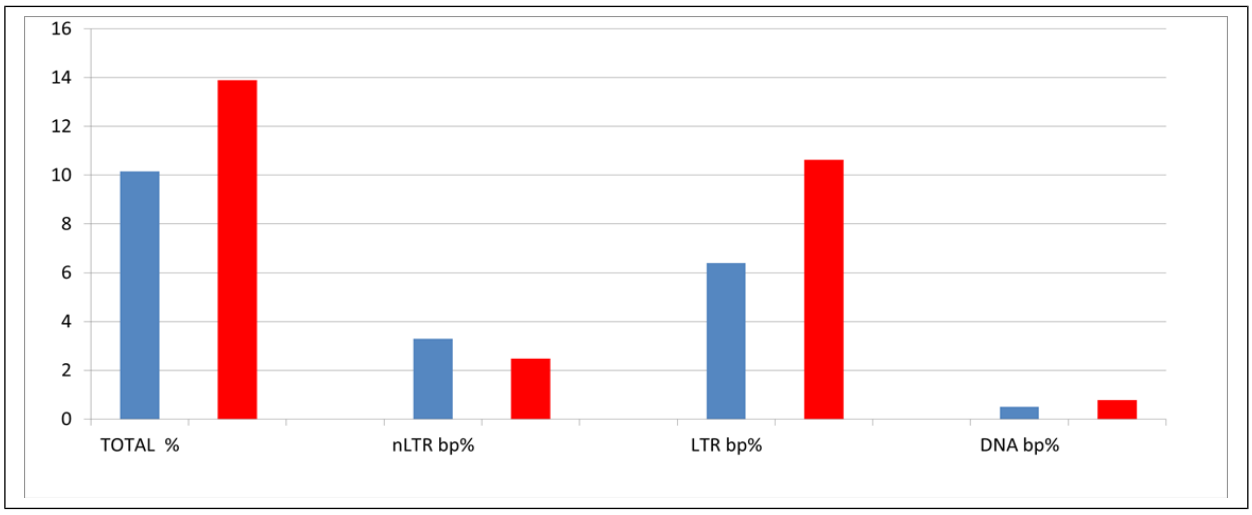


Figure 3.19. X axis: blue bars; *Drosophyla virilis*, red bars; *Drosophila grimshawii*. Y axis: percentage of genome occupied by transposable elements.

3.3 Discussion

In paleontology, the term “evolutionary radiation” defines events with a large number of fossil forms, which appeared in a relatively short period of time as the Cambrian explosion. In zoology, one speaks of taxa in which the radiation is taking place at present time, as in cichlids fish of Lake Malawi (Joyce et al. 2005). In this term is qualitative not quantitative. In this work I define rigid criteria, and suggest a system that allows more accurate comparisons (See fig 3.2). If a taxon has a lower age and more species than another, I am confident in the fact that in recent times speciation events were more frequent. With this system we can evaluate if a taxon possess a relative rate of speciation (“RRS”) higher or lower than another. Sometimes comparisons are not possible. For example, the family Hominidae has fewer species than Cercopitecidae; on the other hand Cercopitecidae is older than Hominidae. In this case we can establish if a difference in the rate of speciation occurs, because the two families could have the same rate of speciation and the higher age of Cercopitecidae could result in a higher number of species. Among the 31 species of mammals analyzed by Jurka et al. (2011), I selected pairs of mammalian species belonging to families of the same order. Species with short generation time present favorable conditions to spread in subpopulations. This could affect the rate of speciation (Makarieva and Gorshkov 2004). For this reason I have started from the analysis of a proxy for generation time. This analysis allowed me to check a factor unrelated to the TE content of. So I expect that a higher rate of speciation corresponds to a shorter generation time. The values give only nine of 19 comparisons corresponding to the expected trend (Fig. 3.3). At least for the mammalian species analyzed, I can therefore conclude that factors affecting the speed of reproduction and consequently the population size do not appear correlated with the rate of speciation. A quite different situation emerges when genomes are analyzed. The number of TE families belonging to the 5% dataset is related to the rate of speciation in 14 comparisons out of 19 (Fig. 3.4). The trend is confirmed when the families of TE belonging to the 1% dataset are concerned, with 15 out of 19 comparisons that follow the expected trend (see Fig. 3.5). This datum demonstrates that in the light of the rate of speciation the amount of

families of TEs recently mobilized is a feature not randomly distributed in Placentalia. If the TE activity has an impact on the genome sufficient to increase the likelihood of speciation, it is reasonable to presume that the number of single TE insertions occurred in a certain period of time is a more important feature with respect to the number of families to which they belong. In fact, we can have a recent big spread of TE belonging to a unique family, as observed f.e. for the SINEs_Fc family of *Felis catus* with 2.312 elements from the 1% dataset (Jurka et al. 2011). The comparison *Canis lupus-Felis catus* fails to give the expected trend with four families of TE from the 1% dataset in *Canis lupus* and only three in *Felis catus.* If we look at the number of elements of the same group, *Felis catus* embodies 2372 elements, while *Canis lupus* only 462. In addition, we should take in account that the same number of insertions has a higher impact in smaller genomes. So I correlated the number of recently mobilized TE to the genome size, in order to have the density of insertion. This new parameter allowed me to make comparisons with all species without the interference of genome size. While the density of TE insertions from the 5% dataset (Fig 3.6) shows a correlation with the rate of speciation lower with respect to the number of TE families, the density of insertion of TE from the 1% dataset shows the highest correlation among TE and rate of speciation. For all parameters used in this work the comparisons *Erinaceus europaeus-Sorex araneus*, *Microcebus murinus-Callithrix jacchus*, *Otolemur garnettii-Callithrix jacchus* seem to be an exception to the rule. Other factors contributing to the rate speciation must therefore be involved in some cases with a greater effect with respect to the TE content. As confirmation of the effectiveness of this parameter, the species of the families Muridae: *Mus musculus-Rattus norvegicus* and Hominidae: *Pan troglodites-Homo sapiens* showed that the variation in TE content inside the same taxon don't change the result of comparison with the other species. The statistical validation of data obtained in the analysis of the 19 pairs of mammals state that only the four parameters used to test the activity of transposable elements are related to the rate of speciation. Among them the lowest W value was reached by evaluation of the density of insertions of elements belonging to the 1% dataset (Table 3.3). In addition this parameter display 16 out of 19 pairs

following the expected trend (Fig 3.7). Judging from these data we infer that the density of insertion of transposable elements diverging less than 1% from their consensus sequence could be the most related to the rate of speciation. The evaluation of average density of TE insertion for Laurasiatheria, Euarchontoglires, Xenarthra and Afrotheria allowed me to make a large scale analysis of the whole subclass Placentalia. Applying the same rules to obtain the Relative Rate of Speciation, it is possible to carry out four comparisons: Laurasiatheria-Xenarthra, Laurasiatheria-Afrotheria, Euarchontoglires-Xenarthra, Euarchontoglires-Afrotheria (see Fig 3.1). Also in this case the starting hypothesis is confirmed. For both TE belonging to both 1% and 5% datasets, Laurasiatheria and Euarchontoglires exceed by more than three-fold the density of insertion respect to Xenarthra and Afrotheria (Fig. 3.12,3.13). In this analysis, the assessment of the confidence interval tells us that the data related to transposons from the 1% dataset are not statistically supported. This is probably due to a too small sample of species available (two for Xenarthra and three for Afrotheria). We evaluated the percentage of total TES in *Latimeria chalumnae*, one of the 8 species of Sarcopterygii still alive and *Danio rerio* belonging to Actinopterygii that are present with more than 31.000 living species. The common ancestor of these species lived more than 420 mya ago when Sarcopterygii and Actynopterigii separated (Coates 2009). *Latimeria chalumnae* is a well known living fossil; the 400 mya ago fossils don't show morphological differences respect to the species living at present days. The genome size of *Latimeria chalumnae* is over two gb whereas while *Danio rerio* one is 1,4 gb. It is now established that the amount of transposable elements is directly correlated to the genome size (Kidwell 2002). Paradoxically *Latimeria chalumnae* has a total TE content of only 14%, while in zebrafish TEs accounted for 48% of the genome. It is likely that throughout its history, the genome of *Latimeria chalumnae* had an inflation originated by TE bursts that have led the current size. If TE have ceased their activities, it is possible that the oldest elements have accumulated many mutations and are not more detectable by Repeat Masker. On the contrary, zebrafish presents a genome with a fraction of TE more than threefold higher.

The extreme differences in the rate of speciation among the two species may coincide with their TE content as in mammals. Whereas most of the studies on the relationship TE-speciation have been performed in vertebrates, I felt it necessary to increase the spectrum of the species analyzed in order to determine whether this mechanism is generalizable in all phyla. I have tried to test the hypothesis in two insects of the Diptera order: *Drosophila virilis* and the Hawaiian *Drosophila grimshawii* (see fig. 3.19). I classified Hawaiian species as a group with a higher rate of speciation considering that it had a radiation of about 1000 species becoming from a common ancestor that colonized the islands about 25 Mya ago (O'Grady et al. 2011). *Drosophila virilis* was selected as the nearest relative from the north American continent. Notwithstanding that this comparison is made between two closely related species, there is an appreciable difference in the fraction of total TE (i.e. 10 % in *Drosophila virilis* and 14% in *Drosophila grimshawii*). The higher content of total TE in the Hawaiian species could be a first clue that the dynamics observed in vertebrates may also be present in insects. The analysis of the total TE percentage was performed also in some mammalian genomes. In the comparison *Lama pacos-Bos taurus* (fig. 3.8) the total TE percentage shows the same trend of the most recently inserted elements. (fig. 3.9.a.b.c). The Relative Rate of Speciation in Camelidae family is dramatically lower in the comparison with Bovidae, the ratio of species number being more than 1/25 while the ratio among the ages of the taxa being over 4/1. These extreme differences could reflect the differences in both recent and ancient TE bursts. In the comparison *Dypodomis ordii-Rattus norvegicus*, an higher density of TE belonging to the 5% dataset in the species with lower rate of speciation was observed. (See fig. 3.11.b). The fraction of total TE and the density of insertion of elements belonging to the <1% dataset follow the expected trend (see fig. 3.11a.c.). Those data reveal that within an intermediary period there was greater proliferation of TE in the family Heteromidae then in Muridae. At this stage, we selected three species of mammals belonging to different orders: *Ornythoryncus anatinus* (order Monotremata), *Dasypus novemcinctus* (order Cingulata) and *Myotis lucifugus* (order Chiroptera) (Fig.3.14). The three taxa are listed following a growing Relative Rate of Speciation. Data about the density of

insertions reveal that in *Ornitorincus anatinus* and *Dasypus novemcinctus* elements from the 1% dataset are lacking (Fig 3.15.a). The elements belonging to the 5% dataset are directly correlated with the Relative Rate of Speciation: 116 ins/Gb *Ornithorhynchus anatinus*, 2767 *in Dasypus novemcinctus*, 59290 in *Myotis lucifugus*. (see Fig 3.15.b.) Surprisingly the trend changes when we analyze the percentage of total TE: 30% in *Ornithorhynchus anatinus*, 47% in *Dasypus novemcinctus*, 22% in *Myotis lucifugus* (see Fig. 3.15 c.). Probably, past episodes of TE bursts in *Ornithorhynchus anatinus* and *Dasypus novemcinctus* are unrelated with the actual rate of speciation. I tried to represent three models that allow an approximate classification of the genome type on the basis of its TE contents (See Fig 3.20). The majority of mammal genomes analyzed by Jurka are of type C. Some genomes are totally free of TE belonging to 1% dataset and therefore pertain to type B. This was observed in *Dasypus novemcinctus, Lama pacos, Echinops telfairi, Loxodonta africana*, (Jurka *et* al. 2011). These species belong to families with an estimated low rate of speciation, so we could define them as having "cold genomes". In the case of *Ornitoryncus anatinus* in addition to lacking the TE 1% deviating from theyr consensus sequences, the fraction of TE belonging to the 5% dataset is so low that we can classify it as type A. The low TE activity reduces the formation of new phenotypes and decreases the likelihood that two sub-populations reach reproductive isolation.

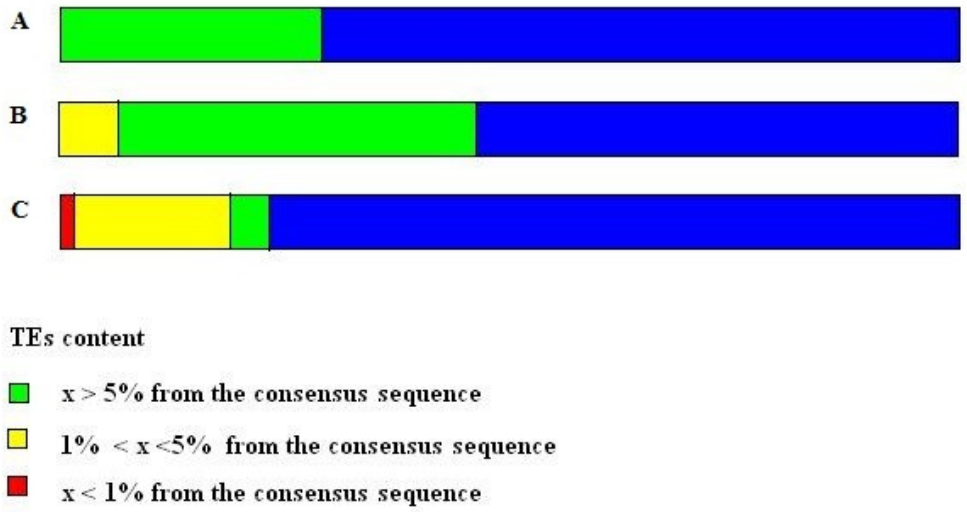


Figure 3.20. Modeling of genomes from the point of view of content in transposable elements, type A: genome devoid of TE <5% deviating from their consensus sequence, type B: genome devoid of TE <1% deviating from their consensus sequence, type C: genome equipped of TE <1% divergent from their consensus sequence.

All the works that tried to prove the link between speciation and Transposable Elements have a weak point. There are in fact taxa apparently under evolutionary radiation but with a very low content of TE. The small content of TE in genomes of birds let argue that their evolutionary radiation is dependent by other factors (Jurka et al. 2011). Apparently Aves is one of those taxa that are an exception to the rule. I therefore tried to test the evolutionary framework on Aves. Recent phylogenies of birds agree in considering the order Galliformes as basal to the others, while Passeriformes is the most derived. Galliformes has 219 living species while Passeriformes has 5829 (Catalogue of Life). The Relative Rate of Speciation is higher in Passeriformes respect to Galliformes. The four species of birds I analyzed show a total TE content ranging from a maximum of 8% in *Meleagris gallopavo* to a minimum of 5% in *Gallus gallus.* (Fig. 3.17). Among all species

analyzed in this work, birds showed the lowest TE fraction. Nevertheless genome analysis has revealed that LTR retrotransposons are poorly represented in the basal order: 0,9%, in *Meleagris gallopavo*, 0.1% in *Gallus gallus.* Otherwise Passeriformes show a higher content: 3,5% in *Taenyopigia guttata*, 2,8% in *Geospiza fortis*. Repeat Masker has revealed that the LTR superfamily is composed exclusively of Endogenous Retro Viruses. In a recent work it has been found that the total number of endogenous retroviruses is 1221 in *Taenyopigia guttata*, 150 in *Meleagris gallopavo* and 450 in *Gallus gallus* (Bolisetty et al. 2012). These data suggest a higher activity of the endogenous retrovirus in Passeriformes than in Galliformes. If Passeriformes confirms a higher TE activity respect Galliformes, the mechanism will be confirmed in Aves that were previously considered an exception to the rule (Jurka et al. 2011). In recent years authors that face this scientific question are increasing. The most important models developed are two: the Epi-transposon hypothesis (Zhe et al. 2009) and the Carrier Sub Population hypothesis (Jurka et al. 2011) (see fig 3.21).

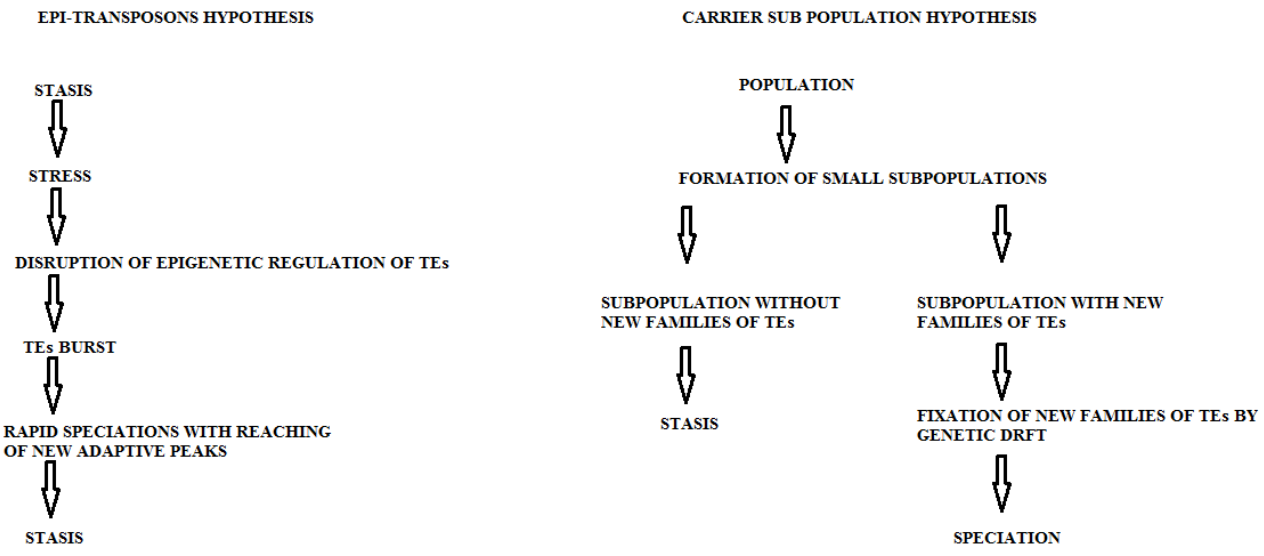


Figure 3.21.

Schematization of the two proposed mechanism that relate TE activity and speciation EPI-trasposon hypothesis (Zhe et al. 2009), Carrier Sub Population hypothesis (Jurka et al. 2011).

It is my opinion that both mechanisms can be present contemporarily. TE bursts triggered by disruption of epigenetic regulation allow the formation of new insertions with high frequency if the population have the skill to spread in a big number of subpopulations; this will increase the probability of fixation of new insertion and eventually of new families of transposable elements by genetic drift. The most intriguing aspect is the possibility that the alternation between TE bursts and long periods of stability could be related with the dynamics of punctuated equilibria (Eldredge and Gould 1977). Selecting three species belonging to the three order of Mammals under biggest radiation, it is possible to notice that the transposition bursts currently in progress are driven by three different superfamilies: i) in *Homo sapiens* the biggest TE burst is mediated by Alu family, a SINE retrotransposon; ii)) in *Mus musculus* by L1, a non-LTR retroposon; iii) in *Myotis lucifugus* by hAT, an autonomous DNA transposon (Jurka et al. 2011). This demonstrates that the relationship between speciation and TE activity is not linked to a specific superfamily but mutations caused by different replicative mechanisms have similar impact on genomes. From Linnaeus to the present day, a number of different species definitions have been put forward. At the beginning only the morphological data were considered. Afterwards, other aspects were taken into consideration. For example, Theodosius Dobzhansky tried to make the correlation an ecological niche: a species. Then Ernst Mayr and the same Dobzhansky gave the decisive impulse by launching the biological species concept: “Species is that stage of the evolutionary process at which the once actually or potentially interbreeding array of forms becomes segregated in two or more separate arrays of forms which are physiologically incapable of interbreeding“ (Dobzansky 1937) or “Species are groups of actually or potentially interbreeding natural populations, which are reproductively isolated from other such groups” (Mayr 1942). We must consider that in the animal kingdom exceptions have been discovered so that a universal model of speciation cannot be postulated. For example non-canonical reproductive strategies allowed in certain taxa saltationist speciation events. In stick insects of the *Bacillus* genus (Insecta Phasmida) 1.06-0.53 mya ago the hybridization between

Bacillus rossius and *Bacillus grandii* occurred (Mantovani et al. 2001); this event was followed by the onset of apomictic parthenogenesis that led to the birth of the diploid hybrid *Bacillus whitei* (Bullini and Nascetti 1990). In this case, the genomes of the two parental species should have been too much divergent to allow meiosis and would be defined as reproductively isolated. Indeed what was initially a "hopeful monster" has become a new species. In the light of those kind of exceptions, future steps will try to define the most common mechanism of speciation. Studies on comparative genomics reveal that TEs facilitate the rearrangement of chromosomes because of their involvement in unequal crossover, duplications, deletions, translocations, and inversions. As observed in the comparison of the sister clades *Homo sapiens* and *Pan troglodites*, the 44% of inversions occurred from the separation of the two lineages are retrotransposon recombination-mediated by L1 and Alu families (Lee et al. 2008). In addition to promoting the emergence of new phenotypes, in the future it will be interesting to see whether the genomic disorder caused by these events is a factor promoting incompatibility among genomes of populations separated from a certain time.We know that in some cases different phenotypes are not sufficient to have a complete post zygotic isolation. The north american sister species *Drosophyla arizonae* and *Drosophyla mojawensis* are one of the most exploited model for the study of speciation. When *Drosophyla arizonae* mothers cross with *mojavensis*, the males are sterile, while in a cross between female *mojavensis* and male *arizonae* all hybrids are fertile (Bono and Markow 2009). In birds of the order Passeriformes we have several examples of accidental crosses between species that give fertile progeny. In the Galapagos islands hybridization events with relative gene flow between Darwin finch of the genus *Geospiza* were observed (Grant and Grant 2010). Considering the reproductive isolation as a necessary and sufficient condition for the formation of a new biological species, it will be necessary to understand whether transposable elements are involved in the mechanisms that allow to reach reproductive isolation. In *Drosophila melanogaster* the hybridization between laboratory strains lacking transposable elements of P and I families and wild-caught strains where these elements are present cause a breakdown of their regulation resulting in the production of sterile hybrids when the TEs are

transmitted by males (Brennecke et al. 2008). What we can guess with the data available is: the lower the density of Transposable Elements recently inserted, the more a genome will be cold and will be much less prone to new speciation events. Transposable Elements under this view become a weapon of defense, promoting the formation of new species to favour the conditions for an increase in biodiversity and reducing the probability of extinction of a taxon. The continuous production of biodiversity provides new possible adaptations in response to changes of the environment, otherwise the tendency to stasis should led to the extintion of taxa.

Chapter 4

CONCLUSIONS

Transposable elements are the biggest source of mutation in eukaryotic genomes. Like all kinds of mutation their activity produce two possible effects: the deleterious mutations reduce the fitness and are negatively selected while the non deleterious may be fixed in a population and potentially arise advantageous phenotypes. In this thesis I face both aspects. It is widely thought that one of the reasons of the wide diffusion of bisexual reproduction in animals is that it should allow a easier management of deleterious mutations through purifying selection mediated by mechanisms as meiotic recombination and gene conversion. The comparison of TE dynamics in bisexual and unisexual taxa try to add efforts to this theory. Comparing genomes of the three parental species of genus *Bacillus* I added data belonging to a new model of study presenting both bisexual and unisexual reproduction. I found that the repression of TE activity observed in other taxa (Schaack et al. 2010; Arkipova and Meselson 2000) is not ever confirmed. The obligately parthenogenetic *Bacillus atticus* seems to have a content of transposable elements similar to the gonochoric *Bacillus rossius* and *Bacillus grandii.* This datum make the *Bacillus* model an exception to the rule maybe because of a recent origin of the unisexual taxon as observed in the wasp *Leptopilina clavipes* (Kraaijeveld et al. 2012). It is probable that all the defense systems evolved to control the TE activity as Pi RNAs and transcriptional repression mediated by methylation of the genome need a certain time to have a efficient control of the whole mobilome. In the second part of the thesis I focused on the relationship among transposable elements and speciation. The old idea that Transposable Elements are mere parasites (Orgel and Crick 1980) is actually overcomed. The

aspects of evolution in which TE are involved is enough to state that TE are not only the biggest source of mutations of genomes but also the biggest source of evolutionary innovation (Jurka 2012). This awareness brings the evolutionary genomics to ask if TEs are involved in the mechanisms of speciation. In this work starting from advances obtained by previous Authors (Zhe et al. 2009; Rebollo at al. 2010; Jurka et al. 2011; Oliver and Greene 2011,2012), I tried to find a method to evaluate the rate of speciation considering age and number of species of a certain taxon and comparing it with the other. This system gave me a Relative Rate of Speciation by which I observed that in the class Mammalia, the activity of TE is bigger in taxa with bigger rate of speciation. Despite the exceptions detected, the statistical data validation let me argue that an high activity of transposable elements coincide with high rate of speciation. The preliminary analysis of the TE content in non mammalian taxa seems to confirm the relation established in mammals. The total landscape showed by the evolutionary genomics let me suppose that, in the long term, advantages brought by active Transposable Elements are bigger than the harmful effects. My study contributes to better understand the genome and its functioning. It is no longer a rigid memory support but an ecosystem composed of numerous factors in motion. In particular way, eukaryotic genomes should be considered as the result of a symbiosis between genes and Transposable Elements.

References

Arkhipova, I., and Meselson, M., 2000. Transposable elements in sexual and ancient asexual taxa. Proc. Natl. Acad. Sci. U.S.A. 97(26): 14473–14477.

Arkhipova, I., and Meselson, M., 2005. Deleterious transposable elements and the extinction of asexuals. Bioessays. 27(1): 76–85.

Bailey, J.A., et al., 2003. An Alu Transposition Model for the Origin and Expansion of Human Segmental Duplications. The american journal of human genetics. 73(4): 823-834.

Belancio, V.P., et al., 2009. LINE dancing in the human genome: transposable elements and disease. Genome med. 1(10): 97.

Bolisetty, M., et al., 2012. Unexpected diversity and expression of avian endogenous retroviruses. mBio. 3(5): e00344-12.

Bono, J.M., and Markow, T.A., 2009. Post-zygotic isolation in cactophilic Drosophila: larval viability and adult life-history traits of D. mojavensis ⁄ D. arizonae hybrids. J. Evol. Biol. 22(7): 1387–1395.

Brennecke, J., et al., 2008. An Epigenetic Role for Maternally Inherited piRNAs in Transposon Silencing. Science. 322(5906): 1387–1392.

Brookfield, J.F.Y., and Badge, R.M., 1997. Population genetics models of transposable elements. Genetica. 100(1-3): 281–294.

Bull, J.J., et al., 1991. Selection of benevolence in a host-parasite system. Evolution. 45(4): 875-882.

Bullini, L., and Nascetti, G., 1990. Speciation by hybridization in phasmids and other insects. Canadian Journal of Zoology. 68(8): 1747–1760.

Cesari, M., et al., 2003. Polymerase chain reaction amplification of the *Bag320* satellite family reveals the ancestral library and past gene conversion events in *Bacillus rossius* (Insecta Phasmatodea). Gene. 312: 289-295.

Charlesworth, B., et al., 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature. 371(6494): 215-220.

Coates, B.S., et al., 2009. Repetitive genome elements in a European corn borer, *Ostrinia nubilalis*, bacterial artificial chromosome library were indicated by bacterial artificial chromosome end sequencing and development of sequence tag site markers: implications for lepidopteran genomic research. Genome. 52(1): 57-67.

Coates, B.S., et al., 2011. A novel class of miniature inverted repeat transposable elements (MITEs) that contain hitchhiking $(GTCY)_n$ microsatellites. Insect. Mol. Biol. 20(1): 15–27.

Coates, M.I., 2009. Palaentology: Beyond the Age of Fishes. Nature. 458: 413-414.

Dobzansky, T., 1937. Genetics and the origin of species. Columbia Univ. Press, New York.

Dolgin, E.S., and Charlesworth B., 2006. The fate of transposable elements in asexual populations. Genetics. 174(2): 817–827.

El-Murr, N., et al., 2012. *MiRNA* genes constitute new targets for microsatellite instability in colorectal cancer. PLoS One. 7(2): e31862.

Feschotte, C., and Pritham, E., 2007. DNA transposons and the evolution of eukaryotic genomes. Annu. Rev. Genet. 41: 331–368.

Feschotte, C., et al., 2002. Miniature inverted-repeat transposable elements (MITEs) and their relationship with established DNA transposons. *In* Mobile DNA II. *Edited by* N. Craig, R. Craigie, M. Gellert and A. Lambowitz. A.S.M. Press. Washington D.C. pp. 1147-1158.

Fine, P.E.M., 1975. Vectors and vertical transmission: an epidemiologic perspective. Ann. NY Acad. Sci. 266: 173–194.

Fugmann, S.D., 2010. The origins of the Rag genes—From transposition to V(D)J recombination. Seminars in Immunology. 22(1): 10–16.

Galindo, C.L., et al., 2011. Sporadic breast cancer patients' germline DNA exhibit an AT rich microsatellite signature genes chromosomes. Cancer. 50(4): 275–283.

Gould, S.J., and Eldredge, N., 1977. Punctuated equilibria: the tempo and mode of evolution reconsidered. Paleobiology. 3(2): 115-151.

Grant, P.R., Grant, B.R., 2010. Conspecific versus heterospecific gene exchange between populations of Darwin's finches. Phil. Trans. R. Soc. Lond. B. Biol. Sci. 365(1543): 1065–1076.

Harmit, S., et al., 1999.The Age and Evolution of Non-LTR Retrotransposable Elements. Mol. Biol. Evol. 16(6): 793–805.

Joyce, D.A., et al., 2005. An extant cichlid fish radiation emerged in an extinct Pleistocene lake. Nature. 435(7038): 90-5.

Jurka, J., et al., 2011. Families of transposable elements, population structure and the origin of species. Biology direct. 6:44.

Kazazian, H.H., 2004. Mobile elements: drivers of genome evolution. Science. 303(5664): 1626–1632.

Kidwell, M.G., 2002. Transposable elements and the evolution of genome size in eukaryotes. Genetica. 115(1): 49-63.

Kohany, O., et al., 2006. Annotation, submission and screening of repetitive elements in Repbase: Repbase Submitter and Censor. BMC Bioinformatics. 25(7): 474.

Kraaijeveld, K., et al., 2012. Transposon proliferation in an asexual parasitoid. Mol. Ecol. 21(16): 3898–3906.

Kramerov, D.A., and Vassetzky, N.S., 2011. Origin and evolution of SINEs in eukaryotic genomes. Heredity. 107(6): 487–495.

Lee, J., et al., 2008. Chromosomal Inversions between Human and Chimpanzee Lineages Caused by Retrotransposons. PLoS ONE. 3(12): e4047.

Leese, F., et al., 2012. Exploring Pandora's box: potential and pitfalls of low coverage genome surveys for evolutionary biology. PLoS One. 7(11): e49202.

Luchetti, A., et al., 2003. Unisexuality and molecular drive: *Bag320* sequence diversity in *Bacillus* taxa (Insecta Phasmatodea). J. Mol. Evol. 56(5): 587-596.

Luchetti, A., and Mantovani, B., 2009. *Talua* SINE biology in the genome of the *Reticulitermes* subterranean termites (Isoptera, Rhinotermitidae). J. Mol. Evol. 69(6): 589–600.

Luchetti, A., and Mantovani, B., 2011. Molecular characterization, genomic distribution and evolutionary dynamics of Short INterspersed Elements in the termite genome. Mol. Genet. Genomics. 285(2): 175-184.

Makałowski, W., et al., 2012. Transposable elements and their identification. Methods Mol. Biol. 855: 337-359.

Makarieva A. M. and Gorshkov V. G., 2004. On the dependence of speciation rates on species abundance and characteristic population size. J. Biosci. 29: 119–128.

Mantovani, B., and Scali, V., 1992. Hybridogenesis and androgenesis in the stick insect Bacillus rossius–grandii benazzii (Insecta, Phasmatodea). Evolution. 46(3): 783–96.

Mantovani, B., et al., 1997. The *Bag320* satellite DNA family in *Bacillus* stick insects (Phasmatodea): different rates of molecular evolution of highly repetitive DNA in bisexual and parthenogenetic taxa. Mol. Biol. Evol. 14(12): 1197-1205.

Mantovani, B., et al., 1999. Genomic evolution in parental and hybrid taxa of the genus Bacillus (Insecta, Phasmatodea). It. J. Zool. 66(3): 265– 72.

Mantovani, B., et al., 2001. The mitochondrial cytochrome oxidase II gene in *Bacillus* stick insects: ancestry of hybrids, androgenesis, and phylogenetic relationships. Mol. Phylogenet. Evol. 19(1): 157-163.

Mayer, C., 2006 - 2010. Phobos 3.3.11. Available from http://worldwidewords.rub.de/spezoo/cm/cm_phobos.htm [accessed 22 April 2013].

Mayr, E., 1942. Systematics and the origin of species. Columbia Univ. Press, New York.

McClintock, B., 1950a. The origin and behavior of mutable loci in maize. Proceedings of the National Academy of Sciences. 36(6): 344-355.

Meglécz, E., et al., 2007. Microsatellite flanking region similarities among different loci within insect species. Insect. Mol. Biol. 16(2): 175-185.

Meredith, R.W., et al., 2011. Impacts of the Cretaceous Terrestrial Revolution and KPg Extinction on Mammal Diversification. Science. 334(6055): 521-524.

Mogil, L.S., et al., 2012. Computational and experimental analyses of retrotransposon-associated minisatellite DNAs in the soybean genome. BMC Bioinformatics. 13(suppl. 2): S13.

Moran, J.V., et al., 1999. Exon Shuffling by L1 Retrotransposition. Science. 283(5407): 1530-1534.

Muotri, A.R., et al., 2007. The necessary junk: new functions for transposable elements. Human Molecular Genetics. 16(2): 159-167.

Nishihara, H., and Okada, N., 2008. Retroposons: genetic footprints on the evolutionary paths of life. Methods Mol. Biol. 422: 201–225.

O'Grady, P.M., et al., 2011. Phylogenetic and ecological relationships of the Hawaiian Drosophila inferred by mitochondrial DNA analysis. Molecular Phylogenetics and Evolution. 58 (2): 244–256.

Oliver, K.R., and Greene, W.K., 2011. Mobile DNA and the TE-Thrust hypothesis: supporting evidence from the primates. Mobile DNA. 2: 8.

Oliver, K., Greene, W., 2012. Transposable elements and viruses as factors in adaptation and evolution: an expansion and strengthening of the TE-Thrust hypothesis. Ecology and Evolution. 2, 11: 2912 - 2933.

Orgel, L.E., and Crick, F.H., 1980. Selfish DNA: the ultimate parasite. Nature. 284(5757): 604–607.

Pace, J.K., Feschotte, C., 2007. The evolutionary history of human DNA transposons: Evidence for intense activity in the primate lineage. Genome research. 17: 422-432.

Plohl, M., et al., 2008. Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. Gene. 409(1-2): 72–82.

Ponicsan, S.L., et al., 2010. Genomic gems: SINE RNAs regulate mRNA production. Curr. Opin. Genet. Dev. 20(2): 149–155.

Pritham, E.J., et al., 2007. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene. 390(1-2): 3-17.

Rasmussen, D.A., and Noor, M.A.F., 2009. What can you do with 0.1x genome coverage? A case study based on a genome survey of the scuttle fly *Megaselia scalaris* (Phoridae). BMC Genomics. 10: 382.

Ray, D.A., et al., 2008. Multiple waves of recent DNA transposon activity in the bat, Myotis lucifugus. Genome research. 18: 717-728.

Rebollo, R., et al., 2010. Jumping genes and epigenetics: Towards new species. Gene. 454(1-2): 1-7.

Richard, G.F., et al., 2008. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol. Mol. Biol. Rev. 72(4): 686-727.

Scali, V., et al., 2003. Linkage between sexual and asexual lineages: genome evolution in *Bacillus* stick insects. Biol. J. Linn. Soc. 79: 137–150.

Schaack, S., et al., 2010. DNA transposon dynamics in populations of Daphnia pulex with and without sex. Proc. Biol. Sci. 277(1692): 2381-2387.

Schmitz, J., and Brosius, J., 2011. Exonization of transposed elements: A challenge and opportunity for evolution. Biochimie. 93(11): 1928-1934.

Sharma, A., et al., 2013. Tandem repeats derived from centromeric retrotransposons. BMC Genomics. 14: 142.

Singer , T., et al., 2010. LINE-1 retrotransposons: Mediators of somatic variation in neuronal genomes?.

Trends Neurosci. 33: 345 –354.

Smit, A.F.A., et al., 1998. RepeatMasker Open–3.0. 1996–2010. <http://www.repeatmasker.org/> .

Sullender, B.W., and Crease, T.J., 2001. The behavior of a *Daphnia pulex* transposable element in cyclically and obligately parthenogenetic populations. J. Mol. Evol. 53(1): 63-69.

Tamura, K., et al., 2011. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance and maximum parsimony methods. Mol. Biol. Evol. 28: 2731-2739.

Thomas, C.A., Muotri, A.R, 2012. LINE-1: creators of neuronal diversity. Front Biosci

(Eliteed). 4: 1663-8.

Wicker, T., et al., 2007. A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. 8(12): 973-982.

Wright, S.I., and Schoen, D.J., 1999. Transposon dynamics and the breeding system. Genetica. 107(1-3): 139–148.

Wright, S., and Finnegan, D., 2001. Genome evolution: sex and transposable elements. Curr. Biol. 11: R296-R299.

Zeh, D.W., et al., 2009. Transposable elements and an epigenetic basis for punctuated equilibria. Bioessays. 31(7): 715-726.