Comparative genomics of transposable element evolution and their evolutionary impacts in fish and other vertebrate genomes

Domitille Chalopin

THÈSE

en vue de l'obtention du grade de

Docteur de l'Université de Lyon, délivré par l'École Normale Supérieure de Lyon

Science de la Vie

Institut de Génomique Fonctionnelle de Lyon

École Doctorale Biologie Moléculaire, Intégrative et cellulaire

présentée et soutenue publiquement le 23 Mai 2014

par Madame Domitille Chalopin

Comparative genomics of transposable element evolution and their evolutionary impacts in fish and other vertebrate genomes

---

Directeur de thèse : Monsieur Jean-Nicolas Volff

Devant la commission d'examen formée de :

Monsieur Cedric FESCHOTTE, University of Utah, Rapporteur

Madame Delphine GALIANA, ENS de Lyon, Co-encadrante

Monsieur Olivier PANAUD, Université de Perpignan, Rapporteur

Monsieur Manfred SCHARTL, Universität Würzburg, Examinateur

Madame Cristina VIEIRA, Université de Lyon, Examinateur

Monsieur Jean-Nicolas VOLFF, ENS de Lyon, Directeur de thèse

... from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved.

- Charles Darwin

ACKNOWLEDGEMENTS - REMERCIEMENTS - DANKSAGUNG

- Before anything else, I would like to greatly acknowledge the many kind people who have been by my side throughout this process. Without their ideas, help, and support this thesis would not have been possible -

First of all, I would like to express my sincere gratitude to my supervisor, Jean-Nicolas Volff, who believed in me at the end of my bachelor, gave me the opportunity to realize my PhD project in his team, and without whom I would probably not be facing you today for the defense. I am very grateful for his advice, discussions, encouragements, rigour and humour.

I would like to address a particular acknowledgement to Delphine who has always been present during my thesis. I kindly thank her for her endless encouragement, for our long scientific discussions about crazy hypotheses, for her support during difficult periods, for her transmissible laughter and for staying by my side during all kinds of stressful moments (paper submission, talk trainings, manuscript prints…).

I would like to thank members of my thesis committee for taking the time to read and judge my work and for having travelled long distances: Cristina Vieira, Manfred Schartl, and particularly my two reviewers Cédric Feschotte and Olivier Panaud.

A very special thanks goes out to Manfred Schartl, without whom I would have never been able to participate in so many international projects. I would also like to address my sincere acknowledgement to the person who was like a mentor and was always present for advice and new scientific interactions: Ein grosses Dankeschön.

I would like to address my kindest thanks to the members of the monitoring committee for thesis: Marie Sémon and Emmanuelle Lerat, for their helpful comments, their interactive discussions and their kind support.

I would like to express many thanks to all my collaborators from the different consortium for giving me great opportunities to participate in such amazing projects: Yann Guiguen and Hugues Roest-Crollius for the rainbow trout, as well as Carine Genet for our interactive exchanges; Manfred Schartl for the platyfish, coelacanth and Amazon molly projects; Ron Walter and his team, Tzuni and Yingjia, who kindly hosted me in Texas for the platyfish project; John Postlethwait and Ingo Braasch for the spotted gar project; Jessica Alföldi and Chris Amemiya for the coelacanth project; Wes Warren for the platyfish and the cave fish projects, as well as Suzanne McCaugh for active interactions concerning the cave fish annotation; Patrick Lemaire for the multi-Urochordate project; Christophe Terzian for the helpful comments regarding foamy

virus study; Elisabetta Colluccia for a mapping project on eel chromosomes; and Mariko Forconi for our interesting interactive coelacanth project.

I am also grateful to current and former lab members: Magali for her bioinformatic help and advice, but also for her kind support and for the moments we shared in Montpellier and Lyon; Fred who helped me to obtain a Master internship in Chicago; Emilien for his evolutionary discussions and musical exchanges; Perrine and Catherine for their helpful contributions to molecular domestication experiments; Astrid and Tina with whom I spent a lot of time during my masters; but also Amandine, Bernard, Marta and Luis-Antony.

I have many people from the IGFL to thank for sharing all these years, for cheerfulness and for the multiple interactive discussions (scientific or not). I am thinking about Cyrielle, a colleague but also a friend and my neighbour: a lot of thanks for her support. Special thanks to Eric, Cyril, Romain, Sandrine, Benjamin and Laure for our burning evening and peaceful moments we shared (couscous, poulet coco, Eric's thesis). Thanks to the people who shared the LR4 basement with us, Violaine, Julie and Juliette. Thanks to Karine and Fred for the climbing sessions during lunch. Thanks to Pauline for her eternal cheerfulness and support. I will not forget Juliana, Mélisandre, Sandrine, Thomas, Guillaume, Laurent and all the nice people with whom I took my lunch or shared a coffee.

I have to notably thank people from the different IT supports, Thomas Bellembois from the IGFL, Emmanuel Quemener from CBP and Hervé Gilquin from PSMN. They have all been very patient and helpful in starting and rotating optimally repeat software.

I must also acknowledge administrative assistance, in particular Fabienne Rogowski, Martine Chapier, Sonia Celard and Joanne Burden for their efficiency and kindness.

I learned a lot during my monitoring session and thus I am grateful to the teaching department of the ENS, which gave me the opportunity to participate in genetic and bioinformatic practical courses. My thanks go to Nathalie Alazard, Stéphane Vincent, Déborah Prévot and Patrick Ravel-Chapuis.

Two particular thanks are dedicated to financial support and grants from "Ministère de la Recherche et de l'Enseignement Supérieur" and "Ligue contre le cancer". I have to recognize that this work would not have been possible without them.

My adventure at the ENS started during my bachelors. Since then, my scientific way was made with a few people that I particularly thank: Chloé for her support and help to catch the theoretical delay I had, but also for our funny and studious weekend, and other numerous memorable moments; Floriane for her solid friendship, support, wine evening and statistical help; Yann for his cheerfulness and shark defense.

I also would like to gratefully thank the people who accompanied and supported me during these last years, outside the laboratory. These people allowed me to maintain a fundamental equilibrium. Thanks to my childhood and sport friends, Lila and Emmanuelle, for a girly weekend and vacation as well as their unwavering support. Thanks to Marine for friendly support and climbing evenings. Also special thanks to Stéphanie, Nelly, Thom, Stéphane, Morgane for the nice moments we shared.

I have no words to express all the necessary acknowledgements to my parents, my little brother and sister, and to say how grateful I am for their endless love and support. Thank you for your presence at each step of my life and your unwavering trust.

Last but not least, I would like to express my lovely thanks to my husband, Jeff, who supported me during my thesis, day after day, both in difficult and happy moments. His daily presence encouraged me to always go further.

- Warm thanks to all whose, who helped me build myself and helped me to have the strength to do this work -

TABLE OF CONTENTS

Summary - Résumé 1

Chapter 1: INTRODUCTION 3

I- From genetics to genomics: toward the understanding of genome structure and evolution 3

a. A little bit of history: molecular biology and transposable element discoveries 3

b. Architecture, size and composition of genomes 9

c. The mystery of transposable elements 10

II- Classification of Transposable elements in Eukaryotes 12

a. Class I Retrotransposons 13

b. Class II DNA transposons 21

c. Other DNA transposons 25

III- The evolutionary impact of transposable elements on host genomes 26

a. Genomic rearrangements induced by TEs and phenotype-associated modifications 26

i. Impact on gene function and/or regulation by TE-insertion

ii. Large-scale rearrangements

iii. Source of new genetic materials

b. Genome size evolution 35

c. Speciation 36

IV- Activity and success of transposable elements in host genomes 37

a. Life cycle of TEs 37

b. Horizontal transfer events 39

c. Detecting an active TE 41

V- TE maintenance versus host defense 42

VI- TE-based genetic tools in vertebrates 45

VII- TEs in vertebrate genomes 46

a. TE diversity 46

b. TE-derived sequences in vertebrate genomes 48

VIII- Aim of the thesis 52

Chapter 2: ANNOTATION OF TEs IN SEQUENCED FISH GENOMES **53**

I- Methods and presentation of the projects **55**

a. Presentation of the fish models 55

b. Methods to annotate and analyze transposable elements 57

c. Presentation of the different fish genome projects 59

II- Poeciliid genome projects **64**

III- The coelacanth project **68**

a. Genomic analyses of TEs in the genome of the African coelacanth, *Latimeria chalumnae* 68

i. TE content and diversity in coelacanth to study TE evolution in vertebrate genomes 68

ii. Comparative analysis with a second genome assembly of the African coelacanth 70

b. Transcriptomic TE analyses in the two coelacanth species, *Latimeria chalumnae* and *Latimeria menadoensis* 71

c. Detection of TE insertion polymorphisms in the two coelacanth species, *L. chalumnae* and *L. menadoensis* 72

Chapter 3: LARGE-SCALE COMPARATIVE GENOMIC ANALYSIS OF TE CONTENT AND EVOLUTION IN VERTEBRATE GENOMES **111**

I- Original article: Comparative analysis of transposable elements in fish highlights mobilome diversity and evolution in vertebrates **113**

II- Further results in addition to the comparative genomic analysis **155**

a. Inclusion of the non-publicly available fish genomic data into the comparative analysis 155

b. Endogenous retrovirus diversity: reflect of past infection 157

Chapter 4: GOLD TEs, MOLECULAR DOMESTICATION IN VERTEBRATES **163**

I- Integrase-derived genes: the case of Gin genes **165**

a. *In silico* analyses of *Gin* genes 166

b. Expression analyses of *Gin-2* 168

Chapter 5: DISCUSSION

JUNK OR NOT JUNK, THIS IS THE QUESTION **171**

I- Comparative genomic analyses of TEs in vertebrate genomes **173**

II- Dynamics of TEs in asexual-species genome: the case of the Amazon molly **179**

III- Evolution of TEs in "slow-evolving" genomes? **182**

IV- History and function of *Gin* genes **186**

V- Molecular domestication in vertebrate genomes: is this phenomenon rare? **190**

Bibliography **193**

Abbreviations **215**

List of Figures **217**

List of publications **219**

Annexes **221**

SUMMARY

Transposable elements (TEs) are mobile genetic elements - able to move and to multiply within genomes - identified in almost all living organisms including bacteria. Considered as junk DNA for long, nowadays they are undeniably major players of gene, genome and host evolution. TEs can be deleterious causing diseases at the individual scale, but they are strong evolutionary agents involved in genome plasticity at a larger scale. Furthermore, these "parasites" can also be source of new genetic materials as promoters or even new genes bringing new functions for hosts. The main objectives of my thesis was to determine the presence or not of the different TE families in fish genomes, as well as their respective content to understand the evolutionary history of these families in fish genomes compared to other vertebrate genomes. Being involved in various genome sequencing consortium projects (including the coelacanth one), I performed a large-scale comparative analysis to highlight the various evolutionary strategies of TEs. I showed that TE content is highly variable in vertebrate genomes, the smallest and the largest being found in fish (6% in tetraodon and 55% in zebrafish), and may contribute to their genome sizes especially in fish that present more TEs than compared sarcopterygian genomes of same sizes. Analyses of TE superfamilies also present a large variation between species. While some superfamilies are widespread (ERV, Penelope, L1, CR1-like retrotransposons, Tc-Mariner, hAT transposons), some have been completely lost in particular lineages, such as CR1 in teleost fish, and others present very patchy distribution, as Merlin, also suggesting that several events of horizontal transfer occurred. These superfamilies underwent differential waves of activity in vertebrate species highlighting TE dynamics and predicting the potential current activity in each genome. On another hand, I particularly focused on the study of a vertebrate-specific TE-derived gene, named *Gin-2*, to understand its origin, evolution, and its potential function in vertebrates that is completely unknown. *In silico* analyses showed that *Gin-2* is a very ancient gene (around 450 My) deriving from a GIN transposon and is absent from placental mammals. Further analyses present a particular expression in brain and gonads during adulthood, while a strong expression during gastrulation suggests a potential role of *Gin-2* in zebrafish development. All together, the different analyses performed during my thesis contribute to a better view of TE evolution and their evolutionary impacts in vertebrate genomes (content, diversity, dynamics, activity, exaptation, functions…).

RESUME

Les éléments transposables (ETs) sont des éléments génétiques mobiles capables de se déplacer et de se multiplier au sein d'un génome. Identifiés dans la plupart des espèces vivantes incluant les bactéries, mais longtemps considérés comme de l'ADN poubelle, les ETs sont aujourd'hui indéniablement des acteurs majeurs impliqués dans l'évolution des gènes, des génomes et des organismes. Si à l'échelle des individus les ETs peuvent avoir des effets délétères pouvant entrainer des maladies, à plus grande échelle ils sont de puissants agents évolutifs impliqués dans la plasticité génomique. Ces « parasites » peuvent également être sources de nouveaux éléments génétiques comme des promoteurs ou même de nouveaux gènes avec de nouvelles fonctions pour l'hôte. Les objectifs majeurs de mon travail de thèse ont été de déterminer les différentes familles d'ETs présentes dans les génomes de poissons, la part que chacune d'entre elles occupe dans ces génomes et enfin de comprendre l'histoire évolutive de ces familles d'ETs dans les génomes de vertébrés. Cette comparaison à grande échelle permet de comprendre les différentes stratégies évolutives des ETs. Ainsi, on peut observer que les génomes de vertébrés présentent des contenus en ETs très variables, en terme de quantité (de 6% pour le tetraodon à 55% pour le poisson zèbre), jouant également sur l'évolution de la taille des génomes, et en terme de diversité mettant en évidence la présence de familles d'ETs dans l'ensemble des lignées ou bien des familles spécifiquement perdues ou gagnées. D'autre part, j'ai particulièrement étudié un gène de vertébrés dérivé d'ETs, *Gin-2*, dans le but de comprendre son origine et son évolution, ainsi que d'émettre des hypothèses quant à sa fonction moléculaire potentielle qui est encore inconnue. Pour cela, des études *in silico* ont permis de mieux comprendre son origine, un transposon GIN. Des analyses d'expression suggèrent que *Gin-2* pourrait jouer un rôle lors du développement embryonnaire chez les poissons. Ces travaux contribuent de manière générale à une meilleure compréhension de l'évolution des ETs, ainsi que leurs impacts évolutifs, dans les génomes de vertébrés (quantité, diversité, dynamique, activité, fonctions,…). Par ailleurs, l'ensemble des résultats obtenus quant à la diversité des ETs peut à présent servir de matrice, pouvant être continuellement complétée avec l'apport de nouveaux génomes séquencés.

CHAPTER 1 : INTRODUCTION

I- From genetics to genomics: toward the understanding of genome structure and evolution

a. A little bit of history: molecular biology and transposable elements discoveries

Molecular biology is a branch of the biology field, which tends to understand the interactions between the different fundamental molecules of life - namely DNA, RNA and proteins - and how these interactions are regulated. It overlaps the important disciplines of genetics and biochemistry. Even under perpetual evolution, the central dogma (CD) where the DNA is transcribed into RNA, which is then translated into proteins (Initial CD in Figure 1) is a good starting point to understand this field. It took more than 90 years between the beginnings of genetics with Mendel's works (1865) to the establishment of the central dogma by Watson & Gamov (1956) (Figure 2).

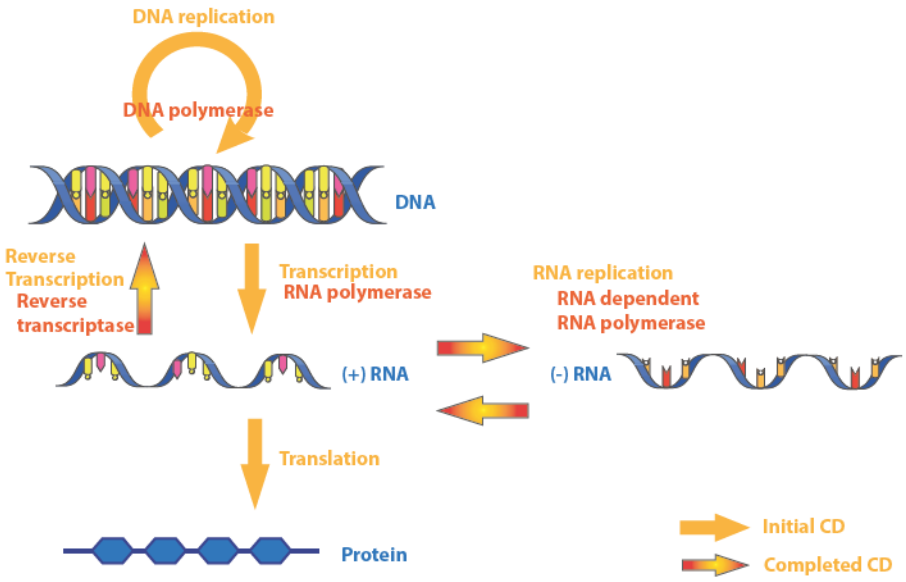


Figure 1: The central dogma established by Watson and Gamov and revisited after the discovery of the reverse trancriptase protein. The various orange and degraded orange arrows follow the central dogma (CD) process, as initially established and aslater completed by the reverse transcription of RNA molecules into DNA molecules, as well as by the RNA replication process. This last one was first discovered in plants.

Molecular biology has been marked by successive discoveries, hypotheses and technical improvements. First of all, in the 19th century Darwin introduced the fact that a transmission of characters (genetics) with changes allows adaptation over time, space and environment. At the same time, Mendel showed that "traits" are transmitted from parents to offspring, introducing the important notion of dominant and recessive traits. It is only in early 20th, with the introduction of the "gene" term and with the chromosomal theory of heredity by Sutton that Bateson coined the term "genetics". Since then, Morgan showed that chromosomes carry genes, Muller used X-rays to generate mutants, and Beadle & Tatum proved that genes code for proteins.

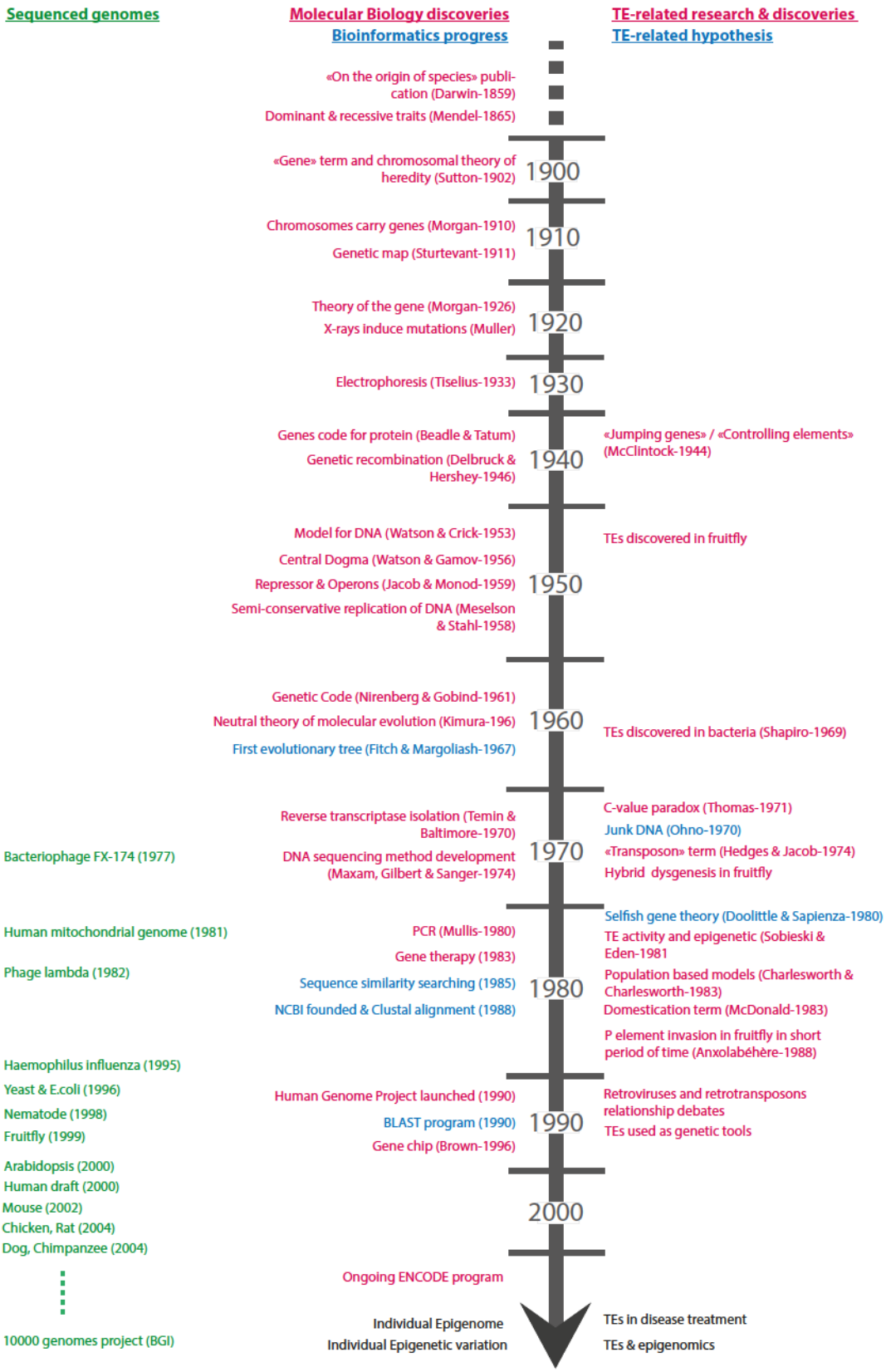


Figure 2: Timeline showing important molecular biology (on the left) and transposable element linked discoveries (on the right side). Sequenced genomes and large sequencing projects are developed on the left side of the timeline (written in green color).

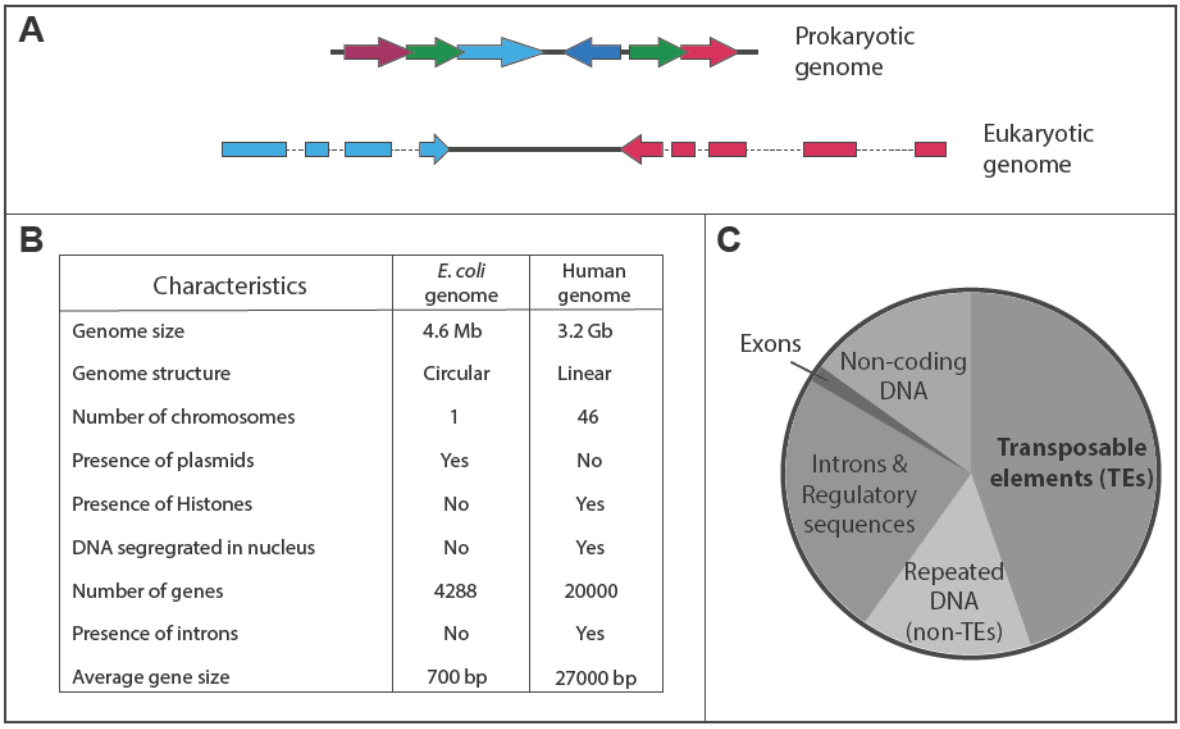
However, fundamental strides were realized when Watson & Cricks discovered the structure of DNA, allowing the establishment of the central dogma (Figure 1) and the cracking of genetic code. In 1964, the central dogma has been challenged when Temin & Baltimore discovered that the direction of DNA to RNA could be reversed using RNA viruses. During the 70's, Maxam, Gilbert & Sanger developed the first sequencing method and the entire genome of a bacteriophage was obtained, marking the beginning of the genomic era. At this time (80's), only small genomes (mitochondria, phages, bacteria) are sequenced, followed by larger one (nematode and fruitfly). The human genome project - representing an important step in term of size, complexity as well as a hallmark for medicine - was launched in 1990. Ten years later the first draft is achieved giving way to start many human projects such as HapMap (catalog of genetic similarities and differences in human beings), dbGaP (Database of genotypes and phenotypes), ENCODE Pilot project and 1000 human genomes, but also other vertebrate genomes. With the high demand for low cost and rapidity, many technologies including several methods of high-throughput (next generation) sequencing, such as pyrosequencing, real-time sequencing or ligation sequencing were developed in the last ten years. The improvement of sequencing was accompanied with the development of bioinformatics tools allowing to store data, to easily analyze them.

In the 40's, Barbara McClintock did another fundamental discovery when she observed colour variation in maize kernels. She showed that "jumping genes" induced this variation, mobile DNA now known as transposable elements (TEs) (Figure 2). Each kernel is an embryo produced from a single fertilization, making the maize an interesting model to study genetics and inheritance of characters. She worked with the now so-called *Ac/Dc* (autonomous "Activator" and non-autonomous "Dissociation") system to demonstrate that "jumping genes" or "controlling elements" were responsible for this colour pattern by breaking specific colour loci. Her work was published in 1951 (McClintock 1951) and she was awarded the Nobel Prize in 1983.
TEs were then discovered in fruitfly (Kidwell *et al.* 1977; Picard *et al.* 1978), bacteria (Shapiro 1969) and human in the 70's. In the 70's, they were mostly considered as junk (Ohno 1972). Susumu Ohno proposed that "junk" DNA, referred to pseudogenes, which does not produce proteins, corresponded to genetic fossils. He stated: "The earth is strewn with fossil remains of extinct species; is it a wonder that our genome too is filled with the remains of extinct genes?". Following the idea that eukaryotic genomes contain a large proportion of non-coding DNA, described as selfish DNA by Dawkins (Dawkins 1976), was independently supported in 1980 with two papers (Doolittle and Sapienza 1980; Orgel and Crick 1980), i.e. they only have detrimental effects on hosts. This large fraction of non-coding DNA is in agreement with the C-value paradox, i.e. the observation that genome size does not correlate with apparent complexity of organisms (Thomas 1971). The interest for TEs continued to grow up with the identification of P and I elements that induce sterility and increase mutation and recombination rate in

Drosophila (PICARD *et al.* 1978; KIDWELL AND NOVY 1979; CROZATIER *et al.* 1988). However, these experiments still supported the fact that TEs are selfish and mainly deleterious for their host.

The number of studies dealing with TEs increased from the 80's, showing the growing interest of the scientific community into a genetic DNA that could play an important role in genome evolution. The first discovery that made scientists think TEs could be of interest was the demonstration that P and I elements have invaded *Drosophila* genomes in less than 50 years (ANXOLABEHERE *et al.* 1988) after a transfer from *D. willistoni* to *D. melanogaster* (DANIELS *et al.* 1990). The short time of these events was very convenient for geneticists to study population genetics of TEs, by measuring the rates of invasion, deletion or maintenance in Drosophila species. Since then, population genetic analyses increased.

Since it was shown that TEs can represent a high proportion of genomes, one of the question appeared: why such an abundant junk and selfish DNA has been conserved in genomes if they are only deleterious and useless for the host? Improvement of technical experiments in the 70's, allowed to assess wider questions concerning the impact of TEs on genome evolution, TE origin, TE amplification mechanisms, host defenses and potential roles of TEs in their hosts. These questions were abundantly raised from the late 80's. Nowadays, it is undeniable that TEs are powerful players of evolution, being both deleterious or inducers of phenotypic changes at short time scales and drivers of genome evolution leading to novelties at larger time scales.



Characteristics	*E. coli* genome	Human genome
Genome size	4.6 Mb	3.2 Gb
Genome structure	Circular	Linear
Number of chromosomes	1	46
Presence of plasmids	Yes	No
Presence of Histones	No	Yes
DNA segregrated in nucleus	No	Yes
Number of genes	4288	20000
Presence of introns	No	Yes
Average gene size	700 bp	27000 bp

Figure 3: Comparison of prokaryotic and eukaryotic genomes. A- Comparison of gene organization in both prokaryotic and eukaryotic genomes. Each colored arrow represents a coding region, interrupted by introns (dashed line inside blue and pink arrows) in eukaryotic genomes B- Comparison of the principal characteristics of *E. coli* and human genomes. C- Human genome composition represented as a camembert.

b. Architecture, size and composition of genomes

The improvements of molecular biology experiments and sequencing methods combined with the accumulation of genomic data from various organisms (viruses, bacteria, archaea and eukaryotes), allowed to better understand the architecture of genomes. Among the different domains of life, two types of genomes can be distinguished (Figure 3).

Prokaryotes, which include archaea and bacteria, have very compact genomes (Figure 3A and B). They are characterized by a relatively small circular genome, which is free in the cytoplasmic compartment (not segregated into a nucleus). In *Escherichia coli*, the number of genes is around 4288 with an average size of 700 base pairs (bp). Long overlaps between genes are possible in prokaryotic genomes and many of them are organized into co-transcribed groups, called operons (but also found ineukaryotes, such as *C. elegans*), in which multiple genes are under the control of a sole promoter. Eukaryotic organisms can be unicellular or multicellular. Except few examples, as yeasts, they both present a genome organization with long intergenic regions. Gene structure is organized in introns-exons (Figure 3A), where introns are removed during splicing, while coding regions (exons) are transcribed to messenger RNA, leading to the protein after translation. Only a small fraction of genes contains introns in unicellular organisms, which generally have more compact genomes than multicellular organism. That is not an exclusive view, but due to the presence of introns and long intergenic regions, eukaryotic genomes are generally much larger than prokaryotic genomes. As comparison, human genome is 3200 Mb long with only 2% corresponding to exons (Figure 3C), while *E. coli* genome is 4.6 Mb, i.e. 100 times smaller (Figure 3B).
Among living organisms, genome sizes vary enormously and are not always related to organism apparent complexity (C-value paradox; Figure 4). In prokaryotes, genome size and number of genes are correlated. Due to the large intergenic regions, the similar correlation is weaker in eukaryotes, where non-coding DNA may play an important role in genome architecture. Indeed, in human, while coding regions (exons) only correspond to ~2% of the genome (Figure 3C), more than 55% of the genome consists in repetitive DNA (TEs and non-TEs). The smallest reported eukaryotic genome belongs to *Encephalitozoon intestinalis*, a parasitic microsporidium (0.0023pg – 2.25 Mb), while the largest one is attributed to *Amoeba dubia* (700pg – 684.6 Gb) (Animal genome size database, http://www.genomesize.com/index.php). Among animals, the smallest corresponds to a plant-parasitic nematode, *Pratylenchus coffeae* (0.02pg – 19.5 Mb) and the largest os the marbled lungfish, *Protopterus aethiopicus* (132.8pg – 129.9 Gb).

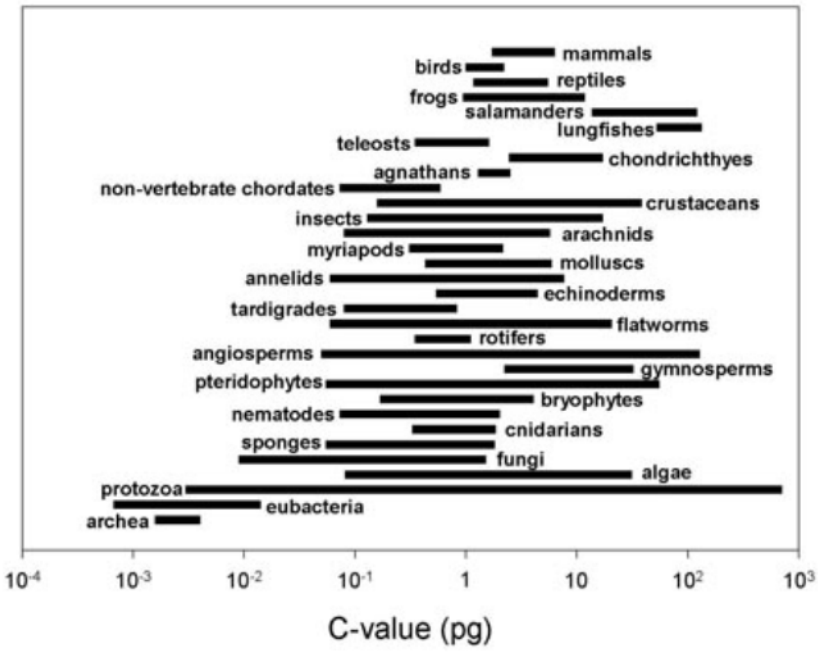


Figure 4: C-value (haploid DNA content) variation among the tree of life including the maximum and minimum reported genome size (in 1pg= 978 Mb). Picture from the « Genome Size Database ». Genome sizes can be determined by flow cytometry or densitometry of the nuclei, and now by complete genome sequencing.

Because sequencing technologies are faster and cheaper, the number of genome sequencing projects strongly increased during the last decade. The 2000's marked the beginning of the genomic era with the sequencing of the human genome and other animal models, such as mouse, rat, fruitfly, or chicken. Nowadays, single genome projects are multiplying but also multi-genome projects such as the 1000 human genomes (GENOMES PROJECT *et al.* 2010). This accumulation of data increases the possibility to understand the eukaryotic genome complexity and to perform large-scale comparative genomics. Beyond the study of genic and regulatory regions, the availability of complete genomes allowed also to analyze non-coding intergenic and intronic sequences, such as transposable elements.

c. The mystery of transposable elements

The purpose of my thesis is to understand how TEs evolved in vertebrate genomes and impacted genome diversity, reorganization and size. As the rest of the introduction fully deals with transposable elements, I only briefly introduce TEs in this subpart.

TEs are repeated mobile genetic elements found in almost all eukaryotic genomes (PRITHAM 2009) and can constitute a large fraction of them. In some plant genomes, they can contribute to more than 90% of all DNA sequences (FLAVELL *et al.* 1974; FLAVELL *et al.* 1994; BENNETZEN 2000; KUMAR AND BENNETZEN 2000), making them a major genomic component. TE composition in genomes can differ not only in term of abundance but also in term of diversity and age of TE families, suggesting that they have differential

invasive success species. Because TEs have the ability to transpose, to mobilize genomic sequences and to recombine, they are important players of gene, chromosome and genome evolution.

Considering these abilities, TEs were long considered as junk, selfish DNA and mutation-inducing elements. From this deleterious point of view, evolutionary forces against transposition events (reducing transposition rate, increasing DNA elimination rate, counter-selection of insertions in coding regions) were described (Charlesworth 1994; Vieira and Biemont 1996). At this stage, no advantage for host genome was considered. However, can such repeated sequences compose such a large fraction of DNA if they are only deleterious? TEs have been found in all genomes from bacteria to vertebrates, except few unicellular small eukaryotic genomes, as the red alga *Cyanidioschyzon merolae*, six apicomplexans (Carlton *et al.* 2002; Gardner *et al.* 2002; Abrahamsen *et al.* 2004; Xu *et al.* 2004; Bishop *et al.* 2005; Brayton *et al.* 2007) and the Unikont *Encephalitozoon cuniculi* (Katinka *et al.* 2001). The reasons why these genomes are TE-free are not elucidated. It has been proposed that their maintenance as TE-free organisms to maintain their small genome size (and thus their small cell size) was crucial for their survival (Pritham 2009).

The recent emergence of data allowed reconsidering the role of TEs in genomes. Indeed, it has been shown that TEs can serve to acquire new sequence information or modify coding and regulatory regions through DNA integration, rearrangements and molecular domestication. Furthermore, it has also been proposed that TEs may play important role in evolutionary transitions and speciation.

II- Classification of transposable elements in Eukaryotes

The most basal classification of TEs is based on their transposition mechanisms, differentiating two classes (FINNEGAN 1989; CAPY 2005; WICKER *et al.* 2007; KAPITONOV AND JURKA 2008). Class I elements, called retrotransposons, transpose via the reverse transcription of an RNA intermediate, which is commonly called "copy-and-paste" mechanism. In contrast, class II elements, called DNA transposons, directly transpose from DNA to DNA similar to a "cut-and-paste" procedure for the majority. At a lower level, based on Wicker's classification (WICKER *et al.* 2007), classes are divided into subclasses, orders, superfamilies and families (Figure 5).

As mentioned before, classes are defined depending on the presence or not of an RNA intermediate during the transposition mechanism. Classes are further divided into subclasses, orders and superfamilies depending on structure, organization and transposition mechanism. Class I, is subdivided into two subclasses, LTR and non-LTR retrotransposons. In class II, subclasses 1 and 2 are defined depending on the number of DNA strands that are cut at the TE donor site (genomic location before transposition). Orders are defined according to differences in structure, insertion mechanism and enzymology, thus separating LTR, DIRS, LINE, SINE and Penelope orders in class I and TIRs, Crypton, Polinton and Helitron in class II. Superfamilies have the same transposition strategy among an order but are characterized by the presence or not of target site duplication (TSDs, small direct repeat formed upon insertion) that can also present a specific length. They are phylogenetically separated, and they do not share, or very few, similarities at the nucleotide level. Finally, the last classification level is composed of families that are highly conserved at the protein level and share strong similarities in restricted conserved domain at the nucleotide level. At a last level, different elements can be distinguished if they do not share more than 80% of identity over 80% of the sequence length.

Moreover, in both classes, autonomous and non-autonomous elements have been identified. Non-autonomous elements are TE sequences that do not encode for any protein and therefore are not able to transpose by themselves. To transpose they use the transposition machinery of autonomous elements from the same subclass.

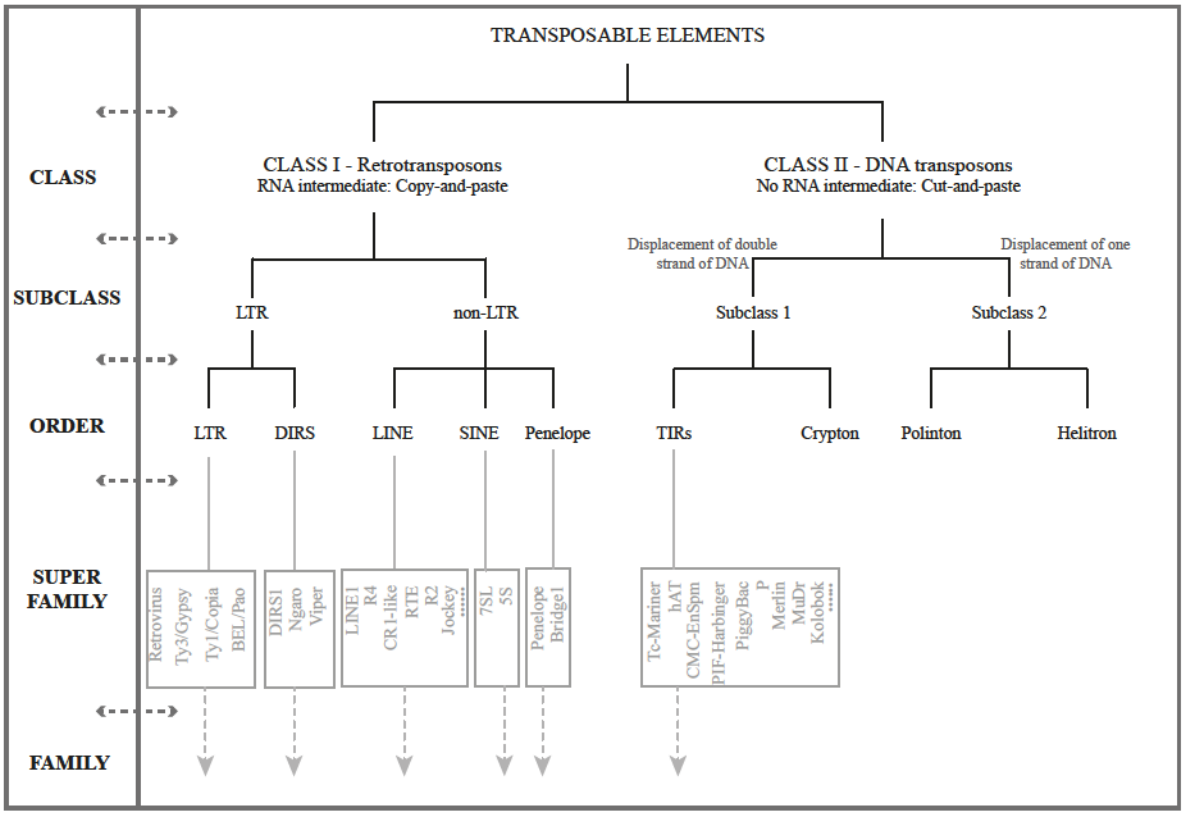


Figure 5: Universal classification of transposable elements. Five levels of organization are defined: classes, subclasses, orders, superfamilies and families, according to structure, organization and transposition mechanisms. The list of superfamilies is not exhaustive, especially in the LINE and TIR orders, which contain 15 and 19 superfamilies, respectively. For instance we also find R1 or Rex-Babar in the LINE order.

a. Class I Retrotransposons

Apart from non-autonomous elements, i.e. elements that do not encode the proteins necessary for their own transposition (SINE, LARD and TRIM), all functional retrotransposons code for a reverse transcriptase. Several classifications were proposed including different criteria. First, the class I can be divided into two subclasses, the LTR retrotransposons and the non-LTR retrotransposons or retroposons, based on the presence/absence of Long Terminal Repeats (LTR) essential for the retrotransposition mechanism in the first case. Among this two subclasses, the LTR retrotransposons are classified based on the presence/absence of an ORF encoding an integrase (INT) (enzyme catalyzing the insertion of the reverse-transcribed DNA into host DNA) with the specific DDE catalytic domain signature, or based on the order of protein domains among the pol polyprotein, the presence of RNA priming (tRNA or mRNA) and the presence of an envelope. The non-LTR retrotransposons are classified according to the presence or not of coding regions (LINEs are autonomous while SINEs are non-autonomous). Based on Wicker's classification, the retrotransposon class contains five orders, namely LTR, DIRS, SINE, LINE and Penelope that are described below.

LTR Retrotransposon order

LTR Retrotransposons and retroviruses can be confounded in the same order since the same features characterize them and since they encode the same proteins with the exception that retroviruses encode a supplementary envelope necessary for the external cell phase. However, some LTR retrotransposons, for instance some elements from the Gypsy superfamily, also encode an envelope. The main characteristic of LTR retrotransposons is the presence of the long direct repeats (LTRs), structured as [U3-R-U5] (Figure 6)(U3: contain enhancers and promoters, R: Repeated sequence, U5: first portion of the retrotranscribed genome) located at both extremities of the element and usually surrounded by the dinucleotides (5'-TG ... CA 3') (Voytas and Boeke 1992; Kumar and Bennetzen 1999). Downstream to the 5' LTR, are present Primer Binding Site (PBS), which are complementary to a specific zone of a tRNA used as primer by the reverse transcriptase (RT) to initiate the reverse transcription. Upstream to the 3' LTR, LTR retrotransposons present a Polypurine Tract (PPT), which serves as a primer to synthesize a new U3-R-U5 fragment of DNA during the retrotransposition process. Autonomous LTR retrotransposons also contain an internal region with the characteristic Gag and Pol (and Env in retroviruses and occasionally in other superfamilies) ORFs, which encode proteins necessary for the replication and transposition of the element (Figure 6). The Gag polyprotein is processed into a matrix protein (MA), which is necessary for the targeting of the cell membrane and for capsid assembly, a capsid protein (CA), which forms the hydrophobic core of the virion, and a nucleocapsid protein (NC), which is involved in RNA packaging. The Pol polyprotein is a single ORF that encodes several proteins (protease, RT, ribonuclease H, INT) that do not have the same organization in the different superfamilies. The protease (PR) plays a major role in the maturation process allowing to cut several peptides. The RT is essential to synthesize DNA from a single strand of RNA. The ribonuclease H (RH) hydrolyzes the original RNA from the RNA/DNA hybrid generated after the retrotransposition process. The integrase catalyzes the insertion of the linear double-stranded viral DNA copy into the new location. Finally, in some elements of this order, a chromodomain for "Chromatin Organization Modifier Domain" is involved in chromatin remodelling, as for Ty3/Gypsy for example that I will present later.

Four superfamilies are numbered in the autonomous LTR retrotransposon order: Ty3/Gypsy, retroviridae (Hull 2001; Gifford and Tristem 2003), BEL/Pao and Ty1/Copia. They were all found in a wide range of eukaryotes including plants, fungi and metazoa. A fifth superfamily that I will not present further, the caulimoviridae was suggested to be part of retrotransposons due to its Gag-PR-RT-RH structure. However, this superfamily corresponds to a double-strand DNA un-enveloped retrovirus with no LTR (Bousalem *et al.* 2008). The four superfamilies differ by the presence of the envelope, the order of encoded-protein inside the Pol ORF and by their phylogenetic position based on a reverse transcriptase alignment (Xiong and Eickbush 1990). Even

among the superfamilies, families can differ by the presence of an envelope, by the presence of supplementary ORFs (for instance encoding a chromodomain in the Gypsy family or accessory genes in retroviruses) but also by frameshifts occurring between Gag and Pol ORFs.

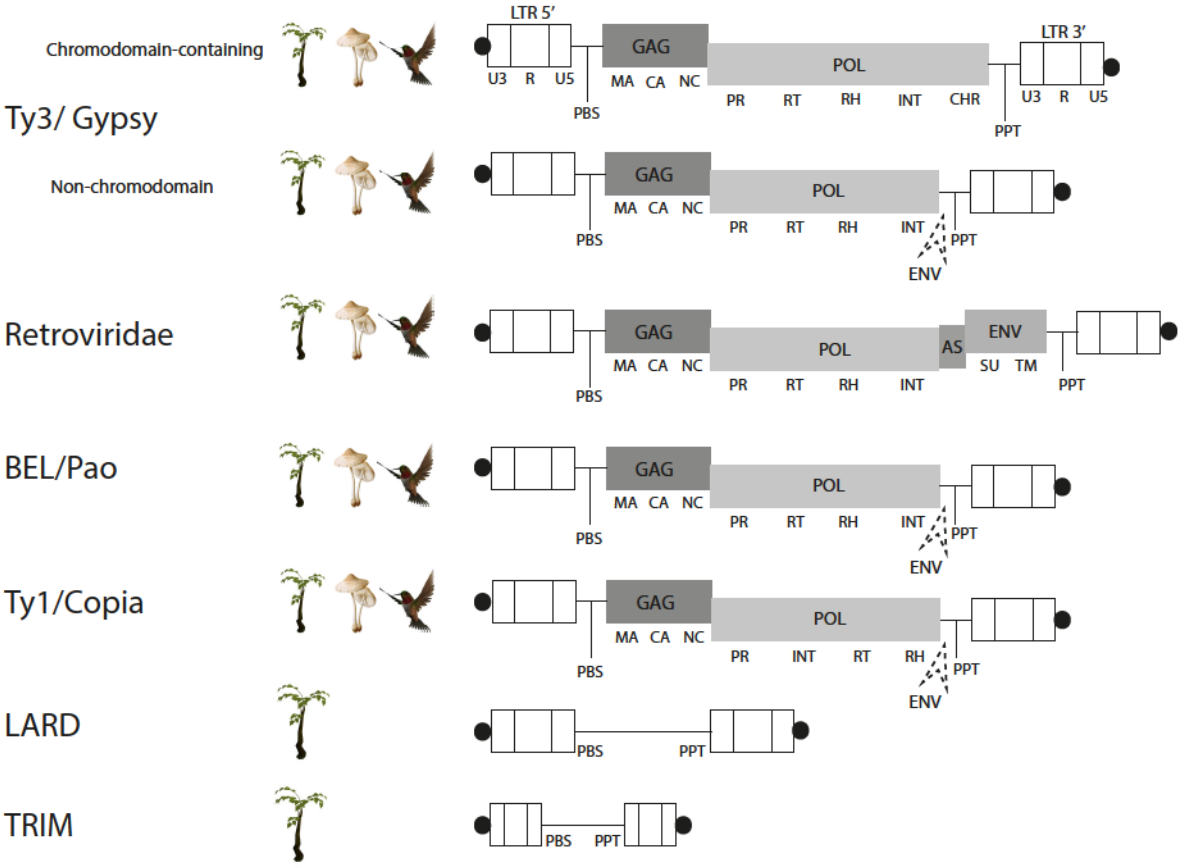


Figure 6: Structure of the different superfamilies from the LTR Retrotransposon orders. Host species are indicated for plants, fungi and metazoans (left to right order). Grey boxes represent Open Reading Frames (ORFs). Black circles represent TSDs. Abbreviations : PBS, primer-binding sites ; PPT, polypurine tract ; MA, matrix ; CA, capsid ; NC, nucleocapsid ; PR, protease ; RT, reverse transcriptase ; RH, RNaseH ; INT, integrase ; CHR, chromodomain ; ENV envelope ; AS accessory genes. Boxes are not at scale.

Within the Ty3/Gypsy superfamily, the INT domain is always located after the RH except in the Gmr1 family (Figure 6) (Goodwin and Poulter 2002). Ty3/Gypsy includes both retrotransposons and retroviruses according to the presence of an envelope. Retroviridae constitutes a retrovirus superfamily with a Gag-Pol-Env organization and can contain accessory genes useful for the transmission of the retrovirus from a host into another. Retroviridae are classified in seven genera based on their replication mechanism: Alpha-, Beta-, Gamma-, Delta-, Epsilon-, Spuma- and Lentiviridae (Gifford and Tristem 2003). BEL-Pao retrotransposons have a similar organization than Ty3/Gypsy and can also contain an envelope (de la Chaux and Wagner 2011). Finally Ty1/Copia superfamily differs in the position of the integrase, which is located directly after the protease (Figure 6) (Flavell 1992).

Non-autonomous LTR retrotransposons have only been described in plants so far. They only present LTRs, PBS and PPT structures, but do not encode any protein. Two different types of non-autonomous elements were described in particular in plants: Large Retrotransposons Derivatives (LARDs) and Terminal-repeat Retrotransposons in Miniature (TRIMs). LARDS are long elements, from 5.5kb to 8.5kb, and it has been shown that they could derive from Ty3/Gypsy or Ty1/Copia autonomous elements (KALENDAR *et al.* 2004). TRIMs are shorter elements no longer than 540 bp. They have small LTRs flanking PBS and PPT domains. No associated autonomous elements could yet be identified (WITTE *et al.* 2001).

Origin of retrotransposons is a major question, which is still debated. Indeed, one of the interesting questions is the origin of retroviruses. Did retroviruses appear before retrotranposons or retrotransposons before retroviruses? It has been suggested that retroviruses are retrotranposons that acquired an envelope ORF (EICKBUSH AND JAMBURUTHUGODA 2008).
Furthermore, from an evolutionary point of view, it is now accepted that integrase-encoding LTR retrotransposons arose from the combination of a DDE-transposase from a DNA transposon and a reverse transcriptase from a non-LTR retrotransposon (CAPY *et al.* 1997; MALIK AND EICKBUSH 1999; MALIK AND EICKBUSH 2001).

DIRS Retrotransposon order

Dictyostelium Intermediate Repeat Sequence (DIRS) is an order of retrotransposons that contains a tyrosine recombinase gene (GLOCKNER *et al.* 2001; GOODWIN AND POULTER 2001) instead of an integrase, making them different from LTR retrotransposons by their of replication mechanism. These enzymes allow the recombination between two double-stranded DNA molecules, using a catalytic tyrosine residue. DIRS have been initially discovered in *Dictyostelium discoidium* as new transposons containing non-identical and inverted LTR delimited by a trinucleotide TTT (except the recently discovered Viper superfamily), and three different ORFs (Figure 7) (CAPPELLO *et al.* 1985). Nowadays, the order is composed of three superfamilies, namely DIRS1, Ngaro and Viper. These three superfamilies differ in term of structure and organization, for instance the number and orientation of terminal repeats or by the number and/or the organization of ORFs that can entirely overlap (GOODWIN AND POULTER 2004; WICKER *et al.* 2007).

DIRS1 elements have been identified in many lineages as Amoeba, Metazoa, plants and fungi (CAPPELLO *et al.* 1985; DE CHASTONAY *et al.* 1992; RUIZ-PEREZ *et al.* 1996; GOODWIN AND POULTER 2001; DUNCAN *et al.* 2002; GOODWIN AND POULTER 2004; POULTER AND GOODWIN 2005). The organization of these elements can vary depending on the host. ORF2 and ORF3 can completely overlap or can be separated by internal complementary regions believed to play a role during the replication cycle. Furthermore, some DIRS1 elements

encode a supplementary methyltransferase located in the ORF that encodes RT and RH (Figure 7).

Ngaro was the second superfamily of DIRS order to be discovered. It is phylogenetically distinct from DIRS1 (Goodwin and Poulter 2001; Goodwin and Poulter 2004). This superfamily was detected in various lineages of animals and fungi but not in plants (Goodwin and Poulter 2004; Muszewska *et al.* 2013). Ngaro sequences contain diverse repeats that could be important for transposition. In contrast to DIRS1, none of the Ngaro elements contains a methyltransferase domain, which supports the clear distinction between DIRS1 and Ngaro elements. Some species present sequences with additional ORFs similar to ORF1 of CR1 LINE elements (see below in the section LINE). It has been suggested that these ORFs may belong to the large hydrolase family that has diverse ranges of functions as esterases, lipases or proteases, among others (Huang *et al.* 2001). These proteins may have a function in the penetration of cell membranes during the process of horizontal transfer (Kapitonov and Jurka 2003) (Figure 7).

Viper is the last superfamily belonging to the DIRS order. It was discovered in *Trypanosoma cruzi* in association with short interspersed sequences (Lorenzi *et al.* 2006). Apart from the inversion of ORF2 and 3, Viper elements present a similar organization than Ngaro ones. Nowadays, it has only been found in *Trypanosoma* species. This last superfamily is still poorly studied (Figure 7).

Some studies suggest that DIRS should not be included within retrotransposons since they have a particular transposition mechanism that uses a tyrosine recombinase (Boeke 2003; Kazazian 2004). However, phylogenetic analysis based on RT sequences classified them among the LTR subclass (Duncan *et al.* 2002; Goodwin and Poulter 2004; Lorenzi *et al.* 2006).

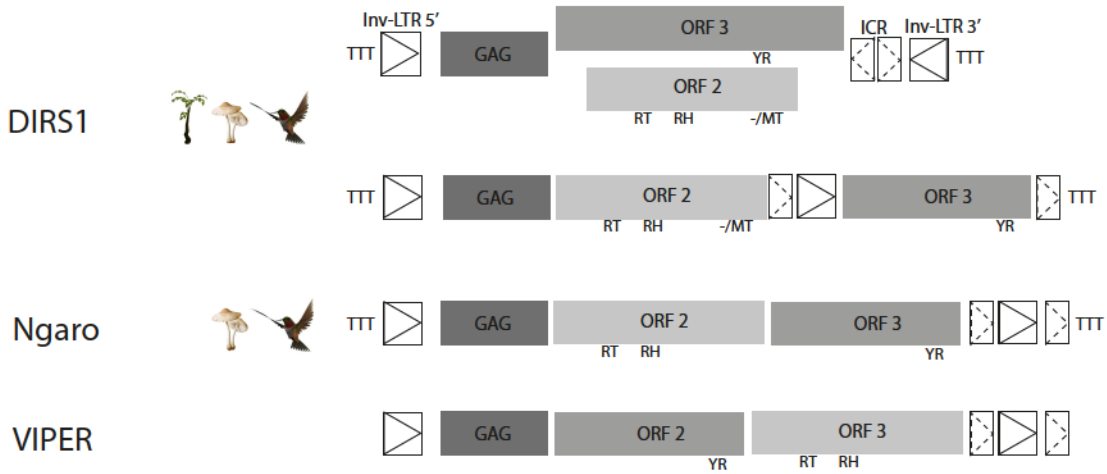


Figure 7: Schematic representation of DIRS, Ngaro and Viper superfamilies. Abbreviations : MT, metyltransferase ; YR, tyrosine recombinasse ; RH, Ribonuclease H ; RT, Reverse Transcriptase ; ICR, internal complementary region.

LINE Retrotransposon order

Long Interspersed Nuclear Elements (LINEs) are shorter than LTR retrotransposons that can make up to several kilobases. Instead of having LTRs, LINEs carry a promoter sequence that mediates synthesis of polyadenylated RNA by the RNA polymerase II of the host. Many superfamilies have been classified into the LINE order based on structural features and RT phylogenies (Xiong and Eickbush 1990; Tu *et al.* 1998; Malik *et al.* 1999). The LINE subdivision comprises 15 superfamilies including CRE, NeSL, R4, R2, L1, RTE, RTEX, Tad, LOA, I and Ingi, Jockey, CR1, L2 and Rex1-Babar (Figure 8) (Malik *et al.* 1999; Lovsin *et al.* 2001; Burke *et al.* 2002). These clades can be separated in two different types. The first type is composed of a single ORF encoding both a RT and a restriction enzyme-like (REL) endonuclease in its C-terminal part (R2, R4, NeSL and CRE superfamilies). Instead of a REL endonuclease, the second type encodes an apurinic-apyrimidic endonuclease (APE) located upstream to the RT. Moreover, a high diversity of structure can be observed among the second type. Some superfamilies present a single ORF with both the APE endonuclease and the RT while others have a first supplementary ORF with unknown precise function. The encoded protein is only known to bind mRNA/DNA molecules (Kapitonov and Jurka 2003) and it has been suggested that this first ORF could have similar function to Gag protein in LTR retrotransposons. In some elements, esterase and PHD (for plant homeodomain, finger motif also found in some chromodomains) domains have also been identified (Figure 8) (Coffin *et al.* 1997; Dawson *et al.* 1997; Kapitonov and Jurka 2003). The discovery of these last domains may suggest that ORF1 proteins are involved in protein-protein interaction related to chromatin remodeling. Finally, few superfamilies code for a RH (cleavage of an RNA/DNA complex) located downstream to the RT (Figure 8). However, whatever their belonging, all LINEs frequently end with either a poly-A tail, tandem repeats or an A-rich region. Moreover, most of the LINE copies identified in genomes are truncated in 5' by the host probably. As 5' parts are essential to transposition, the proliferation of the LINE elements is certainly strongly limited when truncated (more described later).

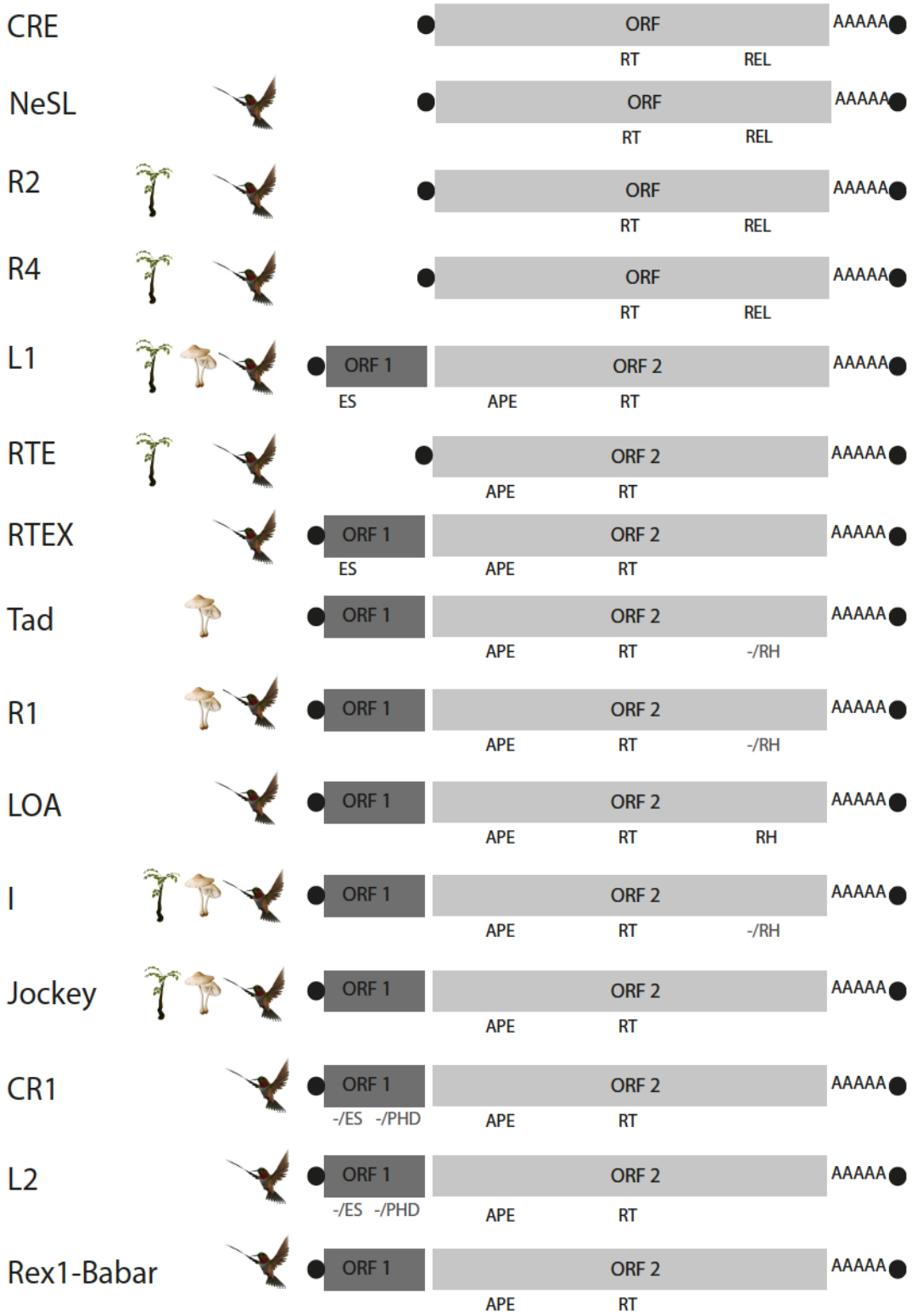


Figure 8: Diversity and structure of the different LINE superfamilies. They can contain one or two ORFs encoding an esterase (ES) with or without a PHD domain, an endonuclease (APE or REL), a reverse transcriptase (RT) and sometimes a ribonuclease H (RH). They also harbour a poly-A tail and TSDs (black circle).

During the LTR retrotransposition process, a dsDNA is directly reintegrated into the genomic DNA thanks to the integrase. Non-LTR retrotransposons use a different retrotransposition mechanism, named target-site-primed reverse transcription (TPRT). In this process, an encoded endonuclease protein creates a single-stranded nick in the genomic DNA, generally within AT-rich regions, leading to a free 3'OH strand, which is used to prime the reverse transcription. The second strand of the genomic DNA is then cleaved and used to prime the second strand synthesis. The reverse transcription is not complete in most of the case generating 5'-truncated elements. Hallmarks of the integration include a poly-A tail in the 3' part and TSDs between 2 to 20 bp length (Curcio and Derbyshire 2003; Cordaux and Batzer 2009; Levin and Moran 2011).

SINE Retrotransposon order

SINE Retrotransposon is an order of non-autonomous elements; they do not encode for any protein and thus have to use the LINE retrotransposition machinery. Due to their transposition mechanism, which copies those of LINEs, they are considered to belong to class I, but they actually have a different origin. Indeed, they do not result from a deletion in LINE sequences, but originate from accidental retrotransposition of different polymerase III transcripts (Kramerov and Vassetzky 2005). SINEs are short repetitive elements from 80 to 500 bp generally flanked by TSDs. They are generally composed of three parts: the 5' head, the central body and the 3' tail. The 5' head corresponds to a part derived from cellular RNA: tRNA (Daniels and Deininger 1985; Sakamoto and Okada 1985), 7SL RNA (Krayev *et al.* 1980; Ullu and Tschudi 1984; Nishihara *et al.* 2002) and 5S RNA (Murakami and Fujitani 1998; Piskurek *et al.* 2009) but not exclusively, and this part contains a promoter for Polymerase III. The body is composed of a first region with unknown function, CT-rich, and in some cases of a second region that shares similarities with the LINE partner, essential for the reverse transcription of the SINE RNA. However, many SINEs do not show any similarity with their respective partner. In this case, the 3' tail adopts the function of the LINE region (Kajikawa and Okada 2002; Dewannieux *et al.* 2003). The tail is a A- or AT-rich region and ends with T residues corresponding to a termination signal for Polymerase III.

Among the different SINE families, some families are well known as MIR or Alu families. In the tRNA superfamily, many families and elements have been specifically studied such as the MIR sequences from mammals (Smit and Riggs 1995), the zebrafish DANA from V-SINE family or salmonid Hpa (Takasaki *et al.* 1994; Ogiwara *et al.* 2002; Piskurek and Jackson 2011), the zebrafish mermaid (Shimoda *et al.* 1996), the salmonid AFC from CORE-SINE (Kido *et al.* 1991; Gilbert and Labuda 1999; Perez *et al.* 1999; Takahashi *et al.* 2001; Matveev and Okada 2009). Among 7SL SINE, the most known elements are the human Alu and the rodent B1 (Ullu and Tschudi 1984; Rowold and Herrera 2000).

Penelope Retrotransposon order

Penelope elements have been discovered in *Drosophila virilis*, as an inducing-agent of hybrid dysgenesis (ARKHIPOVA *et al.* 2003). Since then, they have been found in numerous eukaryotes (EVGEN'EV AND ARKHIPOVA 2005). Penelope is an interesting order of retrotransposons sharing structural features with both LTR and non-LTR retrotransposons. As non-LTR retrotransposons, Penelope elements probably use a TPRT mechanism to transpose and are frequently truncated in 5' and if present, their TSDs show variable sizes. As LTR retrotransposons, they can have, but not systematically, LTR-like structures that can be either direct or inverted (Figure 9). They contain a single ORFs encoding a RT, which is more closely related to telomerase or RT from bacterial and organellar group II introns than to RT from LTR retrotransposons (ARKHIPOVA *et al.* 2003; GLADYSHEV AND ARKHIPOVA 2007), and an endonuclease with an Uri (GIY-YIG) domain (Figure 9). As a rare signature in TEs, Penelope sequence can contain introns, which can vary in number.

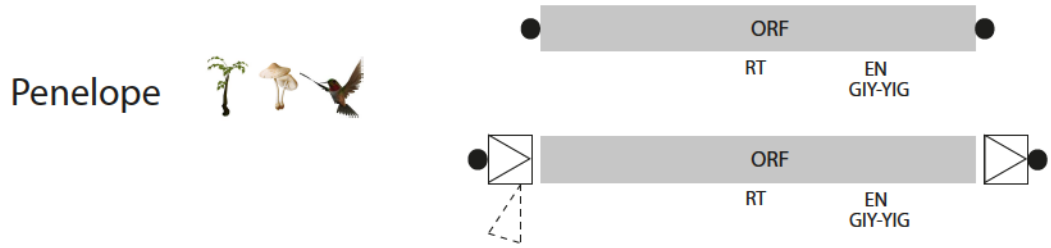


Figure 9: Schematic organization of Penelope element. Two possible but not exclusive structures are shown: sequences can contain direct or inverted repeats (square boxes at the extremities), TSDs, (black circle), intron in the 5' repeat (dotted triangle).

b. Class II DNA transposons

All DNA transposons transpose via a DNA form. If all retrotransposons require the presence of the reverse transcriptase to transpose, class II elements use variable enzymes to catalyze their transposition. DNA transposons are divided in two subclasses distinguishable by the number of DNA strands that are cut during transposition: two strands of DNA are displaced in subclass 1 while only one strand is displaced in subclass 2. Among both subclasses, elements are grouped together into orders depending on their sequence, structural features and transposition mechanism, leading to the distinction between cut-and-paste transposons, which are the most common, rolling-circle transposons (Helitrons) and self-replicating transposons (Polintons, also called Mavericks). Finally, Miniature Inverted Terminal Repeats (MITEs) are small non-autonomous DNA transposons.

Terminal Inverted Repeats (TIRs) or Cut-and-paste transposon order

These DNA transposons are characterized by their terminal inverted repeats (TIRs) that are of variable sizes – as well as their TSDs - and contain the transposase with DNA-binding domain. Nowadays 19 TIR superfamilies have been described (Figure 10) (KAPITONOV AND JURKA 2008; BAO *et al.* 2009; BAO *et al.* 2010; MENG *et al.* 2011) (Repbase), each of them being characterized by a superfamily-specific transposase core and presenting different preferential insertion sites. Six of these superfamilies are only listed in Repbase and consequently not well described (Mirage, Novosib, Rehavkus, ISL2EU, Kolobok, Academ) (JURKA *et al.* 2005a). Most of the transposases encoded by elements belonging to this order are also called DDE/DDD transposases due to the presence of three conserved acidic catalytic residues: DDE or DDD (YUAN AND WESSLER 2011). In four superfamilies (P, PiggyBac, CACTA and Sola), the encoded transposase do not contain the DDE triad, however they might contain either a catalytic aspartate or glutamate residue (HICKMAN *et al.* 2005). Most of the superfamilies are composed of a single ORF, but few of them can contain a second ORF, such as Pif-Harbinger that encodes a DNA binding protein, or CACTA that has a second unclear ORF. Several exons have been identified for few elements, such as in Mirage element (in *C. elegans*; Repbase).

A recent study shows that some superfamilies are closely related and can be grouped together. This is the case of MULE and Rehavkus, but also Harbinger and ISL2EU. Moreover, EnSpm, CACTA, Mirage and Chapaev can be classified together in the so-called CMC superfamily (YUAN AND WESSLER 2011).

From an evolutionary point of view, it is interesting to note that DDE/D transposases and integrases are both characterized by the same catalytic triad (HAREN *et al.* 1999), suggesting a common origin or a converging function.

On the contrary to retrotransposons, that easily increase their copy number by a "copy-and-paste" mechanism and multiply at each transposition cycle, the increase of the copy number is not as obvious for DNA transposons given that they excise to insert in a new location during transposition. To get around the "cut-and-paste" mechanism and amplify, DNA transposons transpose during chromosome replication from a position that has been already replicated to another position where the replication fork has not passed, yet (GREENBLATT AND BRINK 1962).

This order of DNA transposons forms a reservoir of genetic tools. It is interesting to underline that many of the elements belonging to these superfamilies are actually developed as transgenesis tools (see in the last part of the introduction for more details).

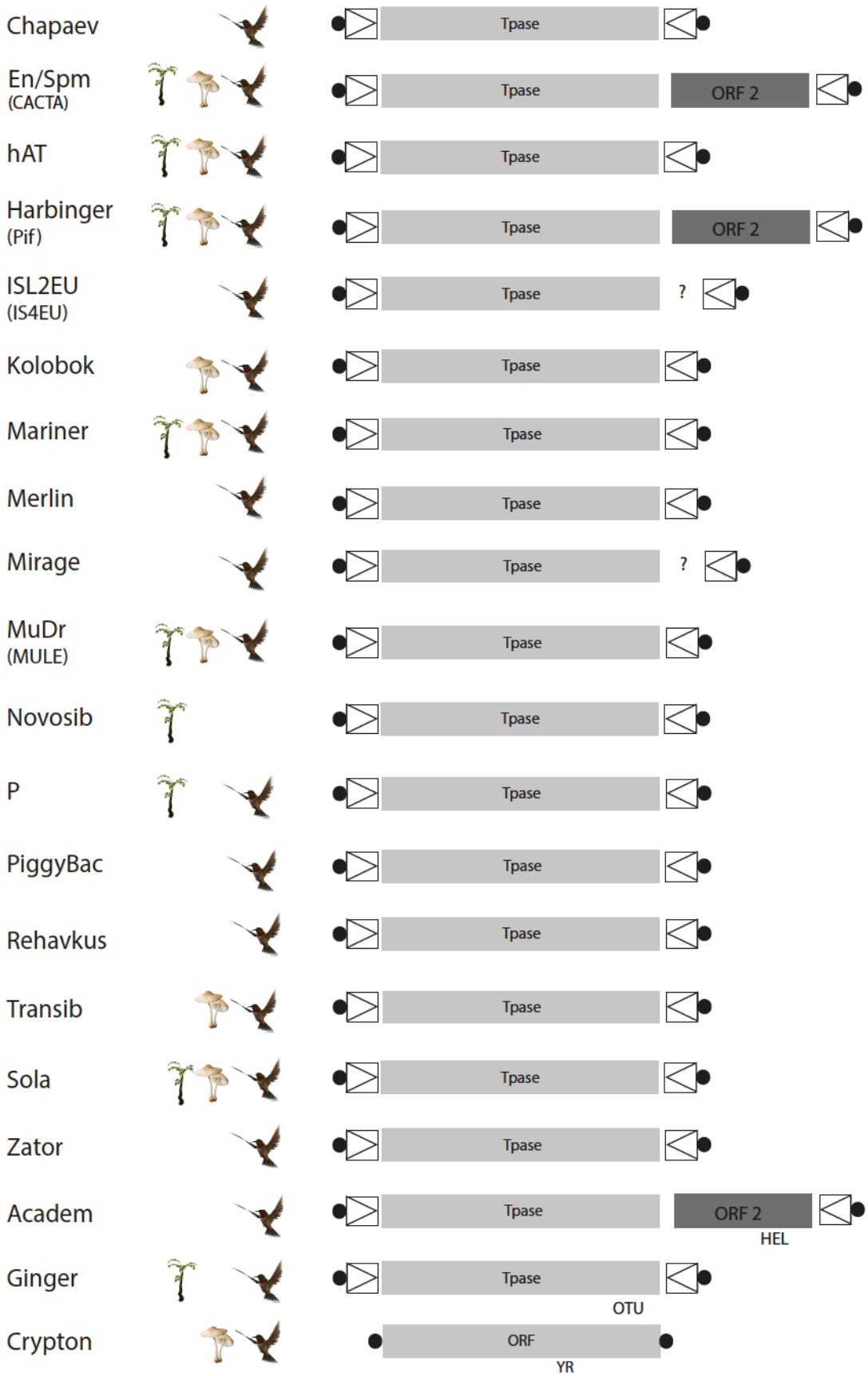


Figure 10: Diversity and structure of DNA transposons from subclass I. DNA transposons mainly contain a single ORF, which encodes a transposase (Tpase), in rare cases completed with a second one encoding a helicase (see Academ element). Many superfamilies, such as Mirage, Kolobok or ISL2EU, have to be better characterized.

Crypton order

Crypton is a poorly described order (Goodwin *et al.* 2003; Kojima and Jurka 2011), which has initially been found in fungi, but has been also, since then, identified in animals (Figure 10). Crypton elements have a single long ORF interrupted by introns, coding for a tyrosine recombinase and containing a DNA-binding domain. The elements do not present any direct or inverted repeats but can be bordered by TSDs (Figure 10).

As no RNA intermediate could be found, Crypton elements have been classified among DNA transposons. However, it has been suggested that they can be related to DIRS1 and Ngaro retrotransposon superfamilies (Figure 5), due to the presence of a tyrosine recombinase also encoded by these retrotransposons. Crypton order diversity, structure and distribution have to be further investigated.

Helitrons

Helitron elements are widespread DNA transposons included in the second subclass II. They replicate via a rolling-circle mechanism without generating any TSD (Feschotte and Wessler 2001; Kapitonov and Jurka 2001). As a consequence, instead of having repeats at their ends, Helitrons are delimited by "TC" or "CTRR" nucleotides (where R corresponds to a purine) and present a hairpin structure that lies at the 3' end (Figure 11). They encode a different type of tyrosine recombinase (Y2-recombinase such as that found in some rolling-circle bacterial transposons), similar to the one encoded by bacterial rolling circle elements, with a helicase domain and nuclease/ligase in two different ORFs (Figure 11).

Helitrons have mostly been studied in plants for gene trapping systems. Indeed, it has been shown that they can carry and mobilize gene fragments within genomes (Morgante *et al.* 2005). For example, 20,000 fragments of genes have been demonstrated to have been picked up and shuffled by these elements in the maize genome (Yang and Bennetzen 2009). Furthermore, many recent studies focused on Helitron horizontal transfer in animals, especially in bats (Thomas *et al.* 2011).

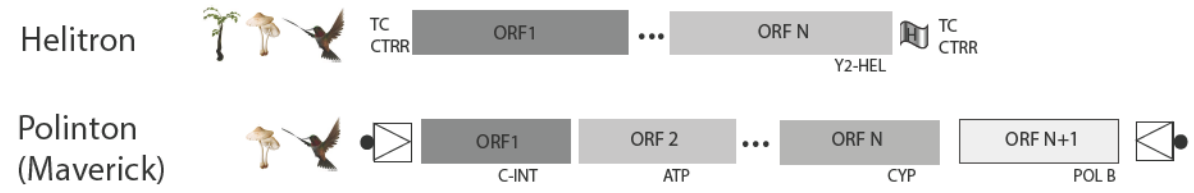


Figure 11: Helitron and Polinton structures, from subclass II. These superfamilies are composed of a minimum of two and four ORFs, respectively, however they can contain a much higher number of ORFs. Helitron are delimited by TC or CTRR nucleotides at both extremities and by a hairpin structure at the 3' end (illustrated by a small flag).

Polinton (Maverick) order

Polintons are very long (15-20 kb) self-synthesizing DNA transposons found in protists, fungi and animals but not in plants (Figure 11) (Kapitonov and Jurka 2006; Pritham *et al.* 2007). These elements can encode up to 10 different proteins that always include an

integrase similar to those of retrotransposons (C-INT), an ATPase (ATP), a cysteine protease (CYP) and a polymerase B (POL B). They are delimited by TIRs and TSDs at both extremities.

c. Other DNA transposons

Miniature Inverted Terminal Elements (MITEs)

Miniature Inverted Terminal Elements (MITEs) are short (<600 nt) non-autonomous DNA transposons that contain TIRs forming a secondary hairpin structure (Fattash *et al.* 2013). It has been proposed that they use the machinery of autonomous TIR elements (Zhang *et al.* 2001) by recognizing similar TIRs. However for many MITE sequences, such as Stowaway in rice, no corresponding TIR element could be found, even if the copy number of the MITE is high, suggesting a still unknown efficient mechanism of transposition. Mobilization of distant autonomous TIRs has been proposed (Feschotte *et al.* 2003). Other models of amplification were proposed linked to the observation that MITE copy number can be much higher than TIR elements. MITEs could be able to utilize the basic cellular machinery, by taking advantage of their hairpin structure, which could perturb the process of DNA replication (Izsvak *et al.* 1999): the newly produced strand forms a stable secondary structure that contains a specific DNA recognition motif to be reinserted into host genomic DNA.

MITEs are generally found around genic regions (Bureau & Wessler 1994). The discovery of cis-acting domains in their sequence suggests that they may regulate the expression of nearby genes.

By now, MITEs have mostly been studied in plant genomes and classified in two families, *Tourist* and *Stowaway* (Bureau and Wessler 1992; Bureau and Wessler 1994; Feschotte *et al.* 2002). MITEs have been identified in several animal genomes, for example Angel in fish (Morgan 1995; Tu 1997; Izsvak *et al.* 1999). The recent accumulation of discovery of MITEs in varying genomes suggests that they might be ubiquitous in eukaryotes.

Other types of DNA transposons: Zisupton

Zisupton is a recently described superfamily of DNA transposons encoding a new type of transposase/integrase with a SWIM domain and an Ulp1 SUMO protease, which can be included or not in the mRNA due to alternative splicing events, but no reverse transcriptase or other typical TE proteins (Bohne *et al.* 2012). Zisupton has been initially discovered on the sex chromosomes of the platyfish *Xiphophorus maculatus* and then detected in other teleost fish such as medaka or zebrafish, as well as in urochordates, cephalochordates, hemichordates, fungi and algae. From the evolutionary point of view, this element probably experienced very interesting stories, as possible horizontal transfers in several *Xiphophorus* populations have been proposed. Moreover, authors proposed that Zisupton might have been domesticated in other vertebrates, as they showed its relationship with *HMGXB3* genes.

III- The evolutionary impact of transposable elements on host genomes

a. Genomic rearrangements induced by TEs and phenotype-associated modifications

Specific features of TEs, such as mobility (change of genomic location), abundance (large fraction of genomes), universality (found in almost all genomes) and old age (found in both eukaryotes and prokaryotes), make them interesting agents of genetic, genotypic and phenotypic evolution at both individual and lineage time scales. These individual and lineage changes can be illustrated by mutational modifications, which can both impact individuals by inserting into a coding sequence for example, or influence the lineage by being maintained over long time and bringing novelties. These events have been listed and associated to specific phenotypic changes induced by TEs (Kidwell and Lisch 1997; Kidwell and Lisch 2000; Bowen and Jordan 2002; Bohne *et al.* 2008; Oliver and Greene 2009; Lisch 2013).
TE-induced genomic rearrangements can be separated in three categories: 1- rearrangements with an impact on gene function or regulation; 2- large-scale rearrangements; 3- rearrangements providing a source of new genetic material. These different events are developed below and summarized in Figure 12 and 13. The processes that are presented here are not exclusive and some of them can be classified in several of the proposed categories.

i. Impact on gene function and/or regulation by TE insertion

The movement of TEs can alter gene structure or gene expression. Alterations of genic environment can be generated through several processes summarized in Figure 12. First, TEs can insert into exons, introns or regulatory regions (such as enhancers or repressors), thus modifying either the coding sequence or the expression of the considered gene. Inserted nearby gene, but not into exons, introns or regulatory sequences, TEs can also have strong consequences on the expression of this closely located gene. Through their capacity to recombine, TEs can also lead to the deletion of exons or even complete genes. For most of these processes, examples of phenotypic changes have been observed.

Insertion of TEs into exons of host genes
The insertion of a TE into an exon represents one of the most drastic changes since it disrupts the open reading frame of the concerned gene leading to a longer protein, a truncated protein or to the non-production of the protein. Many mutational insertions have been listed in human diseases (Deininger and Batzer 1999; Belancio *et al.* 2008).

One of the well-known cases is the insertion of an Alu SINE into the *BRCA* gene, leading to breast cancer in women, or similarly colon cancer induced by a L1 insertion into the *AFC* gene (MIKI *et al.* 1992). A lot of such insertions historically lead to the discovery of TE families. This is the case of the P element in Drosophila discovered in white-eye mutants, or the Ds maize system discovered by Barbara McClintock. Among vertebrates, natural population of *Xiphophorus* mutants have a TX1 retrotransposon inserted in the *Xmrk* oncogene, resulting in its inability to induce melanoma formation (SCHARTL *et al.* 1999).
These insertions into exons do not necessary lead to lethality. Indeed, white-eye fly mutants have been maintained in laboratory. This is particularly interesting to note that TEs can also serve as mutators to decipher gene function, such as insertions of P element in *Drosophila*, which have been intensively used to annotate fly genes.

Insertion into introns of host genes

Most of the TEs inserted into introns have the chance to survive because they are probably less targeted by natural selection and are spliced out during mRNA processing. However, there are some cases where these insertions might alter both gene regulation and/or gene splicing. Indeed, regulatory sequences can be located into introns and the insertion of TEs near these sequences might change the regulation. To note, ectopic expression of *plena* (*ple*) gene in flower of an *Antirrhinum* mutant results from a different orientation of a transposon in the intron of *ple* gene (BRADLEY *et al.* 1993).
It has also been proposed that insertions in very large introns could be beneficial for the host. Indeed, stem loop structures are formed, facilitating the matching of distant splice sites and thus easing the splicing of the large introns (SHEPARD *et al.* 2009).
Finally, TEs that are located in introns can be recruited as new exons: a phenomenom called exonization. In human for example, some Alu-inserted sequences have been expated as novel protein-coding domains via alternative splicing (KRULL *et al.* 2005).

Insertion into regulatory regions of host genes

TE insertions into regulatory regions mainly alter the expression of host genes, provoking the loss of tissue-specific expression for instance, i.e. genes gaining expression in new tissues. They can insert into enhancers, interrupting the expression, or into silencer binding sequences, leading to the expression of the gene instead of its repression. They can also insert into promoters or into TATA boxes. Up to now, well-known examples were described in plants: *Mu* insertion into *knotted1* regulator leading to ectopic expression in maize leaves (BHARATHAN *et al.* 2002); MITE insertion into upstream region of *Vgt1* gene of the maize, which regulates timing of flowering (SALVI *et al.* 2002); *Tam3* transposon into the 5' region of *Antirrhinum niv* gene regulating anthocyanin synthesis (LISTER *et al.* 1993).

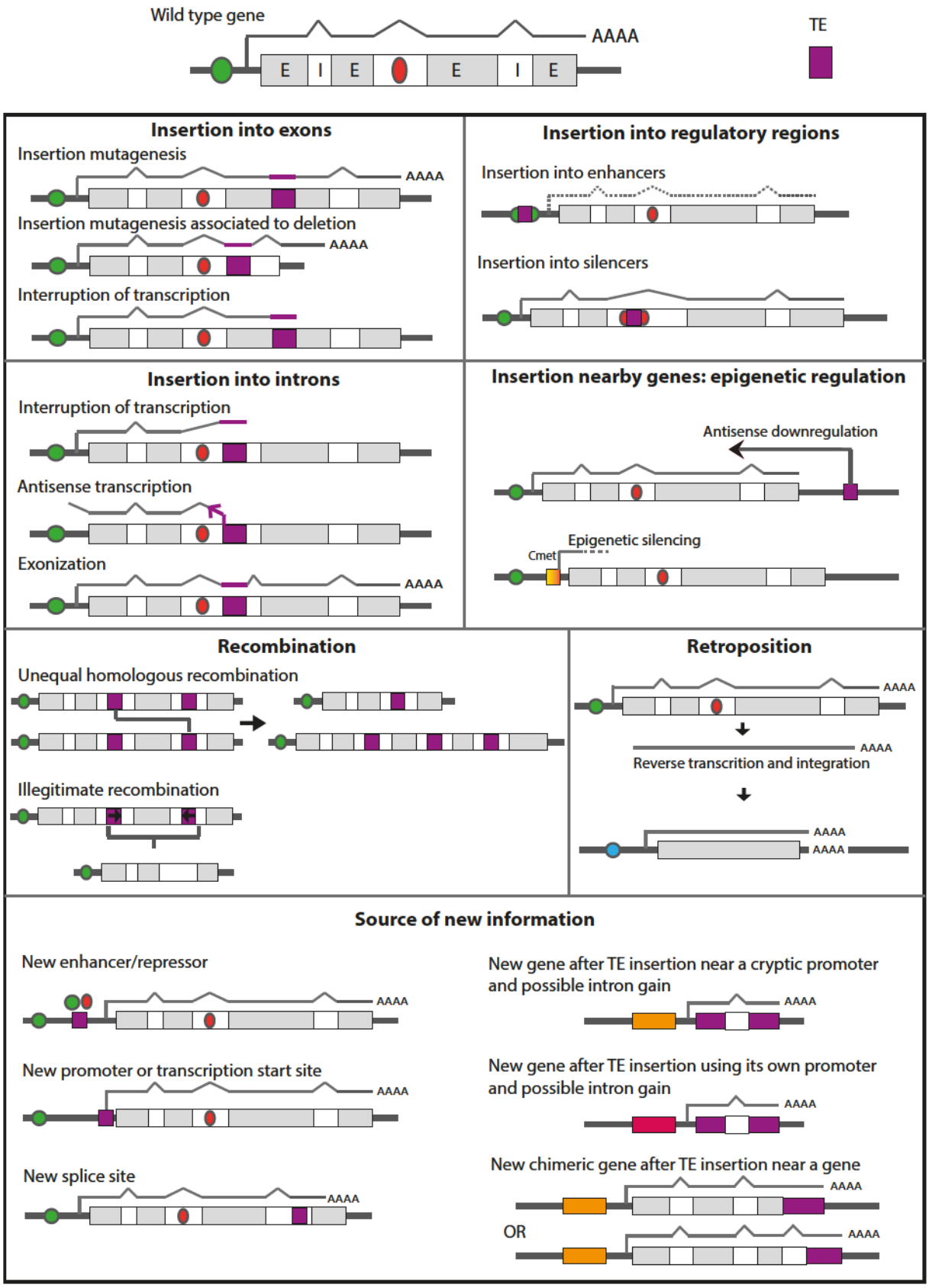


Figure 12: Genomic rearrangements induced by transposable elements in a genic context through insertion, deletion, recombination or retroposition events. The gene is represented by grey (exons) and white (introns) boxes. Purple boxes represent TEs; green and red circles show enhancers and silencers, respectively. Cryptic and TE promoters are represented by orange and pink rectangles, respectively.

Epigenetic regulation

More and more researchers study the impact of insertions of TEs near genes at a different scale, independently of the location of the insertion. We could think that such insertions are harmless for host, however they are targeted by host defense mechanisms in order to silence them. The inserted TE sequence and the region around, including genes, are sometimes methylated leading to an extinction of gene expression (Yoder *et al.* 1997; Matzke and Matzke 1998b; Matzke and Matzke 1998a; Liu *et al.* 2004; Hollister and Gaut 2009). Conversely, unsilenced TEs nearby genes have been found, as LTR retroelements near *Dasheng* gene in rice (Kashkush and Khasdan 2007).

Recombination

TEs have the capacity to recombine and thus to alter their genomic environment by enlarging or shortening genes through both unequal homologous or illegitimate recombination. Process of unequal homologous recombination can convert intact LTR retrotransposons to solo-LTR (Vitte and Panaud 2003). Like insertions and deletions processes, recombination can lead to diseases (Burwinkel and Kilimann 1998; Rossetti *et al.* 2004), as illustrated for instance by a case of hemophilia induced after an alu-mediated recombination event (Vidal *et al.* 2002). From another interesting evolutionary point of view, it has been shown that two retrotransposons play important roles in the mating system of *Neurospora* by inducing unequal recombination in the specific mating locus (Gioti *et al.* 2012).

Moreover, recombination can be a used by the host to regulate the copy number of TEs and thus maintain the integrity of the genome (Charlesworth *et al.* 1986; Charlesworth and Langley 1989). This is a well-known process that might be involved in genome size variation, as explained later (Bennetzen 2005; Bennetzen *et al.* 2005; Tian *et al.* 2009; Sun *et al.* 2012b).

Retroposition and gene capture

As already described, TE insertions can act on gene structure, location or expression through multiple mechanisms such as disruption of coding or regulatory sequences, epigenetic modifications or recombination events. However, these processes are not exclusive and two other can be pointed out. First, retroposition consists of the reverse transcription of a spliced mRNA gene and its reintegration into a new location. This process is driven by a retrotransposon-encoded reverse transciptase. The newly retroposed gene does not contain any intron and is flanked by TSDs. This process can generate new expression patterns, as shown by the retroposed *IQD12* gene at the *SUN* locus leading to new expression in fruits (Xiao *et al.* 2008). This process is probably not only simple genomic mistakes but it could form a positive evolutionary mechanism.

The second interesting consequence of certain TE movements is a process named “gene capture”. During transposition, some TEs are able to take away pieces of close genes or even an entire gene leading to novelties, such as new exons or new chimeric genes in

some cases (JIANG *et al.* 2004; GUPTA *et al.* 2005; LAI *et al.* 2005; MORGANTE *et al.* 2005). Bs1 LTR retrotransposons in maize are the first that were shown to have trapped an ATPase gene (JIN AND BENNETZEN 1994).

ii. Large-scale rearrangements

Chromosomal rearrangements involve the displacement of a part of chromosome. Rearrangements can be intra-chromosomal (inversion) or inter-chromosomal (translocation between two chromosomes) (Figure 13). However, to be viable, these rearrangements have to generate chromosomes with a centromere, essential during mitosis and meiosis, and two telomeres, essential for the end of DNA replication and chromosome protection. Within the genome, repetitive sequences, and so TEs, may act as sites for inequal crossing-overs, which can produce deletions, duplications, inversions and translocations (Figure 13). These rearrangements are possible through two TE-associated mechanisms: homologous recombination and alternative transposition. During classic transposition, complementary ends of a single TE are paired together, while in alternative transposition the pairing is done between different TEs on the same or different chromosomes (GRAY 2000). This has been observed in bacteria with IS10/Tn10 elements (CHALMERS AND KLECKNER 1996), in maize and tobacco with Ac/Ds system (ENGLISH *et al.* 1993; WEIL AND WESSLER 1993) or in Drosophila with P elements (SVOBODA *et al.* 1995; GRAY *et al.* 1996; PRESTON AND ENGELS 1996; PRESTON *et al.* 1996).

Chromosomal rearrangements

Chromosomal rearrangements can be classified in two main groups: balanced and imbalanced. Recombination of TE sequences can lead to these two rearrangements. Even if only these two events will be described, TEs can also induce large deletions or large duplications also changing chromosome organizations. Balanced rearrangements change gene order but do not remove or duplicate DNA, while imbalanced rearrangements modify gene dosage of the affected chromosome. Inversions and reciprocal translocations are balanced rearrangements while deletions and duplications are imbalanced ones. Inversions correspond to a double break of a chromosome followed by a flip of the fragment. Inversions can be pericentric, meaning that it involves the centromere, or paracentric (Figure 13). In reciprocal translocations, acentric fragments of two non-homologous chromosomes are exchanged (Figure 13). Non-reciprocal translocations lead to unequal exchanges. Two acrocentric chromosomes sometimes generate one big chromosome and a very short one, which can be consequently lost. It is important to distinguish translocations occurring in gametogenesis during meiosis, which will possibly impact the offspring, and translocations occurring in somatic cells during mitosis, which directly impacts the concerned cell.

These rearrangements have been associated to gene conversion (SLATKIN 1985; FLOT *et al.* 2013; HANSON *et al.* 2013). Gene conversion is a form of homologous recombination that involves the unidirectional transfer of a DNA sequence to the homologous region, which underwent a double break strand. It has been shown that these events can be associated to the presence of TEs (FLOT *et al.* 2013).
These rearrangements may also play important roles in major genome reorganization and restructuration. Indeed, TEs may play a fundamental role after whole genome duplication to stabilize and re-organize the genetic material allowing it to reach a stable state. By provoking these rearrangements, they probably help the return to a diploid state (DE BOER *et al.* 2007).

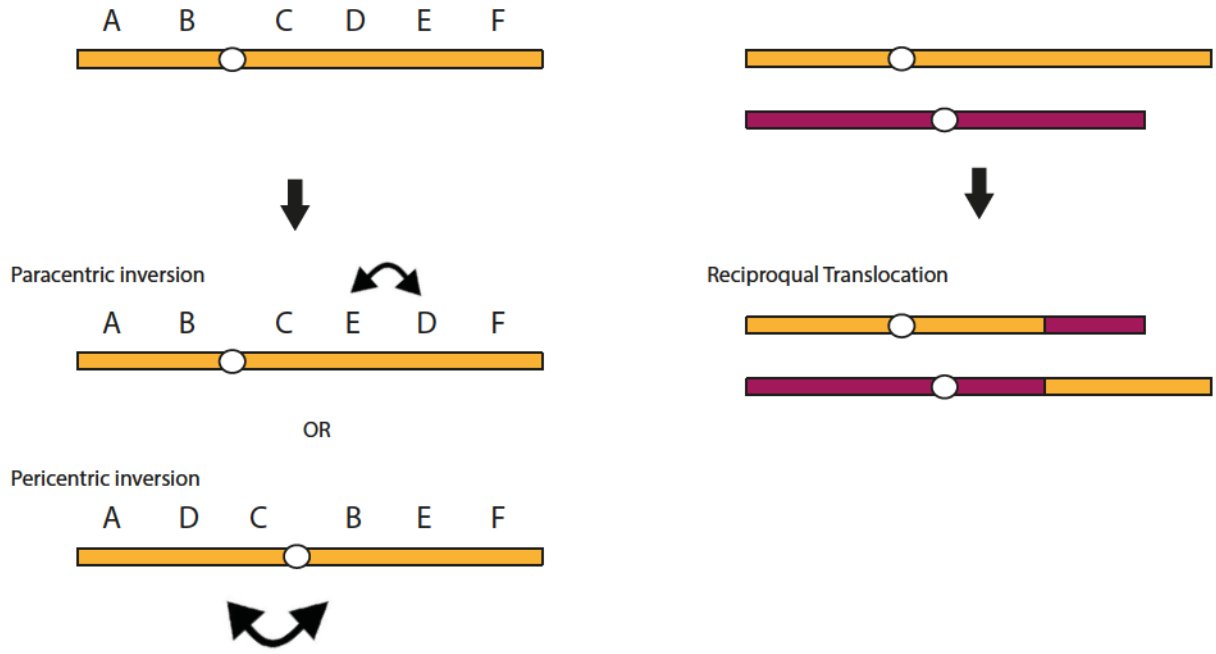


Figure 13: Inversion and translocation are the two main large-scale chromosomal rearrangements that transposable elements can induce. Inversions occur within a single chromosome while translocations occur between two different chromosomes. Of note, TEs can also induce large deletions and large duplications.

Preferential insertion sites

It is essential to explain that TEs generally have preferential insertion sites: they do not insert randomly in the genome and thus will not induce such rearrangements randomly. This can lead to particular zones of TE accumulation and particular chromosome shapes from an evolutionary point of view. Most of TEs insert into heterochromatin and gene-poor regions, thus avoiding counter-selection and elimination. Chromoviruses, related to Ty3/Gypsy retrotransposons, are particularly known to target heterochromatin using their chromodomain (MALIK AND EICKBUSH 1999). Some elements preferentially insert into GC-rich regions, such as DNA transposons from plants that target unmethylated GC-rich regions. These insertions are probably favoured due to the open state of chromatin in these regions (BENNETZEN 2000). Similar trends have been observed for few

retrotransposons (Miyao *et al.* 2003; Baucom *et al.* 2009), as well as for Alu sequences (Hellen and Brookfield 2013a). Moreover, these types of elements are themselves GC-rich, meaning that they contribute first to enrich the regions in GC nucleotides and second to promote insertions of other similar sequences (Miyao *et al.* 2003). In the same idea, some elements are AT-rich and preferentially insert into AT-rich regions, such as L1 retrotransposons that target other L1 (Smit 1999; Prak and Kazazian 2000). Similarly, pararetroviruses are found in AT-rich regions of rice genomes (Liu *et al.* 2012). Cases in which TEs are targeted by other TEs are not isolated: this situation has been observed in the fungus *Dictyostelium discoideum* (Glockner *et al.* 2001) and for Alu sequences in the human genome (Levy *et al.* 2010).

Few TE families show a strong bias of insertion in gene rich regions. Many MITEs are found around genes, but also Helitron transposons in the maize genome (Yang and Bennetzen 2009) or P elements in *Drosophila*, which also insert nearby genes and especially in regulatory regions (Haney and Feder 2009).

Centromeres and telomeres are preferential sites for TE accumulation (Baucom *et al.* 2009; Tsukahara *et al.* 2009). Strikingly, telomeres are only composed of TEs in *Drosophila* (Mason *et al.* 2008; Villasante *et al.* 2008) and *Bombyx mori* (Fujiwara *et al.* 2005). Interestingly, some members of the Penelope order that lack the endonuclease domain are specifically located in telomeric DNA of various eukaryotes including fungi and plants (Gladyshev and Arkhipova 2007).

Finally, sex chromosomes are of particular interest regarding preferential TE insertions. Indeed, entire chromosomes can be subject to TE accumulation and this is particularly true for sex chromosomes. A sex-specific transposition activity in male or female germline (depending on the heterogametic sex) leads to a biased representation of TEs between sex chromosomes. For instance, a specific accumulation has led to a TE over-representation on the Y-chromosome and an under-representation on the X-chromosome in certain mammals. This has been observed for young SINEs in male germline in human, mouse and dog (Weiner *et al.* 1986; Jurka *et al.* 2005b). This accumulation in heterogametic sex is clearly linked to the evolution process of sex chromosomes. It is well known that sex chromosomes, at one point, do not recombine anymore and thus evolve differently. The state of non-recombination promotes the accumulation of TEs: they are not eliminated leading to differentiation of the two sex chromosomes (Steinemann and Steinemann 1991; Erlandsson *et al.* 2000). Despite the degrading accumulation observed on one of the two sex chromosomes, a second type of accumulation can be observed. In mammalian X chromosome, an accumulation of L1 elements plays a dosage compensation role, described as the Lyon hypothesis (Lyon 2003).

iii. Source of new genetic materials

Local and chromosomal rearrangements are not necessary deleterious for the host. In some cases, these modifications can be fixed and maintained in the genome because of their beneficial effect. Indeed, TE insertions can be source of innovation: they can provide material for new enhancers or silencers, new promoters, new splice sites or even new coding regions (new exons, new genes or chimeric genes) and consequently new function for hosts can be created (Figure12).

More and more examples and studies demonstrate the implication of TEs in gene regulation (Medstrand *et al.* 2005; Bourque 2009; Rebollo *et al.* 2012; de Souza *et al.* 2013), making real the term "controlling elements" proposed by B. McClintock. Many TE sequences have been exapted as regulatory elements. Some SINE retrotransposons for instance have been exapted as enhancers (Bejerano *et al.* 2006; Santangelo *et al.* 2007; Sasaki *et al.* 2008; Nakanishi *et al.* 2012).
Since 20 years, it has been shown that a large amount of transcription factors binding sites are enriched in TEs (Polak and Domany 2006; Bourque *et al.* 2008; Bourque 2009). For instance, binding sites of Pou5f1-SOX2 are enriched in ERVK. Multi-species genome alignments experiments highlighting conserved elements and ChIP sequencing, revealing transcription factor binding sites, facilitate the discovery of such elements. Localisation of TEs under strong selective pressure is also a way to identify them (binding of MIR SINEs by ESR1 factor for example) (Kamal *et al.* 2006).

The capture of TEs as new exons into pre-existing genes is called exonization (Nekrutenko and Li 2001; Kreahling and Graveley 2004; Schmitz and Brosius 2011). This phenomenon can follow an insertion into an intron. The process is more frequent in vertebrates than in invertebrates, which generally contain small introns (Sela *et al.* 2010). If the newly added sequence, which creates a new splice variant, is advantageous then it can be kept by the host and may bring new function (Schmitz and Brosius 2011).

New genes can be generated from TEs through a process called molecular domestication that "turns junk into gold" (Miller *et al.* 1992; Volff 2006; Sinzelle *et al.* 2009). Molecular domestication corresponds to the creation of new genes deriving from TE sequences, giving rise to new functions. Domesticated genes differ from TEs on several points: 1- there are single copy (except in polyploid genomes or after a duplication), 2- the synteny of the region is conserved in different species, 3- they have lost the ability to transpose and associated features such as TIRs, 4- they probably assume important biological functions since they have been kept by hosts, 5- their TE origin is still recognizable by sequence similarity. The molecular domestication process is still unclear. However, it seems that after TE insertion advantageous mutations accumulate in the interesting region while negative mutations accumulate around the area of

interest until the elimination of the rest of the TE sequence. To be expressed, this new gene needs a promoter, which can either be a cryptic promoter (the TE has inserted nearby a pre-existing promoter) or the TE promoter itself. Gains of introns are then possible thanks to the intronization process (Irimia *et al.* 2008) as observed in the *MART* gene family (Brandt *et al.* 2005).

Molecular domestication is a process that strongly participates in genome and species evolution. The newly acquired genes generally have important functions for hosts. They may play roles related to immunity, illustrated by *RAG* genes involved in the V(D)J recombination, to apoptosis (*THAP0* involved in interferon γ-induced apoptose; (Gale *et al.* 1998)), to transcriptional activity via chromatin modification (*THAP7*; (Macfarlan *et al.* 2005)), to regulation of chromatin structure (*CENP-B*), to cell cycle (member of *THAP* family; (Giangrande *et al.* 2004)), to reproduction/formation of the placenta (*Syncytin,* (Blaise *et al.* 2003; Blaise *et al.* 2005; Dupressoir *et al.* 2005)), or even to the protection of host against themselves (Malik and Henikoff 2005; Miskey *et al.* 2007; Marco and Marin 2009). A more detailed presentation of TE-derived genes in vertebrates will be done later.

As a proof of fundamental process in the evolution, evolutionary convergences have been observed in some lineages (Casola *et al.* 2008; Emera *et al.* 2012; Schmidt *et al.* 2012). For instance, this is the case of the TE-derived promoter of the *prolactin* gene, an essential gene expressed during pregnancy in human. It has been shown that *prolactin* is expressed under the control of promoters that are all different, but all deriving from TEs, in primates, mice and elephant. Other cases of convergence have been noted, as observed for *CENP-B* genes probably deriving from a pogo transposase in *Drosophila* and from a Tigger transposase in human (Tudor *et al.* 1992; Kipling and Warburton 1997), but also for syncytin genes within mammals (Dupressoir *et al.* 2005; Heidmann *et al.* 2009)

In addition to the "creation" of entirely new genes, TEs can also participate to the formation of chimeric genes, which correspond to the fusion between pre-existing coding regions from two genes or from a gene and a TE (Dooner and Weil 2007). In plants, it has been demonstrated that two different genes became a single gene after an alternative transposition event between *Ac/Ds* elements (Zhang *et al.* 2006). SETMAR is another example of chimeric gene, where a SET histone methyltransferase domain merged with the transposase domain of a mariner-like element (Cordaux *et al.* 2006). Five PiggyBac-derived genes are other examples of chimeric genes, which derived from a PiggyBac transposase and a SCAN domain (Sarkar *et al.* 2003).

b. *Genome size evolution*

Genomes vary more than 200,000-fold in size in eukaryotes (the smallest known is the microsporidium *Encephalitozoon cuniculi*, while the largest is *Amoeba dubia*) and 330-fold in vertebrates (smallest and largest being fish species, namely *Tetraodon nigroviridis* (385 Mb) and *Protopterus aethiopicus* (130 Gb)) (Gregory 2001a; Gregory 2001b).
Genome size variation in natural population is driven by different factors, including insertions, deletions, recombinations, whole genome and tandem duplications and selective forces (Petrov 2001). Due to their capacity to increase their copy number, to recombine or to delete DNA, there is no doubt that TEs play a major role in genome size variation and partially explain the "C-value" paradox. The knowledge of the mechanisms that influence genome size is crucial to understand the correlation with phenotypic traits such as cellular or nuclear sizes (simply illustrated by polyploidy), and mechanistic processes such as duration of mitosis and meiosis or rate of basal metabolism (Petrov 2001; Dufresne and Jeffery 2011). More and more studies analyze the correlation between genome size and TEs, but also the correlation between the effective population size and the genome TE content.
The variation of genome sizes has been largely studied in several plant species, showing that variations of LTR retrotransposon content strongly contribute to it. These variations are linked to LTR retrotransposon amplifications, recombination and losses (Vitte and Panaud 2003; Vitte and Panaud 2005; Vitte *et al.* 2007; Tian *et al.* 2009).

Large genome size species are interesting cases to study, but also difficult to analyze because they are too large to be sequenced (Metcalfe and Casane 2013). However, some works done on smaller genomic data (small parts of large genomes) showed that TEs might be involved in these reported cases of gigantisms. The first example is the genomes of various salamander species. Authors showed that LTR retrotransposons played a significant role in the genome size of these organisms due to their high copy number. Combined to the increase of TE copy number, the rate of DNA loss is specifically weak leading to an even greater increase of TE sequences (Sun *et al.* 2012a; Sun *et al.* 2012b). An accumulation of LINE2 retrotransposons might also be involved in lungfish genome gigantism (Metcalfe *et al.* 2012).

c. *Speciation*

By inducing many types of rearrangements, locally or at the chromosome scale, TEs reshape, restructure and "inject" new information, making them undeniable major players of genome evolution. The representation of TEs as drivers of speciation is not recent. Barbara McClintock first proposed that TE-induced rearrangements may be linked to speciation events (McClintock 1984) but this idea has been further investigated only later (Sites and Moritz 1987; Coyne and Orr 1998; Rieseberg 2001; Kraaijeveld 2010). Instead of being geographically separated – a speciation coined "sympatric" –, species are "genomically" separated, meaning that a barrier appears between two species at the genome level. The idea is that TEs might be at the origin of this barrier by inducing rearrangements for example. A recent study demonstrates that TEs might favor post-zygotic separation after reactivation and lead to speciation in lake whitefish (Dion-Cote *et al.* 2014). They may play a role of barrier between individuals, as illustrated by the phenomenon of hybrid dysgenesis between strains of *Drosophila* possessing or not active P elements (Kidwell *et al.* 1977; Bingham *et al.* 1982; Castro and Carareto 2004). These genomic separations can also appear after a genomic shock, which can be induced after hybridization between two closely related species. Following hybridization, a wide activation of TEs was observed in *Drosophila* (Labrador *et al.* 1999; Vela *et al.* 2014). Similarly, accumulation of endogenous retrovirus sequences in centromeric regions was characterized in kangaroo (Metcalfe *et al.* 2007). These two studies show that TEs might be reactivated after a genomic shock, leading in sometimes to new case of speciation. Most of these examples are not direct evidence of speciation but consist of hypotheses.

Recently, it has been proposed that exonization is a process, which can be population-specific, implying that it may enhance divergence in the population and promote speciation (Sela *et al.* 2010). In addition to these consequent rearrangements, regulatory novelties also participate to the formation of different species from an initial one (Jurka *et al.* 2007).

Environmental stress may indirectly drive speciation. Epigenetic controls are diminished during a stress leading to an uncontrolled proliferation of TEs, and contribute to reproductive isolation between some populations (Rebollo *et al.* 2010; Jurka *et al.* 2011).

After a polyploidization event, genomes tend to attempt a stable diploid state through rediploidization processes. This process is generally associated to TE activity, TEs acting as recombinators. The activity of TEs was proposed to have played an important role in salmonid species diversification after the whole genome duplication that occurred in the common ancestor of salmonids (de Boer *et al.* 2007).

IV- Activity and success of transposable elements in host genomes

The presence of transposable elements in a genome, even at high copy number, does not imply that they are active in this genome (Quesneville *et al.* 2003). In the human genome, for instance, only 100 copies of LINE1 are still active (Brouha *et al.* 2003), while in the chicken genome, only few copies of CR1 seem to be still capable of transposition (Wicker *et al.* 2005). A TE copy can be remnant or active, in both cases leading to genome shaping and diversity. An active copy of TE is characterized by the fact that it is still able to move and to insert into a new genomic location. So, the main question is: how do we know that there is activity, and how can we date activity events? In other words, how can we know when activity occurred, recently or not, in a genome?

a. Life cycle of TEs

As detailed above, the process by which TEs move is called transposition. During a mobile element life cycle, the copy number varies positively (by duplication), negatively (by transposition-related excision, recombination) or stays stable (excision followed by insertion). The dynamic of the abundance of a TE superfamily can be modelled using the rates of appearance-disappearance that determine copy numbers in a genome.

To better understand TE life cycle, from birth to death, Figure 14 shows a simplified view of the dynamic of a newly introduced TE in a host (Brookfield 2005a; Hellen and Brookfield 2013b). A new mobile element is generally acquired by horizontal transfer (HT). However, the considered time zero of the introduction can correspond to a new TE horizontally acquired, as mentioned, but also to the burst of a pre-existing element. The newly introduced TE will invade the genome and increase its copy number mainly depending on the rate of duplicative transposition (increase of copy number) versus the rate of deletion (decrease). After the introduction, if the element has a low rate of transposition, it will rapidly disappear through genetic drift, while elements with high rates of transposition will invade the host genome leading in the most extreme case to the extinction of their host (Le Rouzic and Capy 2005). Most of the current population genetic models nowadays tend to determine the mechanisms that lead to a stable equilibrium of copy number: a balance between forces of transposition, deletion and selection. Effective population size is also a factor to consider to study the spread of TEs. Population size has a direct effect on selection efficacy. In a small population, the brewing decreases, increasing the homozygosity and decreasing the recombination rate, altogether influencing the efficacy of selection against TEs and so the frequency of TEs. Thus, small populations rapidly eliminate TEs, while large population tend to maintain active transposition (Le Rouzic *et al.* 2007).

TEs reaching an equilibrium are then subject to different fates: 1- TEs stay at the equilibrium and few copies are still active; 2- in the case of complete inactivity, copies can be either eliminated by deletion, mutation, or genetic drift, or maintained as fossils in exapted elements; 3- the element can sustain a new burst of activity.
As TE dynamics depend on many internal and external factors, multiple population genetic models have been established to understand this dynamic in host genome (Le Rouzic and Deceliere 2005), including the different rates (such as transposition rate), sexual system of host, effective population size, stress, types of repeats and competition between superfamilies of TEs (Brookfield 1982; Charlesworth 1994; Le Rouzic and Capy 2006; Le Rouzic *et al.* 2007).

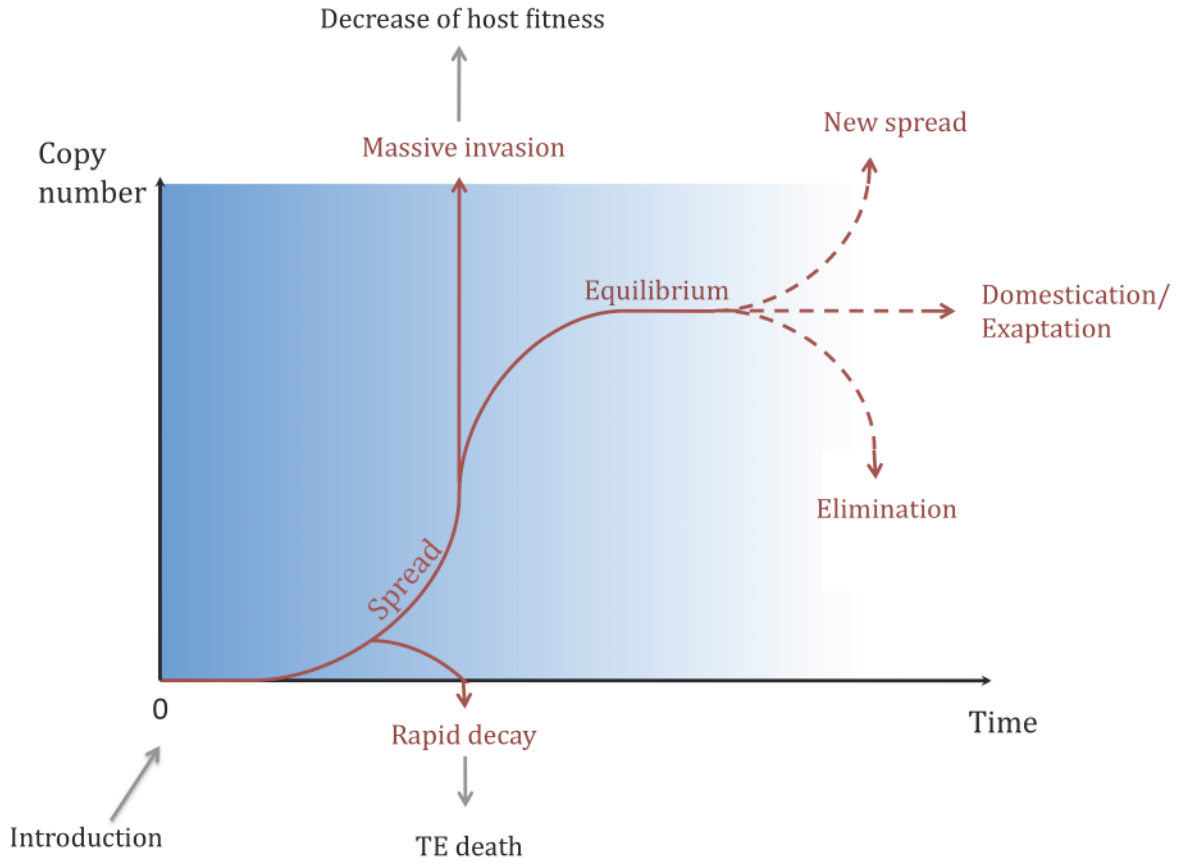


Figure 14: Dynamics of a newly introduced mobile element from birth to death. The number of copies of the elements is represented over time (generally expressed in number of generations). The various possibilities are represented by the red line.

Many factors may influence TE success (spread and maintenance of TEs) and activity, all also interfering with TE diversity (which families are present in a genome). Both host population and environment are important for TE success and survival. Indeed, environment quality (pH, temperature…), resource availability or competition may affect the effective population size but also the breeding system and it is clearly known that sex and recombination are essential for TE maintenance. For example, in asexual species, such as bdelloid rotifers, TEs tend to be purged, which favors the survival of the species (Arkhipova and Meselson 2000; Gladyshev and Arkhipova 2010; Flot *et al.* 2013).

Second, population bottleneck can alter the efficacy of natural selection and genetic drift, opening the door to large TE amplifications. Third, environmental changes may favor HT events by decreasing defense barriers, and potentially lead to the introduction of new mobile elements and subsequently to the creation of new families. Finally, the transposition mechanism itself contributes to TE success: "copy-and-paste" (retrotransposons) and "cut-and-paste" (transposons) mechanisms have differential success, since cut-and-paste transposons have to hijack cell mechanisms to multiply.
It has also been shown that environmental stresses can lead to an over-expression of TEs. It has been recently demonstrated that TEs were re-activated in aging cells (De Cecco *et al.* 2013).

b. Horizontal transfer events

From a simplified point of view, TEs are vertically transmitted from parents to offspring. For TEs that are 100% vertically transmitted, the respective population will be fixed and reach an equilibrium, as explained above (Figure 14), reducing or increasing host fitness by mediating changes (TE-derived sequences) (Hickey 1982; Kidwell and Lisch 2000; Makalowski 2000). The second way for TEs to spread into genomes is the invasion of distant species by cross-passing species barriers and entering into new genomes, a process called Horizontal [Gene] Transfer (H[G]T). HT is the most probable way to bring new TEs into genomes, modifying its TE diversity, content and history. Some newly acquired TEs may remain active after transfer suffering burst of transposition (El Baidouri *et al.* 2014). HT implies the physical transfer of TEs or genetic sequences between two cells, requiring vectors to mediate this transmission. Suitable vehicles for transport and transmission include mostly viruses, parasitoid and parasitic mites (Silva *et al.* 2004; Ivancevic *et al.* 2013), due to their ability to catch DNA sequence, to move between species and to be infectious. Intracellular parasites such as Wolbachia have also been proposed as possible vectors (Heath *et al.* 1999; Kondo *et al.* 2002). It is moreover proposed that few TEs, belonging to particular Gypsy families, are able to move by themselves between species due to their virus-like capacity to infect cells (Mejlumian *et al.* 2002). Once the new TE has been acquired, it enters a new life cycle. If transposition is not efficient enough, this TE will be rapidly eliminated.

For vertically transmitted TEs, we generally expect TE phylogenetic history to be similar to the host species history (but also depends on rate of evolution and other genomic factors). Incongruences between TEs and species phylogenies are one of the ways to infer HT events (Capy *et al.* 1994; Syvanen 2012). To support the HT hypothesis, three kinds of phylogenetic distortion have to be taken into account: 1- elements with high degree of sequence similarity in distant taxa lead to branches shorter than expected. However, this observation can probably not be done considering ancient events; 2- observation of topological differences between TE and host phylogenies, even if this

observation can be biased by the presence of many different families belonging to common superfamilies; 3- observation of patchy distributions among sister species, meaning that one of the two closely related species may possess a TE that the other did not acquired. Of course, these features are not the only clues that have to be considered and also present several biases, such as the rates of mutation or the loss of the considered TE family in a species also leading to mis-interpretations. Moreover, Capy and collaborators (Capy *et al.* 1994) proposed different hypotheses to explain that observed incongruences might originate not necessary from HT. They invoke the evolution rate of TEs depending on the considered host species, the effect of activity level (autonomous versus non-autonomous, retrotransposition versus transposition), and also the ancestral polymorphism between copies of an element in the ancestral genome.

For a long time, it has been considered that only DNA transposons and LTR retrotransposons were subject to HT (Heredia *et al.* 2004; Brookfield 2005b) and that non-LTR retrotransposons were only vertically transmitted (Malik *et al.* 1999). However, there are now evidences of horizontal transfer of non-LTR, for example RTE LINEs (Ivancevic *et al.* 2013; Walsh *et al.* 2013). Nowadays, many TE superfamilies have been identified as horizontally acquired (Schaack *et al.* 2010; Wallau *et al.* 2012). Among retrotransposons, we can note the transfers of LTR retrotransposons of the Gypsy superfamilies in *Drosophila* (Heredia *et al.* 2004) and in plants (Roulin *et al.* 2008), but also of other superfamilies such as Ty1/Copia, as well as non-LTR retrotransposons including many LINE superfamilies, SINEs and Penelope elements.
The many examples of DNA transposon HTs suggest that they are well fitted for the invasion of a range of species. These examples concern a variety of cut-and-paste transposons (Tc1 (Plasterk *et al.* 1999), Merlin (Feschotte 2004), Mariner in salmonids (de Boer *et al.* 2007) , Tol1 in medaka and *C. elegans* (Kodama *et al.* 2008), SPIN transposons (Gilbert *et al.* 2012)) but also rolling-circle transposons (Thomas *et al.* 2010).

If people previously thought that TE HT were rare events in animals, more and more studies have identified and characterized such events in many animal lineages, including vertebrates, showing animal-to-animal transfer as well as bacteria-to-animal transfers (Hotopp 2011; Ivancevic *et al.* 2013). Some superfamilies display a wide range of HT distribution, like the Helitrons, for which HT cases have been repertoried in mammals, reptiles, fish and insects (Thomas *et al.* 2010), or the OC1 transposons found in both vertebrates and invertebrates with a host-parasite association (Gilbert *et al.* 2010). One of the most striking transfers observed that involves very distant species implies a prokaryote and a bdelloid rotifer (Gladyshev and Arkhipova 2009).

c. Detecting an active TEs

Two main analyses can provide evidence for TE activity. The first one is the identification of insertions in offsprings that are absent in parents, thus constituting *"de novo"* insertions. A copy present at a precise position, but which is absent at the parental orthologous positions implies the jump of an active element. This kind of detection can be linked to phenotype changes in offsprings. Diseases induced by L1 or Alu insertions in human are also examples of new spontaneous phenotypes induced by transposons. Secondly, the observation of polymorphic insertions between two individuals of the same population can indicate recent activity. This method has been used in four different strains of mice to demonstrate that L1 elements were responsible for more than 85% of variants from intermediate size in the compared genomes (AKAGI *et al.* 2008).

If there is no possibility to evaluate neither offsprings/parents *de novo* insertion nor insertion polymorphisms between individuals, more indirect evidences must be considered. TE copies containing all intact coding ORFs described above with their complete LTRs/TIRs and TSDs probably transposed recently and are potentially still active if expressed. Potential age of TEs can be inferred assuming that 1- a freshly transposed copy is identical to the parental copy; 2- TEs evolve through a neutral model and 3- older copies accumulate a succession of mutations in both internal and LTRs sequences. Considering all these elements, the age of a copy can be estimated by quantifying its divergence from a consensus sequence (KAPITONOV AND JURKA 1996; BOISSINOT *et al.* 2000) defined from the alignment of older copies, or only divergence between LTRs of a same element (KIJIMA AND INNAN 2010), for LTR retrotransposons.

In term of expression, the presence of TE sequences in transcriptome does not imply that this TE is active. However it is still interesting to analyze this kind of data. Compared to genomes, which are of variable sizes in vertebrates and plants, and sometimes complicated to sequence, transcriptomes are routinely sequenced for many species. Until recently, few studies were focusing on TEs in transcriptomes (SCHONBACH 2004; JIANG *et al.* 2012; COWLEY AND OAKEY 2013). As initial information, the presence of TE superfamilies can be investigated, as well as their respective proportion. This information may be used to study specific enrichment compared to respective proportion in the genome for a defined TE. For example, it has been demonstrated that Alu elements present an enrichment in transcriptome compared to genome, suggesting an important transcriptional activity of this family in the human genome (MANDAL *et al.* 2013). Moreover, comparison of transcriptomes in differential conditions can bring interesting results. Comparison of facultative and obligate asexual bdelloid species showed an increase of TE transcript number in obligate sexual strain (HANSON *et al.* 2013). By RNA-seq analyses of lake whitefish species transcriptomes, authors showed

that TEs were significantly reactivated in embryos after hybridization of two diverging fish lineages, suggesting their potential role in speciation (Dion-Cote *et al.* 2014). Screening of transcriptomes is also fundamental to study small RNA pathways and their interactions with TEs. Czech and colleagues (Czech *et al.* 2013) highlighted genes involved in piRNA-mediated transposon silencing by analyzing drosophila ovarian transcriptomes. It is important to keep in mind that depending on the protocols of RNA preparation (with or without poly-A tails) and sequencing methods, results can be dramatically different.

V- TE maintenance versus host defense

As explained in TE life cycle, TE superfamilies of a host genome tend to reach equilibrium, where TEs can still be active without killing their host. Indeed, host genomes have developed defense mechanisms to avoid TE proliferation, targeting different steps of TE life cycle (Figure 15) (Johnson 2007; Slotkin and Martienssen 2007; Levin and Moran 2011; Rigal and Mathieu 2011). Genome can repress TE transcription using DNA methylation and chromatin modifications. Small RNAs, such as siRNAs, specifically inhibit the translation of TE RNA into TE protein. Finally, the last steps of retrotransposition or transposition can also be blocked, through the deamination of cytosines for example, avoiding the reintegration of the TE into a new location.

Despite the fact that TEs might invade a genome, probably decreasing its fitness as demonstrated for L1 elements in human (Boissinot *et al.* 2006) and might lead to host extinction, they also have negative effect by the simple fact that they are transcribed by the host machinery. Indeed, an active copy of TE is transcribed, then translated, then reverse transcribed (for retrotransposons) and reintegrated in host DNA. All these steps have an intrinsic metabolic cost for host, as TEs use the host cellular machinery (Badge and Brookfield 1997; Hollister and Gaut 2009). For all these reasons, genome defenses actively run to limit TE activity.

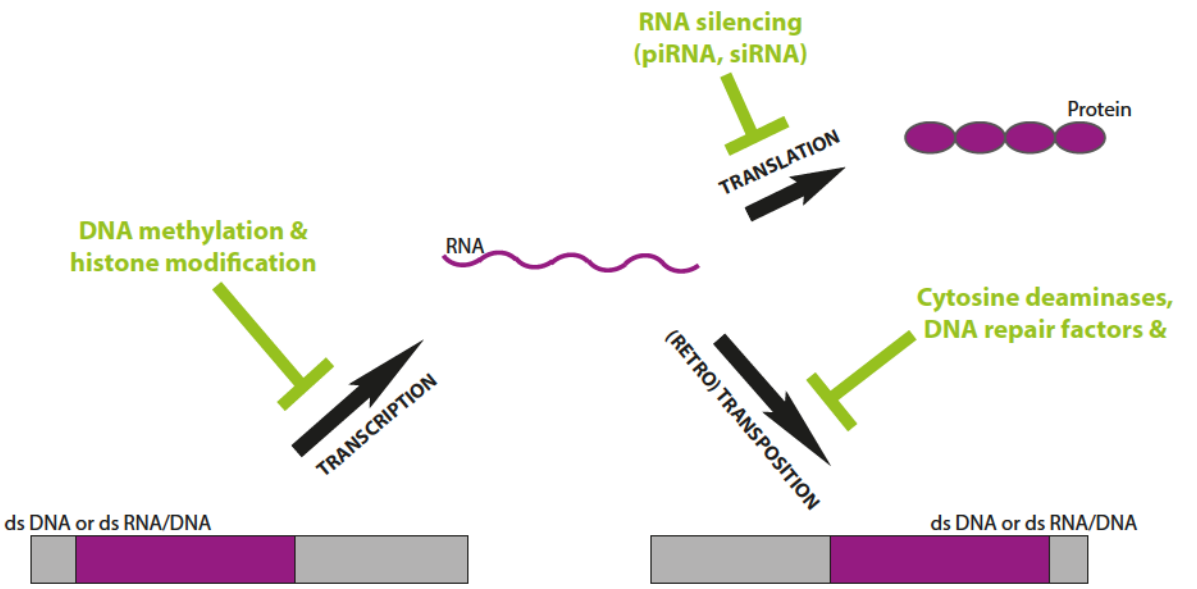


Figure 15: Defense mechanisms acting to restrain TE proliferation in host genomes. Mechanisms can act at different levels: transcription can be inefficient due to DNA methylation or histone modification, small RNAs can destroy the TE messenger RNA to avoid its translation, and finally some factors can act to limit TE insertion into the host DNA. Host DNA is colored in grey and TE related DNA, RNA or protein are purple.

DNA methylation and chromatin modifications: inhibition of TE trancription

DNA methylation commonly corresponds to the addition of a methyl group to cytosine residues. The process is called cytosine methylation (CM). Most of CM occurs within TEs to repress their transcriptional activity in somatic and germline cells (for review, see (Levin and Moran 2011; Rigal and Mathieu 2011)). It has been suggested that CM might be involved in epigenetic silencing of TEs at specific stages during mammalian development (Bourc'his and Bestor 2004). Moreover, it has been demonstrated that TEs, and other types of repeats, represent the most highly methylated sequences in *Arabidopsis*, mouse and human genomes (Zilberman and Henikoff 2007; Lister *et al.* 2008; Rigal and Mathieu 2011). *De novo* DNA methylation, occurring at each generation, is mediated by DNMT3 methyltransferases (and its orthologs in various species). The loss of *Dnmt3l* in mouse leads to the loss of *de novo* CM and a reactivation of TE expression (Webster *et al.* 2005). After establishment, methylation is maintained by DNMT1 methyltransferase. If genes involved in the process of CM have been identified, there is still a lack of information concerning the specific TE targeting and recognition signals. GC methylation view has been recently challenged by the discovery of non-GC methylation (Lister *et al.* 2009), which is probably more common than thought (found in mammals and plants). Methylation can occur in different context as GC (classical view), CHG (H corresponds to A or T) and CHH. In plants, both *CMT3* and *DRM2* genes ensure the persistence of these methylation forms.

Other modification involves histone modifications, especially histone tails and chromatin condensation alterations. Nucleosomes associated with TEs are enriched of histone H3 at lysine 9 (H3K9), which is a signal for transcriptionally repressive and inactive chromatin.

Small RNAs and TE silencing: inhibition of TE translation

Small RNA-based mechanisms include two main pathways that are endogenous small interfering RNAs (endo-siRNAs), generally occurring in somatic cells, and PIWI-interacting RNAs (piRNAs), generally acting in germline cells (Carthew and Sontheimer 2009; Malone and Hannon 2009). These two types of small RNAs differ in their biogenesis, their target and their mode of action, but they both share their small size (20-30 nt), their association with Argonaute proteins and their role in silencing TEs at the post-transcriptional level. Endo-siRNAs have the particularity to stop post-transcriptional activity through the targeting TE mRNA. Reaction starts on a double-stranded RNA (dsRNA), which is processed into 21-24 nt endo-siRNA by members of the Dicer family. The endo-siRNA is loaded on Argonaute (AGO) proteins and a strand is degraded to obtain a single-stranded RNA (ssRNA). The complex formed by AGO and ssRNA is called RISC for "RNA-induced silencing complex". RISC targets the complementary TE mRNA sequence, leading to its degradation. Genomic sources of dsRNA trigger can originate from a bidirectional transcription of complementary strands, but also from inverted repeats of DNA transposon mRNA (TIRs of DNA transposons). Moreover, it has been shown in *C. elegans* that these dsRNAs can be produced by the RNA-dependent RNA polymerase (RdRP) process, which copies ssRNA template into dsRNA. Interestingly, RdRP proteins also participate to TE-mediated heterochromatin production, suggesting that siRNA pathways link both TE silencing and DNA methylation (Hamilton *et al.* 2002). However, the process by which TEs are recognized by the pathway is incomplete. In flies, expression of transposon mRNA increases in Dicer and Argonaute mutants, showing the importance of these two protein families in the RNAi process against TEs. This has first been demonstrated in *C. elegans* (Tabara *et al.* 1999).

Contrary to the siRNA mechanism, which starts with mRNA as template, piRNA can be generated from genomic loci that encode long precursor RNAs. piRNAs, which are 24-31 nt in length, bind Piwi proteins belonging to the Argonaute family (Argonaute proteins are subdivided into Argonaute and Piwi clades). Piwi clade comprises Piwi (or MILI in mice and HILI in human), Aubergine (MIWI and HIWI1) and AGO3 (MIWI2 and HIWI 2 plus HIWI 3 in human). The Piwi proteins are mainly expressed in germ line cells, or in somatic cells in close contact with germ line cells, and are required for the maintenance of the germ line cells in both testes and ovaries of *Drosophila*. In mouse, the three piwi proteins are non-redundantly expressed during spermatogenesis. Beside a developmental role, Piwi pathway was pointed to play also a role in silencing of TEs.
Most piRNA clusters generate small RNA from both genomic strands. Piwi proteins bind to mature antisense piRNA and direct them to complementary sequences in TE-mRNA. Piwi proteins have then the capacity to cleave the TE-mRNA leading to the release of a sense-strand piRNA interacting with other Piwi proteins. The piRNA precursor can then reiterate the cycle and initiate an amplification loop. This cycle process is called Ping-

Pong mechanism (Aravin *et al.* 2007; Slotkin and Martienssen 2007; Levin and Moran 2011). It is important to mention that piRNAs from germ line and from somatic cells use different pathways. The first one is Dicer-independent and amplified through the so-called Ping-Pong process, while the second is Dicer dependent with no amplification. The link between TEs and piRNA defense has been identified in *Drosophila*, in which an increase of TE expression also leads to an increase of piRNA production (Castillo *et al.* 2011).

Cytosine deaminases and DNA repair factors: inhibition of (retro-) transposition

Some cellular processes play a defense role during the last step of TE (retro-) transposition. First, it has been shown that the APOBEC3 protein deaminates cytidines to uracils during the synthesis of the first strand of cDNA during retrotransposition. This leads to cDNA degradation, or mutations in the sequence. This system has been shown to fight against viruses by inhibiting their replication (Aguiar and Peterlin 2008; Chiu and Greene 2008). The products of other genes, such as *TREX1* in human, may limit retrotransposition. However, mechanisms involved are not known yet. In mammals, LINE1 elements are generally 5' truncated, which makes copies unable of autonomous retrotransposition. It has been proposed that host may favour the dissociation of the L1 reverse transcriptase from the nascent cDNA before the end of the process, due to the uncomplete sequence (Gilbert *et al.* 2005; Babushok and Kazazian 2007). Proteins involve in non-homologous end-joining pathway of DNA repair seem to restrict retrotransposition because the TPRT process cannot be completed without the 5' end. In the same idea of getting an element damaged, a so-called RIP (Repeat induced point-mutation) system has been described in *Neurospora crassa* (Selker 2002). RIP is a very efficient mechanism tending to hyper-mutate repetitive DNA and to change the information content.

VI- TE-based genetic tools in vertebrates

The capacities of TEs to move, and sometimes to transport pieces of DNA have led researchers to develop and adapt them for several biological applications. TE-based techniques were initially developed in invertebrates, plants and prokaryotes. Indeed, TE insertion capacities were used for plant transformation for instance to increase resistances to pathogens. Collections of P-element-inserted mutants also exist for fruit fly genetic studies. The random insertion of P-elements allows to characterize gene functions in fruit fly by creating for example null-mutants. Furthermore, P-elements are also used as enhancer and for gene trap experiments (Hummel and Klambt 2008).

Nowadays, the different properties of TEs are more commonly developed to carry out experiments in vertebrates. Among class I, both non-LTR and LTR retrotransposons have been used. LINE sequences have been utilized for gene transfer and mutagenesis

screens in mice (MORAN *et al.* 1996; O'DONNELL *et al.* 2013). Viruses and retroviruses have been developed as vectors for transfection assays due to their infection ability. However, this infectious capacity makes them potentially risky vectors for the host. Among class II, the reactivation of the salmonid DNA transposon "*sleeping beauty*" (SB) has opened wide applications for vertebrate research. SB is a synthetic element from the Tc-Mariner superfamily reconstructed from alignment of non-functional elements reconstructing a functional ancestral sequence. The SB system was initially used in transgenesis and insertional mutagenesis and is now an undeniable powerful tool to establish stable transgenic lines (IVICS *et al.* 1997; IVICS *et al.* 2004). Other DNA transposon-based tools have been developed, as Tol2 in zebrafish, xenope, chicken and mammals, and PiggyBac (WU *et al.* 2006; KAWAKAMI 2007; LI *et al.* 2013).
All these vectors can be used in cells and *in vivo* experiments for somatic and germline transgenesis, involving loss or gain of function. Moreover, the engineering of induced pluripotent stem cells after transposon-genetic reprogrammation is a new medical field with great expectancy. In the medical context, TEs also provide very interesting tracks for gene therapy and disease treatment. The aim is to replace or complement a defective allele by a functional one, or to over-express a protein, for which the activity would have a therapeutic impact.

VII- TEs in vertebrate genomes

a. TE diversity

With about 64,000 identified extant species and by making almost 4% of all described animal, vertebrates represent one of the largest phyla among chordates. Vertebrates are divided in jawless and jawed lineages (Figure 16). Jawless lineages include the extant agnathes with the sea lamprey and the extinct placoderms. Jawed vertebrates are themselves divided into chondrychtians (cartilaginous fish), actinopterygians (ray-finned fish) and sarcopterygians (lobe-finned fish, amphibians, reptiles and mammals) (Figure 16). Vertebrate body size ranges from 7.7mm (a frog) to 33 meters (the blue whale). As comparison, with millions of extant species, land invertebrate sizes range from 1mm to 10 cm. Vertebrates also present a very wide range of genome sizes, ranging from 385 Mb (*Tetraodon nigroviridis*) to 130 Gb (*Protopterus aethiopicus*). This makes them interesting models for comparative genomics.

Vertebrate genomes harbour a wide range of TE superfamilies from all described classes, subclasses and orders, and almost all superfamilies have been identified at least in one vertebrate species. Most of the information describing TE diversity, in vertebrates, results from single genome analyses, as performed for the human genome (PRAK AND KAZAZIAN 2000; LANDER *et al.* 2001; CORDAUX AND BATZER 2009). In the context

of genome sequencing consortium, TEs are identified to mask genomes and to perform gene annotation, but they are not necessary further analyzed. Since few years, many studies have investigated TE evolution and diversity in one given genome, as it was deeply done for human. For instance, LTR retrotransposons dynamics have been studied in salamanders, in order to highlight their potential role in genome gigantism (Sun *et al.* 2012b); non-LTR retrotransposon diversity and elimination was investigated in stickleback (Blass *et al.* 2012), and general TE diversity was approached in several species as lizard (Tollis and Boissinot 2011), opossum (Gentles *et al.* 2007), chicken (Wicker *et al.* 2005) or lungfish (Metcalfe *et al.* 2012).

Combined to single genome analyses, some authors have provided comparative TE composition in various genomes or lineages. For example, the distribution of DNA transposon superfamilies has been investigated within Eukaryotes (Feschotte and Pritham 2007), showing that Tc-Mariner, MuDR, hAT, PiggyBac, PIF, Merlin, P, Transib, Helitrons and Mavericks are each present at least once in vertebrates. Similarly, the evolution and distribution of non-LTR retrotransposons have been described in various eukaryotic lineages but only briefly in vertebrates (Novikova and Blinov 2009).

Among vertebrate lineages, the diversity of retrotransposons has been compared between fish and mammals (Volff *et al.* 2003), suggesting that fish genomes contain a higher diversity of retrotransposons than mammals. Other studies focused on the distribution of particular superfamilies or families, such as Rex1 and Rex3 in fish (Volff *et al.* 2000; Volff *et al.* 2001b), Rex6 in vertebrates (Volff *et al.* 2001a), Mavericks in eukaryotes (Pritham *et al.* 2007), as well as Helitrons in vertebrates (Poulter *et al.* 2003).

Despite the high number of studies focusing of TE diversity in one particular species, or on the distribution of one particular TE family among many lineages, no study has attended to show the diversity and content of all TE superfamilies in all vertebrate lineages yet. A general view of TE content and diversity in vertebrates is important to understand their differential success between and within lineages. It is for instance interesting to try to understand how many TE superfamilies are present in fish genomes successfully invaded them (even at low copy number) while few superfamilies have successfully invaded mammalian genomes where they created a low diversity TE landscape. TE success can also be compared within lineages. Indeed, LINE1 have widely spread out in all mammals, except in megabats (Cantrell *et al.* 2008) and in a group of muroids (Casavant *et al.* 2000; Grahn *et al.* 2005). It has been shown that the extinction of LINE1 in a group of South American rodents has followed the expansion of a retroviral element in the genome (Erickson *et al.* 2011).

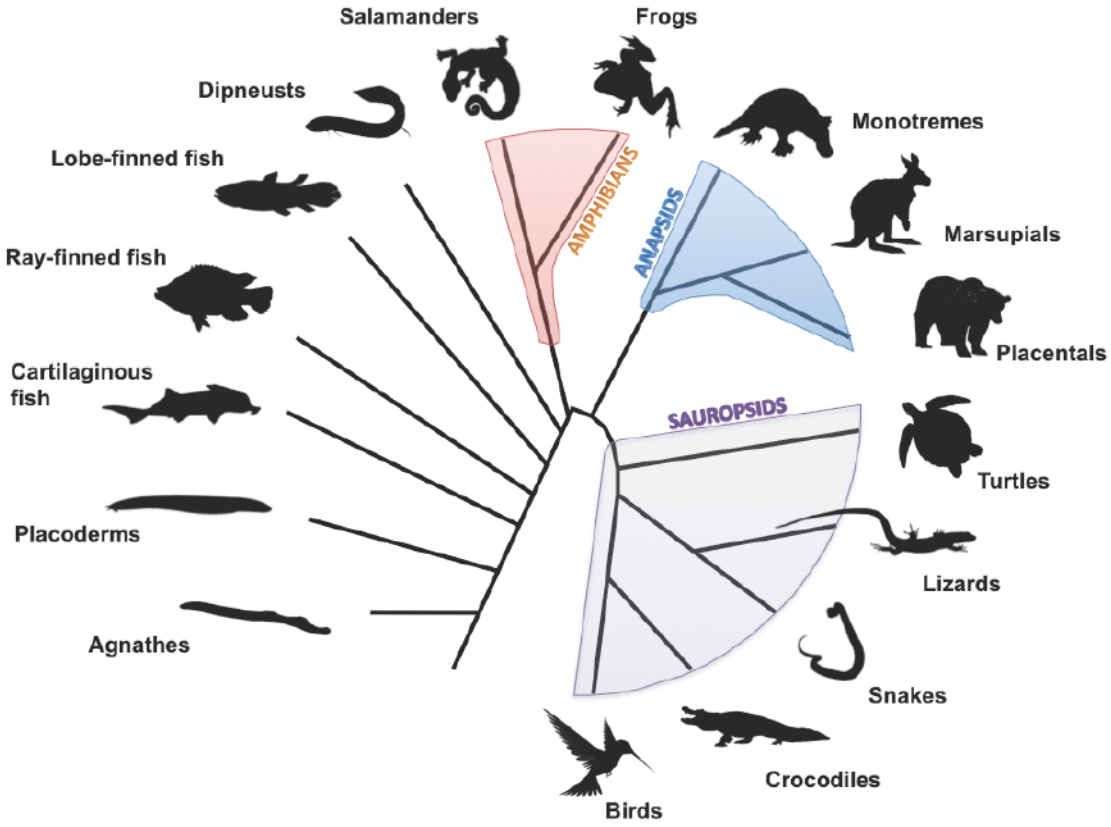


Figure 16: Simplified vertebrate phylogeny that includes extinct and extant jawless lineages (placoderms and agnathes) as well as extant jawed vertebrates: cartilaginous fish with rays and sharks; ray-finned fish mostly composed of teleost fish; sarcopterygians including coelacanth (lobe-finned fish), dipneusts (lungfish), amphibians (salamanders and frogs), anapsids that include mammals, and sauropsids (also named reptiles) that include birds, crocodiles, snakes, squamates (lizards) and turtles.

b. TE-derived sequence diversity

As mentioned in the "evolutionary impact of TEs on host genomes" part, TEs are an important source of genetic novelties. Molecular domestication, or "exaptation", recently became an attractive field, as shown by the multiplication of studies describing TE-derived sequences. Among these TE-derived sequences, regulatory sequences, exon-derived sequences and TE-derived genes have been described. While regulatory sequences, like enhancers or promoters, modulate gene expression, TE-derived genes allow the acquisition of new functions for host, somehow related to the TE domain properties from which it derives. To have an idea of how much TE-derived are important in the human genome, it has been estimated that 25% of the human promoter regions and 4% of human exons contain TE-derived sequences (NEKRUTENKO AND LI 2001; JORDAN *et al.* 2003).

TE-derived regulatory sequences

The advantage of TEs to become regulatory elements in host is that they already contain cis-regulatory elements as promoters, splice sites, termination sites, enhancers and silencers, which are all necessary for the control of their expression. Moreover, thanks to their capacity to spread over genomes, they often present a genome-wide distribution synonym to a high number of new potential regulatory elements. All of this suggests that TEs are well equipped to participate to host gene regulation. In a different manner, LTRs and SINEs retrotransposons prominently harbour TF binding sites compared to other classes (Thornburg *et al.* 2006; Rebollo *et al.* 2012).

It was shown that 18.1% and 31.4% of total transcription start sites, are located within TE sequences in mouse and human, respectively (Faulkner *et al.* 2009). A deep analysis of seven trancription factors, binding over mouse and human genomes shows that five of them preferentially bind to distinct TE families of retroelements and SINEs (Bourque *et al.* 2008). However, the exaptation process has not been demonstrated here. Among interesting cases of exaptation, Bejerano and coworkers (Bejerano *et al.* 2006) identified a LF-SINE co-opted as enhancer and conserved over 410 Mya. This exapted sequence presents regions ultra-conserved between very distant species, such as coelacanth, lungfish, amphibian, birds and mammals.

TE-derived domains found in TE-derived genes

As illustrated in Figure 17, genes can derive from different classes of TEs and from different domains. These domains can be entirely or partially conserved. A TE domain can be merged with an unrelated domain, such as *SETMAR* (Robertson and Zumpano 1997); the derived sequence can conserve two different domains from the same TE element, such as *NYNRIN* (Marco and Marin 2009); or a single TE can be co-domesticated (two different MD events) to create two different genes as *Harbi1* and *Naif1* (Sinzelle *et al.* 2009). Some domains were found to form preferential material for molecular domestication, in particular transposases from DNA transposons, or retrotransposons Gag, protease, INT, RT, RH and envelope genes (Figure 17).

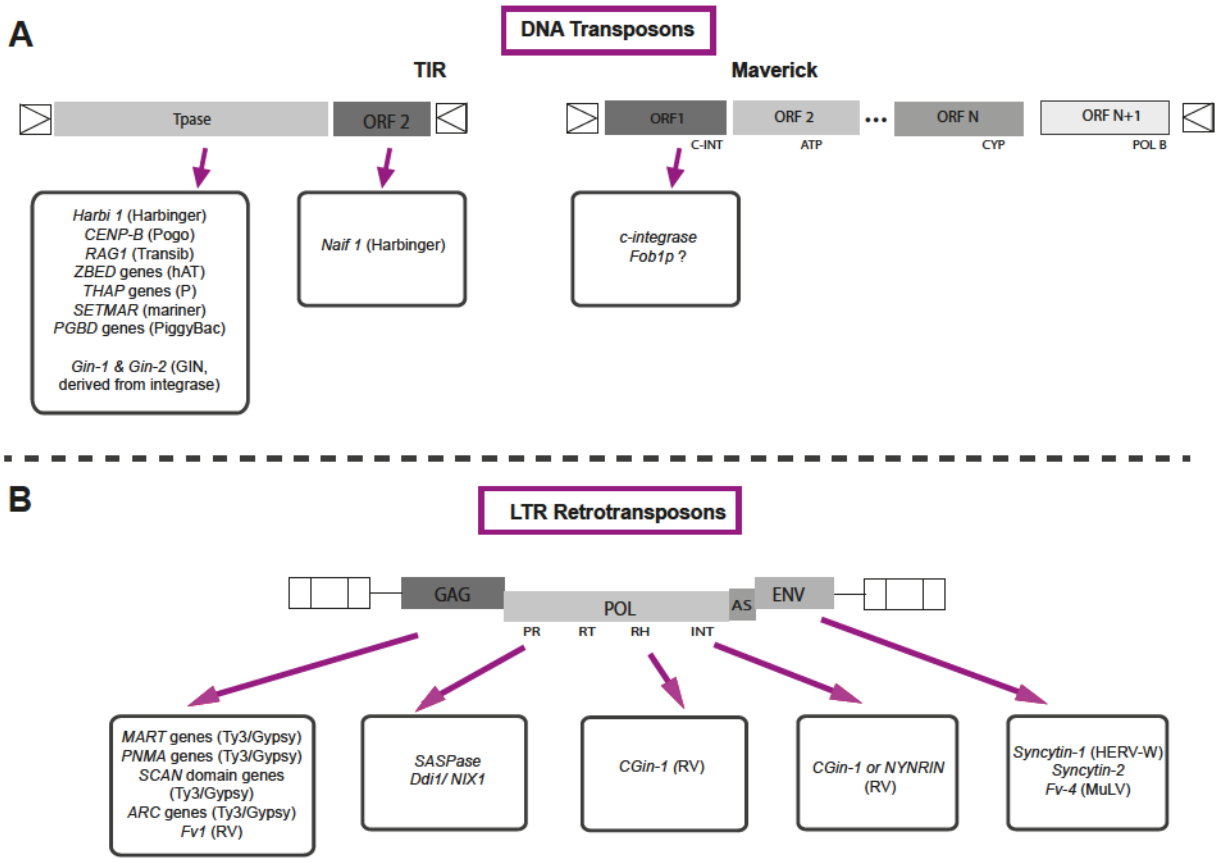


Figure 17: Examples of genes derived from DNA transposon (A) or LTR retrotransposon (B) domains.

TE sequences that are recruited allow to bring new functions for host, probably in relation with the TE-derived domain properties. For example, transposases carry a DNA Binding Domain (DBD) and a catalytic domain that might be helpful for DNA cleavage and joining reactions. All domains of the proteins are however not always conserved. Many TE-derived genes have been conserved over very long periods and are still under strong selection pressure, suggesting that they display fundamental roles. Among transposase-derived genes, *THAP7* still contains a DBD domain and displays regulatory properties mediated by chromatin modifications (MACFARLAN *et al.* 2005). Other *THAP* genes, such as *THAP0* and *THAP1*, have apoptosis-related functions or are involved in cell-cycle control. TE-derived genes that retained the catalytic DDD/E domain might have kept the capacity to mediate transposition. Among them, *Buster*, *Harbi1* and *Rag1* genes still present this domain. However, only the Rag1 protein, which is involved in the V(D)J recombination and thus plays a fundamental role in the adaptive immune system, was shown to be able to achieve transposition (AGRAWAL *et al.* 1998; HIOM *et al.* 1998; KAPITONOV AND JURKA 2005). Another important example of domesticated gene is *CENP-B*, which controls the formation of centromeres via chromatin structure regulation (TUDOR *et al.* 1992; KIPLING AND WARBURTON 1997).

Belonging to DNA transposon domains, transposases and integrases have almost identical functions. GIN transposons even encode a transposase derived from an integrase from LTR retrotransposon (MARIN 2010). Two genes derived from GIN transposons have been identified in different vertebrate lineages, however their respective roles have not been determined yet. Derived from a DNA transposon integrase, Fob1p protein is involved in rDNA metabolism in baker's yeast (DLAKIC 2002; GAO AND VOYTAS 2005). *C-integrase* gene also derived from a DNA transposon integrase, probably from a Maverick element (FESCHOTTE AND PRITHAM 2005). This gene is present in a wide range of species including various vertebrates and worms. Nowadays, integrase-derived genes are poorly described and their functions still remain unknown for most of them.

Numerous genes and multigenic families were identified as derived from LTR retrotransposons (including retroviruses). The majority of them corresponds either to Gag- or Envelope-derived domain and are specific to mammals. The MART family belongs to the Gag-derived multigenic family. It is composed of 11 genes mostly located on the X chromosome (BRANDT *et al.* 2005). Two of these genes, *Peg10* and *Rtl1*, are subject to paternal imprinting and have essential non-redundant roles in placenta development (ONO *et al.* 2006; SEKITA *et al.* 2008). A second Gag-derived large gene family is the *Ma,* or *Pnma* family, for "paraneoplastic Ma antigens". Some *Ma* genes have been proposed to be involved in apoptosis (SCHULLER *et al.* 2005). Other Gag-derived genes might be implicated in cell proliferation control and apoptosis, like *Pnma1,* but also genes from the *SCAN* family (EMERSON AND THOMAS 2011). Some of them might also control replication. Interestingly, mammals anciently turned a viral infection to their advantage by domesticating two *syncytin* genes and the *Fv-4* gene from envelope genes. The *syncytin* genes have independent origins in human, and multiple functions, as important as the formation of placenta (MALIK 2012). Syncytins, which are essential to diverse mammalian lineages, do not have a common origin in the ancestor of mammals: multiple convergent molecular domestication events happened in the different lineages (DUPRESSOIR *et al.* 2005; HEIDMANN *et al.* 2009), such as primates, muroids, cavids or ovis.

Finally, few other genes were identified to derive from proteases, such as the SASPase (BERNARD *et al.* 2005), which plays roles in epidermal differenciation and desquamation.

VIII- Aim of the thesis

With the aim to study the evolution and the evolutionary impact of TEs in vertebrate genomes, especially fish genomes, distinct biological questions were asked at the beginning of the thesis, to conduct to the results presented thereafter. The questions assessed during the thesis are the following:

- What is the TE contribution to genomic and biological diversity, observed between fish species and between vertebrate lineages?
- What are the modes of evolution and transmission of TEs in vertebrates?
- What is the infectious history of retroviruses in vertebrates?
- What are the functions of domesticated (TE-derived) sequences in fish?

All these questions are closely connected and are based on the necessity to acquire a new knowledge on TE diversity and content knowledge in vertebrate genomes. Taking advantages of the booming of fish models and the increase of genome sequencing projects, *in silico* approaches performing large-scale comparative genomics were used to assess the three first questions. TE analyses were initially performed in various fish genomes (presented chapter 2) and then extended for comparison to other vertebrate genomes including tetrapods (mammals, birds, squamates, amphibians) and cartilaginous fish, as well as non-vertebrate chordates (Ciona, lancelet) (presented in Chapter 3). These comparisons were performed to highlight common and/or species or lineage-specific characteristics of TE evolution (content and diversity), but also to retrace the evolutionary history of some TEs (like the LINE order history presented in Chapter 3). Linked to the second and third questions, this analysis was also used to find potential HT events and assess retrovirus diversity. Regarding the last question, I focused my study on one particular gene, called *Gin-2*, which derives from an integrase domain. Both *in silico* and functional analyses were performed to determine its evolutionary history and its expression pattern in fish (presented in Chapter 4).

CHAPTER 2: ANNOTATION OF TEs IN SEQUENCED FISH GENOMES

I- Methods and presentation of the projects

a. Presentation of the fish models

With about 30,000 species, fish almost represent 50% of all extant vertebrate species (NELSON 2006; NEAR *et al.* 2012). Fish is a large term that groups together species with gills able to live under water. It is not a monophyletic group, most of the species are actinopterygians but few are sarcopterygians, like coelacanths and lungfish (Figure 16 and 18). Except 50 species, fish are mostly represented by teleosts including commercially important species such as the Atlantic salmon and the rainbow trout, important models in development, such as the zebrafish and the medaka, genetics and genomics, such as the pufferfishes, cancer research, such as the platyfish. Teleost lineage shows the most impressive radiation in vertebrates, which is highlighted by an exceptional diversity in terms of body shape, body size, coloration, habitat, behaviour, ecology, social system, genome size, mode of reproduction … This diversity offers a good situation to perform comparative genomics. Furthermore, recent advances in genomics allow to better investigate genomic diversity of this lineage. Indeed, fish recently became attractive models in many fields, leading to an increase of sequencing genome projects and so the production of a lot of data helpful for our projects.

The first five fish genomes completely sequenced were: the Japanese pufferfish *Takifugu rubripes* (APARICIO *et al.* 2002), the tetraodon *Tetraodon nigroviridis* (JAILLON *et al.* 2004), the zebrafish *Danio rerio* ((HOWE *et al.* 2013); sequenced but not published before 2013), the medaka *Oryzias latipes* (KASAHARA *et al.* 2007) and the stickleback *Gasterosteus aculeatus* (JONES *et al.* 2012). Four of the five genomes belong to the Percomorph branch (Figure 18), showing a scarcity of genomic data among fish. Thanks to the decrease in cost and time of sequencing, the number of genome projects largely increased. Nowadays, new fish genomes are available and others are in ongoing process. Among the recently available fish genomes, we can find the platyfish *Xiphophorus maculatus* (SCHARTL *et al.* 2013), the tongue sole *Cynoglossus semilaevis* (CHEN *et al.* 2014), the Atlantic cod *Gadus morhua* (STAR *et al.* 2011), the rainbow trout Oncorhynchus mykiss (Berthelot *et al.* 2014, in press) the tilapia *Oreochromis niloticus* (not published), the cave fish *Astyanax mexicanus* (not published), the killifish *Nothobranchius furzeri* (not published) and the spotted gar *Lepisosteus oculatus* (not published). All of them are dispersed over the fish phylogeny (Figure 18). Beside these available data, many other projects are now in progress, such as the Atlantic salmon, the channel catfish, the Amazon molly or the guppy, and we should not forget the genome 10K project launched by the BGI center, which also announces the sequencing of the gulf toadfish, the Atlantic herring, the fathead minnow and nine other fish species.

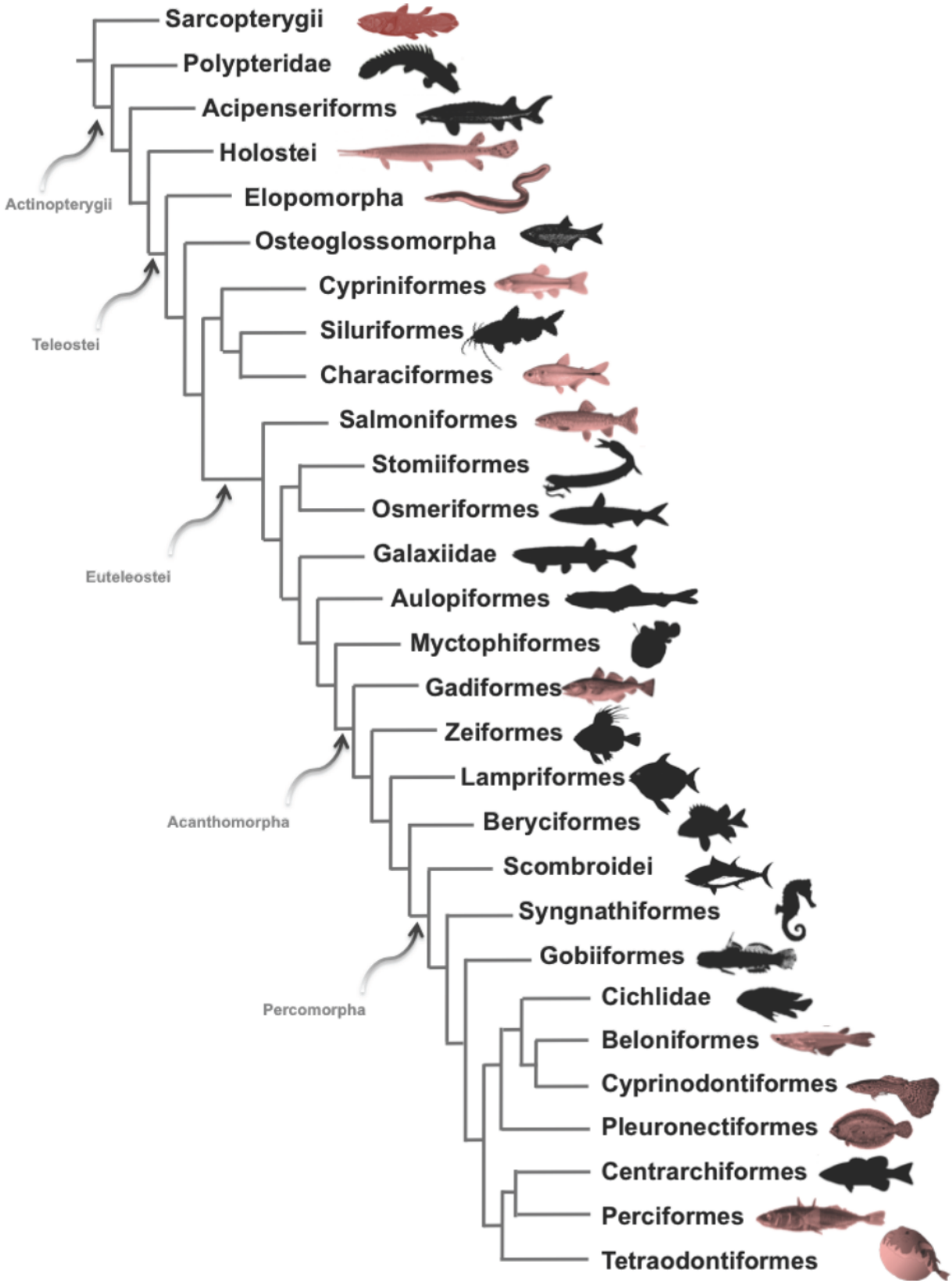


Figure 18: Simplified fish phylogeny including coelacanth and actinopterygians. The phylogeny describes the different fish families. All fish families are represented by a representative icon, which is red balanced for lineages in which at least one species has been sequenced.

b. *Methods to annotate and analyze transposable elements*

For long, TEs were not of interest in genome projects and were poorly studied. It is now obvious that they are major actors in genome evolution. In this context, it is important to evaluate the diversity and the abundance of TEs in genomes. At the moment, many softwares and tools, developed to study TEs, exist and have been tested and listed (LERAT 2010). Softwares can be classified in different types of uses: *Ab initio* detection, classification, feature detection and identification, and folding prediction.

To annotate assembled genomes, we combined both manual and automatic annotations. This strategy allows to avoid the problem of missing families in automatic libraries, and to search for potentially complete sequences characterized by all structural features described in the introduction. Vertebrate genomes are larger than those of insects or bacteria, making the analysis of repetitions highly difficult. Manual annotation, which leads to a library called Lib_v1 (Figure 19) corresponds to the search for the different superfamilies (one by one) using: 1- Blast similarity analysis against known TE protein as reverse transcriptases or transposases, using censor, a software that detects TE similarity against repbase, and 2- by searching for characteristic features such as LTRs or TIRS with Blast2seq. Automatic annotation was performed using RepeatModeler (Smit and Hubley, *RepeatModeler Open-1.0* 2008-2010) (or RepeatScout only in case of no result with RepeatModeler), this leading to a library called Lib_v2 (Figure 19). This library contains consensus sequences from repeated DNA, meaning that all types of repeats are included (low complexity regions, satellites, TEs and non-identified sequences). The two manual and automatic libraries were combined, removing redundancies, and unclassified sequences, i.e. sequences that could not be assigned to any TE superfamily, were subjected to a reannotation. Redundancies can be discarded using BlastN analyses or using the CD-Hit software (LI AND GODZIK 2006). As MITE sequences are not annotated in any of the two annotations, a secondary structure prediction can help to identify them. We considered only small sequences, presenting a folding hairpin structure. To complete the annotation, we also searched for similarities between unclassified sequences and our combined teleost library, which contains TE sequences from other annotated fish genomes. The final library (Lib_v3, Figure 19) was first used to mask the studied genome, i.e. replaced repeated sequences by N's stretches, in order to perform and facilitate gene annotation in the context of a genome project. However, the masking process allows us to do TE analyses. Indeed, we can first evaluate the copy number and the coverage of the respective families in the genome and compare these data to other fish genomes. Then, we estimated the age of the TEs in the genomes. The distance (the number of mutations) between the TE copies in the genome and their corresponding consensus sequences in the TE library indicates the potential age of TEs. To correct for multiple mutations at the same site, we used the Kimura distance estimating the age of TEs (KIMURA 1980). Kimura estimations take into account the

proportion of transversions (corresponding to purine-purine or pyrimidine-pyrimidine mutations) and transitions (purine-pyrimidine mutations).

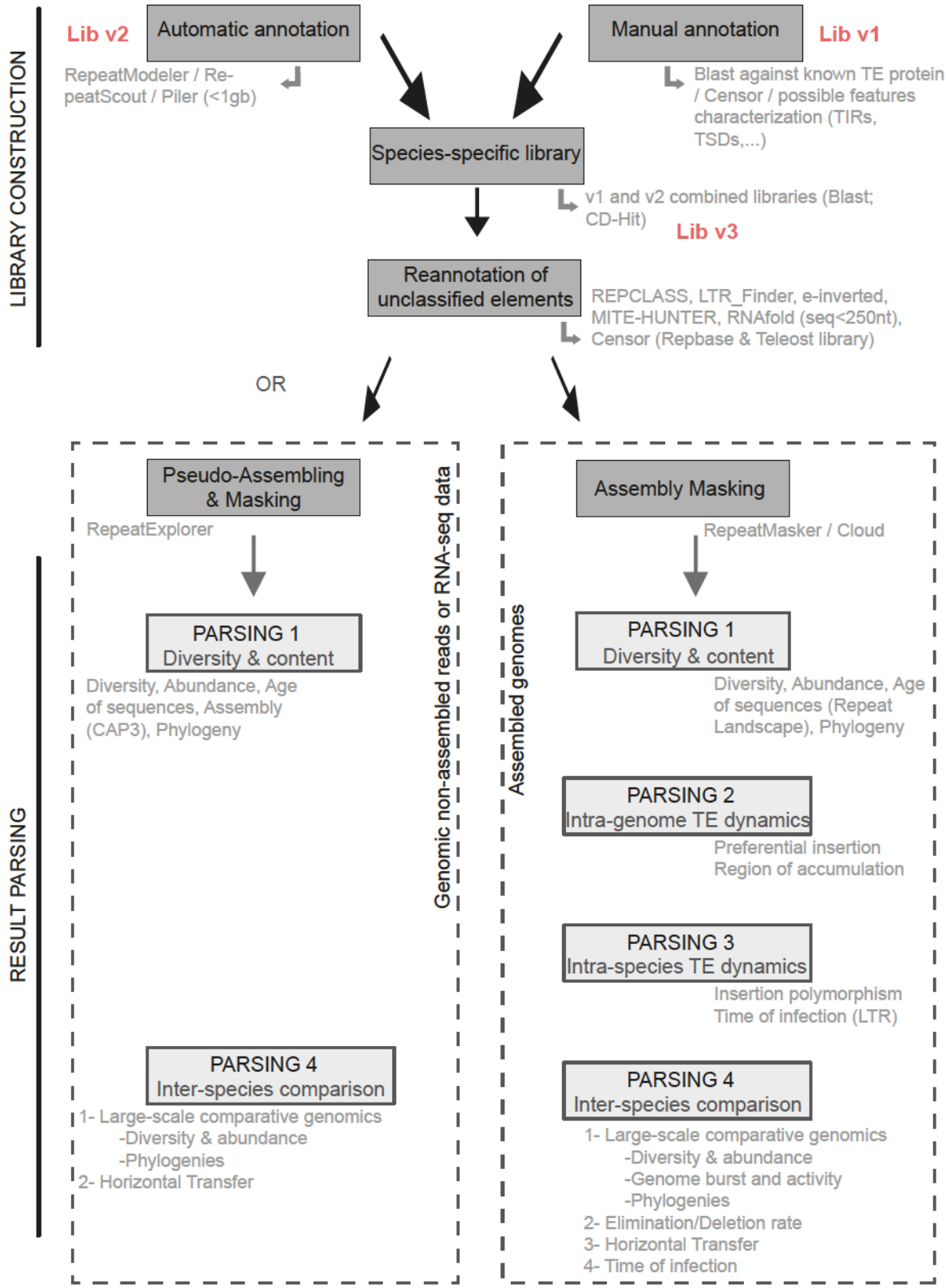


Figure 19: Protocol to annotate and analyze TEs in genomes. Upper panel details methods to annotate TEs using both manual and automatic annotation with description of softwares that we used at each step. Lower panel shows of genome analyses and result parsing, in both assembled and non-assembled genomes.

To summarize, the analyses that we routinely performed to characterize TEs in vertebrate genomes are:

- Library building (automatic, manual)
- Genome masking
- Determination of copy number and genome percentage of superfamilies
- Analysis of repeat landscape using Kimura distances: genome history by evaluating burst of activity
- Phylogenetic reconstructions of TE superfamilies

These analyses can be used to (not exclusive list):

- Perform large-scale comparative genomics
- Determine preferential insertion sites or regions and zones of accumulation in genomes
- Detect insertion polymorphism between individuals or species helping to infer TE activity, age of insertion, rate of elimination, leading to more details about TE dynamics
- Detect HT events

c. *Presentation of the different fish genome projects*

During my thesis, I had the opportunities to participate in several genome projects, considering mostly fish genomes dispersed over fish phylogeny (Figure 18), with the aim to annotate and analyze transposable elements. For some of these projects, my role was to perform the annotation and construct a species-specific TE library, which was necessary for the masking step before assembly and genome annotation. For others, I pushed further the analyses and did comparative genomics, or even transcriptomic analyses in order to evaluate TE quantities in transcriptomes and have an idea on their potential activity. In this chapter, I will describe some of the projects, but not all (to avoid a catalog), in which I have been strongly involved. First of all, following is the list of the different projects with collaborators:

- The southern platyfish, *Xiphophorus maculatus*, Cyprinodontiformes, Poeciliidae (with Manfred Schartl, University of Würzburg, Ron Walter, Texas State University and Wes Warren, the Genome Institute at Washington University) ➔ **Model in several fields such as genetics, cancer, sex chromosomes, speciation**
    - Results are developed in the “Poeciliid genome projects” part and the published paper is available in **ANNEXE 1**.

- The coelacanth *Latimeria chalumnae*, Sarcopterygian, coelacanthiformes (with Chris Amemiya, Benaroya Research Institute, Seattle and Jessica Alföldi, Broad Institute in Boston) ➔ **Species phylogenetically positioned at the fish to tetrapod transition, key location to understand vertebrate evolution**
    - Results including genomic, transcriptomic and insertion polymorphism (three research papers) are developed in the "Coelacanth genome evolution" part and the genome paper is available in **ANNEXE 2** (Figure 20).

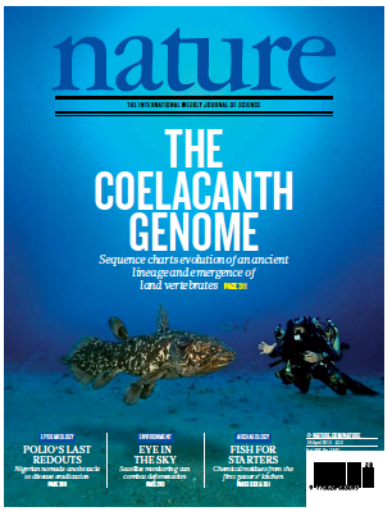


Figure 20: Cover page of the Nature issue containing the coelacanth genome paper.

- The tongue sole *Cynoglossus semilaevis*, Pleuronectiformes (with Manfred Schartl in Germany, and Changwei Shao, Qisheng Tang and Jun Wang, Yellow Sea Fisheries Research Institute, Qingdao and BGI, Shenzhen, China) ➔ **Sex chromosome evolution model with a ZW system**.
    - The published paper is available in annexes (**ANNEXE 3**). A single *de novo* library built using Piler and RepeatScout (but no manual annotation) was used to mask the genome. Results showed that TE diversity is very low compared to other fish genomes available at this time. Many families seem to be absent from the first assembly. The genome of the tongue sole contains about 5.8% of TEs, mainly represented by Tc1 transposons and LINE RTE and Babar retrotransposons. There are few LTR retrotransposons in this genome (0.08% compared to 0.6% in the platyfish and 3.62% in the zebrafish), mainly represented by Sushi elements (from the Gypsy family), which are probably not active anymore. Both APE (RTE, L2 and Rex-Babar) and REL (only one family, R2) endonuclease non-LTR retrotransposons were found, the most abundant elements belonging to the RTE and Babar families. Finally, DNA transposons are mainly represented by Tc1 elements. As the tongue sole is a model to study sex chromosome evolution (ZW system as observed in birds), we further investigated TE abundance on them. The density of interspersed repeats of both Z and W sex chromosomes is much higher (~2.3 and ~6.9 times, respectively) than the average on autosomes. On Z chromosome, the most abundant type of TEs is DNA transposons (36.1% of all TEs); while on the W, LINE elements (31.4% of all TEs) are

predominant. This fish is a very interesting model to study the evolution of sex chromosomes, as observed by the strong accumulation of TEs in particular on the W chromosome. This suggests that the W chromosome will probably continue to accumulate repeats, as observed for the Y chromosomes in mammals.

- The rainbow trout *Oncorhynchus mykiss*, Salmoniformes (with Yann Guigen, INRA de Rennes; Edwige Quillet, INRA Jouy-en-Josas; Hugues Roest-Crollius, ENS de Paris, and Genoscope) ➔ **Tetraploid species under re-diploidization process, of economical importance**.
    - The paper is available in annexes (**ANNEXE 4**). Repeats make up about 38% of the genome, with a large proportion of TEs (about 27.7% of the genome). Both retrotransposons and DNA transposons were identified, covering a large diversity of families. Only few vertebrate TE families, such as Helitron transposons or Copia retrotransposons are absent from this genome. With Kimura analyses, it appears that two or three main bursts of transposition occurred in the genome. The most ancient one seems to be mainly due to a high activity of Tc-Mariner families. In the second, an increase of all families, and particularly CR1 retrotransposons, is observed. Finally, the last one shows a second burst of Tc-Mariner elements. We tried – not successfully - to correlate TE bursts with the 4R (four round of whole-genome duplication (WGD) specific to salmonid lineage) WGD event in order to see if TEs may have played or still continue to play a role in re-diploidization process (**ANNEXE 5**).
    - Moreover, interestingly, retroviruses sequences were identified in the MHC region. One of the retrovirus sequences presents high similarity (91% on 360bp) with a VHSV (Viral haemorrhagic septicaemia virus – Rhabdovirus)-induced mRNA of the rainbow trout (Accession number AF483545). An hypothesis is that the retrovirus sequence is expressed in response to the infection by VHSV virus, suggesting that it might be involved in defense reactions against other viruses.

- The cave fish *Astyanax mexicanus*, Characiformes (with Wes Warren and Suzanne McGaugh, Genome Institute at Washington University) ➔ **Blind cave species of evolutionary adaption interest**.
    - The genome paper, which focuses on QTL analyses, is currently submitted. As I have only been recently involved in this project, I am waiting for a more complete version of the assembly to analyze regions of ageing-related interest in order to try to see if TEs might have been involved in particular deletions or rearrangements. The cave fish is the closest species, for which a genome is available, of the zebrafish. Genomes are

composed of 30% and 52% of TEs in cave fish and zebrafish, respectively, both mostly represented by DNA transposons. Both genomes underwent a similar recent amplification of transposons (Figure 21), even if this was much more important in zebrafish. Furthermore, LTR retrotransposons are almost absent from the cave fish genome.

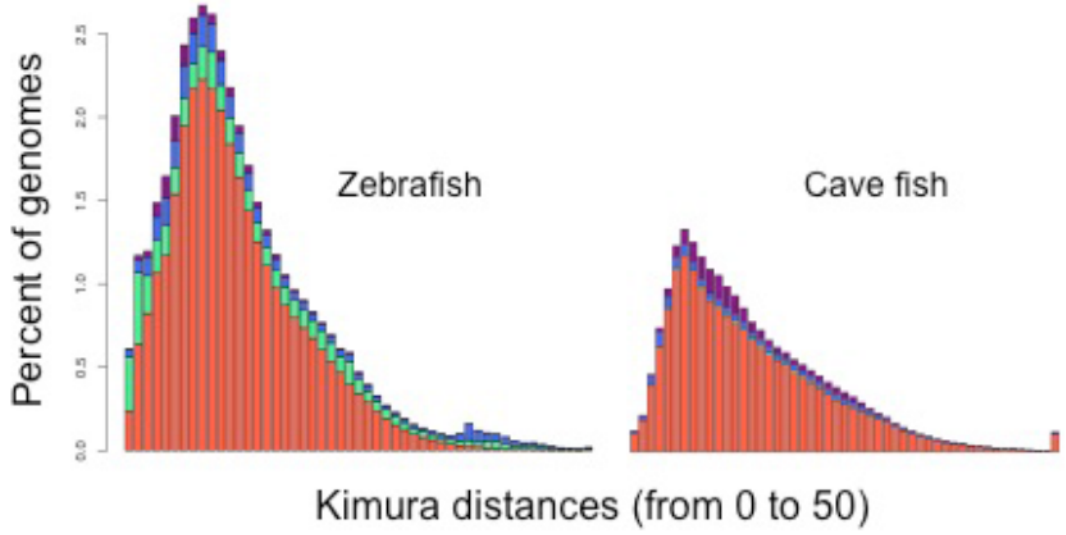


Figure 21: Zebrafish (left side) and cave fish (right side) repeat landscapes. The different classes/orders are represented by orange (DNA transposons), blue (LINE retrotransposons), green (LTR retrotransposons) and purple (SINE retrotransposons).

- The spotted gar *Lepisosteus oculatus*, Lepisosteiformes (with John Postlethwait and Ingo Braasch, University of Oregon, Jessica Alföldi, Broad Institute of MIT and Harvard, Boston) ➔ **"Outgroup" before the teleost-specific whole genome duplication (the teleost-specific WDG is also name 3R WGD)**.
    - The spotted gar is an excellent model by its intermediate phylogenetic position to study the impact of whole genome duplication in teleosts. With about 20% of repetitions (TEs and others), the spotted gar genome has a comparable profile to lobe- and ray-finned fish. As in teleosts, a very high diversity of TEs is observed with almost all previously described TE superfamilies. This genome has undergone two main bursts of TE activity, the most ancient due to Tc-Mariner, CR1 and Deu-SINE elements, and the most recent due to R2, Rex1-Babar and 5S-SINE elements. In recent timing of Kimura analyses, a new burst of Tc-Mariner seems to initiate. Combined to this, transcriptomic data analyses revealed that Tc-Mariner is the most transcribed superfamily followed by the CR1 and Rex1-Babar families.

- The Amazon molly *Poecilia formosa* combined with *Poecilia mexicanus* and *Poecilia latipinna*, Cyprinodontiformes, Poeciliidae (with Manfred Schartl and Wes Warren, Genome Institute at Washington University) ➔ **First asexual vertebrate species to be sequenced and comparative genomics among poeciliids with *Xiphophorus***.
    - Results are developed in the "Poeciliid genome projects" part

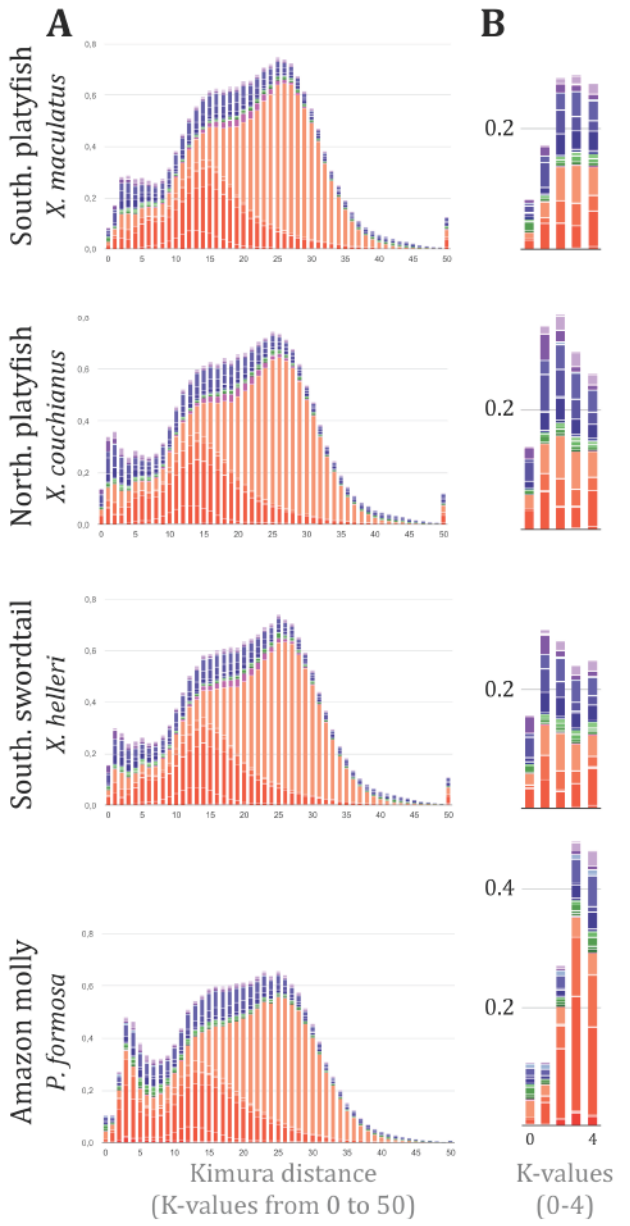
- Two other *Xiphophorus* species: the Northern platyfish *Xiphophorus couchianus* and the Southern swordtail *Xiphophorus helleri*, Cyprinodontiformes, Poeciliidae (with Manfred Schartl, Ron Walter and Wes Warren) ➔ **Comparative genomics between *Xiphophorus* species**
    - Results are developed in the "Poeciliid genome projects" part.

- The killifish *Nothobranchius furzeri*, Cyprinodontiformes (with Christoph Englert and Matthias Platzer, Leibniz Institute for Age Research, Jena, Germany) ➔ **Short-living species developed as ageing model**.
    - This project is a young project, for which I mainly helped to do a complete annotation of TEs, combining automatic annotation on raw reads and on assembly, but also trying to annotate MITE sequences by RNA folding visualisation. I could show that the killifish is composed of almost 40% of repeats. It contains a very high diversity of superfamilies with most of the eukaryotic superfamilies previously described. Moreover, it seems clear that TE superfamilies from several classes have been recently active, in particular hAT and Tc-Mariner transposons, and LINE2, Rex-Babar and RTE retrotransposons.

II- Poeciliid genome projects

We have been involved in several poeciliid genome projects including the Southern platyfish *Xiphophorus maculatus* ((SCHARTL *et al.* 2013) see paper in **ANNEXE 1**), which is a model we used in our lab for sex chromosome evolution studies, and two other *Xiphophorus* species, *X. couchianus* and *X. helleri*, providing us with the occasion to compare diversity, content and evolution of TE families in closely related species that diverged only few million years ago in central America lakes.
We also analyzed the genome of the Amazon molly *Poecilia formosa*, an asexual species that reproduce by gynogenesis (reproductive system in which sperm triggers the developpement of egg cell into an embryo but do not contribute to genetic materials). The Amazon molly is highly interesting as it is the first asexual vertebrate genome to be sequenced. This gynogenetic species, which do not produce any male, is a hybrid species between two other species, *Poecilia mexicana* (female parent) and *Poecilia latipinna* (male parent). The hybridization event has been estimated to occur around 280.000 years ago (LAMPERT AND SCHARTL 2008). The genome of *P. formosa* has been sequenced and assembled, whereas genomes of *P. mexicana* and *P. latipinna* have been only sequenced but not assembled. Having genomic data for these three genomes give us a great opportunity to investigate TE dynamics in an asexual species on the one side, but also to analyze the behaviour of TEs after hybridization event.

I will first present the *Xiphophorus* results, followed by the *Poecilia* results and finally a comparison between *Xiphophorus* and *Poecilia*. First focusing on *Xiphophorus* species, the three genomes of *X. maculatus*, *couchianus* and *helleri* show a very similar diversity and content of TEs, since they are composed of 21.4%, 21.8% and 21.1%, respectively. They mostly contain DNA transposons, with a high diversity of superfamilies (Figure 22D). Their Kimura profiles representing the potential age of sequences, and so superfamily history, also look very similar (Figure 22A), underlying three main bursts of TE activity. Tc-Mariner (K-value 25) and hAT (K-value 14) together with Harbinger (K-value 12) DNA transposons are responsible for the two oldest bursts that make about 0.7 and 0.6% of the genomes, respectively. A focus on low K-values shows differences between the three species, in particular at the two K-values 3 and 4 (different content of the several superfamilies). *X. couchianus* profile presents an increase for K-value 1 and 2, compared to the two others. Finally, *X. maculatus* seems to contain fewer active copies in the current time (K-value 0 and 1), with recently transposed TE sequences representing less than 0.1% of the genome (Figure 22B). Furthermore, the Kimura profiles of the three species present a particular accumulation falling into the K-value 50. This suggests that it might be an accumulation of old sequences, or at least very degenerated sequences.



	AMAZON MOLLY	SOUTHERN PLATYFISH
DNA Transposons	**12,507**	**12,078**
Academ	0,008	0,021
Buster	0,002	0,001
CMC-Chapaev	0,005	0,017
CMC-EnSpm	0,149	0,018
Crypton	0	0
Ginger	0	0,001
IS4EU	0,099	0,052
Kolobok	0,002	0
MULE-MuDr	0,032	0
Novosib	0,008	0
P	0,041	0,04
PIF-Harbinger-ISL2EU	0,567	0,619
PiggyBac	0,339	0,249
TcMariner	6,429	6,495
hAT	3,734	3,286
Helitron	0,031	0,251
Zisupton	0	0,005
Polinton/Maverick	0,052	0,025
MITE	0,036	0,032
Unclassified	0,973	0,966
LINE Retrotransposons	**2,697**	**2,435**
I-Nimb-Jockey	0,213	0,113
LINE1-Tx1	0,064	0,124
LINE2-Maui	1,226	0,905
Proto2	0,01	0
R2	0,006	0,003
R4-Dong	0,074	0,009
RTE-BovB	0,515	0,541
Rex1-Babar	0,545	0,687
Penelope	0,004	0,004
Unclassified	0,04	0,049
LTR Retrotransposons	**0,8**	**0,604**
BEL-Pao	0,034	0,035
DIRS-Ngaro	0,164	0,123
ERV	0,004	0,118
ERV1	0,036	0,01
Gypsy	0,32	0,219
Unclassified	0,242	0,099
SINE Retrotransposons	**0,561**	**0,524**
Hpa	0	0,005
MIR	0,08	0,112
SINE2	0,009	0
SINE3	0	0
V	0,295	0,196
tRNA	0,012	0,067
Unclassified	0,165	0,144
Unknown	6,17	5,588
Total TEs	**22,735**	**21,229**
Simple repeats	1,099	0,907
Low complexity	0,823	0,66
Total repeats	**24,657**	**22,796**

Figure 22: Transposable element superfamilies in the three *Xiphophorus* and the Amazon molly genomes. A- Kimura distances of the different superfamily copies (recent copies on the left, ancient copies on the right). B- Focus on low Kimura distances highlighting recent copies specific to each species. C- TE superfamily contents (percentages) in genomes of *X. maculatus* and *P. formosa*.

Regarding TE content and diversity in *Poecilia* genomes, I only compared TE superfamily diversity between the three *Poecilia* species. Indeed, different methods of analyses and quantification were used because the tools are different for assembled and non-assembled genomes (RepeatExplorer estimation on non-assembled reads for *P. mexicana* and *latipinna*; RepeatMasker on *P. formosa* assembly). RepeatExplorer evaluated a repeat content of around 50% in parental genomes while RepeatMasker estimated 25% of the Amazon molly assembly. The TE content estimation between assembled and non-assembled genomes shows how we probably underestimate TE content in assemblies. However, it is complicated to study non-assembled genomes due

to the high amount of raw sequences that are generally very small. Concerning TE diversity in the three *Poecilia* genomes, the same superfamilies were found with the exception of few missing ones in parents, such as BEL, Nimb, R2 retrotransposons and Crypton, MuDr, Novosib transposons. However, except Nimb retransposons (4200 copie), these superfamilies are present at low copy numbers (less than 1000 copies) in *P. formosa* genome. The fact that few superfamilies are not found in the non-assembled genomes can be due either to the fact that the analyses were performed only on 1% of the raw reads of *P. mexicana* and *P. latipinna* leading to potential information, or to acquisition by horizontal transfer in the hybrid species.

As *P. formosa* has been analyzed with the same methods than *Xiphophorus*, the comparison of *Xiphophorus* and *Poecilia* genomes was possible. Kimura profiles resemble pretty much in the ancient timing but differ in the most recent burst (Figure 22 A, B and C). Indeed, a strong increase in the number of sequences from K-value 0 to 5 can be observed (Figure 22 A and B). A closer look at the K-values 3 and 4 shows that the total TE content can exceed 0.4 % of the *P. formosa* genome. However, for K-value 0 and 1, the activity strongly decreases. The total content of TEs in *P. formosa* genome is 2% higher than Xiphophorus genomes (Figure 22C). This increase might result from the recent observed burst (K-value 3 and 4, Figure 22B). If this increase is species-specific, this might be linked to the hybridization event or to the asexuality system of the Amazon molly.

To analyze the specificity of the recent TE bursts occurring in the four genomes, I performed TE analysis by inverting the libraries: the *X. maculatus* genome was masked and analyzed with the *P. formosa* TE library, while *P. formosa* genomes were masked with the *X. maculatus* TE libraries. Results are presented in Figure 23. In the upper panel, both genomes were analyzed with their specific libraries, while in the lower panel they were analyzed with library from the other species. When inverting the libraries, we observe that the shapes of the landscapes are conserved after K-values higher than seven. However, we lose the recent bursts, suggesting that the old bursts reflect the history in the common ancestor genomes and that the recent bursts are species- or genus-specific. This result reflects the importance to use a specific library but it also shows how *Xiphophorus* and *Poecilia* genomes have diverged recently.
As mentioned in the *Xiphophorus* analysis paragraph, the *Xiphophorus* Kimura profiles present a high content of sequences with K-values around 50 (Figure 22 and 23). These sequences can be observed in the *X. maculatus* genomes masked with the two libraries, independently. The *P. formosa* masked with its own library does not have these sequences, but when the genome is masked with the platyfish library, the sequences located around K-value 50 are recovered.

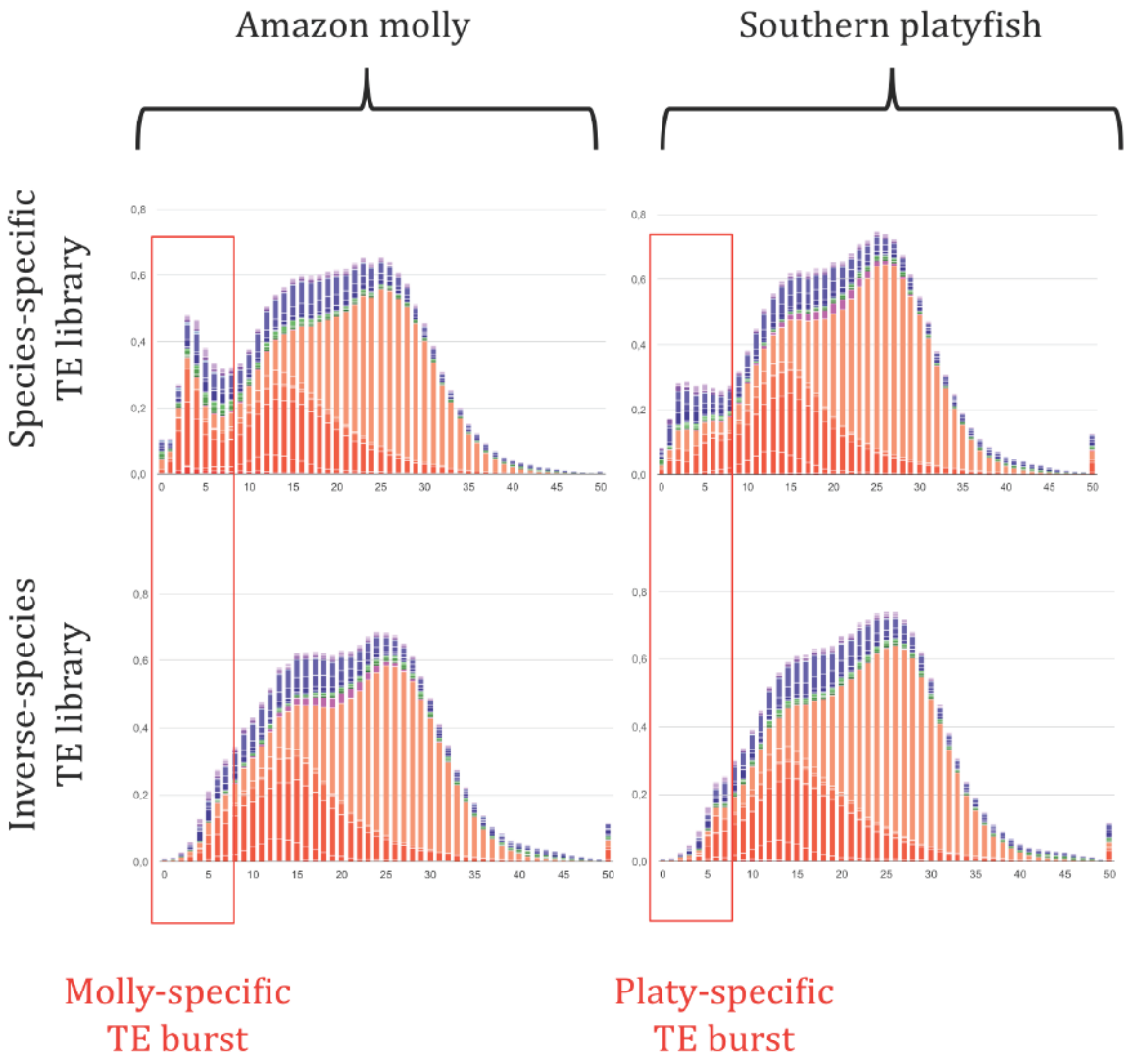


Figure 23: Analysis of TE profiles in *X. maculatus* and *P. formosa* by crossing species-specific TE libraries. The left panel shows the Kimura TE profile in the genome of the Amazon molly *P. formosa*, masked with its own specific library (upper) and with the Southern platyfish library (lower). Right panel represents the Kimura TE profile in the genome of the Southern platyfish, *Xiphophorus maculatus*, masked with its own specific library (upper) and with the Amazon molly library (lower).

III- The coelacanth project

As a member of the international consortium coelacanth genome project, we have been in charge of the annotation of TEs, in the genome of *Latimeria chalumnae* (AMEMIYA *et al.* 2013). We have then used the data produced to better understand TE evolution in vertebrates (CHALOPIN *et al.* 2013). As the coelacanth genome presents a high proportion of CR1 elements, this paper also revisits the classification of the CR1-like superfamily. In a second paper, we analyzed TE expression in coelacanth (FORCONI *et al.* 2013). In the third paper, we looked for proofs of TE recent activity by searching for polymorphic TE insertions between the two species *Latimeria chalumnae* and *Latimeria menadoensis* (Naville *et al.* in preparation).

a. *Genomic analyses of TEs in the genome of the African coelacanth, Latimeria chalumnae*

i. TE content and diversity in coelacanth to study TE evolution in vertebrate genomes

ORIGINAL ARTICLE: "Evolutionary active transposable elements in the genome of the coelacanth" (CHALOPIN *et al.* 2013): J Exp Zool B Mol Dev Evol, doi: 10.1002/jez.b.22521.

The genome of the African coelacanth has been sequenced by the Broad Institute and analyzed by a consortium of 13 international groups, and published in 2013 (see **ANNEXE 2**). The fascination for the coelacanth, a deeply living fish highly looking like old fossils, renders this organism very attractive for many researchers. From the evolutionary point of view, the coelacanth has a key phylogenetic position "between" ray-finned fish and tetrapods. Among others, genomic information from this species brought important insights into the water-to-land transition, but also into the general evolution of vertebrate genomes. From an evolutionary point of view, the hypothesis of a slow-evolving genome has been raised, regarding the phenotypic stasis between fossils from extinct species, and extant species.
In the context of the genome project, I participated to the construction of a TE library combining manual and automatic annotations. We then performed large-scale analyses of TE families and their evolutionary history. With about 25% of TEs, the coelacanth genome presents a TE content higher than most of the fish but lower than mammals and birds. However, regarding TE diversity, it contains more superfamilies than other tetrapods but less than actinopterygians, reflecting its intermediate position among vertebrates. The Blast analyses of transcriptomes from three different tissues (testis,

liver and muscle) showed that the CR1, Deu, DIRS and RTE superfamilies might be particularly strongly expressed in the coelacanth.

We pushed the analyses of CR1 retrotransposons, which are the major type of TEs in the coelacanth genome. Interestingly, it is also one of the main families present in many but not all other vertebrates. In this context, we analyzed sequences from three related LINE clades: CR1, LINE2 and Rex1-Babar to trace back their evolutionary history. Surprisingly, the CR1 clade, which is present in tetrapods and ancestral vertebrates, is absent from teleosts. In contrast to many TE families that have been lost in mammals but are still active in teleosts, CR1 might represent the first example of a major clade of TEs eliminated in fish but maintained in mammals and other tetrapods.

ii. Comparative analysis with a second genome assembly of the African coelacanth

In parallel to the sequencing of an individual of *L. chalumnae* by the BROAD Institute, a Japanese team sequenced different individuals of the same species and produced a second assembly (with a different assembler program) of the coelacanth genome (Nikaido *et al.* 2013). Analyses of this genome, performed by the Japanese group, showed that more than half of the genome is composed of TEs. This result strongly differs from our analyses (about 25%). To better understand the factors explaining such differences, I performed comparative analyses in both genomes using the same methods. The aim of this work was to understand if differences (25% versus 50%) were due to genome assembly, TE library construction or genome masking steps. I masked and compared both genomes with the library that I built.

Total repeats make about 37% (comprising 28% of TEs) and 60% (46% of TEs) of the Broad and Japanese assemblies, respectively. While both assemblies contain the same quantity of SINE and LTR retrotransposons, the content of the two other classes are different, i.e. the Japanese assembly contains 13 % of LINEs, 6% of DNA transposons and 18% of unclassified elements, while the Broad assembly shows 5% of LINEs, 1% of DNA transposons and 12% of unclassified elements. Regarding the distribution of TE sequences with Kimura distances (Figure 24), the main burst is similarly located in both graphs (K-value around 11). However, we can observe clear differences regarding 1- the percent of TEs in genomes, which is higher in the Japanese assembly, 2- DNA transposons distribution (higher recent burst in "Japanese" genome, K-value 5, Figure 24) and 3- a strong accumulation of small or degenerated sequences in "Japanese" genome (K-value 50). These results suggest that the two assemblies do not contain the same quantity of repetitive sequences as already noticed and mentioned.

Furthermore, in the published article of the Japanese genome (Nikaido *et al.* 2013), within the 23% of DNA transposons, 9% are found to be LatiHarb1 elements. LatiHarb1 is a recently discovered DNA transposon that composes a large part of coelacanth genome (Smith *et al.* 2012), but we were not able to find a so high content in the Broad assembly we used. The lack of LatiHarb1 in the Broad assembly might be due to a lack of scaffolds/contigs not included in the assembly, or a bias in the masking step.

Altogether, these results highlight how assemblies from a same species can be different depending on the methods of sequencing and assembly. Indeed, with the two current assemblies of the African coelacanth genome, the content of TEs doubles. We would not expect such differences between individuals from a same species, which is considered as a slow-evolving genome. An important point to consider is how assemblies were performed. Indeed, different methods of sequencing and different assemblers were used, which certainly also integrates a strong bias in TE distribution analyses. Analyses of non-assembled reads might give us preliminary answers concerning these differences.

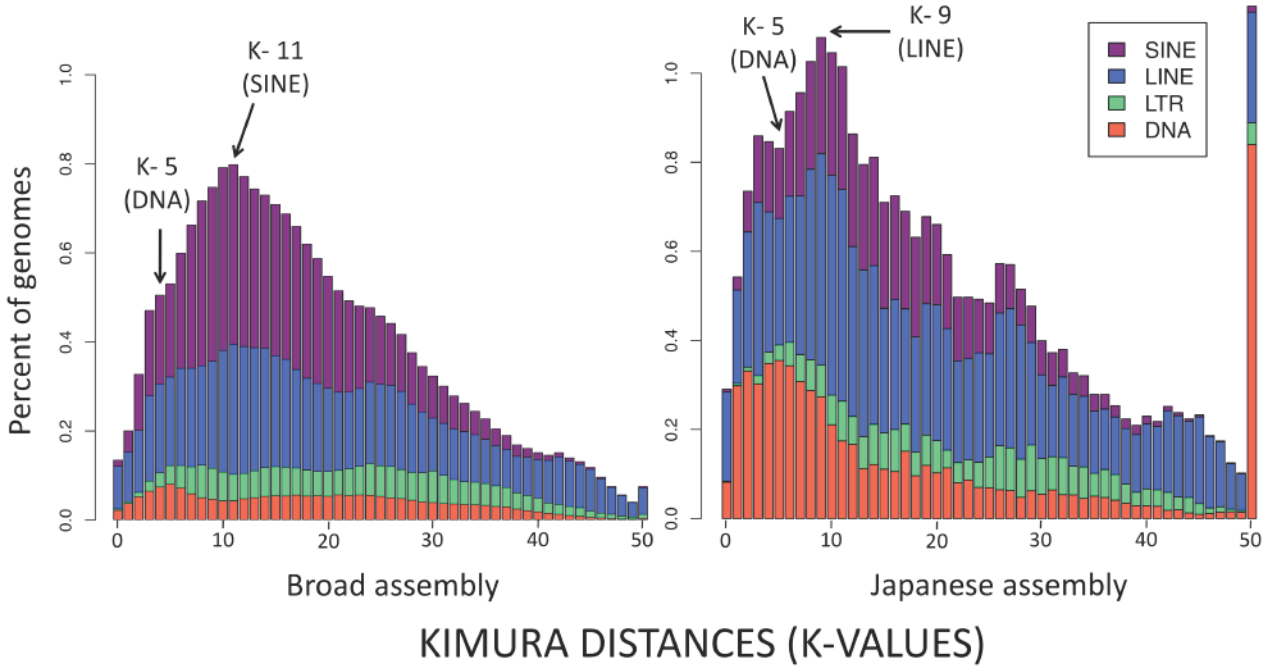


Figure 24: Comparison of the repeat landscapes from the two African coelacanth assemblies, masked with the same library generated from the Broad Institute genome.

b. *Transcriptomic TE analysis in the two coelacanth species, Latimeria chalumnae and Latimeria menadoensis*

ORIGINAL ARTICLE: "Transcriptional activity of transposable elements in coelacanth" (FORCONI *et al.* 2013): J Exp Zool B Mol Dev Evol, doi: 10.1002/jez.b.22527.

Using the three different transcriptomes, from testis, liver and muscle, which have been sequenced during the African coelacanth genome project, we investigated TE content and diversity to predict their potential activity in coelacanth tissues. TE expression was measured like expression of genes, with FPKM (Fragments per kilobase of exons per million mapped fragments) values taking into account fragments length and quantity. No difference of expression of superfamilies was observed between the three tissues, except a higher expression of DNA transposons in muscle. Even if TEs are much less expressed than housekeeping genes, we were also able to determine that few families seem to be highly expressed compared to others, such as CR1 and LF-SINE retrotransposons.

Moreover, by plotting TEs according to their expression value and their copy number in the genome, we examined potentially interesting cases of exaptation or non-identified superfamilies. Furthermore, thanks to this analysis, we identified a new family of SINE, specific to coelacanths, that we called Coeg-SINE.

c. *Detection of TE insertion polymorphisms in the two coelacanth species, L. chalumnae and L. menadoensis*

ORIGINAL ARTICLE: "Insertion polymorphism analysis reveals recent activity of transposable elements in the two extant species of coelacanth" (Naville *et al.* submitted)

The coelacanths are rare species, only two extant species *Latimeria chalumnae* (African coelacanth) and *Latimeria menadoensis* (Indonesian coelacanth) have been identified. In order to search for proofs of recent transposition activity in the coelacanth genomes, which might challenge the hypothesis of slow evolution, we looked for TE insertion polymorphisms between both *Latimeria* species.

We found that several TE superfamilies have been active after the separation between the two species, estimated around 10 Mya. Among these superfamilies, CR1 retrotransposons were the most active superfamily, followed by SINE elements (Coeg and LF-SINEs). Many CR1 sequences are complete and code for two ORFs, including the presence of endonuclease, reverse transcriptase and zinc finger domains. Combined with previous papers, this study brings a supplementary proof of a recent CR1 retrotransposon activity. Moreover, we detected evidence of homologous recombination between LTRs from an epsilon retrovirus. Altogether, we showed that TEs have been active after species divergence, and found clear evidence of insertions, deletions and homologous recombination events, suggesting a non-stasis of the coelacanth genome. However, a study at a higher level has to be performed when the genome of *L. menadoensis* will be completely sequenced.

Insertion polymorphism analysis reveals recent activity of transposable elements in the two extant species of coelacanths

Magali Naville[1,2], Domitille Chalopin[1,2], Jean-Nicolas Volff[1]*

[1] Institut de Génomique Fonctionnelle de Lyon, Ecole Normale Supérieure de Lyon, Lyon, France

[2] These authors contributed equally to this work.

* Corresponding author: Jean-Nicolas Volff (jean-nicolas.volff@ens-lyon.fr)

ABSTRACT

Coelacanths are lobe-finned fish represented by two extant species, *Latimeria chalumnae* in South Africa and Comoros and *L. menadoensis* in Indonesia. Due to their intermediate phylogenetic position between ray-finned fish and tetrapods in the vertebrate lineage, they are of great interest from an evolutionary point of view. In addition, extant specimens look similar to 300 million year old fossils: due to this apparent slowly-evolving morphology, coelacanths have been often described as « living fossils ». As an underlying cause of such a morphological stasis, several authors have proposed a slow evolution of the coelacanth genome. Accordingly, sequencing of the *L. chalumnae* genome has revealed a globally low substitution rate for protein-coding regions compared to other vertebrates. However, genome and gene evolution can also be influenced by transposable elements (TEs), which form a major and dynamic part of vertebrate genomes through their ability to move, duplicate and recombine. In this work, we have searched for evidence of transposition activity in coelacanth genomes through the comparative analysis of orthologous genomic regions from both *Latimeria* species. Comparison of 5.7 Mb (0.2 %) of the *L. chalumnae* genome with orthologous Bacterial Artificial Chromosome clones from *L. menadoensis* allowed the identification of 27 species-specific TE insertions, with a strong relative contribution of CR1 non-LTR retrotransposons. Species-specific homologous recombination between the long terminal repeats of a new coelacanth endogenous retrovirus was also detected. Our analysis suggests that transposon activity is responsible for at least 0.6% of genome divergence between both *Latimeria* species. Taken together, this analysis demonstrates that coelacanth genomes are not evolutionary inert: they contain recently and possibly still active transposable elements, which have significantly contributed to post-speciation genome divergence in *Latimeria*.

INTRODUCTION

The coelacanth is a lobe-finned fish that has been considered extinct since the Late Cretaceous period about 70 million years (my) ago, until a first living specimen, *Latimeria chalumnae,* was discovered in 1938 in South Africa by Marjorie Courtenay-Latimer [1]. From an evolutionary point of view, coelacanths occupy like lungfishes a key phylogenetic position between ray-finned fish and tetrapods at the basis of the sarcopterygian lineage. Their fleshy fins, which resemble the limbs of land animals, make them a pertinent model to study the water-to-land transition. A second coelacanth species, *Latimeria menadoensis*, was subsequently discovered in 1997 in Indonesia, with the capture of two individuals [2]. While coelacanths formed a highly spread taxonomic group during the Devonian [3,4], both extant species are nowadays endangered, with only few inventoried individuals (about 300 for *L. chalumnae* [5]). Despite their geographical remoteness, *L. chalumnae* and *L. menadoensis* present a high degree of nucleotide identity at the genomic level [98.7% based on the comparison of 20 Bacterial Artificial Chromosomes (BACs) from *L. menadoensis* with their orthologous sequences in the *L. chalumnae* genome], as well as for exons (>99.7% based on the comparison of liver and testis transcriptomes from both species) [6,7]. The recent genome analysis performed by Nikaido et al. even proposed a genetic divergence as low as 0.18% between the nuclear genomes of both species [8]. Such an identity rate of 98.7% at the genomic level is similar to that measured between human and chimpanzee. Considering the faster evolution in the primate lineage, the divergence time between the two coelacanth species was approximated at slightly more than 6-8 million years [6].

With fossils dating back to 300 million years that look very similar to extant animals, coelacanths have been placed by some authors in the arguable class of “living fossils”, which are characterized

by a long stasis in their phenotypic evolution [1,9]. The careful analyses of paleontological data, however, has recently challenged this picture [10]. Morphological stasis has often been proposed to rely on a very slow genomic evolution [10-12]. While several studies based on the analysis of particular gene families such as *Hox* or protocadherins already suggested a slow evolutionary rate [13-15], the recent availability of genomic data allowed to address this question in a more systematic way. By analyzing 251 protein-coding genes, which form the most constrained part of the genome, Amemiya et al. showed that these sequences evolve more slowly in coelacanth than in lungfish, chicken and mammals, with a substitution rate being half of that in tetrapods [6]. This is corroborated by the slow rate of nucleotide substitution demonstrated by Nikaido et al. based on the calculation of Ka/Ks ratios between 4,531 genes of *L. chalumnae* and *L. menadoensis* [8]. Analysis of coding sequences seems thus to sustain the idea of a slowly evolving genome.

Transposable elements (TEs) constitute a major source of genome diversity and evolution. These sequences, which are generally repeated, are able to integrate into new locations in genomes. TEs are sorted in several classes, orders and families according to their structure and mode of transposition [16]. Retroelements (class II elements) retrotranspose though the reverse transcription of an RNA intermediate into a cDNA copy, which is inserted somewhere else in the genome (copy-and-paste mechanism). Transposition of class I elements (DNA transposons) does not require any reverse transcription: these elements generally excise and reinsert into a new locus ("cut-and-paste" mechanism). Both classes are further subdivided into orders and superfamilies. TEs can be autonomous or non-autonomous: non-autonomous elements, such as class I MITEs (Miniature Inverted–repeat Transposable Elements) and class II SINEs (Short Interspersed Nuclear

Elements) do not encode the enzymes necessary for their transposition, but instead use the machinery of an autonomous element to achieve transposition.

Originally, TEs have been relegated to parasitic "junk DNA", with occasional negative effects on host genes such as insertional disruption and silencing [17,18]. More recently, converging studies have uncovered major roles of TEs in the evolution of genes, genomes and organisms [19]. TEs are driving forces of genome plasticity: their copies, interspersed along the chromosomes, can recombine and promote genomic rearrangements such as deletions, duplications, inversions and translocations [20,21]. In addition, TEs can duplicate or shuffle host coding sequences, and provide material for new regulatory elements (promoters, enhancers and splicing sites), new exons and even new genes – an evolutionary process called molecular domestication [22-25].

Considering their important impact on genome evolution, a strongly reduced TE activity has been proposed for the "living fossil" coelacanths [21]. It was recently shown that the *L. chalumnae* genome contains 25-50% of TEs, including retrotransposons (with a high proportion of non-Long Terminal Repeat [non-LTR] retrotransposons), endogenous retroviruses and DNA transposons [6,8,26]. Four major bursts of transposition that principally involved LINE1, LINE2, CR1 and Deu non-LTR retrotransposons were detected through copy divergence analysis in the genome of *L. chalumnae*, supporting activity of TEs in the *Latimeria* lineage [6,26]. Analysis of RNA-seq data showed that 14 TE superfamilies are expressed in coelacanth tissues, with a high representation of the CR1 LINE and LF-SINE families [6,27]. Although these results suggested TE activity in coelacanths, a direct evidence of recent transposition was still missing.

In this study, we have looked for the presence of TE insertion polymorphisms in orthologous regions from the genomes of the two extant coelacanth species, *L. chalumnae* and *L. menadoensis*.

Identification of species-specific insertions for the CR1 LINE family as well as for other types of TEs indicated recent transposition activity in the genus *Latimeria* and showed that TEs have significantly contributed to genome divergence between both coelacanth species.

MATERIALS AND METHODS

– *Origin of genomic sequences*

The *L. chalumnae* genome was downloaded from the Ensembl server (http://www.ensembl.org/; accession LatCha1 GCA_000225785.1). *L. menadoensis* BAC sequences were obtained from NCBI (http://www.ncbi.nlm.nih.gov/) with following accession numbers: GI:164698640, GI:170514516, GI:189459217, GI:190886531, GI:193083250, GI:237406519, GI:239735715, GI:239835822, GI:239835823, GI:239835824, GI:239835825, GI:239835826, GI:239835827, GI:239835828, GI:239835829, GI:239835830, GI:305644147, GI:305644148 [Birren 2009, NCBI direct submissions], GI:166987259 [28], GI:220898172 (*HoxA* gene cluster), GI:220898186 (*HoxB* gene cluster), GI:220898198 (*HoxC* gene cluster), GI:220898210 (*HoxD* gene cluster) [13], GI:296011776 [29], GI:407080572 (*IgW2* locus), GI:407080573 (*IgW1* locus) [Saha 2012, NCBI direct submissions], GI:40789109, GI:50253612, GI:50253613, GI:50284579, GI:50284580, GI:50284581, GI:52077680 [Grimwood 2004, NCBI direct submissions], GI:66912372 [Lau 2010, NCBI direct submission].

– *Identification of TE insertions*

In order to determine orthology relationships between *L. menadoensis* BAC clones and the *L. chalumnae* genome, sequence comparison was performed using MegaBlast [30]. Best hits were selected and the main alignment diagonals were used to define the maximum match coordinates along with their orientation. Orthologous fragments are listed in supplementary table S1. *L. chalumnae* and *L. menadoensis* TEs were localized using the RepeatMasker software [31] with a TE library specifically built for the *L. chalumnae* genome [6,26]. TEs were selected according to the following criteria: length ≥ 100 nucleotides (nt) and divergence to the consensus sequence from the

123 library ≤ 20%. Low complexity sequences, simple repeats as well as tRNA and rRNA
124 (pseudo)genes were discarded. Remaining elements located in corresponding *L. menadoensis* BACs
125 and *L. chalumnae* genome fragments were then listed "face to face" as shown in supplementary
126 figure S1, and further manually aligned to visualize orthologous insertions in both species.
127 Species-specific insertions were inspected manually by extracting and comparing both "empty" and
128 "filled" sites using the Muscle alignment software [32].

129 – *Counting of shared TE insertions*

130 The RepeatMasker procedure can artifactually split one unique insertion into several pieces in case
131 of local mutations. For this reason, the simple counting of RepeatMasker annotations along the
132 chromosomes of the two species can provide different results for shared insertions. We thus
133 evaluated the number of common insertions as the minimal value between *L. chalumnae* and *L.*
134 *menadoensis* annotations that filled the filtering criteria of minimal length (100 nt) and maximal
135 percentage of divergence compared to the RepeatMasker consensus (20%).

136 – *Annotation of species-specific TE insertions*

137 TE insertions were assigned to known TE superfamilies based on Wicker's classification by
138 combining comparative and predictive approaches [16]. In order to look for putative coding regions,
139 TE sequences were submitted to the *de novo* gene prediction program Genscan [33] and to a BlastX
140 search [34] on the NCBI website with default parameters against the Genbank non-redundant
141 protein sequence database. Insertions were also analyzed with the Censor software [35] that
142 compares sequences with the Repbase repeat database [36]. Structural features such as Long
143 Terminal Repeats (LTRs) and Terminal Inverted Repeats (TIRs) were identified manually and with
144 the Blast bl2seq utility [34]. Target Site Duplications (TSDs), which correspond to the duplication

of few nucleotides from the insertion site, were searched at the extremities of insertions. TE expression was analyzed by sequence similarity analysis of a transcriptome of *L. menadoensis* testis [6] using the Blast algorithm with default parameters.

– *Sequence alignments and phylogenetic reconstructions*

Sequences not unambiguously associated to particular TE families were classified using phylogenetic analysis. TE sequences were extracted from BACs and genomic segments and aligned using Muscle [32] with default parameters. Phylogenetic trees were constructed with PhyML [37] using Maximum Likelihood and aLRT values (non-parametric bootstraps) on different types of sequences depending on the element analyzed: reverse transcriptase (RT) and integrase core domain for Endogenous RetroViruses (ERVs), RT for CR1/L2 non-LTR retrotransposons, and nucleotide sequence for SINE elements.

RESULTS

We searched for the presence of species-specific TE insertions in *Latimeria* by comparing 36 BAC clone sequences from the Indonesian coelacanth *L. menadoensis* with orthologous regions from the recently published genome of its African congener *L. chalumnae* [6]. *L. menadoensis* BAC clones have a median length of 170 kb and correspond to loci of particular interest including the *IgW1* genes (immunoglobulin heavy chain) [Saha 2012, NCBI direct submission] and the *Hox* genes [13]. The analyzed sequences, which cover ca. 5.7 Mb (0.2%) of the draft of the *L. chalumnae* genome, contain about 1,370 insertions of identifiable transposable elements common to both species (15.7% of the fraction of the genome analyzed).

Identification and characterization of species-specific TE insertions

Using a comparative approach, we searched for species-specific insertions, i.e. for sites "filled" by a TE insertion in one species orthologous to "empty" sites (i.e. without insertion) in the other (Table 1, Table 2 and supplementary Table S2). Of note, for DNA transposons, what we define here as an "insertion" in one species could alternatively be the result of the excision in the other species of an element inserted in the last common ancestor of both species. Furthermore, insertion polymorphism might reflect transposition in one species after speciation, but also insertion polymorphism at allelic positions in the last common ancestor. In any case, we consider these different possibilities as evidence for relatively recent (<10 my old) transposition events.

Manual inspection of candidates for polymorphic insertions allowed to exclude a number of false positives that corresponded to stretches of "N" in the *L. chalumnae* draft genome. Generally the length of such N stretches matched almost exactly the length of the insertion at orthologous

178 positions in *L. menadoensis*, suggesting that they have been produced during the assembly phase.
179 Pairwise alignment of "empty" and "filled" sites allowed to define insertion boundaries, as shown
180 in Figure 1. Several insertions presented evidence of degeneration (short size, truncated open
181 reading frames, etc.) (Table 2).
182 Comparison between both coelacanth species led to the identification of 27 species-specific TE
183 insertions, 13 in *L. chalumnae* and 14 in *L. menadoensis* (Tables 1 and 2). Insertion length ranged
184 from 225 to 5,091 nt, with a mean of 1,363 nt. Insertions were classified according to TE ontology
185 [16] using different specific characteristics of the superfamily, in particular the similarity with
186 known TE-encoded proteins and the presence of specific structural features such as LTRs and TIRs.
187 On the whole, identified polymorphic insertions mainly corresponded to CR1 non-LTR
188 retrotransposons (6/13 and 9/14 insertions in *L. chalumnae* and *L. menadoensis,* respectively). A
189 reverse transcriptase-encoding region, belonging to CR1 ORF2 [38], was present in 3 and 7 of
190 them, respectively, with only 3 cases (one in *L. chalumnae* and 2 in *L. menadoensis*) where it was
191 apparently complete. The two longest *L. menadoensis* CR1 insertions with complete RT sequences
192 showed 95% of nt identity between each other (insertions 7 and 8; Table 2). They further contained,
193 upstream of the RT, an endonuclease domain also belonging to ORF2, and downstream of it, a short
194 and truncated ORF maybe corresponding to a partial ORF1 (93 and 89 amino acids for insertion 7
195 and 8 respectively, against ca. 420 amino acids for the complete expected sequence [38]; figure 1).
196 However, ORF1 is located upstream of the RT in sequences described in the literature [38,39], and
197 the sequence found here did not show any significant similarity with known ORF1 sequences.
198 ORF1 protein function is still poorly understood [38]; it non-specifically binds single-stranded
199 mRNA/DNA molecules and could be dispensable for transposition [38]. While ORF2 of insertion 8

presented several frameshifts and two stop codons, insertion 7 ORF2 did not show such degeneration of the sequence but only a 5' truncation of approximately 140 amino acids, deleting 2 amino acid sites thought to be important for endonuclease activity [39] (figure1). If this sequence showed only 46% of identity with chicken ORF2, most of the other putative active sites were conserved (17/18 amino acids) [39]. Considering the 5' truncation of the endonuclease domain, it suggests that the element might have been recently transposition-competent. The small ORF downstream of the RT was also identified in 9 other CR1 insertions (4 in L. *chalumnae a*nd 5 in L. *menadoensis, r*espectively), where it is always truncated in its 3' part.

Polymorphic insertions of other non-LTR retrotransposons from the L1 (two in *L. menadoensis*) and L2 superfamilies (one in *L. chalumnae*) were additionally detected, as well as SINEs (three in *L. chalumnae* and two in *L. menadoensis*), including the recently described CoeG-SINE family [27]. LTR retroelements were represented by a strongly corrupted copy of a Gypsy-like retrotransposon in *L. chalumnae* and an endogenous retrovirus in *L. menadoensis* (see below). For class I elements, only two insertions with a palindromic structure reminiscent of that of Miniature Inverted Transposable Elements (MITEs) were identified, one of them possibly derived from a hAT DNA transposon (as predicted using Censor). Finally, insertion 27 presented a composite structure comprising, sequentially, (i) a partial CR1 non-LTR retrotransposon with RT domain, framed by an "AAGT" TSD, (ii) a possibly novel SINE element flanked by an "AAGT" TSD, (iii) half of a LF-SINE and (iv) a tRNA-derived SINE. TSDs, which are hallmarks of insertions consisting in few duplicated target site nucleotides (Figure 1), could be clearly identified in 20 out of 27 insertions. Two other insertions occurred in poly-A or AT-rich regions. Twelve out of 17 CR1 insertions were flanked by TSDs, with no obvious sequence specificity for integration.

Altogether, polymorphic TE insertions covered a total of 13.3 kb in *L. chalumnae* and 23.5 kb in *L. menadoensis,* corresponding to approximately 0.23% and 0.41% of the genomic regions analyzed in the two coelacanth species. Hence, transposon activity is responsible for ca. 0.64% of genome divergence between both *Latimeria* species in the regions considered. Only insertions 7, corresponding to a CR1 non-LTR retrotransposon, presents a structure and sequence potentially compatible with an autonomous transposition activity.

Among the 27 polymorphic sequences identified, seven were found to be intronic, the others being located in intergenic regions (Table 2). While half of insertions were located more than 6 kb away from gene exons, five were closer than 1 kb from the next exon; in *L. menadoensis*, a CR1 element was inserted in an intron of the *SRA1* gene, about 200 bp next to the closest exon. Hence, some TE insertions are closely linked to coelacanth genes, with possible influence on their functions and evolution.

Structural and evolutionary analysis of new coelacanth endogenous retroviruses

The ERV insertion in *L. menadoensis* (insertion 24) corresponded to the largest of all identified polymorphic insertions (5,091 nt), and showed a significant BlastX similarity (alignment of 113 amino acids with E-value = 2.2e-10) with an integrase protein encoded by elements from the ERV1 family, which belongs to Epsilon retroviruses [40]. This ERV copy was delimited by two almost identical LTRs (462 nt, 99% of identity) and flanked by TSDs (AGAT) (Figure 2B). The orthologous region in *L. chalumnae* corresponded to a sequence of 462 nt almost identical to the LTRs of the ERV in *L. menadoensis* and flanked by the AGAT TSD sequences (Figure 2A). Hence, the *L. chalumnae* sequence corresponds to a so-called solo-LTR, formed through homologous

recombination between the LTRs of the original retrovirus element, this eliminating one copy of the LTR and the intervening retrovirus sequence.

In order to better characterize the new coelacanth endogenous retrovirus, which was called CoeERV1-1, a consensus sequence was reconstructed from different copies found in the *L. chalumnae* genome (Figure 2C). This sequence contains LTRs (475 nt long) and a central region encoding Gag (viral capsid), Pol (polyprotein responsible for the synthesis of the viral DNA and its integration into host genome, including protease [Pro], reverse transcriptase [RT], ribonuclease H [RH] and integrase [Int]), and Env (envelope). Compared to this reconstructed sequence, the identified insertion (number 24 in Table 1) is strongly truncated, lacking major parts of the Gag and Pol domains (Figure 2 B-C). The internal part of the insertion does not share any similarity with known TE sequences. Seventy-five copies of CoeERV1-1 longer than 300 nt (identity > 90%, E-value < 10e-120) were detected in the *L. chalumnae* genome, including four copies (ranging from 6,804 to 8,045 nt) with LTRs of variable size (from 444 to 551 nt) and all ORFs including the envelope (this ORF being truncated in three of them). One copy lacks only the RT domain (total length 7,192 nt); two copies present all ORFs but exhibit truncated LTRs; and the other present variable truncations of one or several ORFs. Three copies were found to share 92% of identity, at least, over more than 7,900 nt. A wider blast search showed that more than 500 sequences share similarities (identity > 70%, E-value < 1e-07, length > 60 nt) with this ERV, but they are mostly corresponding to LTR parts.

Among the copies identified, at least 10 corresponded to solo-LTRs (8 being 475 nt long and presenting TSDs, and 2 being truncated) sharing more than 98% of identity with LTRs of the

reference complete element. Similarity search with relaxed parameters could not allow to identify more divergent copies of this ERV element, suggesting that the family is represented by only one group of sequences in *Latimeria*.

Inspection of sixteen identified TSD sequences (ranging from 3 to 5 nt) suggested preference of insertion into target sites containing the "GT" (in 5/16 TSDs), "AC/G" (in 8/16 TSDs), both (in 2/16 TSDs) or any (in 1/16 TSDs) nucleotide motifs (Figure 2C).

Retroviruses (RV) are classified in seven genera including Alpha-, Beta-, Gamma-, Delta-, Epsilon-, Lenti- and Spuma-viruses [40]. With six identified genera, the mammalian lineage presents the largest diversity of RV among vertebrates. To better understand the origin and evolution of the coelacanth ERV identified in this work, we performed phylogenetic reconstructions based on both RT (ca. 210 amino-acids) and integrase core domain sequences alignments (ca. 132 amino-acids) (Figure 3 and supplementary figure S2). Coelacanth sequences were found to cluster in the Epsilon-virus group, one of the most spread branches of RV in vertebrates. Interestingly, CoeERV1-1 sequences were closely related to turtle and crocodile RV sequences, a result supported by both RT and integrase-based reconstructions (Figure 3 and supplementary figure S2). Coelacanth ERV sequences share more than 3,000 nucleotides with over 64% of identity with alligator, crocodile and turtle sequences. This strong relatedness could suggest horizontal transfer (HT) between reptiles and coelacanths or infection of both lineages by a same subgroup of related retroviruses.

Copy number and expression of transposable elements in coelacanth

In order to get more insight into the relative transposition activity of TEs identified as polymorphic in this work, we determined their copy number including common insertions by similarity search against the 5.7 Mb of orthologous sequences analyzed in both coelacanth species (filter: length ≥ 80 % of the considered insertion length and sequence identity ≥ 80%; cf methods and Table 2).

In 11 cases, including 4 CR1 non-LTR retrotransposons, 3 CoeG-SINEs, both LTR elements, one MITE and the composite insertion, we were only able to retrieve the query sequence with these filtering parameters, indicating the absence of other related copies of approximately the same size in the regions analyzed (Table 2). In contrast, other elements were found to be reiterated, with copy numbers ranging from 2 to 110 (39 on the average). Particularly, most CR1 elements were repeated (11/15, 73%), as well as LF-SINEs (48 hits). In the case of CR1, hits obtained for all the different polymorphic insertions were overlapping, indicating that these insertions correspond to copies of the same element. In the whole, 105 common insertions of this CR1 element were identified by Blast in the compared regions. L1 and L2 non-LTR retrotransposons as well as a MITE-like element showed a more modest level of reiteration. The two L1 specific-insertions of *L. menadoensis* matched against two sets of sequences mutually non-overlapping, indicating that these two L1 correspond to distinct elements. Nine CR1 and one L1 elements presented at least another copy with a very high level of nt identity (≥ 98%), suggesting rather recent events of transposition.

In order to determine if some of the elements identified in this work might be expressed and are therefore potentially active, TE insertions were used as queries against a *L. menadoensis* testis transcriptome (see methods). Table 2 presents for each insertion the number of Blast hits obtained with length ≥ 80 nt and identity ≥ 95%. Nine elements (2 in *L. chalumnae* and 7 in *L. menadoensis*) matched at least 10 times in the transcriptome: five CR1s, both L1s, one CoeG-SINE and the ERV,

which presents the highest number of hits (41), probably because it is also the longest insertion, the hits being scattered all along its sequence. All nine elements are intergenic, apart from the CoeG-SINE, which is located in an intron. The sets of matches obtained for the different CR1 elements were overlaping, with a total number of 32 different sequences in the transcriptome. This observation is congruent with the similarity search against genomic sequences. Also in agreement with this search, the two L1 matches against distinct sets of transcriptomic sequences. The 5 CR1 elements represented in the transcriptome (according to our filtering parameters) included insertions 7 and 8, which correspond to the same CR1 element and constitute the most complete copies of the retrotransposon identified in the regions analyzed.

DISCUSSION

Comparative analysis of orthologous regions covering 5.7 Mb of the genome of the two extant coelacanth species strongly sustains the recent activity of transposable elements in this lineage, with the identification of 13 and 14 species-specific insertions in *L. chalumnae* and *L. menadoensis,* respectively. Insertions observed specifically in one or the other species suggested that these TE copies transposed after speciation, i.e. approximately within the last 6-8 million years. Alternatively, they might also correspond to insertion polymorphisms that predated the split between both species. Interestingly, with the exception of two MITE-like elements, most polymorphic insertions are retrotransposons. This indicates that DNA transposons are currently probably less active than retrotransposons in coelacanth genomes. CR1 LINEs represent most of recent insertions, with a more marginal contribution of tRNA-SINEs (CoeG- and LF-SINEs) and L1 and L2 LINEs. Hence, our results support relatively recent activity of CR1 retrotransposons and

other LINE and SINE elements in coelacanth genomes. Retrotransposition of non-coding tRNA-SINEs identified in this work might be catalyzed by autonomous CR1-LINEs or other LINEs. However, no significant similarity could be detected between the 3' part of LINE and SINE elements (data not shown). The more discrete presence of one ERV and one Gypsy element indicated that LTR retrotransposons also contribute to insertion polymorphisms. Other types of repeats such as satellite sequences represent an additional form of genome divergence that was not considered in this study.

Target site duplications are hallmarks of insertions for most TEs. Many polymorphic insertions identified in this analysis showed recognizable TSDs, in particular many CR1 elements, suggesting that they transposed recently. Other arguments in favor of a recent/current transposition activity are the presence of very similar copies of a same element in the genome, indicative of recent bursts of transposition, and the representation of TE sequences in transcriptomes, even if the latter does not necessarily imply functionality of the element. We have shown a particular enrichment in CR1 elements in coelacanth genomic sequences, as well as, to a lesser extent, the presence of SINEs and L1 and L2 retrotransposons. Similarity search against a testis transcriptome of *L. menadoensis* uncovered a high number of hits for CR1 elements as well as for L1 LINEs, CoeG-SINEs and an endogenous retrovirus. These results confirmed major activity of CR1 retrotransposons and the contribution of minor other types of TEs, mostly retroelements.

We also showed that homologous recombination between the LTRs of an endogenous retrovirus can contribute to genome divergence in coelacanth. This analysis led to the identification of CoeERV1-1, a so far unknown coelacanth Epsilon retrovirus, which is present in the genome of *L. chalumnae* under the form of elements with two LTRs and partial or complete *gag*, *pol* and *env*, or

353 as solo LTRs. Interestingly, phylogenetic analyses revealed a close relationship of coelacanth
354 CoeERV1-1 with turtle and crocodile Epsilon retroviruses. This might be due to horizontal transfer
355 between coelacanths and reptiles, or to infection of both lineages by related retroviruses. The
356 frequent high degree of nt identity between both LTRs of a same copy, the high similarity between
357 different copies and the presence of numerous copies with TSDs together suggest recent
358 introduction into and/or recent transposition of CoeERV1-1 in the genome of coelacanths.
359 This study allows to estimate the total number of species-specific insertions in the genomes of the
360 two *Latimeria* species and to evaluate the impact of TEs on genome divergence in coelacanths. Our
361 analysis, based on interspecific comparison of 5.7 Mb of orthologous genomic sequences (0.2% of
362 the genome), indicated an average of 13.5 species-specific TE insertions. Hence, each *Latimeria*
363 species might contain 6500-7000 TE insertions not found in the other species. This means that ca.
364 15,000 TE insertions are differentially present in both species, which diverged 6-8 mya. Strikingly,
365 this value is similar to that reported for human and chimpanzee, which show 11,000 differentially
366 present TE insertions (divergence 6 mya; [41,42]). In term of DNA amount, TEs are responsible for
367 about 0.6% of genome divergence, i.e. for about 20 Mb of difference at the scale of the whole
368 genome, between both coelacanth species. Importantly, this analysis particularly included gene-rich,
369 euchromatic regions, which are generally rather poor in TE insertions compared to gene-poor
370 heterochromatic regions. Hence, our estimation of TE contribution to coelacanth species-specific
371 genome divergence is probably an underestimation.
372 To conclude, this work demonstrates that coelacanths possess active transposable elements that
373 significantly contribute to post-speciation genome evolution. Hence the apparent morphological
374 stasis of coelacanths might not be due to reduced TE activity, as previously proposed [21]. Our

375 results also suggest that, beside transposition, other mechanisms such as ectopic homologous
376 recombination and horizontal transfer might contribute to the plasticity of the coelacanth genome.
377 This raises the question of the low impact of these mechanisms on the evolution of the coelacanth,
378 or call again into question the postulated morphological stasis of *Latimeria*, which might not be
379 supported by paleontological evidence [10].
380

381 **Acknowledgments**
382 MN in supported by Ecole Normale Supérieure de Lyon; DC is supported PhD Grants from the
383 French Ministry for Higher Education and Research, and from "Ligue Contre le Cancer".
384

385 **Author Contributions**
386 Initiated the study : JNV. Developed the study, designed the strategy and performed the
387 bioinformatics analyses : MN and DC. Analysed results : MN, DC and JNV. All authors
388 participated to the redaction of and approved the final manuscript.

389 **REFERENCES**

390 1. Smith JLB (1939) A living fish of Mesozoic type. Nature 143: 455-456.

391 2. Holder MT, Erdmann MV, Wilcox TP, Caldwell RL, Hillis DM (1999) Two living species of
392 coelacanths? Proc Natl Acad Sci U S A 96: 12616-12620.

393 3. Cloutier R, Ahlberg PE (1996) Morphology, characters, and the interrelationships of basal
394 sarcopterygians. San Diego, CA: Academic Press.

395 4. Maisey JG (1996) Discovering fossil fishes: New York: Henry Holt & Co.

396 5. Fricke H, Hissmann K, Froese R, Schauer J, Plante R, et al. (2011) The population biology of the
397 living coelacanth studied over 21 years. Marine Biology 158: 1511-1522.

398 6. Amemiya CT, Alfoldi J, Lee AP, Fan SH, Philippe H, et al. (2013) The African coelacanth
399 genome provides insights into tetrapod evolution. Nature 496: 311-316.

400 7. Pallavicini A, Canapa A, Barucca M, Alfoldi J, Biscotti MA, et al. (2013) Analysis of the
401 transcriptome of the Indonesian coelacanth Latimeria menadoensis. Bmc Genomics 14.

402 8. Nikaido M, Noguchi H, Nishihara H, Toyoda A, Suzuki Y, et al. (2013) Coelacanth genomes
403 reveal signatures for evolutionary transition from water to land. Genome Research 23:
404 1740-1748.

405 9. Smith JLB (1956) Old Fourlegs, the story of the coelacanth. London, New York,: Longmans. 260
406 p. p.

407 10. Casane D, Laurenti P (2013) Why coelacanths are not 'living fossils': a review of molecular and
408 morphological data. Bioessays 35: 332-338.

409 11. Friedman M, Coates MI (2006) A newly recognized fossil coelacanth highlights the early
410 morphological diversification of the clade. Proc Biol Sci 273: 245-250.

411 12. Thomson KS (1992) Living fossil: The story of the Coelacanth: W.W. Norton. 252 p.
412 13. Amemiya CT, Powers TP, Prohaska SJ, Grimwood J, Schmutz J, et al. (2010) Complete HOX
413 cluster characterization of the coelacanth provides further evidence for slow evolution of its
414 genome. Proc Natl Acad Sci U S A 107: 3622-3627.
415 14. Higasa K, Nikaido M, Saito TL, Yoshimura J, Suzuki Y, et al. (2012) Extremely slow rate of
416 evolution in the HOX cluster revealed by comparison between Tanzanian and Indonesian
417 coelacanths. Gene 505: 324-332.
418 15. Noonan JP, Grimwood J, Danke J, Schmutz J, Dickson M, et al. (2004) Coelacanth genome
419 sequence reveals the evolutionary history of vertebrate genes. Genome Res 14: 2397-2405.
420 16. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, et al. (2007) A unified classification
421 system for eukaryotic transposable elements. Nat Rev Genet 8: 973-982.
422 17. Deininger PL, Batzer MA (1999) Alu repeats and human disease. Mol Genet Metab 67:
423 183-193.
424 18. Hancks DC, Kazazian HH, Jr. (2010) SVA retrotransposons: Evolution and genetic instability.
425 Semin Cancer Biol 20: 234-245.
426 19. Kazazian HH, Jr. (2004) Mobile elements: drivers of genome evolution. Science 303:
427 1626-1632.
428 20. Bourque G (2009) Transposable elements in gene regulation and in the evolution of vertebrate
429 genomes. Curr Opin Genet Dev 19: 607-612.
430 21. Oliver KR, Greene WK (2009) Transposable elements: powerful facilitators of evolution.
431 Bioessays 31: 703-714.

22. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, et al. (2006) A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature 441: 87-90.

23. Feschotte C (2008) Transposable elements and the evolution of regulatory networks. Nat Rev Genet 9: 397-405.

24. Rebollo R, Romanish MT, Mager DL (2012) Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annu Rev Genet 46: 21-42.

25. Volff JN (2006) Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays 28: 913-922.

26. Chalopin D, Fan S, Simakov O, Meyer A, Schartl M, et al. (2013) Evolutionary active transposable elements in the genome of the coelacanth. J Exp Zool B Mol Dev Evol.

27. Forconi M, Chalopin D, Barucca M, Biscotti MA, De Moro G, et al. (2013) Transcriptional activity of transposable elements in coelacanth. J Exp Zool B Mol Dev Evol.

28. Gwee PC, Amemiya CT, Brenner S, Venkatesh B (2008) Sequence and organization of coelacanth neurohypophysial hormone genes: evolutionary history of the vertebrate neurohypophysial hormone gene locus. BMC Evol Biol 8: 93.

29. Mulley JF, Holland PW (2010) Parallel retention of Pdx2 genes in cartilaginous fish and coelacanths. Mol Biol Evol 27: 2386-2391.

30. Zhang XA, Wang JL, Wu QH (2010) A Greedy Adaptive Sensor Set Selecting Algorithm under Correlated Log-normal Shadowing. Frequenz 64: 127-133.

31. Smit A, Hubley R, Green P (1996-2010) RepeatMasker Open-3.0.

32. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797.

454 33. Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J
455 Mol Biol 268: 78-94.
456 34. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool.
457 J Mol Biol 215: 403-410.
458 35. Kohany O, Gentles AJ, Hankus L, Jurka J (2006) Annotation, submission and screening of
459 repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 7: 474.
460 36. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, et al. (2005) Repbase Update, a
461 database of eukaryotic repetitive elements. Cytogenet Genome Res 110: 462-467.
462 37. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, et al. (2010) New algorithms and
463 methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML
464 3.0. Syst Biol 59: 307-321.
465 38. Kapitonov VV, Jurka J (2003) The esterase and PHD domains in CR1-like non-LTR
466 retrotransposons. Mol Biol Evol 20: 38-46.
467 39. Haas NB, Grabowski JM, Sivitz AB, Burch JB (1997) Chicken repeat 1 (CR1) elements, which
468 define an ancient family of vertebrate non-LTR retrotransposons, contain two closely spaced
469 open reading frames. Gene 197: 305-309.
470 40. Gifford R, Tristem M (2003) The evolution, distribution and diversity of endogenous
471 retroviruses. Virus Genes 26: 291-315.
472 41. Mills RE, Bennett EA, Iskow RC, Devine SE (2007) Which transposable elements are active in
473 the human genome? Trends Genet 23: 183-191.
474 42. Mills RE, Bennett EA, Iskow RC, Luttig CT, Tsui C, et al. (2006) Recently mobilized
475 transposons in the human and chimpanzee genomes. Am J Hum Genet 78: 671-679.

476

25

477 **FIGURE LEGENDS**

478

479 **Figure 1. Example of a polymorphic insertion of a CR1 retrotransposon (element 7 in table 2)**
480 **present in *Latimeria menadoensis* but absent from *L. chalumnae*.**
481 Target Site Duplications (TSDs) are framed in red. CR1 = Chicken Repeat 1; ORF = Open Reading
482 Frame; RT = Reverse Transcriptase; APE = Apurinic/Apyrimidic Endonuclease.

483

484 **Figure 2. Structure of coelacanth endogenous retrovirus CoeERV1-1.** A. Solo-LTR observed in
485 *L. chalumnae*. B. Schematic representation of ERV insertion 24 found at the orthologous position in
486 *L. menadoensis*. C. Reconstructed structure of CoeERV1-1 in the *L. chalumnae* genome. TSD =
487 Target Site Duplication; LTR = Long Terminal Repeats; Gag: ORF encoding protein for the viral
488 capsid; Pol: ORF encoding proteins responsible for synthesis of the viral DNA and integration into
489 host DNA, including protease (Pro), reverse transcriptase (RT), ribonuclease H (RH) and integrase
490 (Int); Env: ORF encoding envelope protein.

491

492 **Figure 3. Phylogenetic relationship between coelacanth CoeERV1 and reptile retroviruses.**
493 Vertebrate retrovirus phylogeny was reconstructed on an alignment of RT (210 amino-acids) using
494 Maximum Likelihood with optimized parameters (best NNI and SPR; optimized invariable sites).
495 Branch values represent supporting aLRT non-parametric statistics. The dashed line highlights the
496 group of Epsilon viruses containing turtle, crocodile, coelacanth and lungfish sequences. Gypsy
497 LTR retrotransposon sequences were used as an outgroup.

498

Table 1. Transposable element insertions in ca. 5.7 Mb of orthologous genomic sequences from the coelacanth species *Latimeria chalumnae* and *L. menadoensis*.

TE classification		TE family	Common insertions	Species-specific insertions	
				L. chalumnae	*L. menadoensis*
Class I (retrotransposons)	LINE	CR1	297	6	9
		L1	11	-	2
		L2	6	1	-
	SINE	Coeg-SINEs	206	1	-
		Others	817	1	2
	LTR	Gypsy	25	1	-
		ERV*	0	- (solo LTR)	1 (element framed by 2 LTRs)
Class II (DNA transposons)		MITE-like	10	2	-
Composite insertion		CR1 / SINEs	-	1	-
Total			1372	13	14

TE = Transposable Element; LINE = Long Interspersed Nuclear Element; SINE = Short Interspersed Nuclear Element; LTR = Long Terminal Repeat; CR1 = Chicken Repeat 1; L1 = LINE 1; L2 = LINE 2; ERV = Endogenous Retrovirus; MITE = Miniature Inverted-repeat Transposable Element.

* The ERV insertion observed in *L.menadoensis* does not strictly correspond to an insertion polymorphism, the solo LTR observed at the orthologous site in *L. chalumnae* probably being the result of a recombination between the two LTRs framing the element (see main text).

Table 2. Structural features of species-specific transposable element insertions in ca. 5.7 Mb of orthologous genomic sequences from the coelacanth species *Latimeria chalumnae* and *L. menadoensis*.

Type of TE	Species with insertion	Insertion identifier	Insertion length (nt)	Target Site Duplication ?	ORF(s) / Domain(s)	Copy number in genomic sequences *	Element representation in the transcriptome **	Genomic position relative to next gene	Distance to the closest exon (kb) and corresponding gene
CR1	L. ch.	1	1622	AT	ORF1 (partial); ORF2: RT	5 (2 with id ≥ 98 %)	17	IGR	5.9 (*HOXD12*)
		2	1060	AT-rich region	ORF2: RT (partial)	1	0	Intron (exon 4 - exon 5)	0.7 (exon 5) (*GRID1*)
		3	1097	GAGTCTTGTT	ORF2: RT (partial)	1	4	IGR	4.0 (*PCDHGC*)
		4	227	CTA	ORF1(partial)	49 (5 with id ≥ 98 %)	1	IGR	9.7 (*ighv14-1 (21)*)
		5	320	TTTAG	ORF1(partial)	37 (5 with id ≥ 98 %)	0	IGR	9.4 (vomeronasal 2 receptor)
		6	303	TATTAGG	ORF1(partial)	1	0	IGR	> 70.9 (*CALCOCO1*)
	Lme	7	2845	ACTCA	ORF1 (partial); ORF2: RT, APE	23 (4 with id ≥ 98 %)	24	IGR	> 9.0
		8	2821	AAT	ORF1 (partial); ORF2: RT, APE	24	31	IGR	3.2 (*PCDHGC5*)
		9	1174	AAGTA	ORF1 (partial) ; ORF2: RT (partial)	4	8	IGR	3.6 (*PCDHGC5*)
		10	1038	CCAT	ORF1 (partial); ORF2: RT (partial)	74 (18 with id ≥ 98 %)	10	IGR	18.3 (protocadherin gamma)
		11	862	GATTAA	ORF1: (partial); ORF2: RT (partial)	86 (19 with id ≥ 98 %)	6	Intron (exon 2 – exon 3)	0.2 (exon 3) (*SRA1*)
		12	1398	TCTA	ORF2: RT (partial)	57 (15 with id ≥ 98 %)	13	IGR	37.5 (*HOXB13*)
		13	1019	Poly-A region	ORF2: RT (partial)	1	0	Intron (exon 2 – exon 3)	1.3 (exon 3) (*PCDHGC*)
		14	385	ND	ORF1 (partial)	110 (14 with id ≥ 98 %)	0	Intron (exon 4 – exon 5)	0.4 (exon 4) (*ighm*)
		15	387	CTATTCC	ORF1 (partial)	109 (12 with id ≥ 98 %)	3	Intron (exon 2 – exon 3)	6.2 (exon 2) (FAT tumor suppressor homolog)
L1	L. me.	16	2168	ACTAATCTTATTTTAA	Endonuclease (PFAM PF02994, "Transposase_22") (partial)	2 (2 with id ≥ 98 %)	20	IGR	41.6 (*hoxc1a*)

		17	1999	ND	RT	4	19	IGR	0.8 (*ighv14-1 (25)*)
L2	L. ch.	18	2219	G	RT	2	0	IGR	3.4 (*PCDHGC*)
CoeG-SINE	L. ch.	19	1362	ND	-	1	16	Intron (exon 4 – exon 5)	0.6 (exon 5) (von Willebrand factor A domain containing 5A)
	L. me.	20	1249	Framed by a duplicated LF-SINE (428 nts)	-	1	0	IGR	> 12.8
		21	1018	ATTTT	-	1	0	IGR	18.0 (*EVX2*)
LF-SINE	L. ch.	22 (inserted within element 23)	391	TG	-	48	0	IGR	33.1 (uncharacterized protein)
Gypsy	L. ch.	23	896	CCCGCAGCGCCC CCCCCAGAGAAT	RT	1	1	IGR	33.1 (uncharacterized protein)
ERV	L. me.	24	5091	AGAT	Gag, Pol, Env (partial)	1	41	IGR	10.6 (*ighv14-1 (21)*)
MITE-like	L. ch.	25	225	CCT	-	2	0	IGR	6.4 (von Willebrand factor A domain containing 5A)
		26	1311	ATTTCAAG	Derived from a hAT transposon	1	5	IGR	2.8 (*CHRNB4)*
Composite insertion	L. ch.	27	2303	T	CR1 (RT, TSD "AGT") / SINE (TSD "AAGT") / LF-SINE / CoeSINE	1	5	IGR	90.7 (*CRHR2*)

ORF = Open Reading Frame; L. ch. = *Latimeria chalumnae*; L. me. = *Latimeria menadoensis*; IGR = Intergenic Region; RT = Reverse Transcriptase; EEP = Exonuclease-Endonuclease-Phosphatase; ND = Not Detected; TSD = Target Site Duplication; *Number of Blastn hits in the analyzed regions, with hit length ≥ 80 % of insertion length and identity ≥ 80 %; ** Number of Blastn hits against *L. menadoensis* testis transcriptome with hit length ≥ 80 nts and identity ≥ 95 %.

518 **SUPPORTING INFORMATION**

519

520 **Figure S1. Protocol for insertion identification.** *L. chalumnae* (Lch) scaffolds and *L.*
521 *menadoensis* (Lme) BACs are represented in blue and pink lines, respectively. Orthology
522 relationships between Lme BACs and Lch genome (A) are determined by sequence comparison
523 using MegaBlast [30] (B), as described in methods. TEs from orthologous fragments are then listed
524 "face to face" (C) and further manually aligned to visualize orthologous insertions between the two
525 species (D). Candidate species-specific insertions are further inspected by extracting and re-aligning
526 corresponding "empty" and "filled" sites.

527

528 **Figure S2. Phylogenetic analysis of vertebrate retrovirus sequences.** Phylogeny is based on both
529 reverse transcriptase (210 amino-acids, left panel) and integrase core domain alignments (132
530 amino-acids, right panel). Reconstruction was performed with the PhyML package [37] using
531 Maximum Likelihood with optimized parameters (best NNI and SPR; optimized invariable sites)
532 and aLRT (SH-like branch supports).

533

534 **Table S1. Coordinates of orthologous fragments in *L. menadoensis* BAC clones and the *L.***
535 ***chalumnae* genome.** Orthology links were determined by similarity search as described in methods.
536 *L. chalumnae* genomic sequences were obtained from Ensembl (http://www.ensembl.org/; accession
537 LatCha1 GCA_000225785.1), *L. menadoensis* BAC sequences from NCBI
538 (http://www.ncbi.nlm.nih.gov/).

539 **Table S2. Coordinates and neighboring genes of the species-specific insertions.** Insertions are
540 numbered as in Table 2. Coordinates in bold correspond to insertions; coordinates in normal font
541 correspond to orthologous empty sites.
542
543

Figure 1

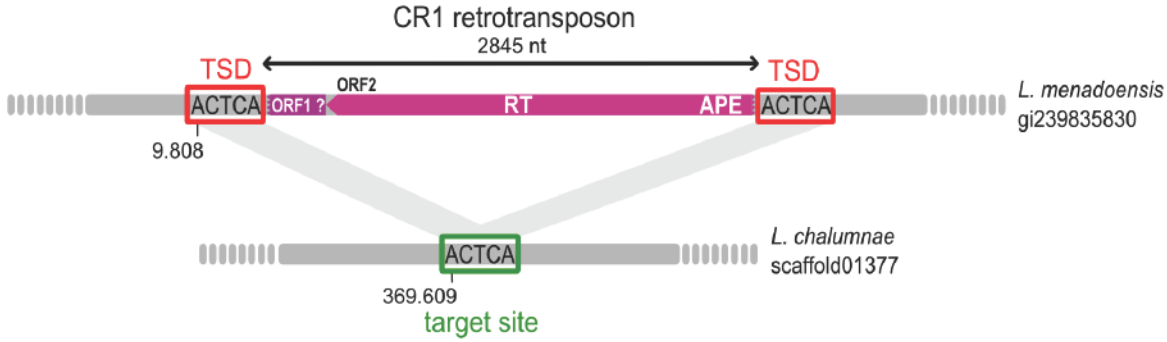


Figure 2

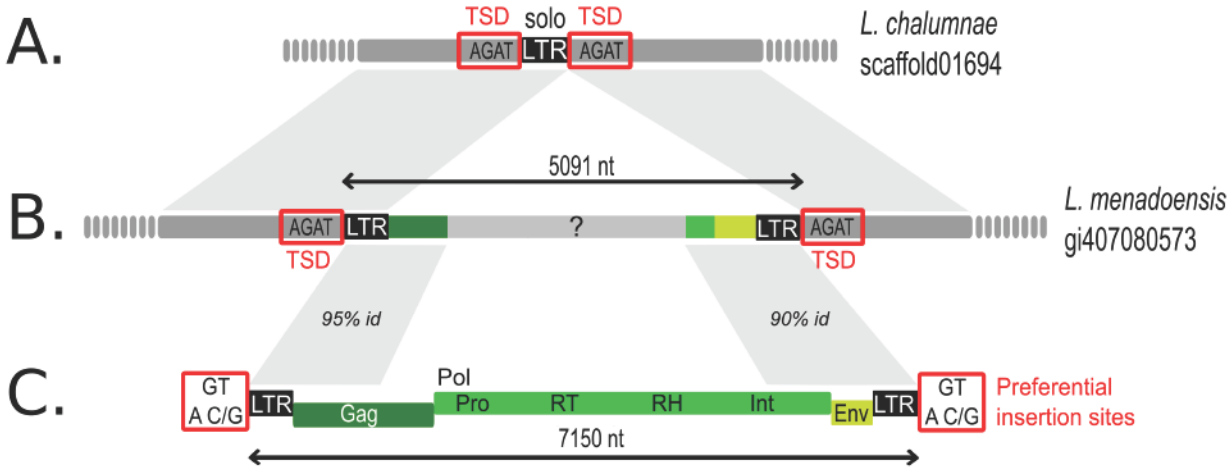


Figure 3

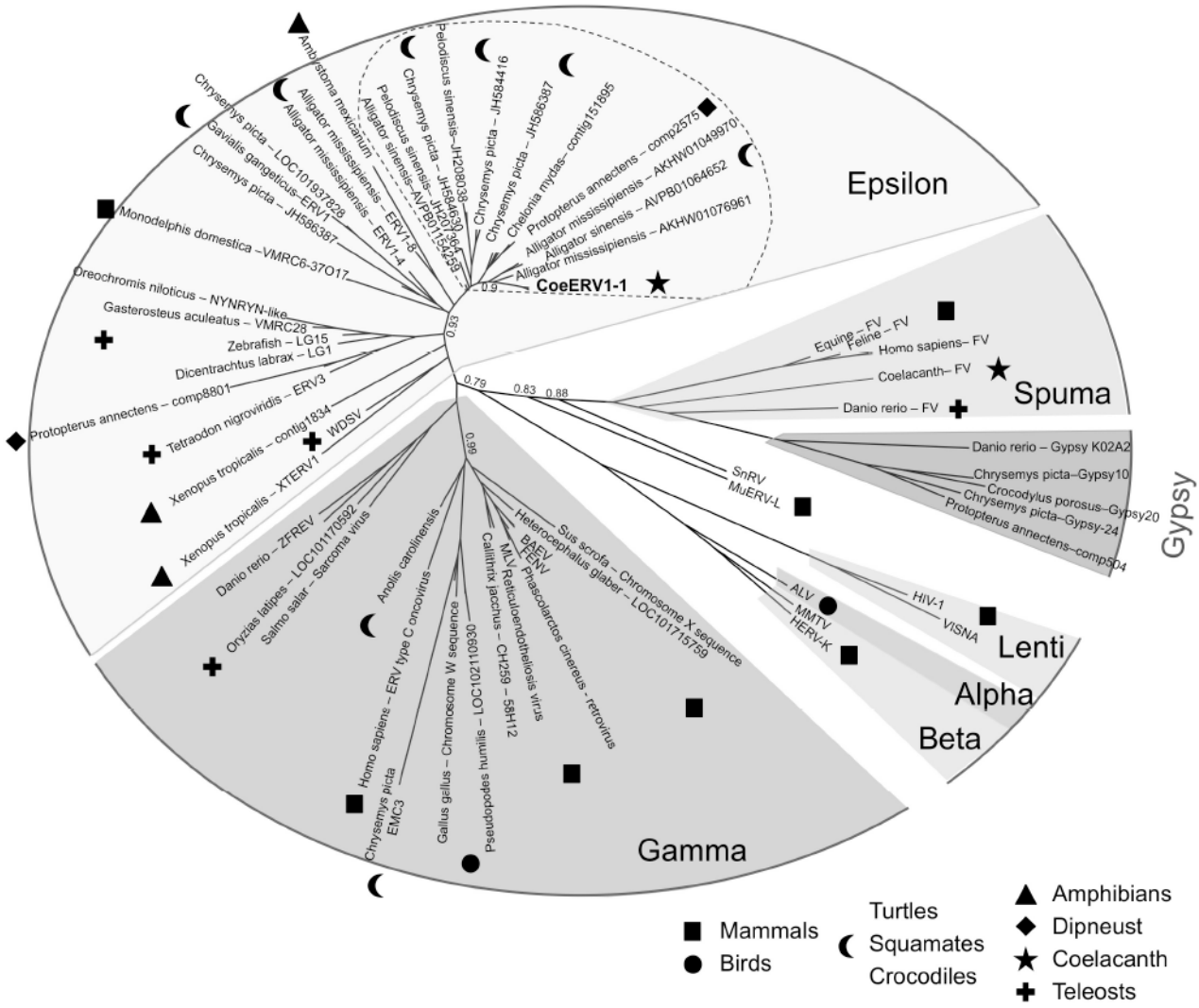

Epsilon
Spuma
Gypsy
Lenti
Alpha
Beta
Gamma
CoeERV1-1
Ambystoma mexicanum
Pelodiscus sinensis–JH208038
Chrysemys picta – JH584416
Chrysemys picta – JH586387
Chelonia mydas– contig151895
Protopterus annectens – comp2575
Alligator mississipiensis – AKHW01049970
Alligator sinensis – AVPB01064652
Alligator mississipiensis – AKHW01076961
Alligator sinensis–AVPB01152253
Alligator mississipiensis – ERV1-5
Alligator mississipiensis – ERV1-4
Chrysemys picta – LOC101937828
Gavialis gangeticus-ERV1
Chrysemys picta – JH586387
Monodelphis domestica –VMRC6-37O17
Oreochromis niloticus – NYNRYN-like
Gasterosteus aculeatus – VMRC28
Zebrafish – LG15
Dicentrachtus labrax – LG1
Protopterus annectens – comp8801
Tetraodon nigroviridis – ERV3
Xenopus tropicalis – contig1834
WDSV
Xenopus tropicalis – XTERV1
Danio rerio – ZFREV
Oryzias latipes – LOC101170592
Salmo salar – Sarcoma virus
Anolis carolinensis
Homo sapiens – ERV type C oncovirus
Gallus gallus – Chromosome W sequence
Callithrix jacchus – CH259 – 58H12
Pseudopodoces humilis – LOC101210690
Chrysemys picta EMC3
Heterocephalus glaber – LOC101715759
Sus scrofa – Chromosome X sequence
BAEV
FENV
MLV Reticuloendotheliosis virus
Phascolarctos cinereus - retrovirus
SnRV
MuERV-L
Equine – FV
Feline – FV
Homo sapiens– FV
Coelacanth– FV
Danio rerio – FV
Danio rerio – Gypsy K02A2
Chrysemys picta–Gypsy10
Crocodylus porosus–Gypsy20
Chrysemys picta–Gypsy-24
Protopterus annectens–comp504
ALV
MMTV
HERV-K
HIV-1
VISNA
0.9
0.83
0.79
0.83
0.88
0.99
Mammals
Birds
Turtles
Squamates
Crocodiles
Amphibians
Dipneust
Coelacanth
Teleosts

Figure S1

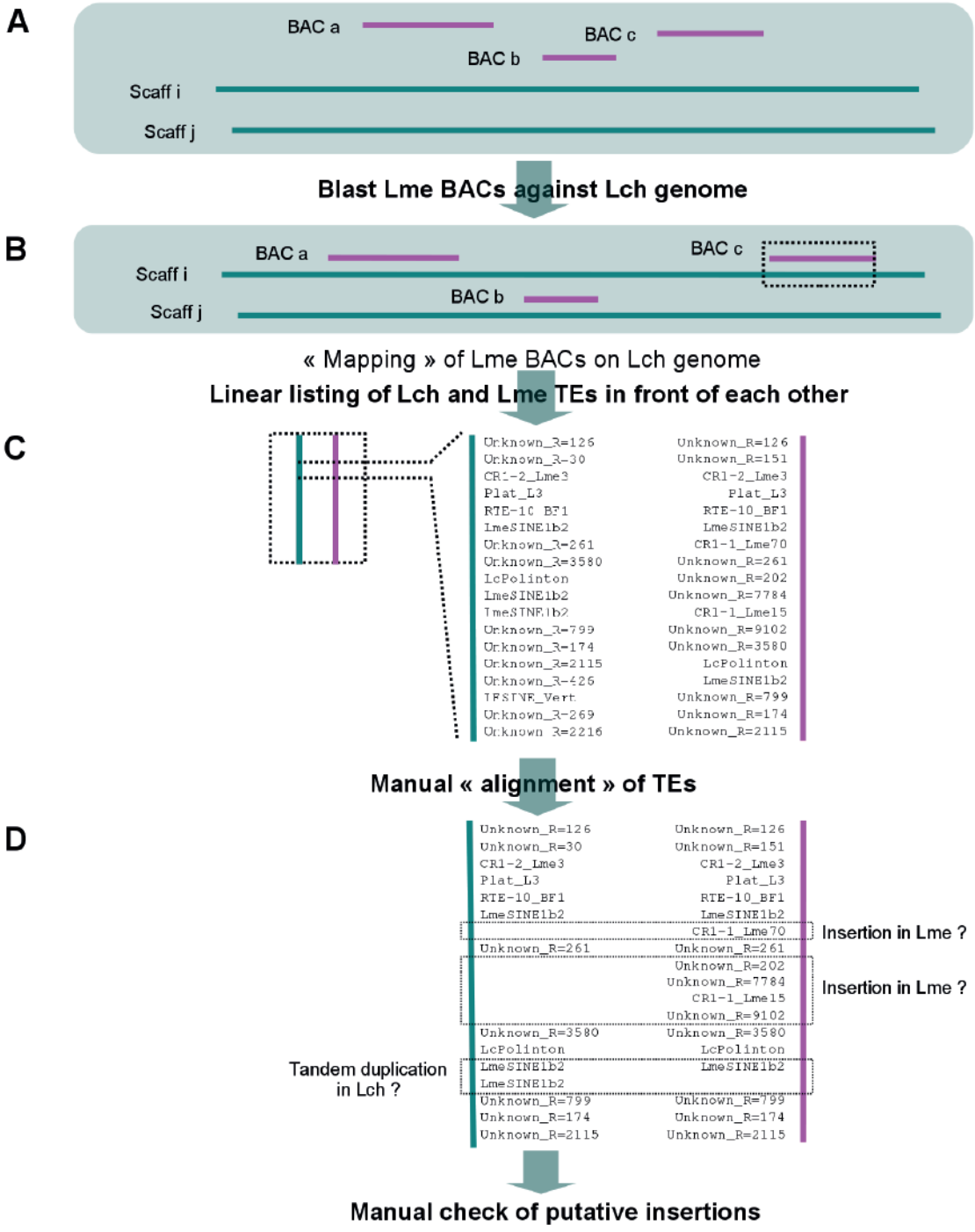

A
BAC a
BAC c
BAC b
Scaff i
Scaff j
Blast Lme BACs against Lch genome
B
BAC a
BAC c
Scaff i
BAC b
Scaff j
« Mapping » of Lme BACs on Lch genome
Linear listing of Lch and Lme TEs in front of each other
C
Unknown_R=126
Unknown_R=30
CR1-2_Lme3
Plat_L3
RTE-10_BF1
LmeSINE1b2
Unknown_R=261
Unknown_R=3580
LcPolinton
LmeSINE1b2
LmeSINE1b2
Unknown_R=799
Unknown_R=174
Unknown_R=2115
Unknown_R=426
TESTNE_Vert
Unknown_R=269
Unknown_R=2216
Unknown_R=126
Unknown_R=151
CR1-2_Lme3
Plat_L3
RTE-10_BF1
LmeSINE1b2
CR1-1_Lme70
Unknown_R=261
Unknown_R=202
Unknown_R=7784
CR1-1_Lme15
Unknown_R=9102
Unknown_R=3580
LcPolinton
LmeSINE1b2
Unknown_R=799
Unknown_R=174
Unknown_R=2115
Manual « alignment » of TEs
D
Unknown_R=126
Unknown_R=30
CR1-2_Lme3
Plat_L3
RTE-10_BF1
LmeSINE1b2
Unknown_R=261
Unknown_R=3580
LcPolinton
LmeSINE1b2
LmeSINE1b2
Unknown_R=799
Unknown_R=174
Unknown_R=2115
Unknown_R=126
Unknown_R=151
CR1-2_Lme3
Plat_L3
RTE-10_BF1
LmeSINE1b2
CR1-1_Lme70
Unknown_R=261
Unknown_R=202
Unknown_R=7784
CR1-1_Lme15
Unknown_R=9102
Unknown_R=3580
LcPolinton
LmeSINE1b2
Unknown_R=799
Unknown_R=174
Unknown_R=2115
Insertion in Lme ?
Insertion in Lme ?
Tandem duplication in Lch ?
Manual check of putative insertions

Figure S2

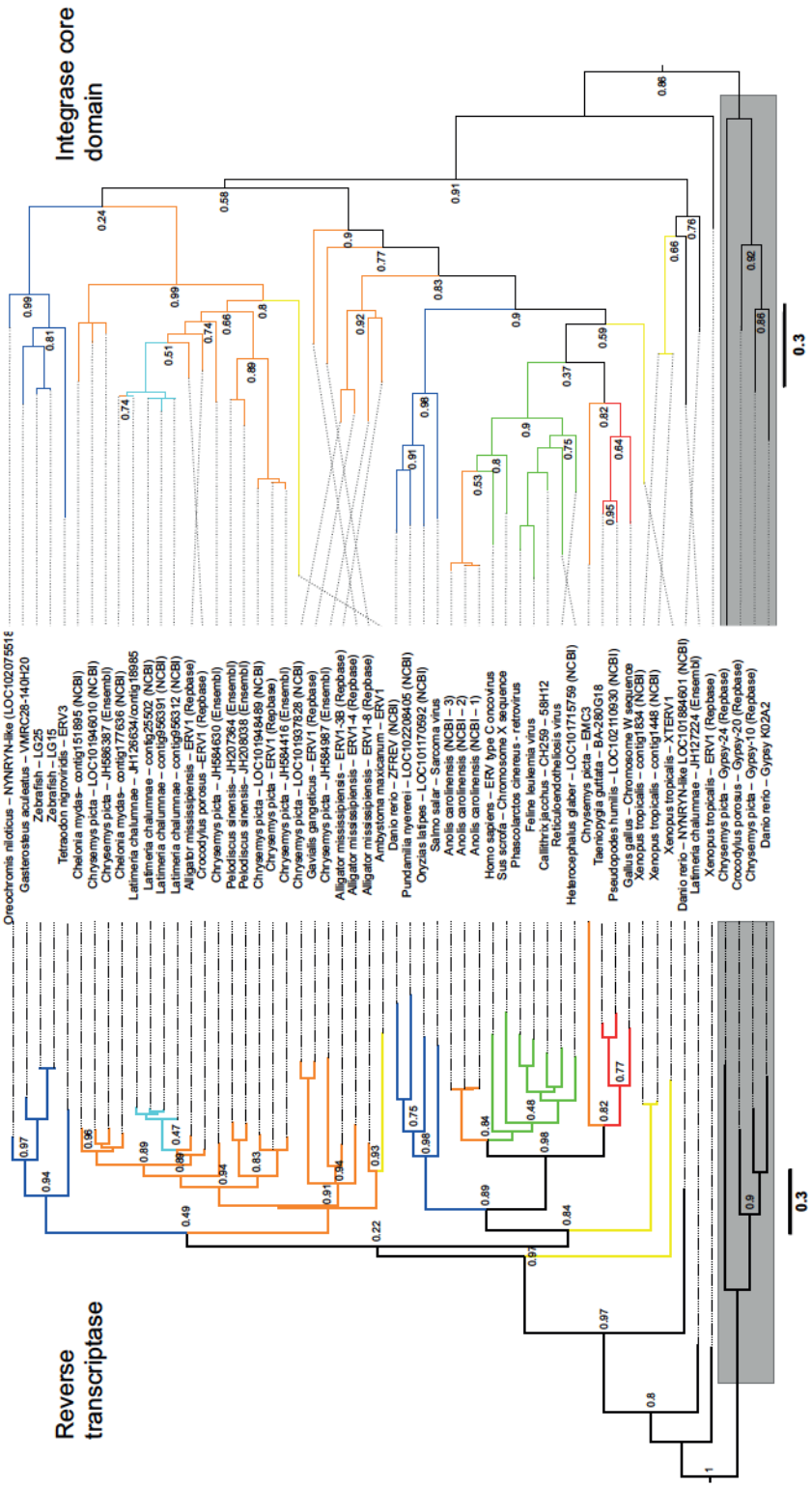

Integrase core domain
Reverse transcriptase
0.3
0.3
Oreochromis niloticus – NYNRYN-like (LOC102075516
Gasterosteus aculeatus – VMRC28-140H20
Zebrafish – LG25
Zebrafish – LG15
Tetraodon nigroviridis – ERV3
Chelonia mydas– contig151895 (NCBI)
Chrysemys picta – LOC101946010 (NCBI)
Chrysemys picta – JH886387 (Ensembl)
Chelonia mydas– contig177636 (NCBI)
Latimeria chalumnae – JH126634/contig18985
Latimeria chalumnae – contig25502 (NCBI)
Latimeria chalumnae – contig956391 (NCBI)
Latimeria chalumnae – contig956312 (NCBI)
Alligator mississipiensis – ERV1 (Repbase)
Crocodylus porosus –ERV1 (Repbase)
Chrysemys picta – JH584630 (Ensembl)
Pelodiscus sinensis– JH207364 (Ensembl)
Pelodiscus sinensis– JH208038 (Ensembl)
Chrysemys picta – LOC101948489 (NCBI)
Chrysemys picta – ERV1 (Repbase)
Chrysemys picta – JH584416 (Ensembl)
Chrysemys picta – LOC101939728 (NCBI)
Gavialis gangeticus – ERV1 (Repbase)
Chrysemys picta – JH584987 (Ensembl)
Alligator mississipiensis – ERV1-39 (Repbase)
Alligator mississipiensis – ERV1-4 (Repbase)
Alligator mississipiensis – ERV1-8 (Repbase)
Ambystoma maculatum – ERV1
Danio rerio – ZFREV (NCBI)
Pundamilia nyererei – LOC102209405 (NCBI)
Oryzias latipes – LOC101170592 (NCBI)
Salmo salar – Sarcoma virus
Anolis carolinensis (NCBI – 3)
Anolis carolinensis (NCBI – 2)
Anolis carolinensis (NCBI – 1)
Homo sapiens – ERV type C oncovirus
Sus scrofa – Chromosome X sequence
Phascolarctos cinereus - retrovirus
Feline leukemia virus
Callithrix jacchus – CH259 – 59H12
Reticuloendotheliosis virus
Heterocephalus glaber – LOC101715759 (NCBI)
Chrysemys picta – EMC3
Taeniopygia guttata – BA-280G18
Pseudopodoces humilis – LOC102110930 (NCBI)
Gallus gallus – Chromosome W sequence
Xenopus tropicalis – contig1834 (NCBI)
Xenopus tropicalis – contig1448 (NCBI)
Xenopus tropicalis – XTERV1
Danio rerio – NYNRYN-like LOC101884601 (NCBI)
Latimeria chalumnae – JH127224 (Ensembl)
Xenopus tropicalis – ERV1 (Repbase)
Chrysemys picta – Gypsy-24 (Repbase)
Crocodylus porosus – Gypsy-20 (Repbase)
Chrysemys picta – Gypsy-10 (Repbase)
Danio rerio – Gypsy K02A2

Table S1

L. chalumnae			*L. menadoensis*			Match orientation (D: direct ; R: reverse)
Scaffold	Start	End	BAC	Start	End	
scaffold00009	3280929	3483227	GI:305644147	1	187889	R
scaffold00056	964812	1334312	GI:220898186	1081	373046	R
scaffold00059	1851179	2022554	GI:190886531	76	162566	R
scaffold00118	2047028	2214133	GI:239835829	1	166048	D
scaffold00119	237991	385773	GI:237406519	4976	147882	D
scaffold00130	383298	541448	GI:164698640	1460	160991	R
scaffold00150	1900790	2213195	GI:220898172	3028	310112	R
scaffold00155	44274	200545	GI:239835824	2666	156486	D
scaffold00254	1	176423	GI:50284580	11746	189301	D
scaffold00254	183290	281494	GI:50284579	64192	150004	D
scaffold00268	1477672	1759694	GI:220898198	120021	403307	R
scaffold00354	106073	392205	GI:407080572	1	277163	D
scaffold00384	1475305	1482115	GI:239835827	149505	156331	R
scaffold00402	700544	867644	GI:305644148	1	48164	R
scaffold00568	506160	684015	GI:239835822	1038	174368	D
scaffold00606	930938	1090454	GI:239835825	1	159151	D
scaffold00623	215249	743013	GI:220898210	2721	510561	R
scaffold00705	768998	874922	GI:239835826	29792	133856	D
scaffold00739	14474	181508	GI:193083250	3996	162395	R
scaffold00744	441927	611074	GI:239835828	1	168867	R
scaffold01111	1028	121776	GI:220898198	1876	110715	R
scaffold01303	366622	554618	GI:189459217	1354	181534	D
scaffold01377	359867	508183	GI:239835830	1	145595	D
scaffold01558	242036	417321	GI:50284581	1	170774	D
scaffold01558	440137	569572	GI:50253612	34992	161955	D
scaffold01681	112207	292021	GI:239835823	2188	173770	R
scaffold01694	223609	452608	GI:407080573	1	210648	R
scaffold01718	346125	497225	GI:296011776	3222	146768	D
scaffold01893	352535	452619	GI:170514516	8680	103197	D
scaffold01950	44085	185500	GI:239835827	1963	139437	R
scaffold01958	308291	432348	GI:239735715	3441	123477	D
scaffold03191	32534	87846	GI:66912372	29226	85540	R
scaffold04994	1	55609	GI:170514516	109303	160968	R
scaffold06915	81	12807	GI:66912372	13200	26009	D
scaffold12374	692	4651	GI:239835826	8731	12694	R

Table S2

Insertion ID	Coordinates of insertion/ empty site in *L. chalumnae*	Coordinates of insertion/ empty site in *L. menadoensis*	Closest gene (*L. chalumnae* ortholog identifier)
1	**Scaffold00623:444,442-446,063**	GI:220898210:285,592	ENSLACG00000009639 (*HOXD12*)
2	**Scaffold00118:2,061,845-2,062,904**	GI:239835829:14,875	ENSLACG00000017224 (*GRID1*)
3	**Scaffold00254:143,688-144,784**	GI:50284580:156,232	ENSLACG00000003645 (protocadherin gamma)
4	**Scaffold01694:371,635-371,861**	GI:407080573:70,685	ENSLACG00000008667 (*ighv14-1 (21)*)
5	**Scaffold01958:325,272-325,591**	GI:239735715:18,223	ENSLACG00000008317 (vomeronasal 2 receptor)
6	**Scaffold01111:49,937-50,239**	GI:220898198:68,977	-
7	Scaffold00254:219,610	**GI:50284579:89,241-92,061**	ENSLACG00000006212 (protocadherin gamma)
8	Scaffold01377:369,609	**GI:239835830:9,808-12,652**	-
9	Scaffold00254:225,605	**GI:50284579:95,826-96,999**	ENSLACG00000007011 (*PCDHGC5*)
10	Scaffold00254:159,054	**GI:50284580:170,519-171,556**	ENSLACG00000003645 (protocadherin gamma)
11	Scaffold01558:304,507	**GI:50284581:55,931-56,792**	ENSLACG00000007697 (*SRA1*)
12	Scaffold00056:1,246,007	**GI:220898186:86,058-87,455**	ENSLACG00000015543 (*HOXB13*)
13	Scaffold00254:99,438	**GI:50284580:111,150-112,168**	ENSLACG00000003645 (protocadherin gamma)
14	Scaffold01694:291,311	**GI:407080573:145,750-146,134**	ENSLACG00000007502 (*ighm*)
15	Scaffold01558:396,181	**GI:50284581:149,127-149,513**	ENSLACG00000008820 (FAT tumor suppressor homolog)
16	Scaffold00268:1,507,311	**GI:220898198:371,470-373,637**	ENSLACG00000016368 (*hoxc1a*)
17	Scaffold01694:~445,237	**GI:407080573:~4,900-6,898**	ENSLACG00000009775 (*ighv14-1 (25)*)
18	**Scaffold00254:219,832-222,050**	GI:50284579:92,281	ENSLACG00000006212 (protocadherin gamma)
19	**Scaffold00354:~368,219-369,580**	GI:407080572:~254,678	ENSLACG00000008465 (von Willebrand factor A domain containing 5A)
20	Scaffold03191:46,696	**GI:66912372:70,080-71,328**	-
21	Scaffold00623:265,988	**GI:220898210:249,373-250,390**	ENSLACG00000010437 (*EVX2*)
22	**Scaffold00402:858,209-858,599**	GI:305644148:8,349	-
23	**Scaffold00402:858,000-859,286**	GI:305644148:8,349	-
24	Scaffold01694:351,066 (solo LTR)	**GI:407080573:91,275-96,365**	ENSLACG00000008667 (*ighv14-1 (21)*)
25	**Scaffold00354:388,393-388,617**	GI:407080572:273577	ENSLACG00000008465 (von Willebrand factor A domain containing 5A)
26	**Scaffold00739:136,634-137,944**	GI:193083250:45,783	ENSLACG00000004838 (*CHRNB4*)
27	**Scaffold01681:185,790-188,092**	GI:239835823:103,074	-

CHAPTER 3: LARGE-SCALE COMPARATIVE GENOMIC ANALYSIS OF TE CONTENT AND EVOLUTION IN VERTEBRATE GENOMES

I- Original article: "Comparative analysis of transposable elements in fish highlights mobilome diversity and evolution in vertebrates" (Chalopin *et al.* submitted)

The paper presented below, combined with supplementary results presented thereafter but not included in the manuscript, represents the most important project of my thesis. It assembles all the independent projects presented in the previous chapters into one large-scale comparative study. Using the previously presented genomes, as well as currently publicly available genomes such as zebrafish, medaka or stickleback, I (re)annotated all fish genomes using the same method to estimate their content and diversity of TE superfamilies. On total, I studied TE dynamics (content, diversity and Kimura distances) in ten actinopterygian fish that I compared to eleven sarcopterygians, two non-bony vertebrates, and three uro- and cephalochordates. This large-scale comparative analysis allowed to highlight lineage specificities, as illustrated by the loss of CR1 retrotransposon family in teleost fish, as well as the poor content and diversity of TEs in bird genomes for example. All together, these results represent the first overview on TE evolution in vertebrates, opening many perspectives that will be described in the discussion.

Comparative analysis of transposable elements in fish highlights mobilome diversity and evolution in vertebrates

Domitille Chalopin[1], Magali Naville[1], Delphine Galiana[1], Jean-Nicolas Volff[1]*

[1] Institut de Génomique Fonctionnelle de Lyon, Ecole Normale Supérieure de Lyon, Centre National de la Recherche Scientifique UMR 5242, Université de Lyon I, 46 allée d'Italie, 69364 Lyon Cedex 07, France

* Corresponding author: Jean-Nicolas Volff (Jean-Nicolas.Volff@ens-lyon.fr)

Running title: Evolution of transposable elements in vertebrates

Keywords: Transposable elements, evolution, diversity, genome size, fish, vertebrates

ABSTRACT

Transposable elements (TEs) are major components of vertebrate genomes, with major roles in genome architecture and evolution. In order to characterize common patterns but also lineage-specific differences in TE content and evolution, we have compared the global mobilome of 23 vertebrate genomes, including 10 actinopterygian fish, 11 sarcopterygians and two non-bony vertebrates. We found important variations in TE content (from 6% in Tetraodon to 55% in zebrafish), with mobile DNA contributing more importantly to genome size in fish than in mammals. Some TE superfamilies were found to be widespread in vertebrates, including endogenous retroviruses, Penelope, L1 and CR1-like retrotransposons, and Tc-Mariner and hAT DNA transposons. Most other elements showed a more patchy distribution, indicative of multiples events of loss or gain. Interestingly, sequential loss of TE families was observed during the evolution of the sarcopterygian lineage, with a particularly strong reduction in TE diversity in birds and mammals. Phylogenetic trends in TE composition and activity were observed: for example, teleost fish genomes are rather dominated by DNA transposons and contain few ancient TE copies, while mammalian genomes have been predominantly shaped by non-long terminal repeat retrotransposons, with more old TE sequences. Differences were also detected within lineages: the mouse genome underwent much recent amplifications of LINE elements than the human genome, as observed in the medaka genome for retrotransposons and DNA transposons compared to the related platyfish. This study allowed to infer the composition of the ancestral vertebrate mobilome, and to identify putative cases of horizontal transfer of transposable elements. Taken together, the results obtained highlight the importance of transposable elements in the structure and evolution of vertebrate genomes, and reveals their major impact on genome diversity between and within lineages.

INTRODUCTION

The genomes of mammals and other vertebrates have been shown to be significantly repetitive, with a strong contribution of transposable elements (TEs) to genome size and architecture (Deininger et al. 2003; Kazazian 2004; Feschotte and Pritham 2007; Böhne et al. 2008; Kordis 2009). TEs are genetic elements able to move within and occasionally between genomes. Due to their mobile and repetitive nature, they are strong mediators of genome plasticity: TEs can insert into and disrupt host sequences, but can also serve as substrates for homologous recombination, this generating DNA rearrangements such as deletions, duplications, inversions and translocations (Burns and Boeke 2012). Such rearrangements can be deleterious for the host through the alteration of gene coding potential and regulation, or modification of other important genomic sequences (Kazazian 2004). TEs are therefore source of genetic diseases in human and other organisms (Vorechovsky 2010; Hancks and Kazazian 2012).

However, TEs cannot be considered solely as selfish and parasitic junk DNA with negative effects on the fitness of their host: there is now convincing evidence that they are important for the functioning and evolution of genes, gene networks, genomes and organisms, with potential roles in biodiversity and speciation (Böhne et al. 2008; Feschotte 2008; Ellison and Bachtrog 2013; Xie et al. 2013). Particularly, many exons and regulatory sequences of host genes, and even new RNA and protein-coding genes are derived from TEs, a phenomenon called molecular domestication (Volff 2006; Rebollo et al. 2012; Jacques et al. 2013; Kapusta et al. 2013). One prominent example of TE-derived gene with important function in vertebrates is the RAG1 protein, which together with RAG2 catalyzes the V(D)J somatic recombination responsible for the diversity of antigen-binding regions in immunoglobulins and T-cell receptors. Both functional and evolutionary analyses have demonstrated that RAG1 and its DNA-binding sites are derived from a DNA transposon (Hiom et al. 1998; Agrawal et al. 1998; Kapitonov and Jurka 2005).

The propensity of TEs to transpose and increase their copy number is controlled by host genome defense mechanisms such as DNA methylation and Piwi-interacting small RNAs (piRNAs) (Levin and Moran 2011). An equilibrium between transposition and elimination can be established, with phases of TE expansion and reduced activity (Le Rouzic and Capy 2005; Goodier and Kazazian 2008; Brookfield 2011). TEs can multiply in genomes either after introduction through horizontal transfer or mutational activation of resident copies until the host becomes able to regulate their activity, for example through the production of

specific piRNA (Le Rouzic and Capy 2004; Evgen'ev 2013). Prolonged reduced TE activity might lead to the elimination of the element.

Based on their mechanisms of transposition, TEs are ordered in two main classes, which are themselves split in orders, superfamilies, families and subfamilies (Finnegan 1989; Wicker et al. 2007). Class I corresponds to retrotransposons, which transpose through a "copy and paste" mechanism involving the reverse transcription of an RNA intermediate into a cDNA that will be inserted somewhere else in the genome. TE class I is composed of five orders (Malik et al. 1999; Eickbush and Jamburuthugoda 2008). Two orders possess Long Terminal Repeats (LTRs): LTR retrotransposons (including retroviruses), which harbor classical flanking LTRs in direct orientation, and *Dictyostelium* Intermediate Repeat Sequence (DIRS) elements, with more complex LTRs that can be inverted and internally repeated. The three remaining orders, LINEs, SINEs (Long/Short Interspersed Nuclear Elements) and Penelope-like elements, are non-LTR retrotransposons. Autonomous retrotransposons encode a reverse transcriptase, while non-coding non-autonomous elements like the SINE sequences are mobilized *in trans* by proteins from autonomous elements. Prominent superfamilies within retrotransposons orders in eukaryotes include Gypsy, BEL and Copia in LTR retrotransposons, Ngaro, DIRS1 and Viper in DIRS, LINE1, CR1-like (including the CR1, L2 et Rex-Babar families), RTE, Jockey, R4 and R2 in LINEs, and tRNA-, 5S- and 7SL-derived in SINEs (Wicker et al. 2007).

Class II transposons (DNA transposons) do not require reverse transcription for transposition. They are subdivided into two subclasses depending on the number of DNA strands that are cut during transposition (Wicker et al. 2007). Subclass I, in which both DNA strands are cleaved, contains TIR (Terminal Inverted Repeat) transposons, which is the most abundant and diverse order, and Crypton elements. Autonomous TIR elements encode a transposase and move through a "cut and paste" mechanism; Crypton elements use a tyrosine recombinase for transposition probably involving recombination between a circular molecule and the DNA target. Subclass II elements, which cut only one strand of DNA to transpose, include Helitrons (which replicate via a rolling-circle mechanism; Kapitonov and Jurka 2001) and Maverick/Polinton transposons (self-synthesizing transposons; Feschotte and Pritham 2005). Non-autonomous class II elements such as MITEs (Miniature Inverted Transposable Elements) use the machinery of autonomous DNA transposons to transpose.

With half of known extant vertebrate species, teleost fish represent a very diverse group of animals at the organismal, ecological but also genomic levels (Volff 2005; Nelson 2006; Ravi and Venkatesh 2008; Sarropoulou and Fernandes 2011). For example, fish show a wide range

of genome sizes (from 0.32 to 133 billion base pairs; Gregory 2001). Different fish models have been developed to study vertebrate development (medaka and zebrafish; Wittbrodt et al. 2002), cancer (platyfish; Schartl et al. 2013), speciation and behaviour (cichlids and stickleback, Jones et al. 2012) and genome structure and evolution (Fugu and Tetraodon, Jaillon et al. 2004). Other species studied, such as the Atlantic salmon, rainbow trout or Nile Tilapia, are also of economical interest. Some studies have suggested that teleost fish genome present a higher diversity of transposable elements than other vertebrate genomes, and that major differences in TE content exist between vertebrate sublineages (Volff et al. 2003; Duvernell et al. 2004; Furano et al. 2004; Böhne et al. 2008; Novick et al. 2009; Kojima and Jurka 2011). However, information on TE diversity and evolution in fish and other vertebrates is still patchy and incomplete. We therefore took advantage of the growing amount of genomic data to perform a systematic comparative analysis of TE content and activity in fish and other vertebrate sublineages. Our study uncovers common TE patterns in vertebrates, but also major differences in TE activity and evolution that very likely contributed to sublineage-specific genomic and organismal diversity in vertebrates.

RESULTS

We analyzed 23 vertebrate genomes including the genomes of 11 sarcopterygians (four mammals: human, mouse, opossum and platypus; two birds: chicken and zebra finch; three reptiles: Mississipi alligator, green anole and Chinese soft-shell turtle; one amphibian: Xenopus, aka the western-clawed frog; one coelacanth), 10 actinopterygians (spotted gar, European eel, zebrafish, cod, medaka, platyfish, tilapia, stickleback, Tetraodon and Fugu), one chondrichthyan (elephant shark) and the jawless sea lamprey. For comparison, two urochordates (Ciona and Oikopleura) and one cephalochordate (amphioxus) were used as non-vertebrate chordate outgroups (Figs. 1-4). For the tetrapods (with the exception of the Chinese soft-shell turtle) and *Ciona*, we used TE composition from pre-masked genomes available in public databases, while for other animals we constructed species-specific TE libraries (see Methods). This allowed evaluating the abundance and diversity of TE superfamilies in vertebrate genomes, as well as their transposition history. To determine the contribution of small sequences in vertebrate genomes, we also quantified TE proportions in genome after eliminating sequences shorter than 80 nucleotides and sharing less than 80% identity with their consensus in the species-specific library (see Supplemental Material).

Diversity of global TE content in vertebrate genomes

The global contribution of transposable elements to vertebrate genomes was analyzed. The analysis was performed taking into account all TE sequences (Fig. 1A) and after elimination of sequences shorter than 80 nucleotides and sharing less than 80% identity with their consensus, in order to visualize the quantity of small sequences in genomes (see Supplemental Material). After such a filtering, the estimated total genome coverage by TEs was reduced particularly in elephant shark, platyfish, European eel, spotted gar, turtle, alligator and all mammals, suggesting that these genomes contain a significant number of very short and/or degenerated elements.

As shown in Fig. 1A, TE content is variable in vertebrate genomes. The genome of "basal" vertebrates (lamprey and elephant shark), some fish species, coelacanth, *Xenopus*, non-bird reptiles and mammals contains a high fraction of TEs (>20% of the genome). In contrast, the genomes of pufferfishes (Fugu and Tetraodon) and birds are poor in TEs. TE content is very variable in sequenced fish genomes, with a ca. 10 fold difference between compact pufferfish genomes and the TE-rich genome of the zebrafish.

Contribution of TEs to vertebrate genome size

It is generally accepted that DNA repeats (TEs, satellites, tandem repeats, simple repeats) contribute to genome size variations in eukaryotes. We tested the relationship between genome size and TE in vertebrates content and observed a positive correlation, statistically supported by Pearson's test ($p=0.001$; Fig. 2 and Supplemental Material). A positive correlation was also detected after independant analysis of both actinopterygian and sarcopterygian lineages. However, a shift was observed between both regression lines. This indicated that for a same genome size, relative TE contribution was more important in actinopterygian fish than in sarcopterygians, or that for a similar TE content, sarcopterygians have larger genomes than actinopterygian fish. The shift between actinopterygian and sarcopterygian regression lines was also observed when not only TEs but also other types of repeats where included in the study (data not shown). Thess results suggest that low copy number or non-repeated sequences much more significantly contribute to genome size in sarcopterygians than in actinopterygian fish.

Relative contribution of different TE types to vertebrate genomes

The relative contribution of major types of TEs, i.e. LTR, LINE and SINE retrotransposons and DNA transposons, was estimated in the genomes analyzed (Fig. 1). Species genomes were classified into four main categories: (i) genomes with predominance of DNA transposons: Amphioxus, Ciona, most teleost fish (tilapia, platyfish, medaka, cod, zebrafish and european eel) and Xenopus; (ii) genomes with predominance of LINEs and SINEs: non-bony vertebrates (lamprey, elephant shark), some actinopterygian fish (Fugu and spotted gar), coelacanth, chicken and all mammals; (iii) genomes with predominance of LTR retrotransposons: Oikopleura; (iv) genomes with no predominance of a particular class of TEs, including some teleost fish (Tetraodon, stickleback and Tilapia), non-bird reptiles and zebra finch. Some genomes are particularly poor in DNA transposons, with a mobilome almost exclusively constituted by retroelements (elephant shark, coelacanth, birds and mammals).

Distribution of TE superfamilies in vertebrates

Some TE superfamilies, including endogenous retroviruses (ERVs), Penelope-like, LINE1 and CR1-like retrotransposons, and Tc-Mariner and hAT DNA transposons, were found to be present in all vertebrate lineages (Fig. 3). Endogenous retroviruses, which are remnants of retroviral past infections, form a very diverse group of retroelements subdivided into seven

families in vertebrates (Gifford and Tristem 2003). ERVs are very abundant in amniotes (more than 5% of the genome of mouse and opossum), but have lower copy numbers in other vertebrate and non-vertebrate lineages. No ERV sequence was detected in *Ciona*. Penelope-like elements were identified in all vertebrates and non-vertebrate species studied in this work, but with modest copy numbers in mammals. The LINE1 superfamily presents contrasted genome coverage between lineages, constituting over 5% of the genome in marsupials and placental mammals but with low copy numbers in monotremes and birds. While the CR1-like superfamily is globally ubiquitous in vertebrates, its constituting families CR1, LINE2 and Rex1/Babar present a more patchy distribution (Chalopin et al. 2013). The CR1 family is completely absent from teleost fish. On the contrary, Rex1 is not present in tetrapods except in the western-clawed frog, which might have acquired this element by horizontal transfer. Finally, Tc-Mariner and hAT are widespread DNA transposons superfamilies, which are both constituted of many families (ex: Tc1, Tc2, Tigger…) with patchy distributions (data not shown). Both Tc-Mariner and hAT superfamilies significantly contribute to all vertebrate genomes, particularly in reptiles, amphibians and fish.

Several other TE superfamilies were detected in the majority of species analyzed but with punctual lineage-specific loss events: Gypsy retrotransposons have been lost in birds, RTE retrotransposons in chicken and western-clawed frog, PiggyBac transposons in platypus (and with very low copy numbers in tetrapods), and Helitron transposons in birds.

Interestingly, many TE superfamilies are present in fish but have been lost in the tetrapod lineage leading to mammals. This is the case for Copia retrotransposons and for Maverick and Harbinger DNA transposons, which have been eliminated in mammals and also in birds. DIRS retrotransposons are absent from alligator, birds and mammals. Jockey retrotransposons were only detected in alligator and anole among tetrapods. R2 retrotransposons, which specifically insert near 28S rDNA genes in a variety of metazoan genomes (Malik et al. 1999; Luchetti and Mantovani 2013), are absent from the vast majority of terrestrial tetrapods, but recent analyses have identified these elements in the Chinese pond turtle and in zebra finch (Luchetti and Mantovani 2013). The EnSpm DNA tranposon superfamily, initially described in plants (Gierl et al. 1985) and zebrafish (Bao and Jurka 2008), was detected in coelacanth but not in most tetrapods, with low copy numbers elements in frog and turtle.

Finally, several superfamilies show a more patchy distribution, revealing multiple events of loss (or gain) of transposable elements. This is the case for BEL/Pao retrotransposons, found in arthropoda and recently identified in other animals (de la Chaux and Wagner 2011): these elements have been lost in mammals, birds/alligator, turtle, European eel and Elephant shark.

R4 retrotransposons, also targeting rDNA genes like R2, were initially found in nematodes (Burke et al. 1995) and later in fish (Volff et al. 2001), lizard (Novick et al. 2009) and coelacanth (Amemiya et al. 2013; Chalopin et al. 2013). While R4 elements were detected in most fish genomes and are strongly represented in the green anole, they were not found in birds/alligator, turtle, frog and eel. Moreover, R4 presence is very weak in mammals, spotted gar, elephant shark and lamprey. Many DNA transposon superfamilies have a patchy distribution in vertebrates, including Sola, MuDr, Merlin, Chapaev, Kolobok, Crypton and P (Fig. 3).

Teleost genomes contain a high diversity of TE superfamilies

With an average of 24 superfamilies present in each species studied, the actinopterygian lineage including teleost fish (9 species in this study) is the lineage showing the highest TE diversity in vertebrates (Fig. 3). All superfamilies were found in at least one actinopterygian species, and most of them are present in all teleost genomes (Gypsy, BEL/Pao, ERV, DIRS, Penelope, R4, R2, LINE1, RTE, LINE2, Rex1/Babar, Jockey, Helitron, Maverick, Zisupton, Tc-Mariner, hAT, Harbinger, PiggyBac and EnSpm). A strong genomic contribution of LINE2 retrotransposons as well as Tc-Mariner and hAT transposons is observed. Teleost fish are the only vertebrates that contain Zisupton transposons (Böhne et al. 2012). However actinopterygian genomes have lost the CR1/LINE3 family but contain the related Rex1/Babar elements, which are absent from tetrapods. Differences between teleost species are visible. With 27 superfamilies, zebrafish and cod present the highest TE diversity. Many losses of DNA transposons have occurred in some species, particularly in Fugu, Tetraodon, stickleback, tilapia and platyfish.

Sequential loss of TE superfamilies in the sarcopterygian lineage

Within sarcopterygians, major lineage-specific differences in TE superfamily content were observed. With 26 TE superfamilies in its genome, the coelacanth presents the highest TE richness, with a diversity similar to that observed in actinopterygian fish (Fig. 3). In contrast, tetrapods show on average only 14 superfamilies, with 21 superfamilies in Xenopus (amphibian), 15-18 in non-bird reptiles, 7-9 in birds and 11-14 in mammals. This suggests sequential elimination of ancestral TE families during tetrapod evolution. Reduction of TE diversity in tetrapods is particularly associated with, but not restricted to, loss of DNA transposon superfamilies.

In amphibians, TE landscape in western-clawed frog is essentially composed of CR1-like retrotransposons, four DNA transposon superfamilies (Tc-Mariner, hAT, Harbinger and PiggyBac) and one supplementary DNA superfamilies specific to this species, named T2 (2.3% of the genome; Hikosaka and Kawahara 2010). As a comparison, in salamanders (47% of TEs in average), many retrotransposon superfamilies are highly repeated, such as Gypsy (from 7 to 20% of the genome, depending on the species), ERV (from 2 to 11%) or CR1-like (from 2 to 8%; Sun et al. 2012a,b). DNA transposons are poorly represented (less than 7%).
CR1/LINE3, Gypsy/Ty3, Penelope, Tc-Mariner and hAT elements are particularly reiterated in non-bird reptile genomes. Birds have lost many types of TEs: only 7-9 TE superfamilies have been maintained in the two species studied, with predominance of ERVs and CR1 retrotransposons, and, to a lesser extent, of Tc-Mariner and hAT DNA transposons.
Finally, in mammals, the same TE superfamilies are found in the three sublineages (monotremes, marsupials and placentals). However, LINE2 elements are predominant in the platypus (monotreme), while LINE1 is the most reiterated non-LTR retrotransposon in opossum and placental mammals (therians). In addition, some low copy number DNA transposons (PiggyBac, MuDr and Merlin) were detected in human but not in the mouse.

Transposable element activity during the evolution of the vertebrate lineage

For the genomes studied, Kimura distances (K) were calculated for all TE copies of each element in order to estimate TE "age" and transposition history (Kimura 1980). Copy divergence is correlated with the age of activity: very similar copies (low K-values) are indicative of rather recent activity (on the left part of the graph), while divergent copies (high K-values) have been generated by more ancient transposition events (on the right part of the graph). Results were grouped for the four different types of TEs (DNA transposons, LTR, LINE and SINE retrotransposons) (Fig. 4).
Mammalian profiles are characterized by a strong predominance of retroelements compared to DNA transposons. In the human genome, one major ancient transposition burst mainly involving LINEs was detected, as well as a more recent important expansion of SINEs that was not associated with an increase in LINE activity. This contrasts with the situation in mouse, where evidence for more recent LINE (and LTR) amplification was observed, but without strong increase in SINE copy number. Two major concomitant LINE/SINE amplifications were detected in opossum (in addition to a more recent LTR burst), and one in the platypus. In contrast to the situation in human and opossum, the identification of LINEs elements with very low K-values suggests the presence of recent, possibly active copies in

mouse and platypus. Ancient divergent LINE elements with high K-values were found in therians but not in platypus (monotreme). Including publicly available mammalian Kimura profiles in the analysis confirmed that mammalian genomes are mostly shaped by non-LTR retrotransposons and a regular activity of LTR retrotransposons over time (see Supplemental Material). Sublineage-specific transposition bursts were observed (SINEs in primates, DNA transposons in hyrax and bats; Pritham and Feschotte 2007).

Two major transposition bursts were detected in birds: one involving LINEs and DNA transposons, and the second LINEs in chicken, and the oldest one with LINEs and the youngest one with LTR retroelements in zebra finch. Ancient TE copies were identified, but few recent elements.

In contrast to the situation observed in birds and mammals, DNA transposons have been very active during the evolution of the three non-bird reptile species analyzed. LINEs have also strongly contributed to the genomes of these species. Profiles are relatively similar in alligator and turtle, with ancient and "middle-aged" bursts of transposition and few recent copies. In contrast, the genome of the green anole has undergone a younger general burst of transposition and contains recent copies from all four TE classes.

The genome of Xenopus, the only representative of the amphibian lineage included in this study, has been predominantly shaped by DNA transposons, and underwent more recently amplification of LINEs. Many recent copies were detected, suggesting transposition activity mostly due to DNA transposons.

The coelacanth genome is dominated by LINEs and SINEs, with at least one major "middle-aged" transposition burst and some recent LINE copies.

In actinopterygian fish, the spotted gar, which is a non-teleost species, has a genome that has been shaped by all four TE classes, with two major bursts of activity and few recent copies. Within teleosts, significant interspecific differences in profiles were observed, with generally one or two general bursts of transposition. Some genomes are dominated by rather recent copies (cod, medaka, stickleback, Fugu), some by ancient copies (platyfish), and some by both (eel, tilapia, Tetraodon), with no clear phylogenetic signal: both pufferfishes show clearly different patterns, as it is also the case for the related medaka and platyfish. Teleost genomes generally contain fewer very ancient copies (K-values > 25) than mammalian genomes, suggesting lineage-specific differences in the dynamics of TE elimination. In contrast to gar, most teleost genomes studied have been strongly shaped by DNA transposons. This is particularly the case for zebrafish, which shows the higher amplification of DNA transposons among the vertebrates studied here. LINEs significantly contribute to the genome

of several species including medaka, tilapia and Fugu, and a significant middle-aged burst of LTR elements (around K-value 19) was detected in Tetraodon. Recent copies suggesting activity were particularly identified in zebrafish, tilapia, stickleback and pufferfishes.

Elephant shark Kimura profile is unique in vertebrates. It is mostly constituted by LINE retrotransposons and relatively recent LTR retrotransposons, with few recent copies. The genome of the lamprey is dominated by DNA transposons and LINE retrotransposons, with many young DNA transposon copies. A very ancient burst of LTR retrotransposons was also detected.

In non-vertebrate species included in this work, Ciona and amphioxus genomes mainly contain DNA transposons and LINE retrotransposons, while Oikopleura is mainly composed of LTR retrotransposons. Active copies are probably present in these species, with an extremely recent strong burst of DNA transposons and LTR retrotransposons in Oikopleura.

DISCUSSION

In this work, we have analyzed the different types of transposable elements present in vertebrate genomes. Using species-specific TE libraries, we have analyzed retrotransposons and DNA transposons in sequenced genomes from species covering major branches of the vertebrate lineage. This study uncovered an important inter- and intra-lineage diversity concerning the nature, genomic contribution, activity and evolution of transposable elements in vertebrate genomes.

Diversity of TE contribution to genome size in vertebrates

TEs and other repeats make up an important part of most vertebrate genomes. However, the global contribution of TEs is variable between lineages: for example, the genome of mammals contains many more TEs than the genome of birds. Variability in TE content is also observed within lineages: in teleost fish, the genome coverage of TEs is 10x higher in zebrafish (55% of the genome) than in the pufferfish Tetraodon (6%). Short TE-related sequences strongly contribute to some vertebrate genomes including those of mammals. It is of course important to note that we focused on already sequenced genomes; other particularly TE-rich (or TE-poor) vertebrate genomes are still to be sequenced, for example the large genomes of salamanders and lungfish. In addition, our evaluation of TE content is certainly an underestimation: we worked on assembled genome drafts, which in principle do not include TE-rich regions of the genomes like centromeres or other heterochromatic regions. Our

methods of analysis were very conservative, and may have missed other types of TEs, or very old and divergent elements. Using alternative methods, it has been for example estimated that TE content in the human genome might be as high as 66-69% (de Koning et al. 2011).

Factors influencing genome size and DNA content variation between species are multiple, including whole genome duplications, segmental duplications, deletions and DNA repeat proliferation (Parfrey et al. 2008). It has been established that TEs and other DNA repeats play an important role in genome size diversity (Petrov 2001; Kidwell 2002; Ågren and Wright 2011). In insects, both satellite sequences and TEs have been implicated in genome size variation (Vieira et al. 1999; Vieira et al. 2002; Kidwell 2002; Bosco et al. 2007). Accordingly, we showed a correlation between TE content and genome size in vertebrates, indicating that larger genomes tend to have more transposable elements than smaller genomes. Such a correlation was also observed after testing separately actinopterygian fish and sarcopterygians. However, the results obtained suggested that TEs contributed much more significantly to genome size in actinopterygian fish than in sarcopterygians. This might also indicate that sarcopterygian genomes contain a significant fraction of very old, divergent sequences that were identified neither as repeats nor as TEs in our study.

Inter- or intra-lineage differences in TE contribution to genomes might be explained by variability in TE activity, which can be influenced by transposition rates of TEs present in the genomes, competition between TEs and variations in host-mediated defense mechanisms against mobile elements (Le Rouzic and Capy 2006). TE elimination rate is also an important parameter, species with a slow rate of DNA loss tending to increase their genome size (Petrov 2001; Sun et al. 2012a). The Kimura distance-based comparative analysis performed in this work indeed suggested strong variability in TE activity between vertebrates, with important differences in the number of recent potentially active elements. For example, the number of recent copies is low in human and opossum but higher in the mouse. The genome of the green anole continues to sustain strong transposition activity while there are almost no active copies in the alligator anymore. In addition, lineage-specific differences in TE elimination rates might be also involved. For instance, large mammalian genomes contain more ancient divergent and fractionated TE copies than fish genomes. Within actinopterygians, differences were even observed between related fish species: the genome of the platyfish contains many more old TE copies than that of the related medaka. Differences in TE elimination between fish species is supported by the genome architecture of the pufferfish: in this compact genome, all types of repeats are excluded from euchromatic gene-rich regions and accumulate

in particular heterochromatic compartments, a structure generally not observed in other fish species (Dasilva et al. 2002; Fischer et al. 2005).

TE landscape diversity in extant vertebrate species

Mobilome diversity in vertebrates is manifested not only by global variations in TE content between lineages and species but also by differences in the types and superfamilies of TEs present in genomes, and by their differential colonization success. In this study we showed that major differences exist between and even within lineages regarding the different types of TEs composing the mobilome. For example, the genomes of mammals, birds, coelacanth and elephant shark have been almost exclusively shaped by retroelements and contain few DNA transposons, while DNA transposons are the most prominent type of TEs in teleost fish and Xenopus genomes. Within mammals, LTR elements constitute a significant part of mobilome in therians (human, mouse and opossum) but not in platypus (monotreme). In fish, TE composition of zebrafish and Tetraodon are extremely different, and each fish species possesses its own Kimura distance-based TE profile in terms of relative contribution of each type of TE.

This diversity is also observed considering the number of TE superfamilies. Many examples of lineage-specific loss (or gain) of TE superfamilies have been identified. Some vertebrate lineages contain many TEs superfamilies, including teleost fish (24 TE superfamilies on average per genome), the coelacanth (26 TE superfamilies) and Xenopus (amphibian, 21 TE superfamilies). In contrast, a strong reduction of TE diversity was observed in mammals (11-14 TE superfamilies) and birds (7-9 superfamilies). The three non-bird reptile species analyzed showed an intermediate TE richness, with 15-18 superfamilies. These results suggest a reduction of TE diversity through sequential elimination of TE superfamilies in the sarcopterygian lineages having led to mammals and birds. In birds, the small genome size and low TE content suggest that loss of certain TE families might be a consequence of general constraints acting toward a reduction of non-coding DNA content in the genome. In contrast, loss of certain TE superfamilies in the large genomes of mammals might be associated with the extreme success of specific families of LINE and SINE non-LTR retrotransposons, for example LINE1 and Alu sequences in primate genomes. As a result of competition for genomic resources, resident successful families might have supplanted and eliminated other types of TEs. Alternatively, extinction of some TE families might have been driven by mutational inactivation or through the development a new specific defense systems by the host, allowing the massive opportunistic expansion of the remaining TE families in genomes.

Even when a same TE superfamily or family is present in different genomes, differences in copy number might contribute to lineage divergence. This is the case even for TE superfamilies present in all vertebrate lineages, including endogenous retroviruses (ERVs), Penelope-like, LINE1 and CR1-like retrotransposons, and Tc-Mariner and hAT DNA transposons. For example, ERVs are very abundant in amniotes but present lower copy numbers in other vertebrates. Penelope-like elements display very modest copy numbers in mammals. The LINE1 superfamily is a major component of the genome of marsupials and placental mammals, but is poorly represented in monotremes and birds.

Toward an inference of the ancestral vertebrate mobilome?

This analysis provides a frame for a first attempt to approach the ancestral vertebrate mobilome, i.e. TE composition in terms of diversity in the last common ancestor (LCA) of the vertebrate species studied. This is a very difficult task, particularly because TEs can be also introduced through horizontal transfer into lineages.

If we assume a major mode of vertical transmission, we can infer that many superfamilies of autonomous transposable elements were present in the genome of the vertebrate LCA. TE superfamilies found in almost or all species studied were probably represented in the ancestral vertebrate mobilome, including Gypsy and ERV LTR retrotransposons, Penelope-like retrotransposons, LINE1, RTE, and CR1-like non-LTR retrotransposons, as well as Helitron, Tc-Mariner, hAT and PiggyBac DNA transposons. Some SINE elements are also widespread, like the V-SINE elements (Ogiwara et al. 2002; Piskurek and Jackson 2011), but many have been formed in specific vertebrate lineages. LCA mobilome probably also included superfamilies present in jawless vertebrates, chondrichtyans, actinopterygians, amphibians and reptiles but lost in birds and mammals, including Copia and DIRS LTR retrotransposons as well as Maverick and Harbinger DNA transposons.

However, the possible acquisition of TEs by horizontal gene transfer (HGT) should not be excluded particularly for DNA transposons, even if HGT events are rather rare in vertebrates (Wallau et al. 2012; Syvanen 2012). HGT has been largely described in insects (Sormacheva et al. 2012). In vertebrates, the SPIN DNA transposon, which presents a patchy distribution in vertebrates, has been transmitted horizontally several times in mammals and other tetrapods (Pace et al. 2008; Gilbert et al. 2012). HGT of the non-LTR retrotransposon BovB has been reported between reptiles and within mammals (Kordis and Gubensek 1999; Walsh et al. 2013). Horizontal transmission of Tc-Mariner elements possibly also occurred between teleosts and lampreys (Kuraku et al. 2012). Other putative cases have been reported (Novick

et al. 2010; Thomas et al. 2011; Oliveira et al. 2012; Gilbert et al. 2013). Retroviruses, which infect vertebrates, have been proposed to serve as vectors for HGT (Yohn et al. 2005; Piskurek and Okada 2007). According to their patchy distribution, several TE superfamilies are candidates for HGT: MuDr in human, Merlin in stickleback, zebrafish, western-clawed frog and human (Feschotte 2004), Chapaev in the green anole, fugu and platyfish, and P in platyfish and coelacanth. Alternatively, these TEs might have been lost repeatedly during the evolution of the vertebrate lineage. Much work is required to determine the modes of acquisition and loss of these elements in vertebrates.

Conclusions

In this work, we present an overview of TE content, diversity, activity and evolution in the main vertebrate lineages. The results obtained highlight inter- and intra-lineage diversity, showing that differential transposable element activity and evolution has strongly contributed to genome divergence in vertebrates. TEs can also mediate diversity through lineage-specific events of molecular domestication, this leading to new gene regulations and functions (Böhne et al. 2008). The functional consequences of lineage-specific TE expansion on genome architecture and regulation remain to be investigated. Further work on individual TE families and subfamilies will uncover new aspects of TE dynamics in vertebrate and allow to discover new cases of horizontal transfer.

METHODS

Genomic datasets

To built TE libraries, we collected genomes of amphioxus (Branchiostoma_floridae_v2.0.assembly.fasta on JGI amphioxus project), Oikopleura (Odioica_reference_v3.fa on Genoscope website), lamprey (Petromyzon_marinus.Pmarinus_7.0.70.dna.toplevel.fa at Ensembl server) (http://www.ensembl.org/index.html), elephant shark (EsharkAssembly at http://esharkgenome.imcb.a-star.edu.sg), fugu (Takifugu_rubripes.FUGU4.66.dna.toplevel.fa at Ensembl server), Tetraodon (Tetraodon_nigroviridis.TETRAODON8.73.dna.toplevel.fa at Ensembl server), stickleback (Gasterosteus_aculeatus.BROADS1.68.dna.toplevel.fa at Ensembl server), tilapia (Oreochromis_niloticus.Orenil1.0.68.dna.toplevel.fa at Ensembl server), platyfish (Xiphophorus_maculatus.Xipmac4.4.2.69.dna.nonchromosomal.fa at Ensembl server), medaka (Oryzias_latipes.MEDAKA1.73.dna.toplevel.fa at Ensembl server), Atlantic cod (Gadus_morhua.gadMor1.73.dna.toplevel.fa at Ensembl server), zebrafish (Danio_rerio.Zv9.66.dna.toplevel.fa at Ensembl server), European eel (draft genome version 1, www.zfgenomics.org/sub/eel), spotted gar (NCBI accession number GCA_000242695), African coelacanth (Latimeria_chalumnae.LatCha1.72.dna_toplevel.fa at Ensembl server) and Chinese soft-shell turtle (Pelodiscus_sinensis.PelSin_1.0.73.dna.toplevel.fa at Ensembl server).

For tetrapods (except for turtle) and Ciona, we directly use pre-masked genomes and RepeatMasker outfiles (".out" and ".align") on RepeatMasker Genomic Datasets (http://www.repeatmasker.org/genomicDatasets/RMGenomicDatasets.html): ciona (ci2), frog (xenTro2), American alligator (allMis0), green anole (anoCar2), zebra finch (taeGut1), chicken (galGal3), platypus (ornAna1), opossum (monDom5), mouse (mm9) and human (hg19).

For pre-masked genomes, genome sizes correspond to the golden path available on the Ensembl server. For others, genome sizes were calculated during the masking process.

Construction of species-specific transposable element libraries

We established species-specific TE library by combining automatic and manual annotations for the following species: amphioxus, lamprey, elephant shark, Fugu, Tetraodon (only manual combined with the fugu library), stickleback, tilapia, platyfish, zebrafish (manual and Repbase sequences) (Jurka 2000) and spotted gar. Manual annotation consists in the search of TE sequences through tBlastN (Altschul et al. 1990) using proteins from different TE

superfamilies as queries. Reverse transcriptases and transposases were "blasted" against genomes to find retrotransposon and DNA transposon sequences, respectively. The longest sequences containing specific TE features such as TIRs or LTRs were kept for further analysis. Censor (Jurka et al. 1996) was also used to test the nature of the sequences. Automatic annotation was performed using the RepeatModeler software (Smit, AFA, Hubley, R. *RepeatModeler Open-1.0*. 2008-2010 <http://www.repeatmasker.org>) with default parameters. For the coelacanth, we used and reannotated the library from Amemiya et al. (2013). We completed the annotation of the "Unknown" sequences using Censor.

Sequence alignments and phylogenetic reconstructions

Consensus TE nucleotide sequences were retrieved from TE libraries, translated into proteins using Augustus (human and chicken models) (Stanke et al. 2004) and Softberry (fish and zebrafish models) (Softberry Inc.), and aligned using Clustal omega (Sievers et al. 2011). Phylogenetic trees were reconstructed using maximum likelihood with optimized parameters and default aLRT (non-parametric branch support) using the Seaview interface (Gouy et al. 2010).

Genome masking

Amphioxus, Oikopleura, lamprey, elephant shark, Fugu, Tetraodon, stickleback, tilapia, platyfish, medaka, cod, zebrafish, spotted gar, coelacanth and soft-shell turtle genomes were locally masked using RepeatMasker version 3.3.0 (Smit, AFA, Hubley, R & Green, P. *RepeatMasker Open-3.0*.1996-2010 <http://www.repeatmasker.org>) software with "-a" and "-lib" default parameters.

Copy number and genome coverage estimation

Copy number and genome coverage were calculated on RepeatMasker outfiles (".out"). Copy number corresponds to listed insertion of masked genomes, removing intra-TEs insertions. Total copy number and coverage for each superfamily were calculated using a home script. Additionally, a second calculation included only sequence insertions longer than 80 nucleotides and sharing more than 80% of identity with the reference sequence from the species-specific library (Supplemental Material). This last calculation eliminates very short and divergent sequences in vertebrate genomes.

Kimura distance-based distribution analysis of TE copies in genomes

Kimura distances were calculated on alignments included in ".align" files after genome masking. The rates of transitions (A<->G or C<->T) and transversions (purine to purine, and vice versa) were calculated on alignments and transformed to Kimura distance (Kimura 1980) by using [K= - ½ ln(1 – 2p – q) – ¼ ln(1 – 2q)] where "q" is the proportion of sites with transversions and "p" is the proportion of sites with transitions.

ACKNOWLEDGEMENTS

DC is supported by a PhD grant from the French Ministry for Higher Education and Research and by a PhD grant from the "Ligue contre le cancer". We thank the CBP and PSMN centers at the ENS Lyon, especially Emmanuel Quemener for technical IT supports. We also acknowledge Emmanuelle Lerat and Clément Goubert from LBBE for assistance and advices. We also thank Marie Sémon for her helpful comments.

AUTHOR CONTRIBUTIONS

DC performed TE libraries, masking and analyses, and drafted the manuscript. MN wrote the home script to generate TE copy number and content data. DG participated in data analysis. JNV coordinated the study, participated in data analysis and edited the manuscript.

DISCLOSURE DECLARATION

We have no disclosure declaration.

FIGURE LEGENDS

Figure 1: Percentages and relative proportions of DNA transposons, LTR, LINE and SINE retrotransposons in vertebrate genomes

The amount of DNA transposons, LTR, LINE and SINE retrotransposons, and unclassified elements (Unknown), as well as their respective proportions (unclassified elements were not included) were estimated based on RepeatMasker outfiles. A. Percentages of TEs in the different vertebrate and non-vertebrate genomes. B. TE relative proportions in genomes. Abbreviations: NV Non-vertebrates; NBV Non-bony vertebrates; AF Actinopterygian fish; LF Lobe-finned fish; A Amphibian; R non-bird Reptiles; B Birds; M Mammals.

Figure 2: Correlation between genome sizes and transposable element contents in vertebrates

Actinopterygian fish are represented by blue dots, sarcopterygians by pink dots, non-bony vertebrates by empty dots. Pearson's correlations were estimated taking into account all vertebrates (black line), only actinopterygian fish (blue dashed line) and only sarcopterygians (pink dashed line). The three person's correlations are significantly (p-value < 0.05) positive ($0.1<R<0.9$). Abbreviations: Hs *Homo sapiens*, human; Mm *Mus musculus*, mouse; Md *Monodelphis domestica* opossum; Oa *Onrithoryncus anatinus*, platypus; Gg *Gallus gallus*, chicken; Tg *Taeniopygia guttata*, zebra finch; Ac *Anolis carolinensis*, green anole; Am *Alligator mississippiensis*, American alligator; Ps *Pelodiscus sinensis*, Chinese soft-shell turtle; Xt *Xenopus tropicalis*, western-clawed frog; Lc *Latimeria chalumnae*, coelacanth; Lo *Lepisosteus oculatus*, spotted gar; Aa *Anguilla anguilla*, European eel; Dr *Danio rerio*, zebrafish; Gm *Gadus morhua*, Atlantic cod; Ol *Oryzias latipes*, medaka; Xm *Xiphophorus maculatus*, platyfish; On *Oreochromis niloticus*, tilapia; Ga *Gasterosteus aculeatus*, stickleback; Tr *Takifugu rubripes*, fugu; Tn *Tetraodon nigroviridis*, Tetraodon; Cm *Callorhinchus milii*, elephant shark; *Petromyzon marinus*, lamprey; Bf *Branchiostoma floridae*, amphioxus.

Figure 3: Diversity and abundance of TE superfamilies in vertebrates

Presence/absence of TE superfamilies was determined using automatic annotation, manual verification and literature information (R2 for zebra finch, BEL for Ciona). Presence of superfamilies is shown by full squares with or without grey gradient. Absence is represented by dashed squares. The pre-masked genome of the western-clawed frog does not provide

details within the CR1-like superfamily. Number of superfamilies present in each species is indicated on the right side (number of superfamilies covering more than 0.001% of the genome and total number of superfamilies). Divergence times between species were estimated using the TimeTree public database (Hedges et al. 2006) and the literature data based on fossil records. The divergence time between zebrafish and other teleost fish is estimated between 160 and 230 million years.

Figure 4: Kimura-distance based distribution of transposable elements in vertebrate and non-vertebrate genomes

The graphs represent the percentage of copies (Y-axis) for each type of TE (DNA transposons, SINE, LINE and LTR retrotransposons) in the different genomes analyzed, clustered according to Kimura distances (X-axis, K-value from 0 to 50). Copies clustering around 0 do not much diverge from consensus sequences of the species-specific TE library (potentially recent copies), while sequences around 50 are very divergent copies (potentially ancient copies).

SUPPLEMENTAL MATERIAL LEGENDS

Supplementary Figure 1: Contribution of small TE sequences to vertebrate genomes

Histograms represent, for each species, the total TE content in genomes (left bars) and TE content discarding sequences smaller than 80 nucleotides and sharing more than 80% of identity with consensus sequences from libraries (right bars).

Supplementary Figure 2: Kimura profiles within mammalian lineage

Kimura profiles were recovered from the RepeatMasker genomic datasets website (http://www.repeatmasker.org/genomicDatasets/RMGenomicDatasets.html) and placed along mammalian phylogeny.

Supplementary Table 1: Genome content for DNA transposons, LTR, LINE, SINE retrotransposons and Unclassified TEs with and without filter

Supplementary Table 2: Data used to generate the plot of the genome sizes versus TEs

Supplementary Table 3: Statistical values of Pearson's correlations.

The tests were performed on three different groups of samples (all vertebrates, only actinopterygian fish and only sarcopterygians), and comprising all repeats (with and without small sequences) or only TEs (with or without small sequences). The estimated values correspond to the p-value of the test (p-value) and to the estimated measure of association (cor) corresponding to Pearson's analysis.

Supplementary File 1: Raw statistics of transposable element family and superfamily copy numbers and contents, for each species analyzed.

FIGURE 1

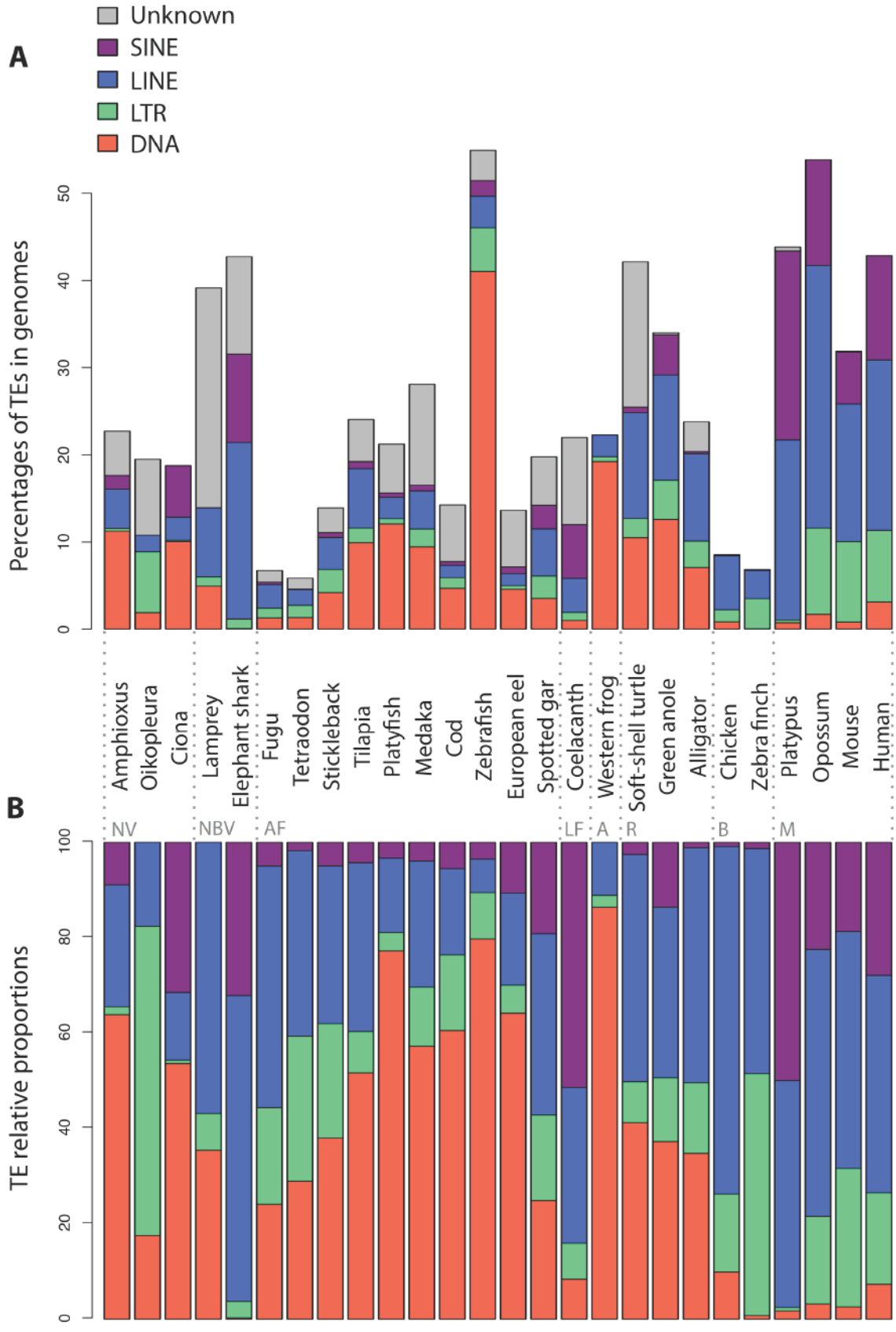

A
Unknown
SINE
LINE
LTR
DNA
Percentages of TEs in genomes
0
10
20
30
40
50
B
TE relative proportions
0
20
40
60
80
100
NV
NBV
AF
LF
A
R
B
M
Amphioxus
Oikopleura
Ciona
Lamprey
Elephant shark
Fugu
Tetraodon
Stickleback
Tilapia
Platyfish
Medaka
Cod
Zebrafish
European eel
Spotted gar
Coelacanth
Western frog
Soft-shell turtle
Green anole
Alligator
Chicken
Zebra finch
Platypus
Opossum
Mouse
Human

FIGURE 2

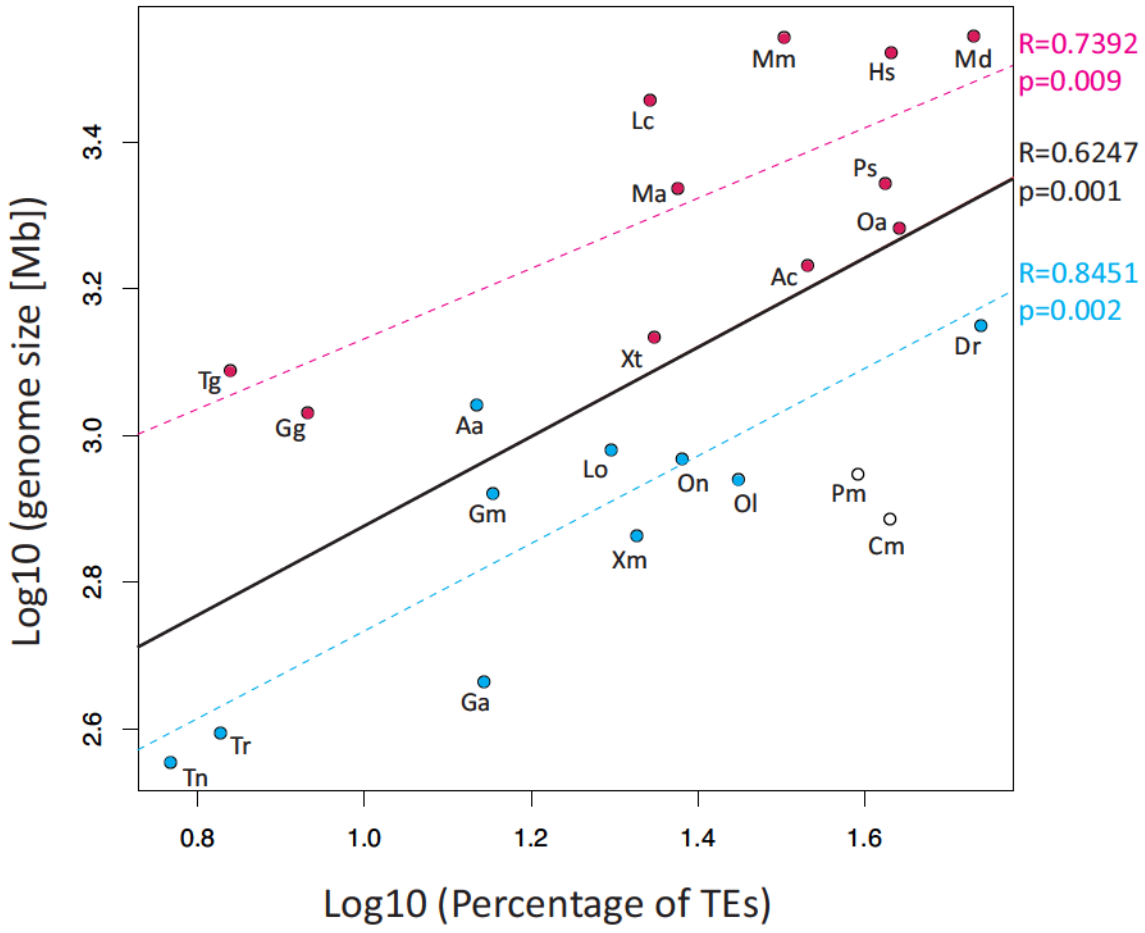

R=0.7392
p=0.009
R=0.6247
p=0.001
R=0.8451
p=0.002
Mm
Hs
Md
Lc
Ps
Ma
Oa
Ac
Xt
Dr
Tg
Gg
Aa
Lo
On
Ol
Pm
Gm
Cm
Xm
Ga
Tr
Tn
Log10 (genome size [Mb])
Log10 (Percentage of TEs)
2.6
2.8
3.0
3.2
3.4
0.8
1.0
1.2
1.4
1.6

FIGURE 3

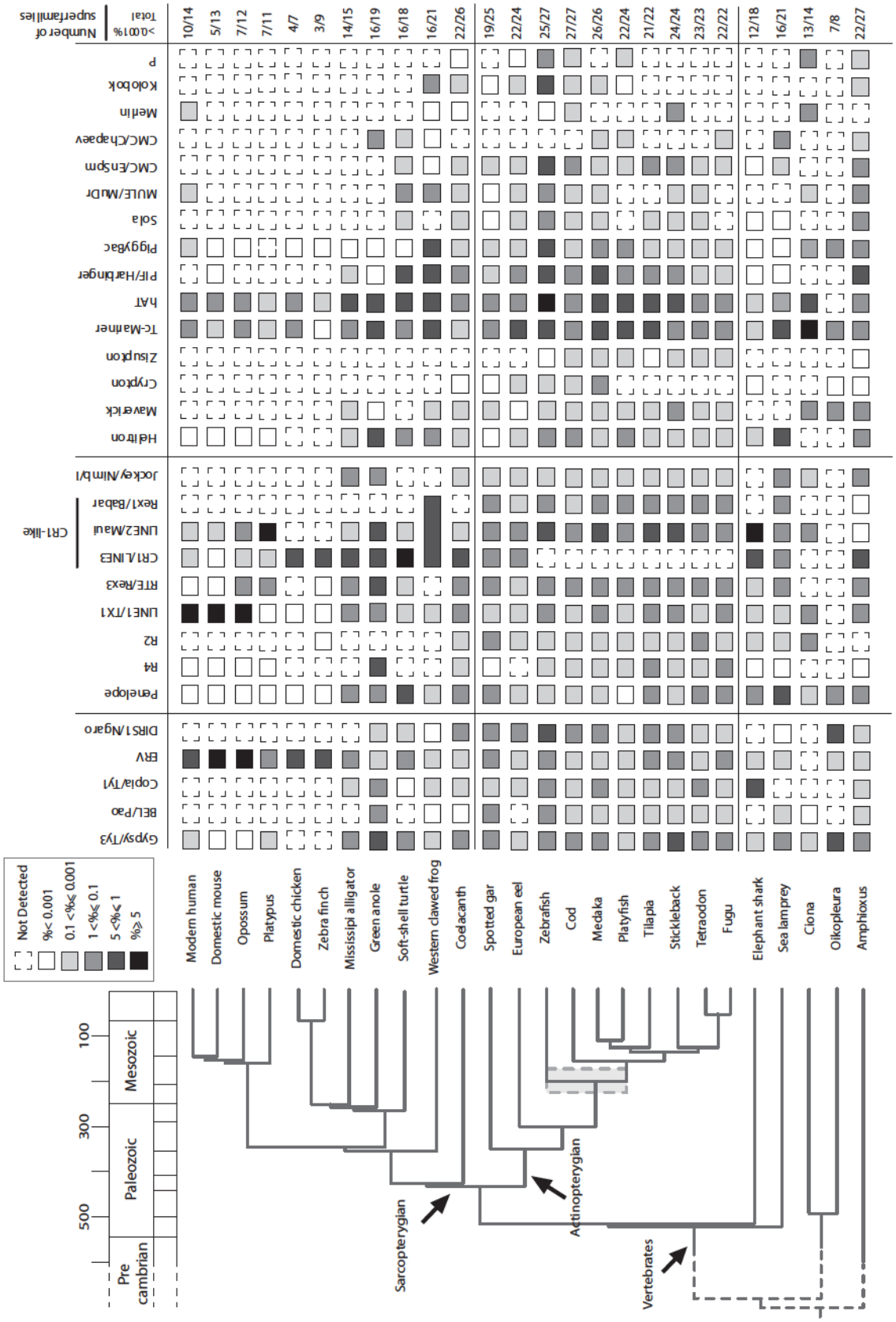
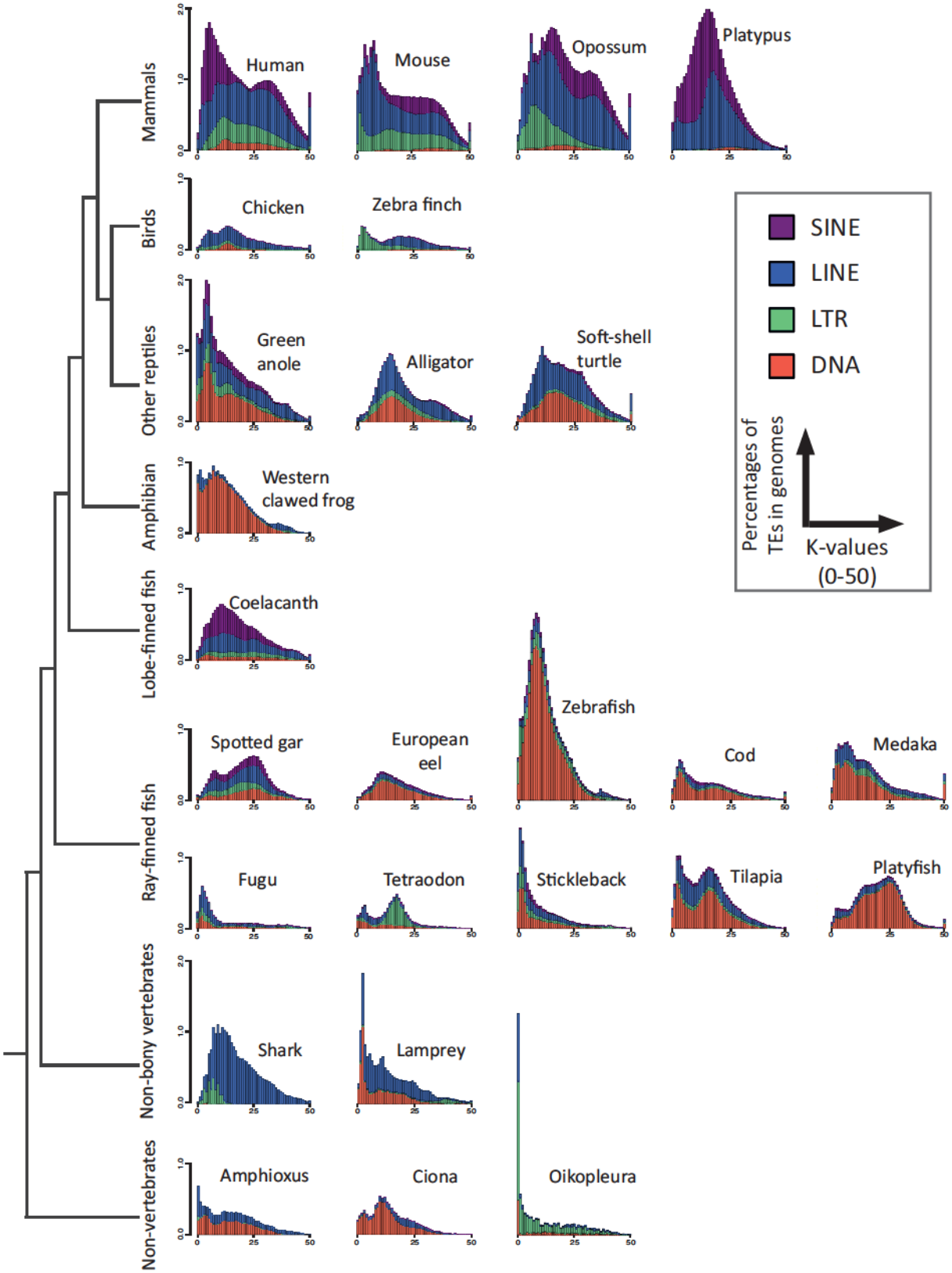

Number of superfamilies
Total >0.001%
10/14
5/13
7/12
7/11
4/7
3/9
14/15
16/19
16/18
16/21
22/26
19/25
22/24
25/27
27/27
26/26
22/24
21/22
24/24
23/23
22/22
12/18
16/21
13/14
7/8
22/27
P
Kolobok
Merlin
CMC/Chapaev
CMC/Ens1pm
MULE/MuDr
Sola
Piggybac
PIF/Harbinger
hAT
Tc-Mariner
Zisupton
Crypton
Maverick
Helitron
Jockey/Nimb/l
Rex1/Babar
UNE2/Maui
CR1/LINE3
RTE/Rex3
UNE1/TX1
R2
R4
Penelope
CR1-like
DIRS1/Ngaro
ERV
Copia/Ty1
BEL/Pao
Gypsy/Ty3
Not Detected
%< 0.001
0.1 <%≤ 0.001
1 <%≤ 0.1
5 <%< 1
%≥ 5
Modern human
Domestic mouse
Opossum
Platypus
Domestic chicken
Zebra finch
Mississipi alligator
Green anole
Soft-shell turtle
Western clawed frog
Coelacanth
Spotted gar
European eel
Zebrafish
Cod
Medaka
Platyfish
Tilapia
Stickleback
Tetraodon
Fugu
Elephant shark
Sea lamprey
Ciona
Oikopleura
Amphioxus
Pre cambrian
Paleozoic
Mesozoic
500
300
100
Sarcopterygian
Actinopterygian
Vertebrates

FIGURE 4


Mammals
Birds
Other reptiles
Amphibian
Lobe-finned fish
Ray-finned fish
Non-bony vertebrates
Non-vertebrates
Human
Mouse
Opossum
Platypus
Chicken
Zebra finch
Green anole
Alligator
Soft-shell turtle
Western clawed frog
Coelacanth
Spotted gar
European eel
Zebrafish
Cod
Medaka
Fugu
Tetraodon
Stickleback
Tilapia
Platyfish
Shark
Lamprey
Amphioxus
Ciona
Oikopleura
SINE
LINE
LTR
DNA
Percentages of TEs in genomes
K-values (0-50)

REFERENCES

Agrawal A, Eastman QM, Schatz DG. 1998. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. Nature 394:744-751.

Ågren JA, Wright SI. 2011. Co-evolution between transposable elements and their hosts: a major factor in genome size evolution? Chromosome Res 19:777-786.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403-410.

Amemiya CT, Alföldi J, Lee AP, Fan S, Philippe H, MacCallum I, Braasch I, Manousaki T, Schneider I, Rohner N, et al. 2013. The African coelacanth genome provides insights into tetrapod evolution. Nature 496:311-316.

Bao W, Jurka J. 2008. EnSpm DNA transposons in Zebrafish. Repbase Reports 8:823.

Böhne A, Brunet F, Galiana-Arnoux D, Schultheis C, Volff JN. 2008. Transposable elements as drivers of genomic and biological diversity in vertebrates. Chromosome Res 16:203-215.

Böhne A, Zhou Q, Darras A, Schmidt C, Schartl M, Galiana-Arnoux D, Volff JN. 2012. Zisupton – A novel superfamily of DNA transposable elements recently active in fish. Mol Biol Evol 29:631-645.

Bosco G, Campbell P, Leiva-Neto JT, Markov TA. 2007. Analysis of Drosophila species genome size and satellite DNA content reveals significant differences among strains as well as between species. Genetics 177:1277-1290.

Brookfield JFY. 2011. Host-parasite relationships in the genome. BMC Biol 9:67.

Burke WD, Müller F, Eickbush TH. 1995. R4, a non-LTR retrotranposon specific to the large subunit rRNA genes of nematodes. Nucleic Acids Res 23:4628-4634.

Burns KH, Boeke JD. 2012. Human transposon tectonics. Cell 149:740-752.

Chalopin D, Fan S, Simakov O, Meyer A, Schartl M, Volff JN. 2013. Evolutionary active transposable elements in the genome of the coelacanth. J Exp Zool B Mol Dev Evol.

Dasilva C, Hadji H, Ozouf-Costaz C, Nicaud S, Jaillon O, Weissenbach J, Roest Crollius H. 2002. Remarkable compartmentalization of transposable elements and pseudogenes in the heterochromatin of the Tetraodon nigroviridis genome. Proc Natl Acad Sci U S A. 99:13636-13641.

De Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet 7:e1002384.

De la Chaux N, Wagner A. 2011. BEL/Pao retrotransposons in metazoan genomes. BMC Evol Biol 11:154.

Deininger PL, Moran JV, Batzer MA, Kazazian HH Jr. 2003. Mobile elements and mammalian genome evolution. Curr Opin Genet Dev 13:651-658.

Duvernell DD, Pryor SR, Adams SM. 2004. Teleost fish genomes contain a diverse array of L1 retrotransposon lineages that exhibit a low copy number and high rate of turnover. J Mol Evol 59:298-308.

Eickbush TH, Jamburuthugoda VK. 2008. The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Res 134:221-234.

Ellison CE, Bachtrog D. 2013. Dosage compensation via transposable element mediated rewiring of a regulatory network. Science 342:846-850.

Evgen'ev MB. 2013. What happens when Penelope comes?: An unusual retroelements invades a host species genome exploring different strategies. Mob Genet Elements 3:e24542.

Feschotte C. 2004. Merlin, a new superfamily of DNA transposons identified in diverse animal genomes and related to bacterial IS1016 insertion sequences. Mol Biol Evol 21:1769-1780.

Feschotte C and Pritham EJ. 2005. Non-mammalian c-integrases are encoded by giant transposable elements. Trends Genet 21:551-552.

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

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. Nat Rev Genet 9:397-405.

Finnegan DJ. 1989. Eukaryotic transposable elements and genome evolution. Trends Genet 5:103-107.

Fischer C, Bouneau L, Coutanceau JP, Weissenbach J, Ozouf-Costaz C, Volff JN. 2005. Diversity and clustered distribution of retrotransposable elements in the compact genome of the pufferfish Tetraodon nigroviridis. Cytogenet Genome Res 110:522-536.

Furano AV, Duvernell D, Boissinot S. 2004. L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish. Trends Genet 20:9-14.

Gierl A, Schwarz-Sommer Z, Saedler H. 1985. Molecular interactions between the components of the En-I transposable element system of Zea mays. EMBO J 4:579-583.

Gifford R, Tristem M. 2003. The evolution, distribution and diversity of endogenous retroviruses. Virus Genes 26:291-315.

Gilbert C, Hernandez SS, Flores-Benabib J, Smith EN, Feschotte C. 2012. Rampant horizontal transfer of SPIN transposons in squamate reptiles. Mol Biol Evol 29:503-515.

Gilbert C, Waters P, Feschotte C, Schaack S. 2013. Horizontal transfer of OC1 transposons in the Tasmanian devil. BMC Genomics 14:134.

Goodier JL, Kazazian HH Jr. 2008. Retrotransposons revisited: the restraint and rehabilitation of parasites. Cell 135:23-35.

Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27:221-224.

Gregory TR. 2001. The bigger the C-value, the larger the cell: genome size and red blood cell size in vertebrates. Blood Cells Mol Dis 27:830-843.

Hancks DC, Kazazian HH Jr. 2012. Active human retrotransposons: variation and disease. Curr Opin Genet Dev 22:191-203.

Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22:2971-2972.

Hikosaka A, Kawahara A. 2010. A systematic search and classification of T2 family miniature inverted-repeat transposable elements (MITEs) in Xenopus tropicalis suggests the existence of recently active MITE subfamilies. Mol Genet Genomics 283:49-62.

Hiom K, Melek M, Gellert M. 1998. DNA transposition by the RAG1 and RAG2 proteins: a possible source of oncogenic translocations. Cell 94:463-70.

Jacques PÉ, Jeyakani J, Bourque G. 2013. The majority of primate-specific regulatory sequences are derived from transposable elements. PLoS Genet 9:e1003504.

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484:55-61.

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, et al. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 431:946-957.

Jurka J, Klonowski P, Dagman V, Pelton P. 1996. CENSOR - a program for identification and elimination of repetitive elements from DNA sequences. Computers and Chemistry 20:119-122.

Jurka J. 2000. Repbase Update: a database and an electronic journal of repetitive elements. Trends Genet 16:418-420.

Kapitonov V and Jurka J. 2001. Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci USA 98:8714-8719.

Kapitonov VV, Jurka J. 2005. RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLoS Biol 3:e181.

Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M, Feschotte C. 2013. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet 9:e1003470.

Kazazian HH Jr. 2004. Mobile elements: drivers of genome evolution. Science 303:1626-1632.

Kidwell MG. 2002. Transposable elements and the evolution of genome size in eukaryotes. Genetica 115:49-63.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111-120.

Kojima KK, Jurka J. 2011. Crypton transposons: identification of new diverse families and ancient domestication events. Mob DNA 2:12.

Kordis D, Gubensek F. 1999. Horizontal transfer of non-LTR retrotransposons in vertebrates. Genetica 107:121-128.

Kordis D. 2009. Transposable elements in reptilian and avian (sauropsida) genomes. Cytogenet Genome Res 127:94-111.

Kuraku S, Qiu H, Meyer A. 2012. Horizontal transfers of Tc1 elements between teleost fishes and their vertebrate parasites, lampreys. Genome Biol Evol 4:929-936.

Le Rouzic A, Capy P. 2004. Theorical approaches to the dynamics of transposable elements in genomes, populations, and species. Genome Dyn Stab, Eds Lankenau DH and Volff JN, Transposons and the dynamic genome.

Le Rouzic A, Capy P. 2005. The first step of transposable elements invasion: parasitic strategy vs. genetic drift. Genetics 169:1033-1043.

Le Rouzic A, Capy P. 2006. Population genetics models of competition between transposable element subfamilies. Genetics 174:785-793.

Levin HL, Moran JV. 2011. Dynamic interactions between transposable elements and their hosts. Nat Rev Genet 12:615-27.

Luchetti A, Mantovani B. 2013. Non-LTR R2 element evolutionary patterns: phylogenetic incongruences, rapid radiation and the maintenance of multiple lineages. PLoS One 8:e57076.

Malik HS, Burke WD, Eickbush TH. 1999. The age and evolution of non-LTR retrotransposable elements. Mol Biol Evol 16:793-805.

Nelson JS. 2006. Fishes of the world. New York, John Wiley and Sons.

Novick PA, Basta H, Floumanhaft M, McClure MA, Boissinot S. 2009. The evolutionary dynamics of autonomous non-LTR retrotransposons in the lizard Anolis carolinensis shows more similarity to fish than mammals. Mol Biol Evol 26:1811-1822.

Novick P, Smith J, Ray D, Boissinot S. 2010. Independent and parallel lateral transfer of DNA transposons in tetrapod genomes. Gene 449:85-94.

Ogiwara I, Miya M, Ohshima K, Okada N. 2002. V-SINEs: a new superfamily of vertebrate SINEs that are widespread in vertebrate genomes and retain a strongly conserved segment within each repetitive unit. Genome Res 12:316-324.

Oliveira SG, Bao W, Martins C, Jurka J. 2012. Horizontal transfers of Mariner transposons between mammals and insects. Mob DNA 3:14.

Pace JK, Gilbert C, Clark MS, Feschotte C. 2008. Repeated horizontal transfer of a DNA transposons in mammals and other tetrapods. Proc Natl Acad Sci USA 105:17023-17028.

Parfrey LW, Lahr DJG, Katz LA. 2008. The dynamic nature of eukaryotic genomes. Mol Biol Evol 25:787-794.

Petrov DA. 2001. Evolution of genome size: new approaches to an old problem. Trends Genet 17:23-28.

Piskurek O, Okada N. 2007. Poxviruses as possible vectors for horizontal transfer of retroposons from reptiles to mammals. Proc Natl Acad Sci USA 104:12046-12051.

Piskurek O, Jackson DJ. 2011. Tracking the ancestry of a deeply conserved eumetazoan SINE domain. Mol Biol Evol 28:2727-2730.

Pritham EJ, Feschotte C. 2007. Massive amplification of rolling-circle transposons in the lineage of the bat Myotis lucifugus. Proc Natl Acad Sci USA 104:1895-1900.

Ravi V, Venkatesh B. 2008. Rapidly evolving fish genomes and teleost diversity. Curr Opin Genet Dev 18:544-550.

Rebollo R, Romanish MT, Mager DL. 2012. Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annu Rev Genet 46:21-42.

Sarropoulou E, Fernandes JM. 2011. Comparative genomics in teleost species: knowledge transfer by linking the genomes of model and non-model fish species. Comp Biochem Physiol Part D Genomics Proteomics 6:92-102.

Schartl M, Walter RB, Shen Y, Garcia T, Catchen J, Amores A, Braasch I, Chalopin D, Volff JN, Lesch KP, et al. 2013. The genome of the platyfish, *Xiphophorus maculatus*, provides insights into evolutionary adaptation and several complex traits. Nature Genet 45:567-572.

Sievers F, Wilm A, Dineen DG, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539.

Sormacheva I, Smyshlyaev G, Mayorov V, Blinov A, Novikov A, Novikova O. 2012. Vertical evolution and horizontal transfer of CR1 non-LTR retrotransposons and Tc1/mariner DNA transposons in Lepidoptera species. Mol Biol Evol 29:3685-3702.

Stanke M, Steinkamp R, Waack S, Morgenstern B. 2004. AUGUSTUS: a web server for gene finding in eukaryotes. Nuc Acids Res 32:309-312.

Sun C, Shepard DB, Chong RA, López Arriaza J, Hall K, Castoe TA, Feschotte C, Pollock DD, Mueller RL. 2012a. LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders. Genome Biol Evol 4:168-183.

Sun C, López Arriaza JR, Mueller RL. 2012b. Slow DNA loss in the gigantic genomes of salamanders. Genome Biol Evol 4:1340-1348.

Syvanen M. 2012. Evolutionary implications of horizontal gene transfer. Annu Rev Genet 46:341-358.

Thomas J, Sorourian M, Ray D, Baker RJ, Pritham EJ. 2011. The limited distribution of Helitrons to vesper bats supports horizontal transfer. Gene 474:52-58.

Vieira C, Lepetit D, Dumont S, Biémont C. 1999. Wake up of transposable elements following *Drosophila simulans* worldwide colonization. Mol Biol Evol 16:1251-1255.

Vieira C, Nardon C, Arpin C, Lepetit D, Biémont C. 2002. Evolution of genome size in *Drosophila*. Is the invader's genome being invaded by transposable elements? Mol Biol Evol 19:1154-1161.

Volff JN, Körting C, Froschauer A, Sweeney K, Schartl M. 2001. Non-LTR retrotransposons encoding a restriction enzyme-like endonuclease in vertebrates. J Mol Evol 52:351-360.

Volff JN, Bouneau L, Ozouf-costaz C, Fischer C. 2003. Diversity of retrotransposable elements in compact pufferfish genomes. Trends Genet 19:674-678.

Volff JN. 2005. Genome evolution and biodiversity in teleost fish. Heredity 94:280-294.

Volff JN. 2006. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays 28:913-922.

Vorechovsky I. 2010. Transposable elements in disease-associated cryptic exons. Hum Genet 127:135-154.

Wallau GL, Ortiz MF, Loreto EL. 2012. Horizontal transposons transfer in Eukarya: detection, bias and perspectives. Genome Biol Evol 4:801-811.

Walsh AM, Kortschak RD, Gardner MG, Bertozzi T, Adelson DL. 2013. Widespread horizontal transfer of retrotransposons. Proc Natl Acad Sci USA 110:1012-1016.

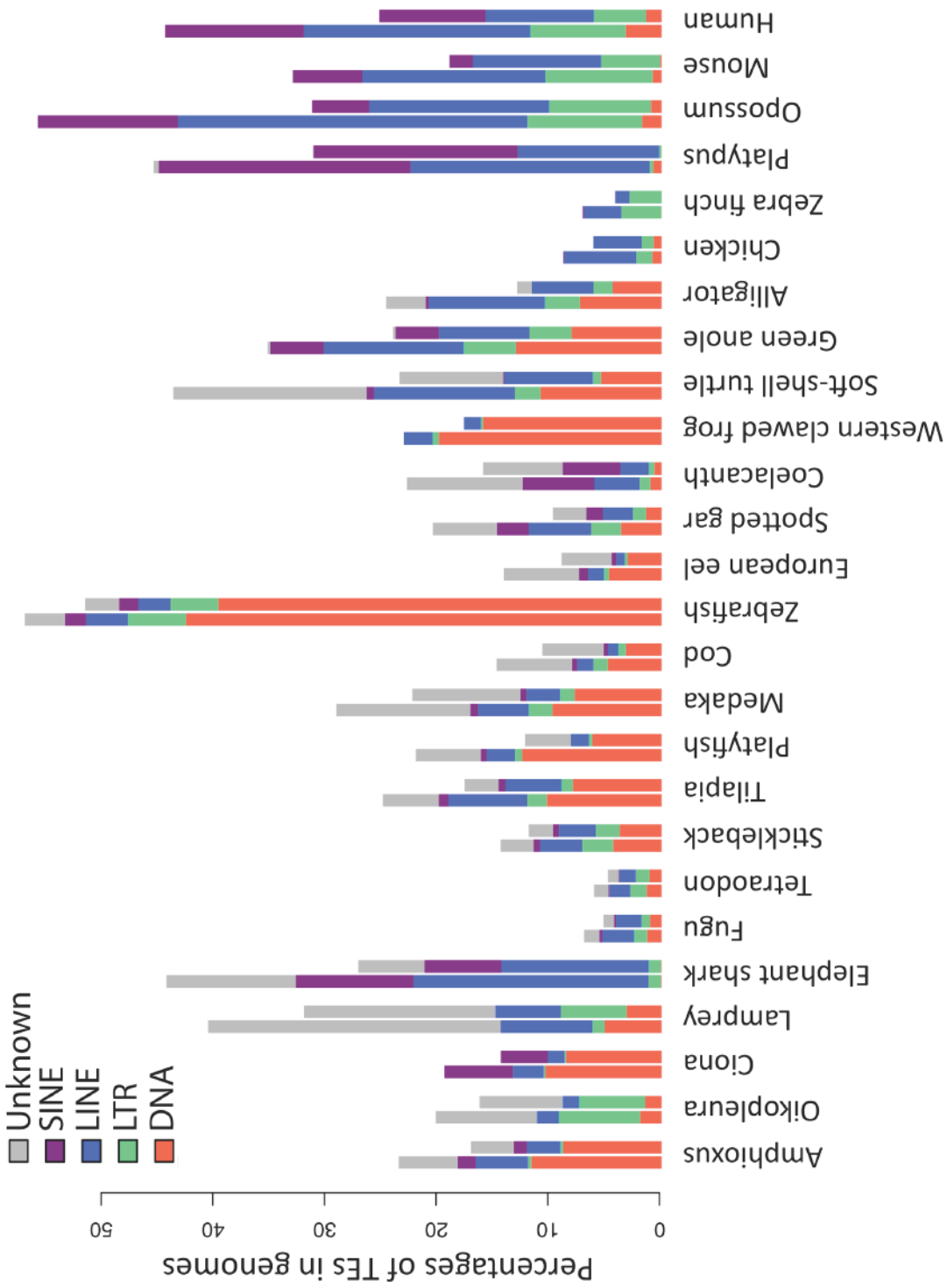
Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al. 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973-982.

Wittbrodt J, Shima A, Schartl M. 2002. Medaka – a model organism from the far East. Nat Rev Genet 3:53-64.

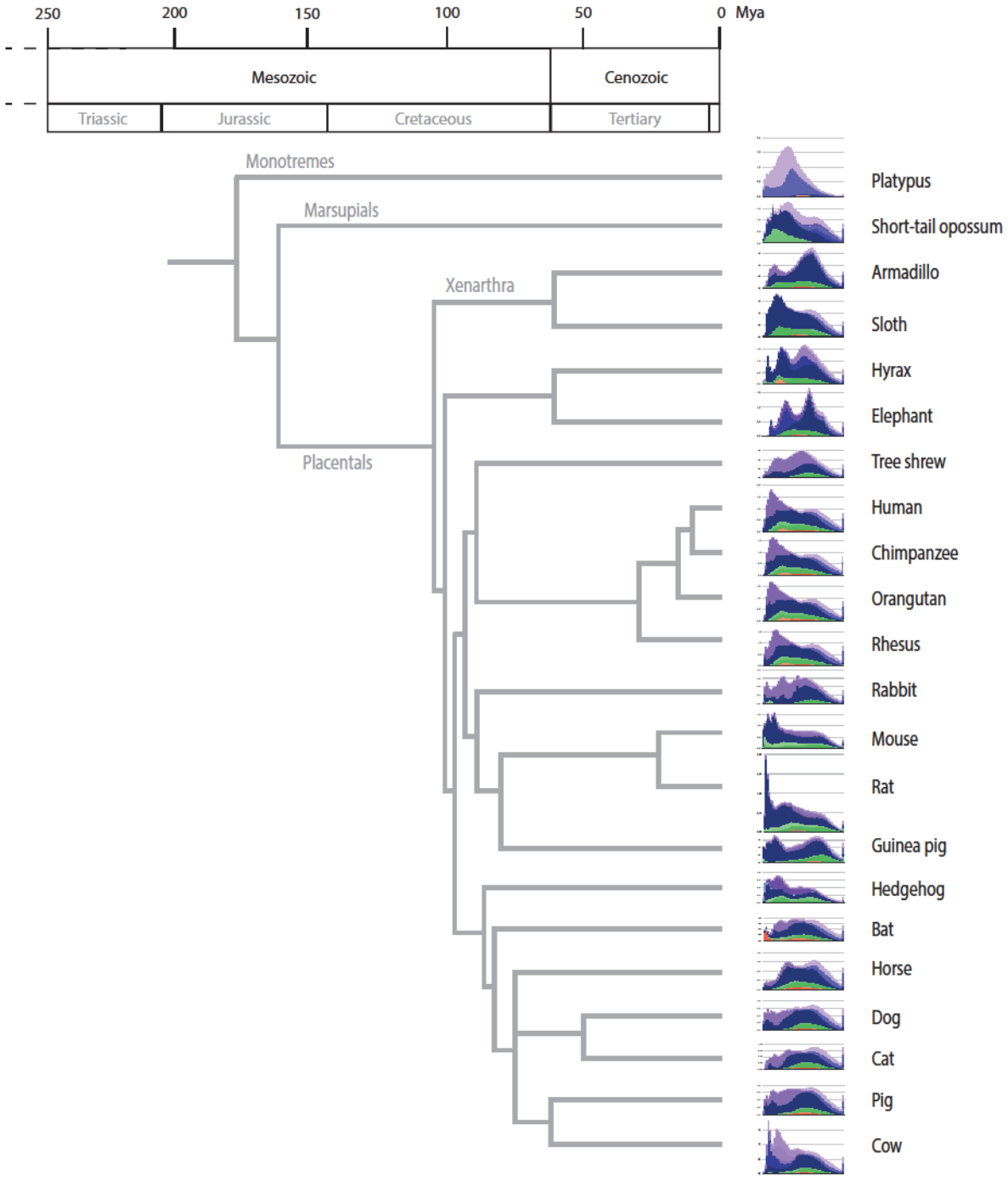
Xie M, Hong C, Zhang B, Lowdon RF, Xing X, Li D, Zhou X, Lee HJ, Maire CL, Ligon KL et al. 2013. DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. Nat Genet 45:836-841.

Yohn CT, Jiang Z, McGrath SD, Hayden KE, Khaitovich P, Johnson ME, Eichler MY, McPherson JD, Zhao S, Pääbo S, Eichler EE. 2005. Lineage-specific expansions of retroviral insertions within the genomes of African great apes but not humans and orangutans. PLoS Biol 3:e110

SUPPLEMENTARY FIGURE 1



SUPPLEMENTARY FIGURE 2



SUPPLEMENTARY TABLE 1

Species	SINE		LINE		DNA		LTR		Unknown		Total Repeat		Total TEs	
	cov-1	Cov (80/80)	cov-1	Cov (80/80)	cov-1	Cov (80/80)	cov-1	Cov (80/80)	cov-1	Cov (80/80)	cov-1	Cov (80/80)	cov-1	Cov (80/80)
Amphioxus	1,57	1,14	4,51	2,9	11,26	8,53	0,29	0,22	5,08	3,68	23,58	16,79	22,71	16,47
Oikopleura	0		1,9	1,45	1,89	1,45	6,99	5,67	8,72	7,17	19,49	16,13	19,5	15,74
Ciona int	5,91	4,04	2,66	1,46	10,05	8,3	0,14	0,09	0	0	21,89	15,09	18,76	13,89
Lamprey	0	0	7,92	5,65	4,93	3,06	1,07	5,65	25,22	16,5	40,5	27,11	39,14	30,86
Elephant Shar	10,15	6,61	20,25	12,72	0,078	0,075	1,094	1,066	11,15	5,72	46,21	27,78	42,722	26,191
Fugu	0,269	0,165	2,732	2,183	1,3	1,023	1,093	0,745	1,329	0,937	8,85	5,75	6,723	5,053
Tetraodon	0,08	0,059	1,79	1,41	1,33	1,08	1,4	1,18	1,25	0,94	8,31	5,17	5,85	4,669
Stickleback	0,55	0,49	3,66	3,19	4,2	3,65	2,66	2,06	2,84	2,11	15,92	11,92	13,91	11,5
Tilapia	0,82	0,59	6,81	4,84	9,93	7,69	1,67	0,97	4,83	2,92	25,54	17,32	24,06	17,01
Platyfish	0,52	0,006	2,44	1,55	12,07	6,02	0,6	0,29	5,59	3,95	22,79	12,47	21,22	11,816
Cod	0,43	0,35	1,4	0,93	4,69	3,12	1,23	0,63	6,51	5,27	21,16	14,04	14,26	10,3
Medaka	0,65	0,51	4,37	2,91	9,45	7,53	2,04	1,25	11,58	9,32	29,38	21,97	28,09	21,52
Zebrafish	1,81	1,63	3,62	2,8	41,04	38,26	5	4,11	3,47	2,93	61,98	55,58	54,94	49,73
European eel	0,76	0,393	1,38	0,74	4,58	2,96	0,42	0,24	6,5	4,33	16,51	9,4	13,64	8,663
Spotted gar	2,73	1,43	5,41	2,56	3,54	1,38	2,55	1,14	5,54	2,89	20,57	9,46	19,77	9,4
Coelacanth	6,173	4,98	3,915	2,463	1,002	0,652	0,907	0,474	10,001	6,845	30,36	20,27	21,998	15,41
Frog	0	0	2,48	1,45	19,23	15,43	0,55	0,19	0,001	0	25,23	18,98	22,261	17,07
Turtle	0,647	0,091	12,141	7,641	10,495	5,27	2,187	0,686	16,67	8,94	42,81	22,77	42,14	22,628
Lizard	4,6	3,75	12,08	7,79	12,58	7,8	4,51	3,63	0,24	0,22	35,63	25,25	34,01	23,19
Alligator	0,24	0,002	10,03	5,32	7,08	4,25	3,02	1,65	3,41	1,28	24,47	12,54	23,78	12,502
Chicken	0,075	0,002	6,2	4,19	0,84	0,72	1,39	1,01	0,042	0,003	9,91	6,27	8,547	5,925
Zebra Finch	0,09	0,02	3,22	1,21	0,05	0,003	3,46	2,81	0,04	0,002	8,14	4,29	6,86	4,045
Platypus	21,69	17,59	20,65	12,21	0,75	0,05	0,3	0,19	0,45	0,08	46,38	30,58	43,84	30,12
Opossum	12,1	4,89	30,14	15,55	1,71	0,94	9,88	8,78	0,014	0,001	56,18	30,99	53,844	30,161
Mouse	5,95	2	15,8	11,06	0,82	0,16	9,22	5,08	0,09	0,05	34,63	18,88	31,88	18,35
Human	11,96	9,15	19,53	9,36	3,13	1,35	8,22	4,5	0,013	0,01	44,73	25,17	42,853	24,37

SUPPLEMENTARY TABLE 2

Species	Genome_Size	All_repeats	All_rep-filter	Only_TEs	Only_TEs_filter	Genome_Size(log10)	All_repeats(log10)	All_rep-filter(log10)	Only_TEs(log10)	Only_TEs_filter(log10)
Amphioxus	521.9	23.6	16.79	22.7	16.5	2.71758729685546	1.37291200297011	1.22505069613805	1.35602585719312	1.21748394421391
Oikopleura	70.4	19.5	16.13	19.5	15.74	1.84757265914211	1.29003461136252	1.20763436738896	1.29003461136252	1.19700472802305
Ciona	115.2	21.9	15.09	18.8	13.9	2.06145247908719	1.34044411484012	1.17868923977559	1.27415784926368	1.1430148002541
Lamprey	885.5	40.5	27.11	39.1	30.9	2.94718856552609	1.60745502321467	1.43312951758049	1.59217675739587	1.48995847942483
Shark	768.5	46.21	27.78	42.72	26.19	2.88564387183576	1.66473596851871	1.4437322414016	1.6306312440205	1.41813549842523
Tetraodon	358.6	8.31	5.17	5.85	4.67	2.55461028522616	0.919601023784111	0.713490543093943	0.76715586608218	0.669316880566112
Fugu	393.3	8.85	5.75	6.72	5.05	2.59472394640975	0.946943270697825	0.75966784468963	0.827369273053825	0.703291378118661
Stickleback	461.5	15.92	11.92	13.91	11.5	2.66417170536193	1.20194306340165	1.07627625540422	1.14332712999205	1.06069784035361
Tilapia	927.4	25.54	17.32	24.06	17.01	2.96726709155979	1.4072208929274	1.23854788768133	1.38129562300383	1.23070431361257
Platyfish	729.7	22.79	12.47	21.22	11.82	2.86314434625267	1.35774432518038	1.09586645347854	1.32674537956532	1.07261747654524
Cod	832.1	21.16	14.04	14.26	10.3	2.92017552201002	1.32551566336315	1.14736710779379	1.15411952551585	1.01283722470517
Medaka	869	29.38	21.97	28.09	21.52	2.93901977644867	1.46805179145424	1.34183005692051	1.44855173920158	1.33284226699435
Zebrafish	1412.4	61.98	55.58	54.94	49.73	3.14995770889106	1.79225157190326	1.74491854244135	1.73988865508454	1.69661845923222
Eel	1101.2	16.51	9.4	13.64	8.66	3.04186620272129	1.21774707326279	0.973127853599699	1.13481437032046	0.937517892017347
Gar	954.8	20.6	9.46	19.77	9.4	2.97991241033472	1.31386722036915	0.975891136401793	1.29600666931367	0.973127853599699
Coelacanth	2860.6	30.36	20.27	22	15.41	3.45645713430378	1.48230176722344	1.30685374869301	1.34242268082221	1.18780263871842
Xenopus	1358.3	25.23	18.98	22.26	17.07	3.13299570069227	1.40191725051757	1.27829620809127	1.34752515999869	1.23223352111473
Turtle	2202.5	42.81	22.77	42.14	22.63	3.34291591708409	1.63154522783431	1.35736303061514	1.62469453127208	1.35468455395473
Anole	1701.4	35.63	25.25	34.01	23.2	3.23080642846888	1.55181582235102	1.40226138245468	1.53160663193272	1.3654879848909
Alligator	2172.4	24.47	12.54	23.78	12.5	3.33693979412059	1.38863396935179	1.0982975364947	1.37621185028267	1.09691001300806
Zebrafinch	1222.8	8.14	6.27	6.9	4.05	3.08735543005405	0.910624404889201	0.797267540830716	0.838849090737255	0.607455023214668
Chicken	1072.5	9.9	4.29	8.55	5.93	3.03039730085676	0.99563519459755	0.632457292184724	0.931966114728173	0.773054693364263
Platypus	1917.7	46.38	30.58	43.84	30.12	3.28278066824913	1.66633074430197	1.4854374810763	1.64187054547631	1.47885496752866
Opossum	3501.6	56.18	30.99	53.84	30.16	3.54426653360506	1.74958173486556	1.49122157623928	1.73110505121592	1.47943133719774
Mouse	3480.5	34.63	18.88	31.88	18.35	3.54164163809681	1.53945249154946	1.27600198996205	1.50351831272407	1.26363606858811
Human	3323.9	44.73	25.17	42.85	24.37	3.52164794949791	1.65059889817266	1.40088321554836	1.63195082625922	1.38685552918472

SUPPLEMENTARY TABLE 3

All bony vertebrates	**p-value**	**cor**
All repeats	0.001545	0.6215714
All repeats - filtered	0.009812	0.5267528
Only TEs	0.001439	0.624714
Only TEs-filtered	0.1178	0.5157164
Actinopterygian	**p-value**	**cor**
All repeats	0.00157	0.8561233
All repeats - filtered	0.01219	0.7515679
Only TEs	0.002082	0.8450686
Only TEs-filtered	0.01262	0.7492376
Sarcopterygian	**p-value**	**cor**
All repeats	0.005847	0.7672616
All repeats - filtered	0.01983	0.6857711
Only TEs	0.009331	0.7392451
Only TEs-filtered	0.02609	0.6632783

II- Further results added to the comparative genomic analysis

a. Inclusion of non-publicly available fish genomic data into the comparative analysis

Taking advantage of being involved in various genome sequencing projects, we included all fish data we had in our possession in our previous analyses. We thus added content statistics of cave fish, rainbow trout, Amazon molly, killifish and tongue sole, bringing the total number of compared fish genomes to 16 (Figure 25) (including the coelacanth).

As already mentioned - and this is striking on the figure 25 - coelacanth TE composition strongly differs from those of actinopterygians, in particular from those of teleosts. Indeed, the coelacanth genome mostly contains SINE and LINE retrotransposons, while DNA transposons are the main components of teleost genomes. The spotted gar is also different from teleosts by the fact that the four classes (DNA, LTR, LINE and SINE) are all well represented in this genome, with a preference for LINE retrotransposons and DNA transposons.
We can observe in Figure 25 that DNA transposons are predominant in teleost genomes, where they often make up half of the total TE content, like in zebrafish, cave fish, platyfish, Amazon molly, tongue sole and stickleback. We also observe that LINE retrotransposons sometimes compose a non-negligible part of teleost genomes, as illustrated by rainbow trout, killifish, stickleback, tilapia and the two pufferfish profiles. On the contrary, LTR and SINE retrotransposons are always in minority in teleost genomes. Teleost genomes still present a high proportion of unclassified elements, which can account for almost half of the total TE content as in cave fish and killifish. For killifish, it might be due to the fact that a first library was established on non-assembled reads, leading to a high quantity of unclassified small sequences of 100 bp long in the library.
Interestingly, in closely related species we are able to detect similarities in TE composition, as for the two poeciliids (platyfish and molly) and the medaka, or for the two pufferfishes. This probably reflects the composition of the genome of their common ancestor. If species diverged not too long ago - than a 100 My in the case of poeciliids and medaka separation (Schartl *et al.* 2013) -, we might be able to reconstruct the TE composition of the ancestral genome.
Not taking into account unclassified elements, the median of TE content in fish genomes is around 15%. In the extreme cases, pufferfish and tongue sole contain a very small amount of TEs (almost 4% in the three genomes). Interestingly, these three genomes are compact (pufferfish genomes < 400Mb; tongue sole genome ~480Mb). They might maintain their small size by efficiently restricting TE amplification. On the contrary, the

zebrafish shows an amazingly high TE content compared to all other fish and even vertebrates, which is mostly due to DNA transposons.

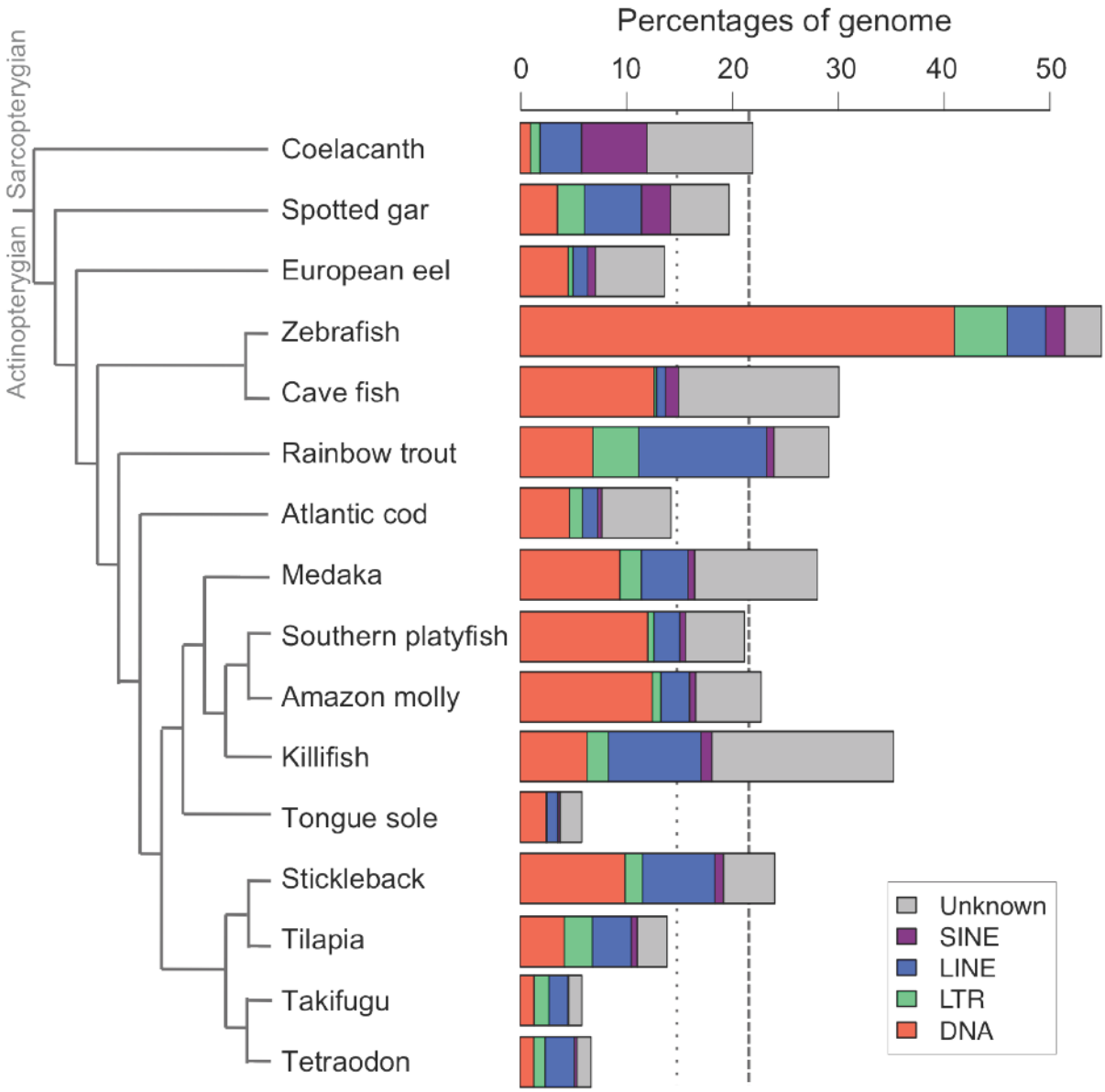


Figure 25: Quantity of DNA transposons, LTR, LINE, SINE retrotransposons and unclassified (Unknown) elements in fish genomes (percentages). The grey dotted and darker dashed lines represent the median (14.6% and 21.6%) of all fish TE contents without or with consideration of unclassified elements, respectively.

b. Endogenous retroviruses diversity: reflect of past infections

On the contrary to DNA and RNA viruses, retroviruses have the capacity to insert into genomes. They can become endogenous viral elements (EVEs: dsDNA copy of the viral genome integrated into the host germline) by using their own machinery. Of note, even if DNA and RNA viruses do not possess any machinery to integrate into genomes, some of them still found parallel way to be integrated and become endogenous, as illustrated by the Bornaviral EVE, which probably used LINE1 machinery (Feschotte 2010; Horie *et al.* 2010; Patel *et al.* 2011). Others, such as Herpes virus did not let any fossil traces in genomes. Retroviruses can move between genomes thanks to their extra-cellular phase, using their envelope proteins, and can also be vertically transmitted once integrated in host genome. For these reasons, we particularly focused on the diversity of vertebrate retroviruses, i.e. retroviridae. Indeed, it has been suggested that they might be one of the vectors driving horizontal transfer event between species. EVE and particularly retroviridae are considered as essential agents of genetic variation in vertebrates, as well as source of novelties, and potential vectors for the transport of sequences between hosts. Retroviruses were the first EVEs to be discovered, certainly because they have an integration phase into genomes: they nowadays constitute the vast majority of described EVEs.

I previously showed that fish presents a high diversity of TEs compared to mammals. What about retrovirus diversity, which seems to be high in mammals? I evaluated the diversity of retroviruses, which are present in all vertebrate lineages. This represents the first step of analysis to retrace the history of past infections in the different vertebrate lineages. Exogenous retroviruses have the capacity to infect new cells and species; they can be found in the extra-cellular environment, while endogenous retroviruses are located only in genomes. Retroviridae (retroviruses in vertebrates) are subdivided in seven families (Figure 26), also called genera: Alpha, Beta, Gamma, Delta, Epsilon, Lenti and Spuma-virus (Gifford and Tristem 2003).

Analysis of retroviridae diversity in vertebrate genomes

Retroviridae diversity was analyzed in vertebrates by combining literature information, results from the comparative genomic article (Chalopin *et al.*, submitted), above, and results from the article dealing with polymorphism in coelacanth (Naville *et al.*, submitted). I also studied foamy viruses (see below), and finally completed information with Blast analyses and phylogenetic reconstructions based on reverse transcriptase alignments (Figure 26). By now, Alpha-retroviruses have been only identified in birds. Beta-retroviruses were initially found in mammals. A phylogenetic class located between Alpha- and Beta-retrovirus, named Beta-like, was further identified in birds and pythons (Jern *et al.* 2005). The Gamma genus has the particularity to be leukemia and sarcoma-associated. It comprises the well known MuLV, FeLV and GaLV (for Murine, Feline and Gibbon ape Leukaemia Virus). It is the first genus for which sequences from

different vertebrate lineages were identified in various mammals, birds, reptiles and amphibians by PCR amplification (BENVENISTE AND TODARO 1973; TRISTEM *et al.* 1996). Furthermore, among reptiles, we identified sequences in turtles and squamates but not in crocodiles. Our recent studies in the coelacanth project (Naville *et al.* submitted) further revealed a new clade of potentially Gamma-retrovirus in teleost genomes. Delta- and Lenti-retrovirus are two mammalian-specific genera. The epsilon genus contains some of the known piscine retroviruses that induce diseases in fish species, as the Walleye epidermal hyperplasia or the snakehead retrovirus. With the exception of a potential endogenous member found in the African clawed toad, the epsilon genus was thought to contain exogenous retroviruses only. We recently found many epsilon retrovirus sequences in various teleosts including zebrafish, tetraodon, tilapia and stickleback; we also found a sequence in opossum, salamander, and many instances in the African lungfish, the two coelacanth species, and several turtles and crocodile species. These results suggest that epsilon retroviruses are widespread in vertebrates, even if we could not detect them in bird genomes. As we found many sequences in coelacanth genomes, we propose that coelacanth epsilon sequences are endogenous. Finally, the spuma genus has been identified in all vertebrate lineages including cartilaginous fish, except in turtle, squamates and crocodile.

	Alpha	Beta & Beta-like	Gamma	Delta	Epsilon	Lenti	Spuma & Spuma-like
Mammals	–	✖	✖	✖	✖	✖	✖
REPTILES Birds	✖	✖	✖	–	–	–	✖
REPTILES T - S - C	–	– ✖ –	✖ ✖ –	–	✖ – ✖	–	✖ – –
Amphibians & Salamanders	–	–	✖	–	✖	–	✖
Coelacanth	–	–	–	–	✖	–	✖
Teleosts	–	–	✖	–	✖	–	✖
Cartilaginous fish	–	–	–	–	–	–	✖

Figure 26: Retroviridae diversity in vertebrates. Presence (crossed) or absence, which can be due to a non-detection of the seven retroviridae genera (Gifford and Tristem 2003) are summarized in the table. Reptiles are composed of birds, turtles (T), squamates (S, including snakes and lizards) and crocodiles (C). Literature data were completed by Blast analyses and phylogenetic reconstructions.

Caulimovirus detection

We further discovered in two genomes (spotted gar and medaka) very small sequences (not longer than 30 nt) of caulimovirus that were only recognizable by automatic annotation. Caulimoviruses, also called pararetroviruses, were identified in plant genomes and encode a reverse transcriptase, just like retroviruses. As no long sequence could be isolated in fish genomes, we could not perform deeper analyses. These sequences could be either the result of contamination or remnants of very old infections. They can also be sequences randomly sharing homology with caulimoviruses, but might belong to a different virus. Nowadays, none of our analyses can resolve the question.

Characterization of Foamy viruses in fish and other vertebrates

Foamy virus (FVs; name derived from foam effects induced in cell culture) is the only family belonging to spuma-retroviruses. The high capacity of these viruses to propagate by staying apathogenic for their host (no associated disease could be identified) makes them interesting tools for gene therapy (Cordonnier *et al.* 1995; Rethwilm 2010; Bodem *et al.* 2011). FVs were initially discovered in mammals (including primates, feline, bovine, ovine and equine species) only as exogenous retroviruses (Gifford and Tristem 2003). However, the discovery of an endogenous sequence in the sloth genome modified this view (Katzourakis *et al.* 2009). In addition to the canonical *gag*, *pol* and *env* retroviral genes, mammalian FVs also encode *tas* (a transactivator) and *bet* (encoding a protein of unknown function but that may surround naive cells; it was detected only in infected cells) accessory genes in the 3' end, making them complex retroviruses. Their viral genomes are the longest known among the retroviridae genera, ranging from 11.956 to 13.246 kb in length (Lecellier and Saib 2000). As pointed out by Llorens and co-authors (Llorens *et al.* 2009), the discovery of a foamy-like sequence in zebrafish genome challenges what we know about FV distribution. Because we also found a FV sequence in the platyfish genome, we further investigated foamy diversity in vertebrate genomes.

The foamy virus sequence of the platyfish was initially detected in the sex-determining (SD) region of X and Y sex chromosomes. Its sequence includes LTRs (1715 nts) and Gag, Pol and Env (containing a transmembrane region) ORFs, with a total length of 17,027 nucleotides. At least two complete copies were identified in this SD region, with more than 95% sequence identity between each other. The whole platyfish genome contains more than 30 copies (> 85% identity), but the SD-region copies are the only complete ones. Looking more precisely at each copy and solo-LTR, a preference for insertion at TG-sites was determined (Schartl *et al.* 2013). The zebrafish sequence is 21,550 nts long, with LTRs (994 nts), the three *gag*, *pol* and *env* genes, and two unknown supplementary ORFs. Teleost foamys appear to be much longer than mammalian ones.

Search for possible endogenous foamy virus sequences in other vertebrate genomes was performed on ENSEMBL versions of the chicken, zebra finch, green anole, Chinese soft-shell turtle, Western clawed frog and all teleost genomes. A new sequence was identified in the cod genome (contig 24163). The predicted protein corresponds to a foamy reverse transcriptase and clearly groups within the foamy clade. No more complete foamy virus sequence could be identified in reason of the short contig lengths of the cod assembly. Other foamy viruses were also identified in the coelacanth with the three canonical ORFs and two small supplementary ones (Han and Worobey 2012) and in salamanders (no description; (Sun *et al.* 2012b)). We also found a RT sequence in the elephant shark genome. This sequence is a single ORF containing only the RT and does not strongly match with any other sequence in the genome. However, this elephant shark sequence is located with spuma sequences in the phylogeny (Figure 26 and 27). Phylogenetic reconstruction of foamy and other retroviruses show that a turtle sequence may be related to MuERV-L retrovirus, close to the spuma-viruses. Focusing on the spuma branch, two different branches seem to separate teleost: fish on one side and mammals, coelacanth and shark sequences on the other side. According to our analyses, spuma retroviruses were found neither in marsupials, monotremes, birds, squamates and crocodiles, nor in amphibians (Figure 27).

If these analyses still need to be completed with the recent genomic data, we showed that spuma-retroviruses are much more widely distributed over vertebrate lineages, being probably exogenous retrovirus in epitherian mammals and endogenous in other genomes for which sequences were identified. Spuma-retroviruses are probably composed of two spuma-like families found in placental mammals, birds, turtles, dipneust and coelacanth, and of the foamy family composed of two subfamilies found in placental mammals, salamander, coelacanth and elephant shark for the first and teleost fish for the second.

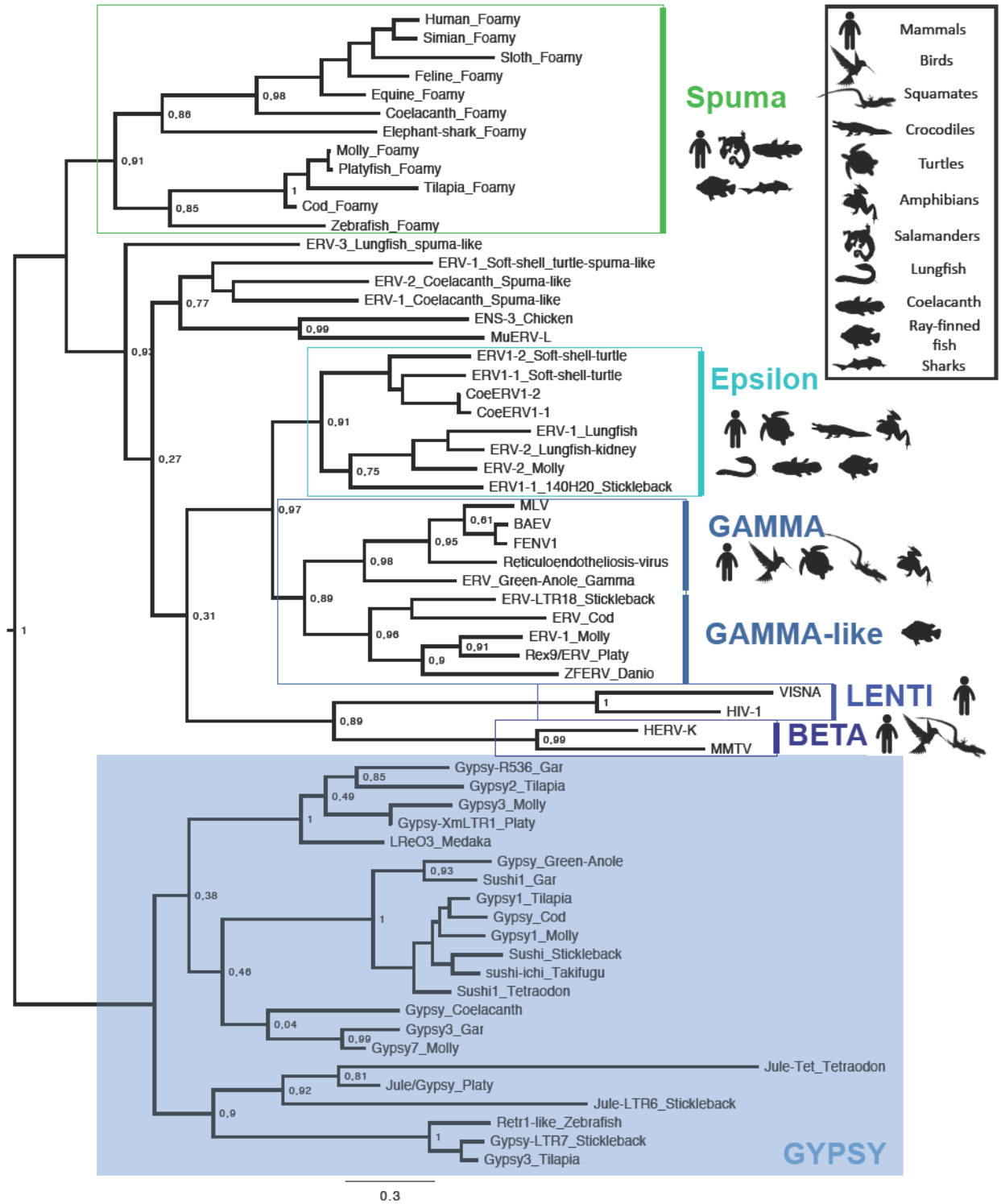


Figure 27: Phylogenetic reconstruction of retroviruses in vertebrates, focusing on the spumaretrovirus genus. Gypsy (or errantiviridae) sequences were added as outgroup. Reconstruction was performed using the PhyML software, based on reverse transcriptase alignment (200 amino acid sites) with optimized parameters and aLRT branch support. Black icons on the right side of the genera represent their respective distribution in vertebrate lineages including mammals, birds, turtles, squamates, crocodiles, amphibians, salamanders, dipneusts, coelacanths, teleosts and cartilaginous fish.

CHAPTER 4: GOLD TEs, MOLECULAR DOMESTICATION IN VERTEBRATES

I- Integrase-derived genes: the case of *Gin* genes

a. *In silico analyses of Gin genes*

ORIGINAL ARTICLE: "Genetic innovation in vertebrates: Gypsy intergrase genes and other genes derived from transposable elements" (CHALOPIN *et al.* 2012): Int J Evol Biol, doi:10.1155/2012/724519.

Beside the deleterious effects frequently associated to TEs, molecular domestication (MD) is an important process, leading to evolutionary innovations such as the "creation" of new exons or new genes. Through this process, TEs can be source of new genes or new regulatory sequences becoming beneficial and helpful for their host.
As shown in the introduction, domesticated genes can derive from different TE ORFs, such as transposases or reverse transcriptases. Nowadays, only four integrase-derived genes have been identified, *c-integrase* in mammals, *Fob1p* in yeast and finally *Gin-1* and *Gin-2* (Gypsy integrase genes) (LLORENS AND MARIN 2001; VOLFF 2006; MARIN 2010; CHALOPIN *et al.* 2012). It was initially thought that *Gin* genes were derived from LTR retrotransposon integrases, but it has now been demonstrated that they in fact derived from GIN transposons, which themselves encode a transposase close to LTR retrotransposon integrases (MARIN 2010). *Gin-1* is present in amniotes, suggesting an important function. To explore the evolutionary history of *Gin-2* and its potential role in vertebrates, we performed both *in silico* and expression analyses (CHALOPIN *et al.* 2012).
We identified *Gin-2* as a very conserved gene (in term of location but also the introns/exon structure) in all extant vertebrate lineages from cartilaginous fish, except in placental and monotreme mammals. Its maintenance over long periods strongly supports that it might have an important role in vertebrates. Due to the presence of an integrase in the coding region of *Gin-2*, the function might be related to DNA or RNA binding.

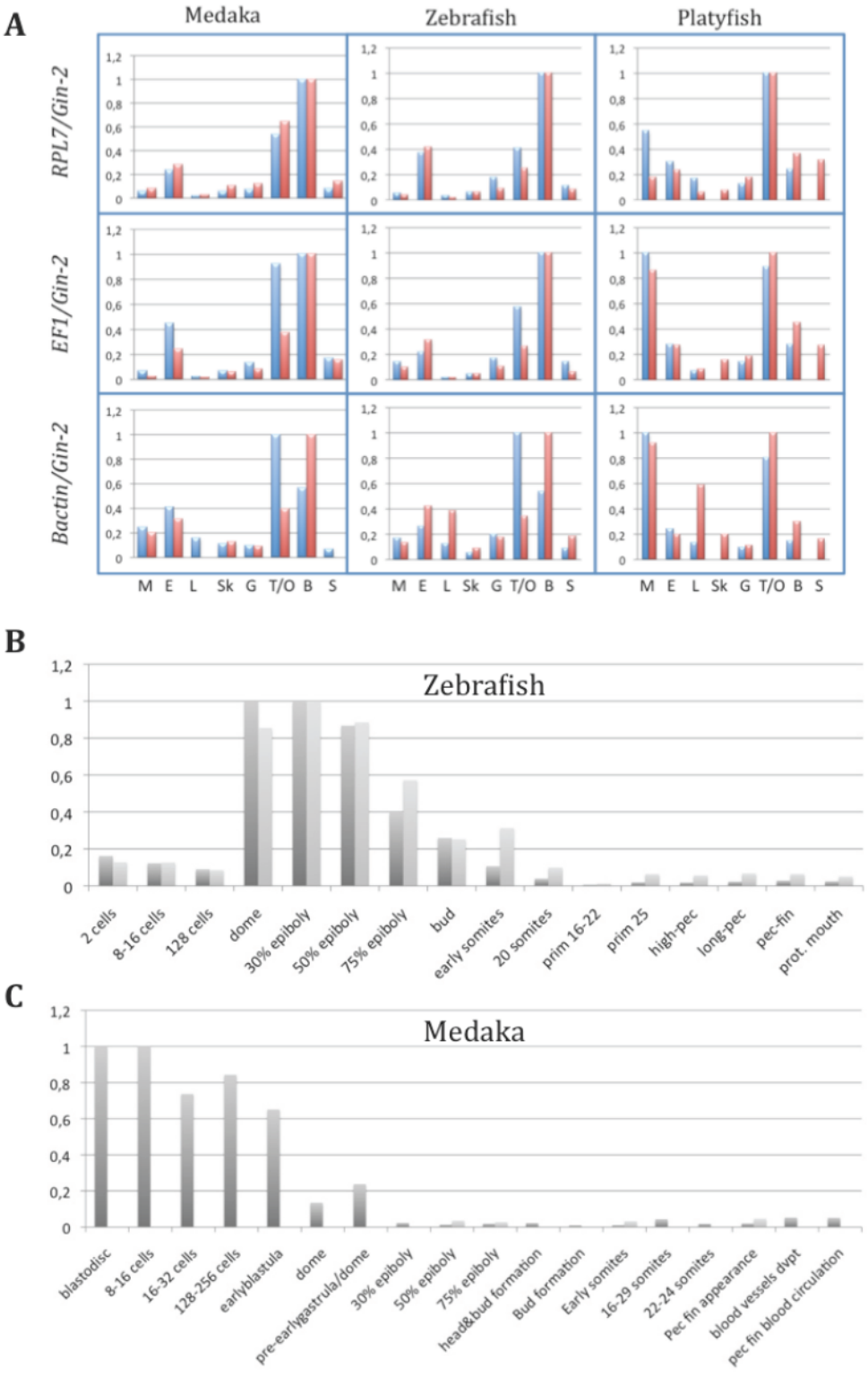


Figure 28: Expression analyses of *Gin-2* in fish using RT-qPCR experiments. A- Expression of *Gin-2* in several adult tissues of medaka, zebrafish and platyfish. Expression of *Gin-2* was normalized independently with three housekeeping genes, *B-actin*, *EF1* and *RPL7* (male in blue and female in red). For each assay, two independent sets and three replicates were used. Abbreviation: M, Muscle; E, Eyes; L, Liver; Sk, Skin; G, Gills; T/O, Testis/Ovary; B, Brain; S, Spleen. B- Expression of *Gin-2* during zebrafish embryogenesis. Experiments were performed with two independent sets of embryos (dark and light grey). C- Expression of *Gin-2* during medaka embryogenesis.

b. Expression analyses of Gin-2

We completed the expression analyses that were previously published (Chalopin *et al.* 2012) by performing RT-qPCR on adult tissues of three fish: zebrafish, medaka and platyfish, and on developmental stages of zebrafish and medaka. Since the platyfish is ovoviviparous, embryo experiments could not be performed without killing the mothers, and for the reason of the high number of fish we should use, we have decided to not perform this analysis in this species.

Expression patterns were not previously assessed in adult tissues of medaka fish. As shown in Figure 28A, *Gin-2* is mostly expressed in brain for males and females, then gonads (testis and ovary), and weakly in eyes. In zebrafish, the higher expression of *Gin-2* is detected in brain, as observed in medaka, and in gonads and eyes to a lesser extent. Similarly in both species, *Gin-2* is expressed in gonads, with a strong bias for testes. Different results were obtained for the platyfish. Indeed, *Gin-2* is mainly expressed in gonads for both male and female, but seems to also present a high expression in muscle and a bit less in brain and eyes (Figure 28A).

Assessing the expression profile of a given gene during embryogenesis might help to determine its potential function. Indeed, genes can have different functions during embryogenesis and in adulthood. We collected embryos from both medaka and zebrafish to obtain at least 10 embryos per stage (Kimmel *et al.* 1995; Iwamatsu 2004), and to extract the whole amount of RNA. Using this technique, we lose precisions because RNAs are extracted from many single entire embryos collected by visual assessment. Furthermore, whole RNA extraction just allows us to simply determine if the gene is expressed in the different stages but with no precision of localization as we obtain with *in situ* hybridization. During zebrafish embryogenesis, *Gin-2* is apparently not expressed until the 128 cells stage but strongly expressed at the dome stage (Figure 28B). We have to mention that between these two stages, there are six intermediate stages occurring in 2 hours (Figure 29B). After the dome stage, *Gin-2* is still strongly expressed until 50% of epiboly and then the expression decreases up to the early somite stages. All together, *Gin-2* seems to be specifically expressed during the gastrulation stage (Figure 29B), during which important developmental changes occur such as epiboly.
A different pattern was observed in medaka embryos. Indeed, *Gin-2* seems to be only expressed during the maternal period (Aizawa *et al.* 2003), when the developmental control of the embryos is handed from maternally provided mRNAs. In medaka, *Gin-2* is expressed up to the early blastula (Figure 28C). Since we have not performed *in situ* hybridization in medaka embryos, we have no answer concerning the difference of expression between the two species.

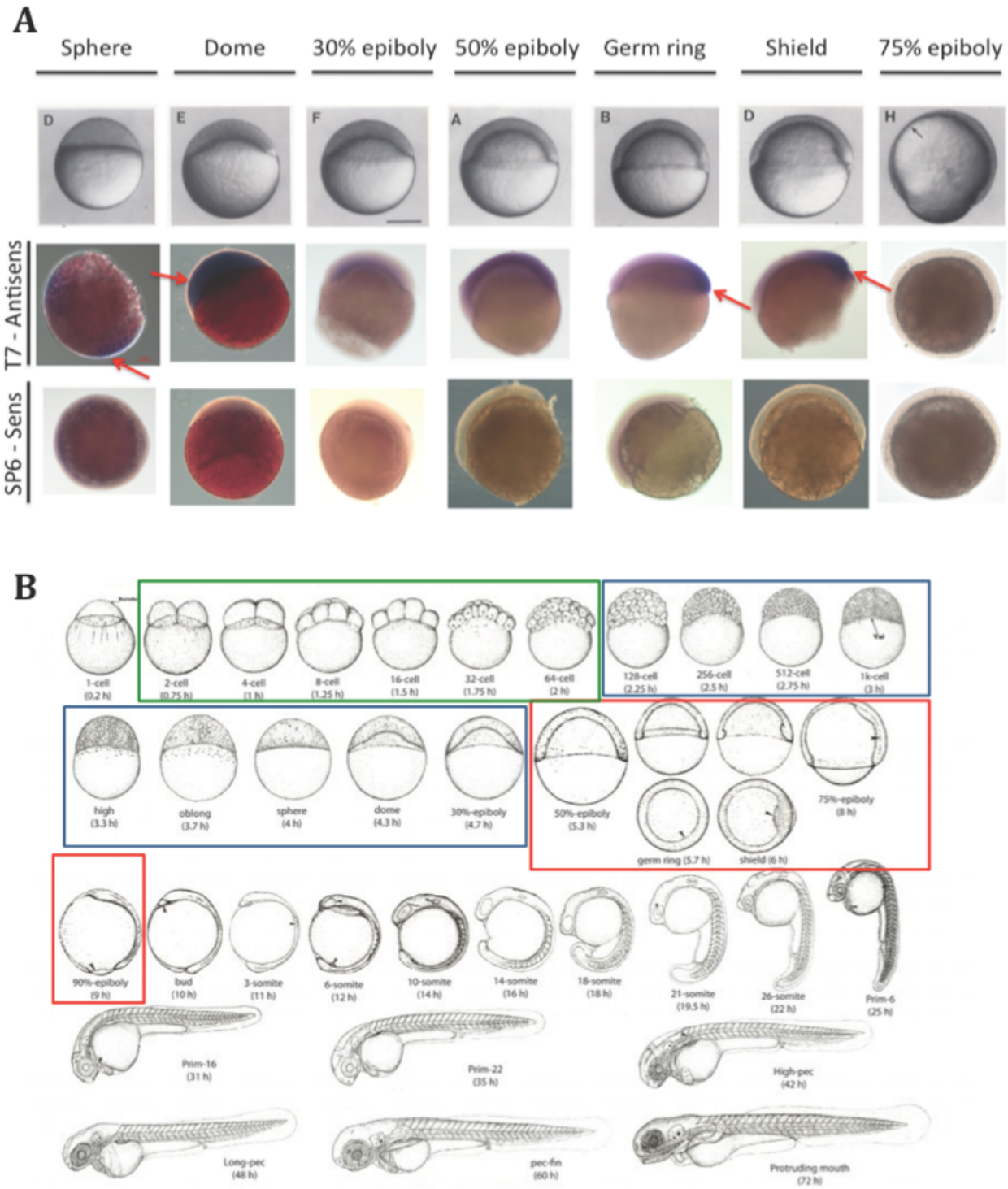


Figure 29: Expression of *Gin-2* during zebrafish embryogenesis determined by *in situ* hybridization. A- Expression from the sphere to 75% of epiboly. The first line represents the different stages of embryogenesis. The second line represents zebrafish embryos hybridized with T7-antisens probes (positive hybridization) and the third line shows embryos incubated with Sp6-Sens probes (negative control). B- Developmental stages of zebrafish. The various periods of developmental changes are highlighted: Green rectangle frames the cleavage period; blue frames the blastula period and red the gastrula period.

To obtain more information on *Gin-2* expression during zebrafish development, we carried out *in situ* hybridization experiments at the stages where expression was expected: from the dome stage to bud stage (Figure 28A and B). *In situ* hybridization uses labelled complementary RNA strand to localize a specific mRNA in a given tissue. Of note, RT-qPCR results do not exclude that the expression pattern may be more expanded (earlier to later stages) and further experiments will have to be performed with the analyses of more intermediate and tighter stages.

At the sphere stage (4 hpf), a general expression of *Gin-2* in embryos is observed. At later stages, between dome and shield stages, intensity of the expression strongly increases, recovering the whole-yolk (Figure 29A). The expression seems to reach a pick with a strong localized pattern between germ ring and shield stages (red arrows, 5.7 and 6 hpf). These results, i.e. high expression after dome and before 75% of epiboly, are quite coherent with the RT-qPCR results presented before (Figure 28B), where a higher expression was detected between dome and 50% of epiboly. Interestingly, at the two germ ring and shield stages, the expression seems to form a gradient with a higher expression on the future dorsal side (red arrows) that fades out toward the future ventral part. Only two hours later, during 75% epiboly stage, no expression was detected (Figure 29A). The expression is abruptly abolished suggesting a precise temporal window of *Gin-2* expression during zebrafish embryogenesis. As expected, when hybridized late embryos (18 somites), no signal was observed (data not shown).

To summarize, *Gin-2*, which derived from GIN transposons themselves deriving from retrotransposons, was identified in all extant vertebrates from cartilaginous fish, except in placentals and monotremes. The presence of *Gin-2* over such long periods (more than 450 My) suggests its importance among vertebrates. Due to the presence of a well-conserved integrase in the coding region, the function might be related to DNA or RNA binding. During fish adulthood, *Gin-2* seems to be particularly expressed in brain and gonads. Embryonic expression analyses show a particularly strong expression of *Gin-2* from the end of blastula and during the gastrulation, with an interesting gradient pattern. The expression abruptly stops, suggesting an efficient post-transcriptional regulation such as microRNA.

CHAPTER 5: DISCUSSION, JUNK OR NOT JUNK: THIS IS THE QUESTION

I- Comparative genomic analyses of TEs in vertebrate genomes

Diversity and content of TEs in vertebrate lineages

Our broad comparative analysis of TE diversity and content within vertebrate genomes in order to retrace TE dynamics and evolution, allowed us to highlight trends that are summarized in Figure 30. Regarding the dominance of the classes in term of quantity, vertebrate genomes mostly contain LINE retrotransposons, especially LINE1 and CR1-like, DNA transposons, mainly from the Tc-Mariner or hAT superfamilies. The coelacanth is an exception, as it is composed for the most part of SINEs. Beside the fact that genomes might show preference for TE superfamily spreading out them (LINE1 in human for example), they also harbour a more or less wide range of other classes and superfamilies of TEs. While mammal and bird genomes present a very poor diversity of TE superfamilies (from 7 to 14), other sarcopterygians such as turtles, squamates and crocodiles harbour a significant higher diversity (from 15 to 21 superfamilies), and finally the remaining of the vertebrates mostly living under water (coelacanth, teleost fish, cartilaginous fish and sea lamprey) shows a wide range of superfamilies (from 22 to 27). On the whole, a decrease in the diversity can be observed from the early vertebrates to mammals and birds (Figure 30). However, a different scenario concerns the retroviridae (RV) genera, described in Chapter 3. Indeed, many more RV genera were identified in mammals and birds than in other vertebrates. Moreover, if they contain a poor TE diversity, mammalian genomes show a very high content of TEs reaching more than half of the genomes. TEs also cover a high proportion of squamate, turtle, shark, lamprey and some teleost fish genomes. However, they are poorly represented in bird genomes. Furthermore, we highlighted a positive correlation between genome size and TE content in vertebrates. This analysis showed that sarcopterygians have bigger genomes than actinopterygians, for a similar TE content; and actinopterygians present a higher TE content than sarcopterygians, for a similar genome size. This shows that sarcopterygians have other sequences that strongly contribute to their genome size. A general rule concerning TE diversity, content and genome size is not so obvious to draw. In the future, it would be interesting to add more genomes from the sarcopterygian lineage in our analysis. Nowadays, many placental genomes are available but this is not the case for marsupials and monotremes. Regarding birds, we have analyzed two genomes, the domestic chicken and the zebra finch, but similar results were published for the turkey (Dalloul *et al.* 2010) and falcons (Zhan *et al.* 2013), suggesting that what we observed is, up to now, the rule for this lineage. Beside, the soft-shell turtle, for which we have also performed the annotation, other turtle genomes are now available such as that of the green sea turtle. Very different results were obtained concerning TE content and diversity (Wang *et al.* 2013), as they shown that TEs make less than 10% of the genome (compared to 40% in our analysis). It will be very important to re-analyze these

genomes using our own protocol as it is not really relevant to compare results obtained from different studies as discussed in the chapter 2 with the comparison of the two coelacanth assemblies. Finally, only few squamate, crocodile, snake and amphibian genomes are sequenced and available.

Including teleost fish in this synopsis is probably confusing by the simple fact that they contain a very high number of species with genome size varying from the smallest to the largest observed in vertebrates and live in extremely variable habitats. Moreover, a supplementary round of genome duplication occurred at the base of the teleosts (named 3R WGD) compared to tetrapods, probably having impacted genome dynamics and organization. The figure 30 shows a very simplified view concerning teleost genomes. In fact, if all analyzed fish genomes are relatively small compared to mammals for instance, the range size is from less than 400 Mb to 1.4 Gb and it exists even much bigger ones. All analyzed teleost genomes harbour a high diversity of TE superfamilies and a poor diversity of retroviridae. However, TE content can be highly variable (very low in pufferfish, <10%, and very high in zebrafish, >50%). As a general tendency, we can also notice that DNA transposons have successfully invaded teleost genomes. However, some genomes also contain a non-negligible proportion of LINEs, like the rainbow trout genome. The rainbow trout is of particular interest as salmonids underwent a fourth round of genome duplication, dating back to approximately 50 My (Berthelot *et al.* 2014, in press, **ANNEXE 4**). Due to this recent event of duplication, salmonid genomes are currently under the process of re-diploidization, involving many genomic changes. Salmonid genomes, including those of the rainbow trout and the Atlantic salmon, are very good models to study the dynamics and the role of TEs in tetraploid animal genomes. Taking the advantage of being involved in the rainbow trout consortium sequencing project, we compared Kimura profiles of TEs and of 4R ohnologs (paralogs from the salmonid specific-duplication) (data not shown, **ANNEXE 5**) with the aim to see if there a correlation between the recent TE burst occurring in the rainbow trout and the genome duplication event. There is a close correspondence between both factors (TEs and 4R event), possibly explaining the burst of CR1-like and Tc-Mariner elements, however some confirmations have to be done. The Atlantic salmon genome, which should be available soon, will probably help to bring answers. We will be able to perform phylogenetic reconstruction of Tc-Mariner elements in order to see if we are able to detect explosion of families in one of the two species. Tc-Mariner elements can be selected for phylogenetic analyses, because they have been strongly active in the past, and a larger recent burst can also be observed. If the ancient burst is common to all salmonids, but not the recent one, families and sub-families should be differentiated through phylogenetic analyses. In the genome paper (Berthelot *et al.* 2014, in press), it is also shown that the re-diploidization process is associated with gene deletion. It would be interesting to target these losses by orthology in order to detect or not traces of TEs making them agents of recombination.

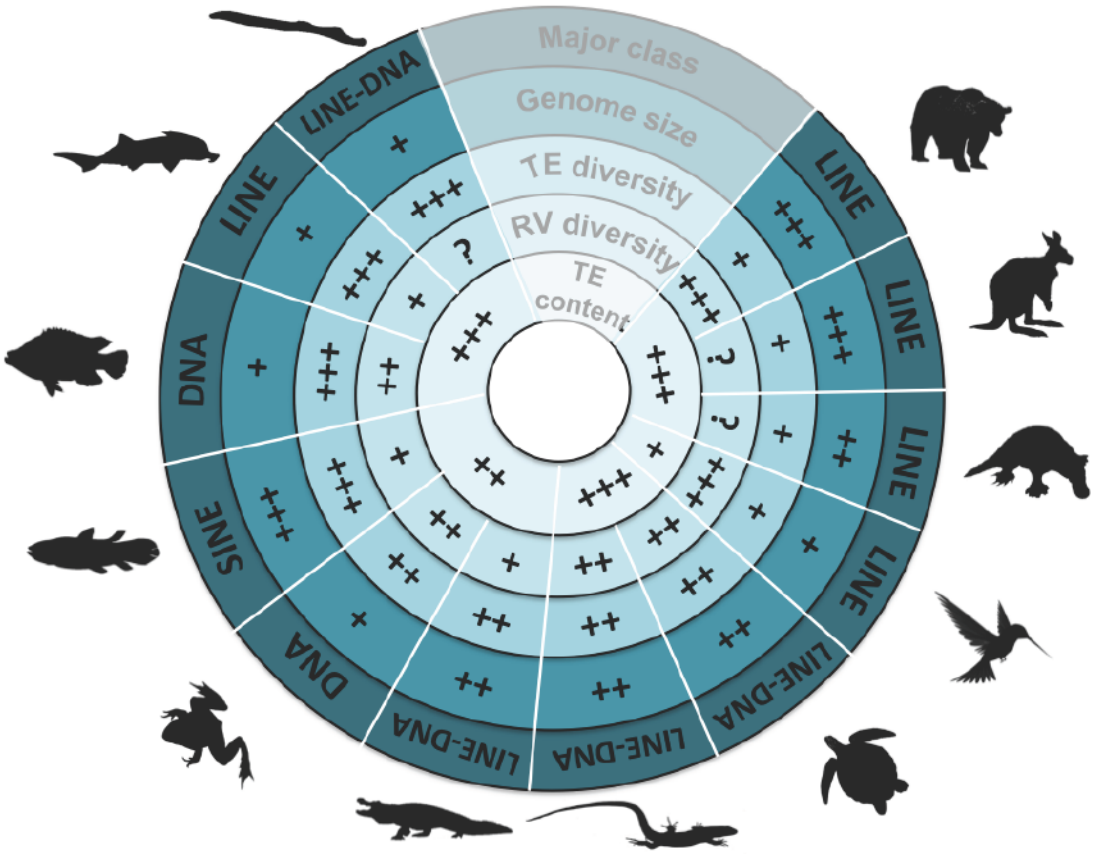


Figure 30: Schematic comparison of TE diversity and content in vertebrate genomes. The figure represents a non-exhaustive view of genome sizes, TE diversity and TE content in the previously analyzed vertebrate genomes. For genome size: +++ bigger than 2.5 Gb; ++ from 1.5 Gb to 2.5 Gb; + less than 1.5 Gb. For all-TE diversity (without SINE superfamilies): +++ more than 18 superfamilies covering at least 0.001% of the genome; ++ between 11 and 17; + less than 10. For RV (retroviridae) diversity (7 genera): +++ 5 to 7 genera; ++ 3 to 4; + less than 2. RV diversity was not investigated in marsupials, monotremes and jawless vertebrates. For TE content (comprising unclassified elements): +++ more than 25% of the genome; ++ from 11 to 24%; + less than 10%.

Retroviridae diversity in vertebrate genomes

Contrary to what we observe concerning the diversity of retrotransposons and DNA transposons, i.e. a rich diversity in early vertebrates decreasing to the amniotes, mammals contain more retroviridae (endogenous or exogenous retroviruses) genera than all other vertebrates and this diversity decreases to chondrychtians (Figure 30). We have not investigated the intra-diversity of retroviridae in jawless vertebrates, however it seems that this vertebrate-specific diversification might be correlated with the appearance of lymphocyte cells (Gifford and Tristem 2003). Indeed, adaptive immunity that depends on lymphocytes, can be found only in jawed vertebrates (Litman and Rast 1996; Flajnik 2002; Litman *et al.* 2010). Mammals that present the higher degree of diversity of RV, have also developed the most diversified repertoire of immunoglobulin isotypes. Furthermore, the adaptive immunity also involved the V(D)J

recombination implying the RAG proteins, which are vertebrate-specific TE-derived genes. All of these specificities of vertebrate genomes correspond to complex systems developing various defense mechanisms on one side and harbouring a high diversity of parasites (TEs and RV) on the other side. In future directions, it will be interesting to further investigate RV diversity within the three mammalian lineages: placentals, marsupials and monotremes, for which this point is absolutely not resolved. It also has to be done in jawless vertebrates in order to check if these organisms can be also subject to various infections by RV even without any developed adaptive system. Moreover, concerning the jawless species, do they lack an evolved adaptive system (they do not have differenciated T- and B-cell essential controlling cellular and humoral immunity) because RV does not infect them, or do they lack RV because species rapidly die after an infection, not allowing us to detect them in genomes?

In a more complete analysis, I studied the organization and distribution of foamy viruses in vertebrate genomes. This type of retrovirus, which belongs to spuma-RV, is found as exogenous form in mammals and endogenous form in fish. Beside these two lineages, foamy sequences were also found in the elephant shark and in the coelacanth. Among the foamy group, all fish sequences are grouped together while mammalian, coelacanth and shark sequences form a second distinct group. By now, we still have no idea if a single infection event occurred at the base of vertebrates, or if several independent infections occurred in the different lineages.

Cold versus warm-blooded species

Regarding TE diversity, birds and mammals present a particularly poor pattern in comparison to other vertebrates. The question that arises is what do we know from these two lineages that differs from others and can help to explain this situation? Bird and mammals are the two vertebrate lineages that are qualified of warm-blooded organisms, meaning that they self-regulate their temperature. They have various mechanisms to ensure that their body temperature is constant. In contrast, cold-blooded animals regulate their body temperature based on the climate of the environment where they are located. While mammals have their body covered with hair or fur for physical protection, insulation to reduce heat loss and for camouflage uses, birds have feathers for protection, insulation, camouflage and flight. Due to the fact that these organisms have to constantly maintain their temperature, the energy cost is higher and their basal metabolism is thus higher. The expression of TE gene products, as viral envelope proteins, may disrupt normal cellular activity and impose a cost of transposition (Brookfield 1996; Nuzhdin *et al.* 1996), which is added to the whole metabolic cost of the host. Due to the fact that warm-blooded species already need more energy, as presented before, than cold-blooded ones to maintain the basal metabolism, these species have probably developed a defense mechanism to thwart this noxious situation.

One possibility is to reduce non-vital energetic cost, such as TE transposition. Indeed, this might be true in birds for which we observe a TE-poor diversity and a poor content (Figure 30). In addition to chicken and zebra finch, some published data support our observations, such as the turkey genome, which contains approximately 7% of TEs, mostly from the CR1 superfamily (Dalloul *et al.* 2010). These analyses might explain the fact that in bird genomes, TE activity seems to have been well controlled, avoiding extra-metabolic cost. In mammalian genomes, the situation is a bit more complex: they present a high TE content, but a poor diversity, mostly represented by LINE1 and Alu retrotransposons. This high TE content might be the reflection of past repeat amplification, reflected in Kimura analyses (see Chapter 3 section). However, only few elements are nowadays still active in human and many other mammalian genomes (suggested by Kimura distance analyses), meaning that even if TEs represent a significant part of mammalian genomes, they are not active and energetic cost linked to TEs is weak. Of note, this hypothesis is a speculation and do not fit to all mammals (many copies are still active in the mouse for instance).
These data might suggest in some cases that the TE-poor diversity observed in some mammals and in birds might be linked to a metabolic stress due to their warm-blooded status, consequently strongly restricted by the genomes.

If the warm-blooded hypothesis might correspond to bird genomes, this is not so obvious for mammalian genomes. A simpler hypothesis to explain the difference of TE dynamics between mammals and fish, for example, can be the competition between TE superfamilies. Fish present a high diversity of TE superfamily that might be auto-regulated, limiting the number of copies, while mammals have a low diversity with few superfamilies that might be able to invade genomes.

It is interesting to underline that both mammals and birds present a higher diversity of retroviridae than other vertebrates, which might seem to be contradictory with their warm-blooded status. Exogenous retroviruses are not targeted by the same defense mechanisms than other TEs. As cited above (Flajnik 2002), it has been shown that mammals and birds present a higher diversity of immunoglobulin isotypes. It is not clear, as already mentioned, if the high RV diversity is due to the high immunglobuline isotypes repertoire, or if the development of such an immune system allowed the host to support a RV explosion. However, we can suppose that in this context, the energetic cost linked to the presence of RV is cancelled and integrated with the energetic cost linked to the defense mechanism that the host has to set up in any case, in order to fight pathogen (as fungi or bacteria) infections and survive.

Large-scale comparative genomics improvements

The large comparative genomic study I have performed on vertebrates during my thesis opens new perspectives for multiple future directions. First of all, the data generated might be used as a base to be completed and compared in the future using newly sequenced genomes. Of note, we need to add more genomes in the analysis if we want to extract robust tendencies among the different lineages. As mentioned above, the study of more birds is absolutely necessary to see if they are all poor in TEs. More turtles, squamates, crocodiles, snakes and amphibians will be of interest to analyze. Furthermore, in a large view, multivariate analyses will be more informative to assess the evolution of vertebrate genomes, and we can imagine to integrate in such study other variables such as genome sizes or effective population sizes.

TE mapping along chromosomes

It would be also interesting to take advantage of this kind of analyses using the genomic coordinates of TEs that have been generated through genome masking to obtain maps of TE localization in vertebrate karyotypes, which could be publicly available as a visual database. However, this can only be done on advanced assembled genomes (chromosome-associated assembly). Indeed, it might be useful to detect specific zone of accumulation, such as centromeres, telomeres or sex chromosomes. This would help to understand TE dynamics within genomes. Sex chromosomes are of particular interest as they show a specific TE accumulation associated with the non-recombinating status of the chromosomes in the heterogametic sex. Sex-linked TE roles have been largely studied in many mammalian species (Lyon 2003; Bachtrog 2006) but not in other vertebrates. If all mammals present a XY sex determination system, with the male being the heterogametic sex, the situation is not as simple in fish. Indeed, the sex determination system in fish is highly variable and can be either environmental (temperature dependent for instance) and genetic, with heterogamety XX/XY or ZZ/ZW (for more details see (Devlin and Nagahama 2002; Mank *et al.* 2006)). For many fish, the sex determination system is not yet identified. One way to help in the identification of putative sex chromosomes, and consequently to help in the characterization of the sex determination mechanism in a species, might be to detect specific zones of accumulation as it can be a feature of sex chromosomes.

TEs as phylogenetic markers

Within certain genus, the phylogenetic relationships between the different species are not completely resolved. These uncertainties might be due to the poor number of differences between orthologous genes, in various species with short time of divergence. This is for example the case of *Xiphophorus* genus, which currently comprises 26 species

(KALLMAN AND KAZIANIS 2006) being phenotypically (platyfish and swordtail) and geographically (Northern or Southern South America) subdivided. However, among these subdivisions, some interrogations still subsist. In this context, it has been shown that TEs might be powerful phylogenetic markers (COOK AND TRISTEM 1997; SHEDLOCK *et al.* 2004; KRAMEROV AND VASSETZKY 2005). Indeed, once inserted in the germline, a TE will be vertically transmitted to all the descendants and thus can be used as landmarks of genome evolution. Because SINEs do not excise and because they are small sequences, SINEs probably represent the most convenient tool to perform this kind of analysis. However, SINE elements tend to be eliminated and it is important to consider that their lifetime is not infinite, making them good tools only within a shortly time expanded genus. The comparison of three *Xiphophorus* (*X. maculatus, helleri and couchianus*) species showed the presence of almost 5000 copies of V-SINEs. A more precise characterization of the sequences and PCR amplification in the 26 *Xiphophorus* species might help to resolve the phylogenetic relationships of unresolved ones (KALLMAN AND KAZIANIS 2006).

TEs as genetic tools

As a final point, I mentioned in the introduction (see section "IV-TE diversity in vertebrates") that some TEs are developed as transgenesis and mutagenesis tools. Indeed, their intrinsic features (such as TIRs) provide interesting opportunities for exogenous DNA integration into genomes. A TE can be a good candidate to be used as a genetic tool if it presents a sufficient level of transpositional activity in a given species and if there is no endogenous copy in the targeted genome to avoid mobilization of active copies (IVICS *et al.* 2009). For these two reasons, it is essential to initially evaluate the diversity and the activity of TE superfamilies in vertebrate genomes for which we can have an interest. In this way, our large-scale comparative genomics provides an initial overview of TE diversity and content in a wide panel of vertebrate species.

II- Dynamic of TEs in asexual-species genome: the case of the Amazon molly

In vertebrates, approximately 80 asexual species mostly reproducing by parthenogenesis have been described including fish, amphibians and reptiles. If parthenogenesis – all-female species – is the major mode of reproduction, gynogenesis (also named sperm-dependent parthenogenesis) and hybridogenesis are also observed in some fish and amphibian species (LAMPERT AND SCHARTL 2008). The maintenance of parthenogenetic species as natural populations raises many questions from the evolutionary point of view. Indeed, these species are considered as dead-ends due to the loss of their ability to recombine during meiosis and generate new combinations of

traits, suggesting that these organisms cannot adapt to potential environmental changes and then might succumb more rapidly to diseases or predation for example. However, the persistence of many asexual species, such as the salamanders from *Ambystoma* genus, which appeared 5 million years ago (Bi and Bogart 2010), questions this simple view. Most of these species originate from inter-species hybridization events. They combined at their inception two different genomes, source of a huge genetic diversity. Nowadays, insights regarding molecular mechanisms implicated in the conservation, in such species, of heterozygosity and genetic diversity have emerged, such as the production of oocytes with twice the number of chromosomes compared to sexual species, or the addition of haploid genome through hybridization during fertilization for example. From an evolutionary point of view, these asexual genomes seem to evolve differently compared to sexual genomes, since no meiotic recombination occurs anymore, for example. Furthermore, TEs also are probably subject to a different dynamic depending on sex, follows the question: what is the fate of TEs in asexual genomes?

It has been largely shown that TE amplification was limited by sexual reproduction and recombination event associated to maintain a stable copy number. With the absence of sex and recombination, asexual population may be driven to their extinction by an unchecked proliferation of TEs (Arkhipova and Meselson 2005a; Dolgin and Charlesworth 2006) and to survive asexual species must be TE-free. In absence of horizontal transfer, which brings new elements, TEs cannot spread in an asexual genome (Hickey 1982). However, asexual species arise from sexual populations that may contain active TEs. The fate of the new hybrid depends on its capacity to counter-select and excise TE copies, in order to limit their proliferation leading to genome invasion and destruction. Dolgin and Charlesworth (Dolgin and Charlesworth 2006) demonstrated that an asexual population might be able to survive even with abandonment of sex, especially in case of large population sizes in which selection is more effective, correlated with a higher level of TE excision. The same authors have tried to estimate the number of generations necessary to reach the stable state (genomes contain the reasonable number of TEs to avoid its extinction): from hundred to tens of thousands of generations. To summarize, a newly formed asexual genome has two fates: either it survives thanks to a complete TE elimination, either it disappears due to uncontrolled TE accumulation.

Among the long-term asexual species that have been sequenced and studied, the bdelloid rotifers have been further investigated. Bdelloid rotifers are asexual species since 40 My and are not known to produce males, hermaphrodites or to undergo meiosis. In rare cases, they can sustain cryptic sex. Interestingly, this long-term asexual species lack deleterious vertically transmitted retrotransposons. Indeed, analyses show that the genome has a very poor content of TEs (only 3%) but an extensive diversity of superfamilies, most of them only represented by one or two full-length copies

(ARKHIPOVA AND MESELSON 2005b; FLOT *et al.* 2013). However, they have demonstrated that many families were horizontally acquired. Further analyses highlighted evidence of gene conversion, a mechanism that limits the accumulation of mutation in absence of meiosis, but also an increase of genes involved in defense against TEs, as Dicer and Argonaute/Piwi. Furthermore, the comparison of sexual and asexual species of Wolbachia parasitoids (the two systems of reproduction exist in Wolbachia, making it an interesting model) indicates that asexual species present an increase of DNA transposon copies, suggesting a faster accumulation of these elements compared to sexual ones (KRAAIJEVELD *et al.* 2012). If authors have not detected a significant decrease of the fitness in asexual species, they suggest that it could be the future of these organisms if the accumulation of TEs continues.

Another interesting model organism to analyze in order to assess TE dynamic in vertebrate asexual species is the Amazon molly. The Amazon molly *Poecilia formosa*, an asexual gynogenetic fish species (requires the intervention of sperm during the reproduction, but the sperm genome does not contribute to the genetic material of offsprings, in general), is the first asexual vertebrate to be sequenced. The Amazon molly is an all-female live-bearing species arising from the hybridization of *P. mexicana* (maternal ancestor) and *P. latipinna* (paternal ancestor), which occurred 280,000 Mya with about 800,000 generations (LAMPERT AND SCHARTL 2008). We investigated the diversity and content of TEs in the genome of the Amazon molly in comparison to its parents and to *Xiphophorus* species, which diverged from *Poecilia* up there 30 Mya (geological estimation). The comparison of TE diversity between *P. formosa*, *P. mexicana* and *P. latipinna* shows that few superfamilies could not be identified in any of the two parental non-assembled genomes, as BEL, DIRS1, Nimb, R2 retrotransposons, Crypton, MuDr and Novosib DNA transposons, while we could find them in the genome of *P. formosa*. Two hypotheses can explain this observation. First, these superfamilies were horizontally acquired by *P. formosa* after the hybridization event. Second, these superfamilies are also present in parental genomes at low copy-number, and thus were not detected by RepeatExplorer analysis due to the small portion of reads analyzed. We then compared the diversity and content between *P. formosa* and three *Xiphophorus* species. The Amazon molly contains 2% more TEs than *Xiphophorus*, mainly due to a small increase of the quantity in each class. If this difference is weak (but significant), a focus on TE potential activity presents more interesting results. Indeed, the comparison of the repeat landscapes in the different species show that the pattern between K-values 6 to 50 are similar, suggesting that the observed amplifications of TEs during the corresponding period occurred before the divergence between *Xiphophorus* and *Poecilia*. On the contrary, what we observe in the recent timing (K-values 0 to 5) seems to be genus-, or even species-specific: the Amazon molly underwent a higher recent TE amplification than *Xiphophorus*, which then strongly decreases. This result suggests that the Amazon molly suffered a rapid and strong burst of TE activity that almost

immediately stopped, followed by a loss of activity. Thanks to these results, we can suggest that the Amazon molly survived almost over 300,000 My supporting a strong amplification of TEs, may be just after the hybridization event. However, the kimura profile suggests that TE activity has been taken under control and finally restrained to limit the deleterious invasion. We can hypothesize that the species has probably not lived enough time still to completely eliminate TE from its genome.
Assemblies of the parental species, *P. mexicana* and *P. latipinna*, might certainly help us to trace back and to date the several amplifications of TEs we observe on Kimura profiles, and to discriminate between either event specific to the Amazon molly or to the genus. Moreover, completing the diversity analyses of the parents will be very helpful in order to identify potential horizontally acquired superfamilies, as suggested before.

Regarding the two possible fates of an asexual genome, namely the survival as a TE-free genome or the extinction of the species due to an invasion of TEs, we have currently no key to store the Amazon molly in one or the other cases. This genome probably suffered a recent amplification that might have occurred after the hybridization, however at this time it seems that TEs are in the process of silencing (no proof of activity). The Amazon molly is a relatively recent hybrid species, compared to bdelloid for instance, explaining why we still observe a high proportion of TEs and the survival of the species.
A more intermediate possibility could be that the Amazon molly applies a stronger control than *Xiphophorus* genomes, without complete elimination of TEs.

III- Evolution of TEs in "slow-evolving" genomes?

The coelacanth is a fish – it lives under water, have gills and fins – that is located at an intermediate phylogenetic position within vertebrates between actinopterygians and tetrapods but belongs to the sarcopterygian lineage. The coelacanth is a relatively rare species comprising only two living species, *Latimeria chalumnae* (in Africa) and *Latimeria menadoensis* (in Indonesia), making them difficult to observe and study. Coelacanths are very interesting organisms from numerous points of view. From an evolutionary point of view, their key phylogenetic position is an advantage to help understanding the transition from water to land in vertebrates; from a physiological point of view, it lives over 100 meters deep, suggesting an important adaptation and questioning about the use of their fins that resemble to terrestrial vertebrate limbs; from an ecological point of view, only few specimens are still alive in restricted geographical locations making them a small animal population to study; from a paleontological point of view, many extinct fossil species allow to date different events, also thanks to the comparison with the current living species; from an immunological point of view, it is the only vertebrate species lacking *immunoglobulin-M* gene, which is considered indispensable for adaptive immunity; and finally from a genomic point of

view, the coelacanth is considered to have a slow-evolving genome due to the strong phenotypic resemblance between extinct and living specimens, suggesting a stasis at the genomic level, confirmed by analysis of protein-coding gene evolution (AMEMIYA *et al.* 2013). For all these reasons, the sequencing of the coelacanth genome represents a great opportunity for researchers to assess many of these interrogations. Being involved in the consortium of sequencing project, we took the chance to exploit the coelacanth genome and perform various studies. Indeed, it represents an important part of this thesis, and is a nice example of how we can use *in silico* analyses and what can we do with TEs.

The first step of analyses is the evaluation of TE diversity and content in the coelacanth genome, in collaboration with German teams (CHALOPIN *et al.* 2013). We found that this supposed "slow-evolving" genome is a relatively rich genome in terms of TE diversity (with about 26 superfamilies detected compared to less than 21 in other sarcopterygian studied species) but also in terms of quantity (with about 25% of the genome compared to less than that in the majority of other fish genome). Even more, the content of TEs might reach 50% depending on the assembly we are looking at. Indeed, a second sequencing and assembly, performed by Japanese teams, gave different results: the genome seems to be composed of a very different content of TEs, but presents a relatively similar diversity. To compare these two genomes, we performed the same analyses on the second genome ("Japanese" genome) that we done on the first one (Broad genome), using the library built from the first one. This study allowed us to highlight the difference in term of quantity. However, a bias remains, as we have not constructed a new library for the second genome, which would of course be interesting to build. Indeed, it is possible that this second assembly might contain other superfamilies, not present in the first library, and it is also possible that consensus sequences might be different if the parameters used to obtain the automatic library are different, influencing the Kimura analyses. Moreover, it is important to note that our current library still contains a lot of non-classified elements, probably leading to a bias in the statistic content we calculated. A more pushed reannotation needs to be done on this library, including for exemple the search for MITE elements using secondary structure analyses that are not in the current version.

The second step of analyses concerns the potential activity of TE elements in the coelacanth genome (FORCONI *et al.* 2013). In collaboration with an Italian group, we scanned the transcriptome of three different tissues, muscle, liver and testis, belonging to specimens from the two species. The muscle transcriptome sequencing was reported in the genome paper (*L. chalumnae*, (AMEMIYA *et al.* 2013)), while liver and testis transcriptome belonging to *L. menadoensis* were obtained and described in a parallel study (PALLAVICINI *et al.* 2013). We showed that non-LTR retrotransposons are the major contributors of TE RNAs in transcriptomes, especially CR1 and LF-SINE in the three

transcriptomes, as well as a LINE2 in testis and a hAT transposon in muscle. Furthermore, regarding the expression values versus the number of genomic copies, we tried to discriminate between TEs poorly represented in the genome but highly expressed, TEs poorly represented in the genome and in transcriptomes, TEs highly represented in genome but poorly expressed and TEs both highly represented and expressed. This analysis represents an interesting way to find cases of exaptation, and thus might be routinely included in our analyses, when transcriptomes are also available. It also allowed us to identify a new family of SINE elements that we called Coeg-SINE for Coelacanth G-rich SINE. All these elements were found to be expressed at a very low level but present in a high copy number in the genome. Many analyses still have to be performed concerning Coeg-SINE. For instance, the associated autonomous LINE if any,has not been discovered, yet. To do this, further alignments and phylogenetic reconstructions using LINE and SINE nucleotide sequences of the coelacanth genomes might help to identify retrotransposition partners through sequence similarities. These transcriptomic analyses represent an initial step allowing to have an idea of TE expression, but they present several biases. Indeed, it has been shown that the vast majority of the human genome is transcribed, thus it is probably the case for the coelacanth also, suggesting that the observed overall expression might be in fact a basal transcription. Moreover, transcriptomic analysis represents a bias because poly-A mRNAs are generally filtered. We thus should have an under-representation of LINE and SINE elements. Finally, the raw expression values, used to visualize the whole genome expression did not take into account the number of copy number within the genome. This can represent a strong bias: for example, in case of a basal transcription, a TE family that underwent a strong burst in the past will be present at a high copy number and as a consequence will be highly detected in transcriptomes. But this high proportion in the transcriptome is not linked to a real high activity by now. It is the reason why we also plotted copy number and expression values together.

As mentioned in the introduction, direct evidence for recent TE activity can be obtained through the identification of polymorphic insertions between parents and offspring, or between individuals from the same species. This is what we tried to assess in the third step of our analysis. Using the few available BAC sequences of *L. menadoensis*, we compared the TE composition and position at orthologous regions between the two species (Naville *et al.*, submitted). The idea was to try to identify orthologous sites, full in one species (with TE) and empty in the other (without TE). By this way, we demonstrated that some TEs have been active after the separation of both closely species, arguing against the genomic stasis hypothesis. Indeed, we found 27 species-specific TE insertions (or excisions) in the analyzed regions, which represent 5.7 Mb (0.2%) of the genome. Most of them belong to the CR1 family, which is in accordance with transcriptomic data showing a high expression. We also highlighted events of homologous recombination, which occurred probably more than once between LTRs of

epsilon-retrovirus sequences. Thanks to this analysis, we proved that TEs are dynamic elements of the coelacanth genome, probably driving its evolution, together with other evolutionary mechanisms.

This study opens many perspectives for further investigation. First, only a small proportion of the genome has been analyzed (0.2% of genome). A larger study at the genome scale might allow to obtain a more precise quantification of insertion polymorphism. Furthermore, the orthologous regions that we analyzed are probably mainly euchromatic regions, which are highly transcribed, and thus under strong selection against TEs. This represents a bias as the evolution and activity of TEs in these regions is certainly different than the ones in less active and less contrained regions. Indeed, the available genomic sequences of *L. menadoensis*, for which no assembled genome exists, are BAC sequences that have been selected for precise gene of interests, such as *Hox* and *Immunglobulin* genes. Knowing that, we probably expect a higher rate of TE activity in other less constrained genomic regions. Another data that we have not in our hand is the rate of TE elimination in the coelacanth genome. Indeed, we compared two specimens from the two different species, meaning that we evaluated TE movement occurring in a time space of 10 My. At the moment, we have no evidence to determine if the insertions we observed arose just after the separation of the two species, more recently, or represent polymorphism insertions from a common ancestor. The identical analysis with two specimens from a same species might help us to compare the content of polymorphism insertions (do we observe more or less polymorphism insertions when looking at individuals from two species, or from the same species?) and to evaluate the rate of TE mutation allowing us to infer the rate of deletion in the two coelacanth species. Japanese researchers have sequenced five specimens from the two species (4 *L. chalumnae* and 1 *L. menadoensis*) (Nikaido *et al.* 2013). Only one of them has been assembled by now, however this assembly differs a lot from the Broad assembly to perform comparison, as explained above. To obtain a proper comparison, assemblies must be performed with similar methods, but it represents a large amount of works and times. With six assemblies (5 *L. chalumnae* and 1 *L. menadoensis*), a more complete analysis of polymorphic insertions might be very interesting to perform in the future.

More recently, the genome of the elephant shark *Callorhinchus milii* has been sequenced and published (Venkatesh *et al.* 2014). The elephant shark is a cartilaginous fish, which has been used in the coelacanth genome studies as an outgroup of the bony vertebrates. Authors of this recent study demonstrated that the elephant shark presents a slow rate of protein-coding gene evolution compared to all other vertebrates, even the coelacanth. From our point of interest (TEs), one main question arises regarding these two genomes considered as “slow-evolving”, namely the coelacanth and the elephant shark, for which the rate of protein-coding gene evolution is lower than in other vertebrate genomes: do TEs undergo the same slow-evolutionary constraints in term of sequences (mutations for example), rate of transposition, and dynamic (slowly eliminated from these genomes

than others)? Indeed, if it is demonstrated that the rate of protein-coding gene evolution is low in these organisms, we still have no clues concerning the evolution of TEs. Even if the coelacanth and the elephant shark present very different TE compositions, we observe in both a relatively high diversity of TE superfamilies (26 for the coelacanth and 18 for the elephant shark) that represent a non-negligible proportion of these genomes (25% and 40% of the genomes, respectively). Furthermore, several TE superfamilies are probably still active, at least in the coelacanth genome (as mentioned above). We have no evidence concerning the elephant shark genome, except an intense recent burst of LINE2, CR1 and Copia retrotransposons, highlighted by Kimura analyses (Chapter 3, Chalopin *et al.* submitted). The way of answering the question dealing with TE evolution in slow-evolving genomes compared to other genomes would be 1- to select elements from vertebrate-widespread TE superfamily (as LINE1 or Tc-Mariner), 2- to collect all these copies corresponding to the choosen element, 3- to construct phylogenies using these sequences. At the end, we might be able to determine (by comparing the length of branches of the phylogenetic trees obtained) if TEs also harbour a slower evolution (shorter branches) in these two genomes compared to others.

IV- History and function of *Gin* genes

The process by which new genes and other genetic novelties are formed from TEs is called molecular domestication (MD). MD is an important mechanism that still needs to be understood and to be deeply studied in term of molecular processes (how a TEs can become a gene?) but also in term of diversity. More and more genes are described as containing TE-deriving domain and many of these proteins have fundamental functions for their hosts, such as THAP proteins involved in apoptosis, Rag proteins as essential agents of the V(D)J recombination, but also syncytins, genes important for placenta formation in mammals. Among all, some TE-derived genes are known to be eukaryote specific (such as *CENP-B*), others are amniote specific (*Gin-1*) and others are vertebrate specific (*Rag1*). However, looking at literature, most of TE-derived genes are described and studied in mammals. Considering the high number of mammalian TE-derived genes described compared to other vertebrates, and considering the large amount of active TEs or traces in vertebrates, I believe that we only observe the tip of the iceberg, and that there is still a lot of TE-derived genes to discover. Moreover, among the already discovered ones, our knowledges concerning their roles and functions are still limited.

I have started working on MD during my master, studying diversity and structure of the *PNMA* and *MART*, two Gag-derived genes families. Beside these two families, I was also interested in integrase-derived genes. Up to now, only five described genes, *c-integrase*, *Fob1p*, two *Gin* (for *Gypsy integrase*) and *CGin-1* (for *Cousin of Gin-1*) compose the group of integrase-derived genes in vertebrates. My interest was especially focused on *Gin*

genes, *Gin-1* and *Gin-2,* as they present an interesting distribution along organisms. It has been shown that they derived from GIN transposons, and they are found in amniotes for *Gin-1* and all vertebrates from cartilaginous fish except in placental and monotreme for *Gin-2*. As Gin-1 and Gin-2 putative proteins still contain a very conserved and easily "alignable" integrase domain, with detectable catalytic domain and zinc finger motif (only for *Gin-1*). As *Gin-2* has been detected in "basal" vertebrates, the question that arises concerning their origin is: did *Gin* genes occur from a single MD event or did they appear from two independent MD events? Phylogenetic reconstructions from published works (Marin 2010; Chalopin *et al.* 2012) separated both *Gin-1* and *Gin-2* in two different branches, each rooted by different GIN transposons, suggesting two independent MD events at the base of amniotes and vertebrate gnathostomes, respectively. If *Gin-1* seems to clearly derive from GINO transposons, the origin of *Gin-2* is more difficult to determine. The vertebrate sequences of Gin-2 proteins are branched with GIN-like sequences from Ciona. This GIN-like sequence seems to be a gene different from *Gin-2*. Finally, GIN transposons root vertebrate Gin-2 sequences and Ciona GIN-like. Nevertheless, Marìn (Marin 2010) showed that both *Gin-1* and *Gin-2* have a structure highly similar to GINO transposons, even if phylogenetic reconstructions were not so clear. Furthermore, in our last recent analysis (Figure 31), branch supports between Gin-2 sequences and other GIN transposons are very weak, probably reflecting the difficulties to resolve the relationships between *Gin-2* and transposon it derives from. Going back to the question of their origin, one of the hypotheses would be that both genes derived from GINO transposons. However, it is possible that, after the MD event, *Gin-2* diverged much more than *Gin-1*, changing the position of the branch. Adding new GIN transposons and *Gin* sequences from recently sequenced genomes may help to resolve the origin of these two genes. Indeed, most Gin analyses were performed two years ago. Since then, many new genomes have been sequenced or re-assembled. A rapid blast search brought new *Gin-2* turtle and *Gin-1* tasmanian devil, armadillo and turtle sequences that would be helpful for such analyses.

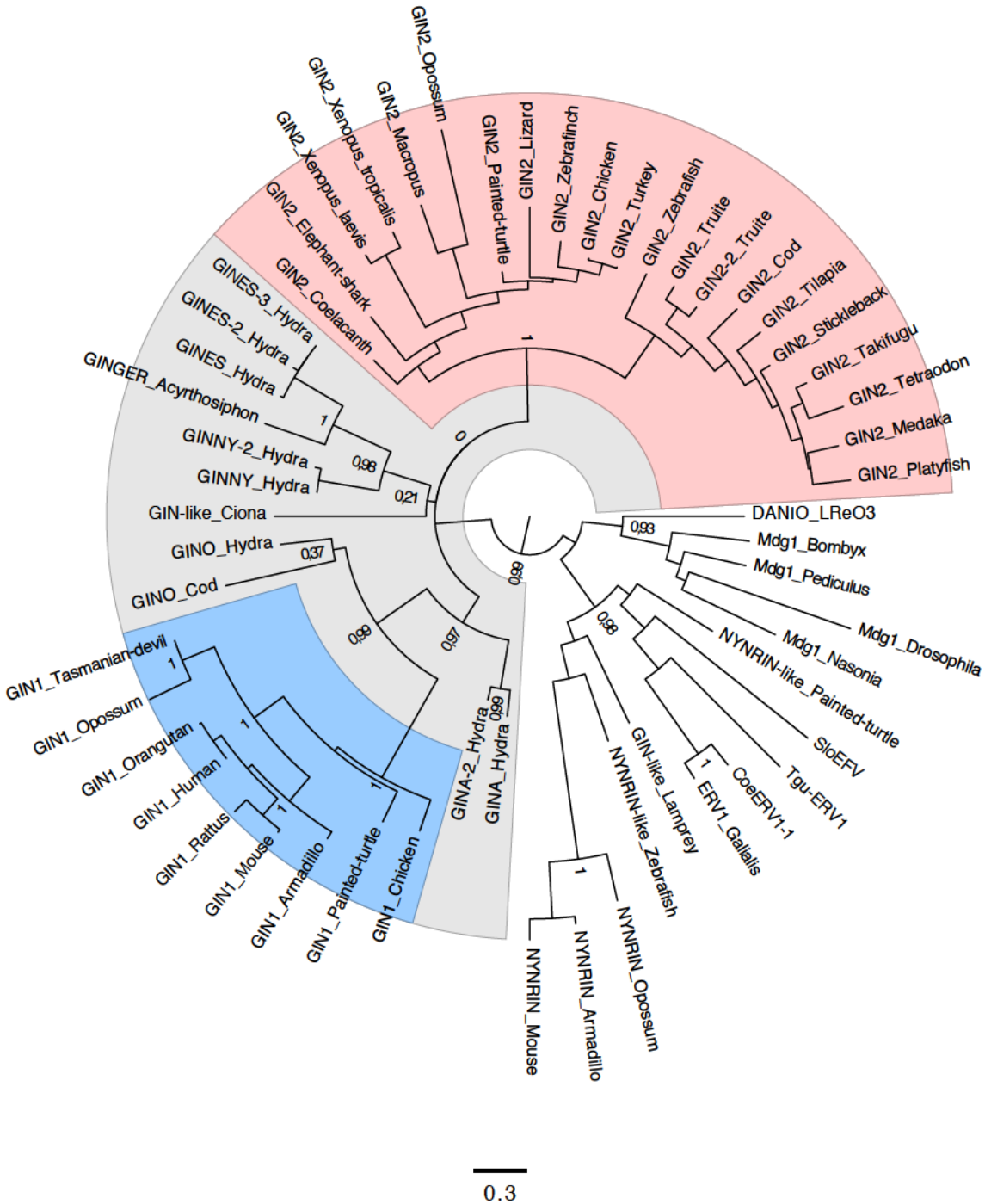


Figure 31: Phylogenetic reconstruction of *Gin* genes and TE integrase domains. Reconstruction was performed from an integrase alignment (188 amino acid sites) using muscle alignment software and maximum likelihood reconstruction, with optimized parameters. *Gin* genes and GIN transposons are surrounded by grey (GIN transposons), blue (Gin-1 sequences) and pink (Gin-2 sequences) colors, CGin-1 (NYNRIN) and ERV sequences were added to analyze NYNRIN origin. Gypsy LTR retrotransposons were added as outgroups.

Considering the potential functions of integrase-derived genes, it has been first suggested that *Gin-1* could be part of a defense mechanism against retrotransposons and retrovirus. As well, Marco and Marìn (Marco and Marin 2009) suggested that *CGIN-1* might be involved in the protection against viral infection by regulating the ubiquitination of viral proteins. It is not the first times that TE-derived genes are proposed to be involved in defense mechanism against TE infection and proliferation (Miskey *et al.* 2007; Newman *et al.* 2008; Bire *et al.* 2013). For example, it has been suggested that SETMAR protein might regulate the expression of Hsmar1 transposase in human cells (Miskey *et al.* 2007). From a different view, *Gin* genes encode a transposase (deriving from an integrase), which might help to DNA or RNA integration. Despite these suggestions, *Gin-1* and *Gin-2* roles are not deciphered yet, but the fact that *Gin-2* has been conserved over 500 My in most vertebrate lineages, suggests its fundamental role in vertebrates. Furthermore, the fact that *Gin-2* has been lost in mammals also rises an interesting question: is Gin-1 proteins counter-balancing Gin-2 function in this lineage? We investigated the expression profile of *Gin-2* in different fish species. In adult tissues, *Gin-2* appeared to be strongly expressed in brain and gonads. Interestingly, many other TE-derived genes present a brain and gonad expression pattern, as illustrated by *MART* (Brandt *et al.* 2005) and *PNMA* genes (Dalmau *et al.* 1999; Rosenfeld *et al.* 2001; Volff 2009). To complete the expression profile analyses, we also carried out experiments in medaka and zebrafish embryos. If *Gin-2* is maternally expressed in medaka embryos, its expression is correlated with gastrula stages in zebrafish. As the gastrulation implies important changes in the embryo, such as cell movements (invagination, involution, ingression, delamination and epiboly) and cell differentiation (endoderme, mesoderme and ectoderme formation), these results lead to the conclusion that *Gin-2* might have a role during embryogenesis, at least in zebrafish. In addition, *in situ* hybridization experiments have been done to precise *Gin-2* expression pattern in embryos.

To summarize about *Gin-2* expression: 1- its expression during precise window of timing during embryogenesis supports the fact that it might have a fundamental role, at least in fish; 2- it represents an interesting case of MD because it is present over very long period and because it might have an important role, but also because of the differential expression observed in zebrafish and medaka. Concerning the two different patterns observed in these two fish, the two possibilities are that *Gin-2* has a similar function in the two species but is expressed at different times, or *Gin-2* has two different roles in zebrafish and medaka during two different windows of expression; 3- Carrying out morpholinos experiments and/or mutagenesis might help to observe *Gin-2*-specific phenotype.

V- Molecular domestication in vertebrate genomes: is this phenomenon rare?

As cited in the previous part, the MD is an important evolutionary event, which can generate several fundamental genetic novelties, such as new regulatory and exonic sequences, or even new genes. In this manuscript, I particularly focused on new genes, so-called TE-derived genes. Associated to new genes, new functions are acquired by the host after an event of MD. The diversity of these TE-derived sequences was investigated more than once in human and mouse (Campillos et al 2006; Volff 2006; Sinzelle et al 2009; Kokosar & Kordis 2013). More than 50 TE-derived genes have been described in human including multigenic families such as *MART*, *PNMA* or *SCAN* families. Given that TEs are widespread in vertebrate genomes, sometimes reaching very high content and because they appear to be a major source of regulatory sequences (Bourque *et al.* 2008; Faulkner *et al.* 2009), we can expect that we currently observe only the tip of the iceberg regarding the repertoire of TE-derived genes in vertebrate lineages. Many TE-derived genes still have to be identified, explaining why such studies are of great interest.

All the studies realized up to now have described genes deriving from DNA transposons and LTR retrotransposons including retroelements. However, regarding the LINE elements, which represent the most abundant elements in human, they have never been identified as source of TE-derived genes. I am not speaking about SINEs since they are non-coding elements even if they can be source of RNA genes (Brosius 1999; Krull *et al.* 2005). The questions that arise from this observation are: 1- is it an artefact due to the fact that many LINE copies are still present in mammals, making their deriving sequences complicated to detect; or 2- are LINE retrotransposons never co-opted as new genes? And if it is the case, why? Likewise, the lack of DIRS-derived genes raises identical interrogations. However, for these two last orders, the lack of corresponding derived-genes might be due to the scarcity of studies in non-mammalian genomes as elements belonging to these orders are mainly found in non-mammalian organisms. Indeed, most of the TE-derived genes were first identified in mammals, and then searched in other vertebrates, but the inverse was not done. Finally, taking into account that most of the TE-derived genes identified are specific to mammals, a third point can be open to debate: are mammals more subject to MD event than other vertebrates? Or is this situation due to the lack of investigation in other vertebrates?

The reasons why we do not have a bigger repertoire of domesticated genes might be due to 1- MD is a rare phenomenon, 2- a large-scale analysis of vertebrate TE-derived diversity has not been investigated yet, 3- the high number of copies of TEs more or less degraded probably confuse and make difficult the detection of TE-derived sequences. In this context, new innovative methods have to be developed and tested in order to

improve TE-derived genes identification. A large-scale analysis of the human genome performed in 2001 (NEKRUTENKO AND LI 2001) crossed 13799 human genes with known TE protein and found out that about 4% of the human genes might contain TE sequences in their coding region. However, as final result, they only found 32 genes containing a TE in an exon.

With the aim to improve TE-derived gene detection (and TE-derived regulatory sequences as well) and therefore to increase the respective repertoire, I developed and tested a large-scale genomic method based on the use of non-specific TE libraries to mask the studied genome. The principle of the method is based on the different genome TE composition in vertebrates: using potentially still active TE sequences in a given genome to detect fossils in other genomes. To evaluate if this method might give interesting results, we performed preliminary tests as follows: 1- we used a "cleaned" human genome (pre-masked with its own TE library) to limit the background noise due to cryptic TEs; 2- the "cleaned" human genome was then masked with the coelacanth-specific TE library; 3- human sequences that shared similarities with coelacanth TEs were listed and filtered to keep only sequences falling into exons and present at low copy number (a domesticated genes is a single copy; or few copies in case of duplication events). At the end, we found sequences related to 13 DNA transposons (from Tc-Mariner, hAT, Harbinger and Maverick superfamilies), 26 LINE (from CR1, RTE, L2 and Penelope) and 9 LTR (from DIRS and Gypsy) retrotransposons, with variable sizes (from 40 nt to several hundred) located at least in one exon of an already annotated gene according to Ensembl. Analyzing in details sequences similar to coelacanth DNA transposons, three genes previously known as derived from TEs were found: *TIGD4* (sharing 430 nt with a Tc-Mariner element), *TIGD5* (sharing 118 nt with a Tc-Mariner) and *HARBI1* (sharing 674 nt with a Harbinger), thus validating the approach. It is interesting to note that all the previously described TE-derived genes in human were not found through this study, probably because they derive from TEs that are not present in the coelacanth. So, it appears necessary, in order to recover all putative TE-derived genes, to do the same analysis using different organisms against one genome or to enhance the detection using a combined library that may allow obtaining more information at once. Overall, these results are extremely encouraging. Now, we have to precisely analyze the different corresponding genes, listed before, especially those containing sequences similar to LINE elements, as no LINE-derived gene has been described so far. Furthermore, we have to test this strategy on other vertebrate lineages, for instance by investigating the coelacanth genome with the human library or with a combined-teleost TE library. In this analysis, we filtered sequences sharing similarities with an exon, it might also be possible to change this filter to recover regulatory sequences located in the first introns of a gene, or upstream near the gene.
This new innovative approach might bring new insights regarding the diversity and distribution of MD event within vertebrate lineages.

Conclusion of the thesis

Since their discovery and for a long period, TEs were considered as junk DNA. Nevertheless, more and more studies underlined the strong beneficial impacts of them in gene and genome evolution. In this context, during my thesis, I studied evolutionary aspects and dynamics of TEs in vertebrate lineages. Involved in several genome consortium projects, including species with particular biological interests (such as the Southern platyfish, the coelacanth or the rainbow trout), I had the opportunity to investigate, in advance, TE population in not published genomes. Taking advantage of the high accumulation of new sequenced vertebrate genomes, in particular fish genomes, I performed a large-scale comparative analysis of TE content, diversity and dynamics within vertebrates. This work gave rise to a large overview of TE evolution in this animal branch. As discussed above, this study opens many perspectives with solid foundations. In parallel, I investigated the innovative power of TEs by analyzing *Gin-2* gene. The results obtained are very encouraging to push the investigation and try to discover the function of this TE-derived gene.

BIBLIOGRAPHY

Abrahamsen, M. S., T. J. Templeton, S. Enomoto, J. E. Abrahante, G. Zhu *et al.*, 2004 Complete genome sequence of the apicomplexan, Cryptosporidium parvum. Science 304**:** 441-445.

Agrawal, A., Q. M. Eastman and D. G. Schatz, 1998 Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. Nature 394**:** 744-751.

Aguiar, R. S., and B. M. Peterlin, 2008 APOBEC3 proteins and reverse transcription. Virus Res 134**:** 74-85.

Aizawa, K., A. Shimada, K. Naruse, H. Mitani and A. Shima, 2003 The medaka midblastula transition as revealed by the expression of the paternal genome. Gene Expr Patterns 3**:** 43-47.

Akagi, K., J. Li, R. M. Stephens, N. Volfovsky and D. E. Symer, 2008 Extensive variation between inbred mouse strains due to endogenous L1 retrotransposition. Genome Res 18**:** 869-880.

Amemiya, C. T., J. Alfoldi, A. P. Lee, S. Fan, H. Philippe *et al.*, 2013 The African coelacanth genome provides insights into tetrapod evolution. Nature 496**:** 311-316.

Anxolabehere, D., M. G. Kidwell and G. Periquet, 1988 Molecular characteristics of diverse populations are consistent with the hypothesis of a recent invasion of Drosophila melanogaster by mobile P elements. Mol Biol Evol 5**:** 252-269.

Aparicio, S., J. Chapman, E. Stupka, N. Putnam, J. M. Chia *et al.*, 2002 Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science 297**:** 1301-1310.

Aravin, A. A., G. J. Hannon and J. Brennecke, 2007 The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science 318**:** 761-764.

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

Arkhipova, I., and M. Meselson, 2005a Deleterious transposable elements and the extinction of asexuals. Bioessays 27**:** 76-85.

Arkhipova, I. R., and M. Meselson, 2005b Diverse DNA transposons in rotifers of the class Bdelloidea. Proc Natl Acad Sci U S A 102**:** 11781-11786.

Arkhipova, I. R., K. I. Pyatkov, M. Meselson and M. B. Evgen'ev, 2003 Retroelements containing introns in diverse invertebrate taxa. Nat Genet 33**:** 123-124.

Babushok, D. V., and H. H. Kazazian, Jr., 2007 Progress in understanding the biology of the human mutagen LINE-1. Hum Mutat 28**:** 527-539.

Bachtrog, D., 2006 A dynamic view of sex chromosome evolution. Curr Opin Genet Dev 16**:** 578-585.

Badge, R. M., and J. F. Brookfield, 1997 The role of host factors in the population dynamics of selfish transposable elements. J Theor Biol 187**:** 261-271.

Bao, W., M. G. Jurka, V. V. Kapitonov and J. Jurka, 2009 New superfamilies of eukaryotic DNA transposons and their internal divisions. Mol Biol Evol 26**:** 983-993.

Bao, W., V. V. Kapitonov and J. Jurka, 2010 Ginger DNA transposons in eukaryotes and their evolutionary relationships with long terminal repeat retrotransposons. Mob DNA 1**:** 3.

Baucom, R. S., J. C. Estill, C. Chaparro, N. Upshaw, A. Jogi *et al.*, 2009 Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. PLoS Genet 5**:** e1000732.

Bejerano, G., C. B. Lowe, N. Ahituv, B. King, A. Siepel *et al.*, 2006 A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature 441: 87-90.

Belancio, V. P., D. J. Hedges and P. Deininger, 2008 Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health. Genome Res 18: 343-358.

Bennetzen, J. L., 2000 Transposable element contributions to plant gene and genome evolution. Plant Mol Biol 42: 251-269.

Bennetzen, J. L., 2005 Transposable elements, gene creation and genome rearrangement in flowering plants. Curr Opin Genet Dev 15: 621-627.

Bennetzen, J. L., J. Ma and K. M. Devos, 2005 Mechanisms of recent genome size variation in flowering plants. Ann Bot 95: 127-132.

Benveniste, R. E., and G. J. Todaro, 1973 Homology between type-C viruses of various species as determined by molecular hybridization. Proc Natl Acad Sci U S A 70: 3316-3320.

Bernard, D., B. Mehul, A. Thomas-Collignon, C. Delattre, M. Donovan *et al.*, 2005 Identification and characterization of a novel retroviral-like aspartic protease specifically expressed in human epidermis. J Invest Dermatol 125: 278-287.

Bharathan, G., T. E. Goliber, C. Moore, S. Kessler, T. Pham *et al.*, 2002 Homologies in leaf form inferred from KNOXI gene expression during development. Science 296: 1858-1860.

Bi, K., and J. P. Bogart, 2010 Time and time again: unisexual salamanders (genus Ambystoma) are the oldest unisexual vertebrates. BMC Evol Biol 10: 238.

Bingham, P. M., M. G. Kidwell and G. M. Rubin, 1982 The molecular basis of P-M hybrid dysgenesis: the role of the P element, a P-strain-specific transposon family. Cell 29: 995-1004.

Bire, S., S. Casteret, A. Arnaoty, B. Piegu, T. Lecomte *et al.*, 2013 Transposase concentration controls transposition activity: myth or reality? Gene 530: 165-171.

Bishop, R., T. Shah, R. Pelle, D. Hoyle, T. Pearson *et al.*, 2005 Analysis of the transcriptome of the protozoan Theileria parva using MPSS reveals that the majority of genes are transcriptionally active in the schizont stage. Nucleic Acids Res 33: 5503-5511.

Blaise, S., N. de Parseval, L. Benit and T. Heidmann, 2003 Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. Proc Natl Acad Sci U S A 100: 13013-13018.

Blaise, S., N. de Parseval and T. Heidmann, 2005 Functional characterization of two newly identified Human Endogenous Retrovirus coding envelope genes. Retrovirology 2: 19.

Blass, E., M. Bell and S. Boissinot, 2012 Accumulation and rapid decay of non-LTR retrotransposons in the genome of the three-spine stickleback. Genome Biol Evol 4: 687-702.

Bodem, J., T. Schied, R. Gabriel, M. Rammling and A. Rethwilm, 2011 Foamy virus nuclear RNA export is distinct from that of other retroviruses. J Virol 85: 2333-2341.

Boeke, J. D., 2003 The unusual phylogenetic distribution of retrotransposons: a hypothesis. Genome Res 13: 1975-1983.

Bohne, A., F. Brunet, D. Galiana-Arnoux, C. Schultheis and J. N. Volff, 2008 Transposable elements as drivers of genomic and biological diversity in vertebrates. Chromosome Res 16: 203-215.

Bohne, A., Q. Zhou, A. Darras, C. Schmidt, M. Schartl *et al.*, 2012 Zisupton--a novel superfamily of DNA transposable elements recently active in fish. Mol Biol Evol 29: 631-645.
Boissinot, S., P. Chevret and A. V. Furano, 2000 L1 (LINE-1) retrotransposon evolution and amplification in recent human history. Mol Biol Evol 17: 915-928.
Boissinot, S., J. Davis, A. Entezam, D. Petrov and A. V. Furano, 2006 Fitness cost of LINE-1 (L1) activity in humans. Proc Natl Acad Sci U S A 103: 9590-9594.
Bourc'his, D., and T. H. Bestor, 2004 Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. Nature 431: 96-99.
Bourque, G., 2009 Transposable elements in gene regulation and in the evolution of vertebrate genomes. Curr Opin Genet Dev 19: 607-612.
Bourque, G., B. Leong, V. B. Vega, X. Chen, Y. L. Lee *et al.*, 2008 Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res 18: 1752-1762.
Bousalem, M., E. J. Douzery and S. E. Seal, 2008 Taxonomy, molecular phylogeny and evolution of plant reverse transcribing viruses (family Caulimoviridae) inferred from full-length genome and reverse transcriptase sequences. Arch Virol 153: 1085-1102.
Bowen, N. J., and I. K. Jordan, 2002 Transposable elements and the evolution of eukaryotic complexity. Curr Issues Mol Biol 4: 65-76.
Bradley, D., R. Carpenter, H. Sommer, N. Hartley and E. Coen, 1993 Complementary floral homeotic phenotypes result from opposite orientations of a transposon at the plena locus of Antirrhinum. Cell 72: 85-95.
Brandt, J., S. Schrauth, A. M. Veith, A. Froschauer, T. Haneke *et al.*, 2005 Transposable elements as a source of genetic innovation: expression and evolution of a family of retrotransposon-derived neogenes in mammals. Gene 345: 101-111.
Brayton, K. A., A. O. Lau, D. R. Herndon, L. Hannick, L. S. Kappmeyer *et al.*, 2007 Genome sequence of Babesia bovis and comparative analysis of apicomplexan hemoprotozoa. PLoS Pathog 3: 1401-1413.
Brookfield, J. F., 1982 Interspersed repetitive DNA sequences are unlikely to be parasitic. J Theor Biol 94: 281-299.
Brookfield, J. F., 1996 Genetic evidence for repression of somatic P element movements in Drosophila melanogaster consistent with a role for the KP element. Heredity (Edinb) 76 (Pt 4): 384-391.
Brookfield, J. F., 2005a The ecology of the genome - mobile DNA elements and their hosts. Nat Rev Genet 6: 128-136.
Brookfield, J. F., 2005b Evolutionary forces generating sequence homogeneity and heterogeneity within retrotransposon families. Cytogenet Genome Res 110: 383-391.
Brosius, J., 1999 RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene 238: 115-134.
Brouha, B., J. Schustak, R. M. Badge, S. Lutz-Prigge, A. H. Farley *et al.*, 2003 Hot L1s account for the bulk of retrotransposition in the human population. Proc Natl Acad Sci U S A 100: 5280-5285.
Bureau, T. E., and S. R. Wessler, 1992 Tourist: a large family of small inverted repeat elements frequently associated with maize genes. Plant Cell 4: 1283-1294.
Bureau, T. E., and S. R. Wessler, 1994 Mobile inverted-repeat elements of the Tourist family are associated with the genes of many cereal grasses. Proc Natl Acad Sci U S A 91: 1411-1415.

Burke, W. D., H. S. Malik, S. M. Rich and T. H. Eickbush, 2002 Ancient lineages of non-LTR retrotransposons in the primitive eukaryote, Giardia lamblia. Mol Biol Evol 19**:** 619-630.

Burwinkel, B., and M. W. Kilimann, 1998 Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease. J Mol Biol 277**:** 513-517.

Cantrell, M. A., L. Scott, C. J. Brown, A. R. Martinez and H. A. Wichman, 2008 Loss of LINE-1 activity in the megabats. Genetics 178**:** 393-404.

Cappello, J., K. Handelsman and H. F. Lodish, 1985 Sequence of Dictyostelium DIRS-1: an apparent retrotransposon with inverted terminal repeats and an internal circle junction sequence. Cell 43**:** 105-115.

Capy, P., 2005 Classification and nomenclature of retrotransposable elements. Cytogenet Genome Res 110**:** 457-461.

Capy, P., D. Anxolabehere and T. Langin, 1994 The strange phylogenies of transposable elements: are horizontal transfers the only explantation? Trends Genet 10**:** 7-12.

Capy, P., T. Langin, D. Higuet, P. Maurer and C. Bazin, 1997 Do the integrases of LTR-retrotransposons and class II element transposases have a common ancestor? Genetica 100**:** 63-72.

Carlton, J. M., S. V. Angiuoli, B. B. Suh, T. W. Kooij, M. Pertea *et al.*, 2002 Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. Nature 419**:** 512-519.

Carthew, R. W., and E. J. Sontheimer, 2009 Origins and Mechanisms of miRNAs and siRNAs. Cell 136**:** 642-655.

Casavant, N. C., L. Scott, M. A. Cantrell, L. E. Wiggins, R. J. Baker *et al.*, 2000 The end of the LINE?: lack of recent L1 activity in a group of South American rodents. Genetics 154**:** 1809-1817.

Casola, C., D. Hucks and C. Feschotte, 2008 Convergent domestication of pogo-like transposases into centromere-binding proteins in fission yeast and mammals. Mol Biol Evol 25**:** 29-41.

Castillo, D. M., J. C. Mell, K. S. Box and J. P. Blumenstiel, 2011 Molecular evolution under increasing transposable element burden in Drosophila: a speed limit on the evolutionary arms race. BMC Evol Biol 11**:** 258.

Castro, J. P., and C. M. Carareto, 2004 Drosophila melanogaster P transposable elements: mechanisms of transposition and regulation. Genetica 121**:** 107-118.

Chalmers, R. M., and N. Kleckner, 1996 IS10/Tn10 transposition efficiently accommodates diverse transposon end configurations. EMBO J 15**:** 5112-5122.

Chalopin, D., S. Fan, O. Simakov, A. Meyer, M. Schartl *et al.*, 2013 Evolutionary active transposable elements in the genome of the coelacanth. J Exp Zool B Mol Dev Evol.

Chalopin, D., D. Galiana and J. N. Volff, 2012 Genetic innovation in vertebrates: gypsy integrase genes and other genes derived from transposable elements. Int J Evol Biol 2012**:** 724519.

Charlesworth, B., 1994 Evolution. How does increased fitness evolve? Curr Biol 4**:** 1146-1148.

Charlesworth, B., and C. H. Langley, 1989 The population genetics of Drosophila transposable elements. Annu Rev Genet 23**:** 251-287.

Charlesworth, B., C. H. Langley and W. Stephan, 1986 The evolution of restricted recombination and the accumulation of repeated DNA sequences. Genetics 112**:** 947-962.

Chen, S., G. Zhang, C. Shao, Q. Huang, G. Liu *et al.*, 2014 Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. Nat Genet 46**:** 253-260.

Chiu, Y. L., and W. C. Greene, 2008 The APOBEC3 cytidine deaminases: an innate defensive network opposing exogenous retroviruses and endogenous retroelements. Annu Rev Immunol 26**:** 317-353.

Coffin, J. M., S. H. Hughes and H. E. Varmus, 1997 The Interactions of Retroviruses and their Hosts in *Retroviruses*, edited by J. M. Coffin, S. H. Hughes and H. E. Varmus, Cold Spring Harbor (NY).

Cook, J. M., and M. Tristem, 1997 'SINEs of the times' - transposable elements as clade markers for their hosts. Trends Ecol Evol 12**:** 295-297.

Cordaux, R., and M. A. Batzer, 2009 The impact of retrotransposons on human genome evolution. Nat Rev Genet 10**:** 691-703.

Cordaux, R., S. Udit, M. A. Batzer and C. Feschotte, 2006 Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. Proc Natl Acad Sci U S A 103**:** 8101-8106.

Cordonnier, A., J. F. Casella and T. Heidmann, 1995 Isolation of novel human endogenous retrovirus-like elements with foamy virus-related pol sequence. J Virol 69**:** 5890-5897.

Cowley, M., and R. J. Oakey, 2013 Transposable elements re-wire and fine-tune the transcriptome. PLoS Genet 9**:** e1003234.

Coyne, J. A., and H. A. Orr, 1998 The evolutionary genetics of speciation. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences 353**:** 287-305.

Crozatier, M., C. Vaury, I. Busseau, A. Pelisson and A. Bucheton, 1988 Structure and genomic organization of I elements involved in I-R hybrid dysgenesis in Drosophila melanogaster. Nucleic Acids Res 16**:** 9199-9213.

Curcio, M. J., and K. M. Derbyshire, 2003 The outs and ins of transposition: from mu to kangaroo. Nat Rev Mol Cell Biol 4**:** 865-877.

Czech, B., J. B. Preall, J. McGinn and G. J. Hannon, 2013 A transcriptome-wide RNAi screen in the Drosophila ovary reveals factors of the germline piRNA pathway. Mol Cell 50**:** 749-761.

Dalloul, R. A., J. A. Long, A. V. Zimin, L. Aslam, K. Beal *et al.*, 2010 Multi-platform next-generation sequencing of the domestic turkey (Meleagris gallopavo): genome assembly and analysis. PLoS Biol 8.

Dalmau, J., S. H. Gultekin, R. Voltz, R. Hoard, T. DesChamps *et al.*, 1999 Ma1, a novel neuron- and testis-specific protein, is recognized by the serum of patients with paraneoplastic neurological disorders. Brain 122 (Pt 1)**:** 27-39.

Daniels, G. R., and P. L. Deininger, 1985 Repeat sequence families derived from mammalian tRNA genes. Nature 317**:** 819-822.

Daniels, S. B., K. R. Peterson, L. D. Strausbaugh, M. G. Kidwell and A. Chovnick, 1990 Evidence for horizontal transmission of the P transposable element between Drosophila species. Genetics 124**:** 339-355.

Dawkins, R., 1976 *The Selfish Gene*. Oxford University Press.

Dawson, A., E. Hartswood, T. Paterson and D. J. Finnegan, 1997 A LINE-like transposable element in Drosophila, the I factor, encodes a protein with properties similar to those of retroviral nucleocapsids. EMBO J 16**:** 4448-4455.

de Boer, J. G., R. Yazawa, W. S. Davidson and B. F. Koop, 2007 Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. BMC Genomics 8: 422.

De Cecco, M., S. W. Criscione, A. L. Peterson, N. Neretti, J. M. Sedivy *et al.*, 2013 Transposable elements become active and mobile in the genomes of aging mammalian somatic tissues. Aging (Albany NY) 5: 867-883.

de Chastonay, Y., H. Felder, C. Link, P. Aeby, H. Tobler *et al.*, 1992 Nucleotide sequence of PAT, a retroid element with unusual DR organization, isolated from Panagrellus redivivus. DNA Seq 3: 251-255.

de la Chaux, N., and A. Wagner, 2011 BEL/Pao retrotransposons in metazoan genomes. BMC Evol Biol 11: 154.

de Souza, F. S., L. F. Franchini and M. Rubinstein, 2013 Exaptation of transposable elements into novel cis-regulatory elements: is the evidence always strong? Mol Biol Evol 30: 1239-1251.

Deininger, P. L., and M. A. Batzer, 1999 Alu repeats and human disease. Mol Genet Metab 67: 183-193.

Devlin, R. H., and Y. Nagahama, 2002 Sex determination and sex differentiation in fish: an overview of genetic, physiological, and environmental influences. Aquaculture 208: 191-364.

Dewannieux, M., C. Esnault and T. Heidmann, 2003 LINE-mediated retrotransposition of marked Alu sequences. Nat Genet 35: 41-48.

Dion-Cote, A. M., S. Renaut, E. Normandeau and L. Bernatchez, 2014 RNA-seq reveals transcriptomic shock involving transposable elements reactivation in hybrids of young lake whitefish species. Mol Biol Evol.

Dlakic, M., 2002 A model of the replication fork blocking protein Fob1p based on the catalytic core domain of retroviral integrases. Protein Sci 11: 1274-1277.

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

Doolittle, W. F., and C. Sapienza, 1980 Selfish genes, the phenotype paradigm and genome evolution. Nature 284: 601-603.

Dooner, H. K., and C. F. Weil, 2007 Give-and-take: interactions between DNA transposons and their host plant genomes. Curr Opin Genet Dev 17: 486-492.

Dufresne, F., and N. Jeffery, 2011 A guided tour of large genome size in animals: what we know and where we are heading. Chromosome Res 19: 925-938.

Duncan, L., K. Bouckaert, F. Yeh and D. L. Kirk, 2002 kangaroo, a mobile element from Volvox carteri, is a member of a newly recognized third class of retrotransposons. Genetics 162: 1617-1630.

Dupressoir, A., G. Marceau, C. Vernochet, L. Benit, C. Kanellopoulos *et al.*, 2005 Syncytin-A and syncytin-B, two fusogenic placenta-specific murine envelope genes of retroviral origin conserved in Muridae. Proc Natl Acad Sci U S A 102: 725-730.

Eickbush, T. H., and V. K. Jamburuthugoda, 2008 The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Res 134: 221-234.

El Baidouri, M., M. C. Carpentier, R. Cooke, D. Gao, E. Lasserre *et al.*, 2014 Widespread and frequent horizontal transfers of transposable elements in plants. Genome Res.

Emera, D., C. Casola, V. J. Lynch, D. E. Wildman, D. Agnew *et al.*, 2012 Convergent evolution of endometrial prolactin expression in primates, mice, and elephants through the independent recruitment of transposable elements. Mol Biol Evol 29: 239-247.

Emerson, R. O., and J. H. Thomas, 2011 Gypsy and the birth of the SCAN domain. J Virol 85: 12043-12052.
English, J., K. Harrison and J. D. Jones, 1993 A genetic analysis of DNA sequence requirements for Dissociation state I activity in tobacco. Plant Cell 5: 501-514.
Erickson, I. K., M. A. Cantrell, L. Scott and H. A. Wichman, 2011 Retrofitting the genome: L1 extinction follows endogenous retroviral expansion in a group of muroid rodents. J Virol 85: 12315-12323.
Erlandsson, R., J. F. Wilson and S. Paabo, 2000 Sex chromosomal transposable element accumulation and male-driven substitutional evolution in humans. Mol Biol Evol 17: 804-812.
Evgen'ev, M. B., and I. R. Arkhipova, 2005 Penelope-like elements--a new class of retroelements: distribution, function and possible evolutionary significance. Cytogenet Genome Res 110: 510-521.
Fattash, I., R. Rooke, A. Wong, C. Hui, T. Luu *et al.*, 2013 Miniature inverted-repeat transposable elements: discovery, distribution, and activity. Genome 56: 475-486.
Faulkner, G. J., Y. Kimura, C. O. Daub, S. Wani, C. Plessy *et al.*, 2009 The regulated retrotransposon transcriptome of mammalian cells. Nat Genet 41: 563-571.
Feschotte, C., 2004 Merlin, a new superfamily of DNA transposons identified in diverse animal genomes and related to bacterial IS1016 insertion sequences. Mol Biol Evol 21: 1769-1780.
Feschotte, C., 2010 Virology: Bornavirus enters the genome. Nature 463: 39-40.
Feschotte, C., N. Jiang and S. R. Wessler, 2002 Plant transposable elements: where genetics meets genomics. Nat Rev Genet 3: 329-341.
Feschotte, C., and E. J. Pritham, 2005 Non-mammalian c-integrases are encoded by giant transposable elements. Trends Genet 21: 551-552.
Feschotte, C., and E. J. Pritham, 2007 DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet 41: 331-368.
Feschotte, C., L. Swamy and S. R. Wessler, 2003 Genome-wide analysis of mariner-like transposable elements in rice reveals complex relationships with stowaway miniature inverted repeat transposable elements (MITEs). Genetics 163: 747-758.
Feschotte, C., and S. R. Wessler, 2001 Treasures in the attic: rolling circle transposons discovered in eukaryotic genomes. Proc Natl Acad Sci U S A 98: 8923-8924.
Finnegan, D. J., 1989 Eukaryotic transposable elements and genome evolution. Trends Genet 5: 103-107.
Flajnik, M. F., 2002 Comparative analyses of immunoglobulin genes: surprises and portents. Nat Rev Immunol 2: 688-698.
Flavell, A. J., 1992 Ty1-copia group retrotransposons and the evolution of retroelements in the eukaryotes. Genetica 86: 203-214.
Flavell, A. J., S. R. Pearce and A. Kumar, 1994 Plant transposable elements and the genome. Curr Opin Genet Dev 4: 838-844.
Flavell, R. B., M. D. Bennett, J. B. Smith and D. B. Smith, 1974 Genome size and the proportion of repeated nucleotide sequence DNA in plants. Biochem Genet 12: 257-269.
Flot, J. F., B. Hespeels, X. Li, B. Noel, I. Arkhipova *et al.*, 2013 Genomic evidence for ameiotic evolution in the bdelloid rotifer Adineta vaga. Nature 500: 453-457.

Forconi, M., D. Chalopin, M. Barucca, M. A. Biscotti, G. De Moro *et al.*, 2013 Transcriptional activity of transposable elements in coelacanth. J Exp Zool B Mol Dev Evol.

Fujiwara, H., M. Osanai, T. Matsumoto and K. K. Kojima, 2005 Telomere-specific non-LTR retrotransposons and telomere maintenance in the silkworm, Bombyx mori. Chromosome Res 13**:** 455-467.

Gale, M., Jr., C. M. Blakely, D. A. Hopkins, M. W. Melville, M. Wambach *et al.*, 1998 Regulation of interferon-induced protein kinase PKR: modulation of P58IPK inhibitory function by a novel protein, P52rIPK. Mol Cell Biol 18**:** 859-871.

Gao, X., and D. F. Voytas, 2005 A eukaryotic gene family related to retroelement integrases. Trends Genet 21**:** 133-137.

Gardner, M. J., S. J. Shallom, J. M. Carlton, S. L. Salzberg, V. Nene *et al.*, 2002 Sequence of Plasmodium falciparum chromosomes 2, 10, 11 and 14. Nature 419**:** 531-534.

Genomes Project, C., G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks *et al.*, 2010 A map of human genome variation from population-scale sequencing. Nature 467**:** 1061-1073.

Gentles, A. J., M. J. Wakefield, O. Kohany, W. Gu, M. A. Batzer *et al.*, 2007 Evolutionary dynamics of transposable elements in the short-tailed opossum Monodelphis domestica. Genome Res 17**:** 992-1004.

Giangrande, P. H., W. Zhu, S. Schlisio, X. Sun, S. Mori *et al.*, 2004 A role for E2F6 in distinguishing G1/S- and G2/M-specific transcription. Genes Dev 18**:** 2941-2951.

Gifford, R., and M. Tristem, 2003 The evolution, distribution and diversity of endogenous retroviruses. Virus Genes 26**:** 291-315.

Gilbert, C., S. S. Hernandez, J. Flores-Benabib, E. N. Smith and C. Feschotte, 2012 Rampant horizontal transfer of SPIN transposons in squamate reptiles. Mol Biol Evol 29**:** 503-515.

Gilbert, C., S. Schaack, J. K. Pace, 2nd, P. J. Brindley and C. Feschotte, 2010 A role for host-parasite interactions in the horizontal transfer of transposons across phyla. Nature 464**:** 1347-1350.

Gilbert, N., and D. Labuda, 1999 CORE-SINEs: eukaryotic short interspersed retroposing elements with common sequence motifs. Proc Natl Acad Sci U S A 96**:** 2869-2874.

Gilbert, N., S. Lutz, T. A. Morrish and J. V. Moran, 2005 Multiple fates of L1 retrotransposition intermediates in cultured human cells. Mol Cell Biol 25**:** 7780-7795.

Gioti, A., A. A. Mushegian, R. Strandberg, J. E. Stajich and H. Johannesson, 2012 Unidirectional evolutionary transitions in fungal mating systems and the role of transposable elements. Mol Biol Evol 29**:** 3215-3226.

Gladyshev, E. A., and I. R. Arkhipova, 2007 Telomere-associated endonuclease-deficient Penelope-like retroelements in diverse eukaryotes. Proc Natl Acad Sci U S A 104**:** 9352-9357.

Gladyshev, E. A., and I. R. Arkhipova, 2009 A single-copy IS5-like transposon in the genome of a bdelloid rotifer. Mol Biol Evol 26**:** 1921-1929.

Gladyshev, E. A., and I. R. Arkhipova, 2010 Genome structure of bdelloid rotifers: shaped by asexuality or desiccation? J Hered 101 Suppl 1**:** S85-93.

Glockner, G., K. Szafranski, T. Winckler, T. Dingermann, M. A. Quail *et al.*, 2001 The complex repeats of Dictyostelium discoideum. Genome Res 11**:** 585-594.

Goodwin, T. J., M. I. Butler and R. T. Poulter, 2003 Cryptons: a group of tyrosine-recombinase-encoding DNA transposons from pathogenic fungi. Microbiology 149**:** 3099-3109.

Goodwin, T. J., and R. T. Poulter, 2001 The DIRS1 group of retrotransposons. Mol Biol Evol 18: 2067-2082.
Goodwin, T. J., and R. T. Poulter, 2002 A group of deuterostome Ty3/ gypsy-like retrotransposons with Ty1/ copia-like pol-domain orders. Mol Genet Genomics 267: 481-491.
Goodwin, T. J., and R. T. Poulter, 2004 A new group of tyrosine recombinase-encoding retrotransposons. Mol Biol Evol 21: 746-759.
Grahn, R. A., T. A. Rinehart, M. A. Cantrell and H. A. Wichman, 2005 Extinction of LINE-1 activity coincident with a major mammalian radiation in rodents. Cytogenet Genome Res 110: 407-415.
Gray, Y. H., 2000 It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. Trends Genet 16: 461-468.
Gray, Y. H., M. M. Tanaka and J. A. Sved, 1996 P-element-induced recombination in Drosophila melanogaster: hybrid element insertion. Genetics 144: 1601-1610.
Greenblatt, I. M., and R. A. Brink, 1962 Twin Mutations in Medium Variegated Pericarp Maize. Genetics 47: 489-501.
Gregory, T. R., 2001a The bigger the C-value, the larger the cell: genome size and red blood cell size in vertebrates. Blood Cells Mol Dis 27: 830-843.
Gregory, T. R., 2001b Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol Rev Camb Philos Soc 76: 65-101.
Gupta, S., A. Gallavotti, G. A. Stryker, R. J. Schmidt and S. K. Lal, 2005 A novel class of Helitron-related transposable elements in maize contain portions of multiple pseudogenes. Plant Mol Biol 57: 115-127.
Hamilton, A., O. Voinnet, L. Chappell and D. Baulcombe, 2002 Two classes of short interfering RNA in RNA silencing. EMBO J 21: 4671-4679.
Han, G. Z., and M. Worobey, 2012 An endogenous foamy-like viral element in the coelacanth genome. PLoS Pathog 8: e1002790.
Haney, R. A., and M. E. Feder, 2009 Contrasting patterns of transposable element insertions in Drosophila heat-shock promoters. PLoS One 4: e8486.
Hanson, S. J., C. P. Stelzer, D. B. Welch and J. M. Logsdon, Jr., 2013 Comparative transcriptome analysis of obligately asexual and cyclically sexual rotifers reveals genes with putative functions in sexual reproduction, dormancy, and asexual egg production. BMC Genomics 14: 412.
Haren, L., B. Ton-Hoang and M. Chandler, 1999 Integrating DNA: transposases and retroviral integrases. Annu Rev Microbiol 53: 245-281.
Heath, B. D., R. D. Butcher, W. G. Whitfield and S. F. Hubbard, 1999 Horizontal transfer of Wolbachia between phylogenetically distant insect species by a naturally occurring mechanism. Curr Biol 9: 313-316.
Heidmann, O., C. Vernochet, A. Dupressoir and T. Heidmann, 2009 Identification of an endogenous retroviral envelope gene with fusogenic activity and placenta-specific expression in the rabbit: a new "syncytin" in a third order of mammals. Retrovirology 6: 107.
Hellen, E. H., and J. F. Brookfield, 2013a Alu elements in primates are preferentially lost from areas of high GC content. PeerJ 1: e78.
Hellen, E. H., and J. F. Brookfield, 2013b Transposable element invasions. Mob Genet Elements 3: e23920.
Heredia, F., E. L. Loreto and V. L. Valente, 2004 Complex evolution of gypsy in Drosophilid species. Mol Biol Evol 21: 1831-1842.

Hickey, D. A., 1982 Selfish DNA: a sexually-transmitted nuclear parasite. Genetics 101: 519-531.
Hickman, A. B., Z. N. Perez, L. Zhou, P. Musingarimi, R. Ghirlando *et al.*, 2005 Molecular architecture of a eukaryotic DNA transposase. Nat Struct Mol Biol 12: 715-721.
Hiom, K., M. Melek and M. Gellert, 1998 DNA transposition by the RAG1 and RAG2 proteins: a possible source of oncogenic translocations. Cell 94: 463-470.
Hollister, J. D., and B. S. Gaut, 2009 Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res 19: 1419-1428.
Horie, M., T. Honda, Y. Suzuki, Y. Kobayashi, T. Daito *et al.*, 2010 Endogenous non-retroviral RNA virus elements in mammalian genomes. Nature 463: 84-87.
Hotopp, J. C. D., 2011 Horizontal gene transfer between bacteria and animals. Trends in Genetics 27: 157-163.
Howe, K., M. D. Clark, C. F. Torroja, J. Torrance, C. Berthelot *et al.*, 2013 The zebrafish reference genome sequence and its relationship to the human genome. Nature 496: 498-503.
Huang, Y. T., Y. C. Liaw, V. Y. Gorbatyuk and T. H. Huang, 2001 Backbone dynamics of Escherichia coli thioesterase/protease I: evidence of a flexible active-site environment for a serine protease. J Mol Biol 307: 1075-1090.
Hull, R., 2001 Classifying reverse transcribing elements: a proposal and a challenge to the ICTV. International Committee on Taxonomy of Viruses. Arch Virol 146: 2255-2261.
Hummel, T., and C. Klambt, 2008 P-element mutagenesis. Methods Mol Biol 420: 97-117.
Irimia, M., J. L. Rukov, D. Penny, J. Vinther, J. Garcia-Fernandez *et al.*, 2008 Origin of introns by 'intronization' of exonic sequences. Trends Genet 24: 378-381.
Ivancevic, A. M., A. M. Walsh, R. D. Kortschak and D. L. Adelson, 2013 Jumping the fine LINE between species: horizontal transfer of transposable elements in animals catalyses genome evolution. Bioessays 35: 1071-1082.
Ivics, Z., P. B. Hackett, R. H. Plasterk and Z. Izsvak, 1997 Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. Cell 91: 501-510.
Ivics, Z., C. D. Kaufman, H. Zayed, C. Miskey, O. Walisko *et al.*, 2004 The Sleeping Beauty transposable element: evolution, regulation and genetic applications. Curr Issues Mol Biol 6: 43-55.
Ivics, Z., M. A. Li, L. Mates, J. D. Boeke, A. Nagy *et al.*, 2009 Transposon-mediated genome manipulation in vertebrates. Nat Methods 6: 415-422.
Iwamatsu, T., 2004 Stages of normal development in the medaka Oryzias latipes. Mech Dev 121: 605-618.
Izsvak, Z., Z. Ivics, N. Shimoda, D. Mohn, H. Okamoto *et al.*, 1999 Short inverted-repeat transposable elements in teleost fish and implications for a mechanism of their amplification. J Mol Evol 48: 13-21.
Jaillon, O., J. M. Aury, F. Brunet, J. L. Petit, N. Stange-Thomann *et al.*, 2004 Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 431: 946-957.
Jern, P., G. O. Sperber and J. Blomberg, 2005 Use of endogenous retroviral sequences (ERVs) and structural markers for retroviral phylogenetic inference and taxonomy. Retrovirology 2: 50.

Jiang, F., M. Yang, W. Guo, X. Wang and L. Kang, 2012 Large-scale transcriptome analysis of retroelements in the migratory locust, Locusta migratoria. PLoS One 7: e40532.
Jiang, N., Z. Bao, X. Zhang, S. R. Eddy and S. R. Wessler, 2004 Pack-MULE transposable elements mediate gene evolution in plants. Nature 431: 569-573.
Jin, Y. K., and J. L. Bennetzen, 1994 Integration and nonrandom mutation of a plasma membrane proton ATPase gene fragment within the Bs1 retroelement of maize. Plant Cell 6: 1177-1186.
Johnson, L. J., 2007 The genome strikes back: The evolutionary importance of defence against mobile elements. Evolutionary Biology 34: 121-129.
Jones, F. C., M. G. Grabherr, Y. F. Chan, P. Russell, E. Mauceli *et al.*, 2012 The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484: 55-61.
Jordan, I. K., I. B. Rogozin, G. V. Glazko and E. V. Koonin, 2003 Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet 19: 68-72.
Jurka, J., W. Bao and K. K. Kojima, 2011 Families of transposable elements, population structure and the origin of species. Biol Direct 6: 44.
Jurka, J., V. V. Kapitonov, O. Kohany and M. V. Jurka, 2007 Repetitive sequences in complex genomes: Structure and evolution. Annual Review of Genomics and Human Genetics 8: 241-259.
Jurka, J., V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany *et al.*, 2005a Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110: 462-467.
Jurka, J., O. Kohany, A. Pavlicek, V. V. Kapitonov and M. V. Jurka, 2005b Clustering, duplication and chromosomal distribution of mouse SINE retrotransposons. Cytogenet Genome Res 110: 117-123.
Kajikawa, M., and N. Okada, 2002 LINEs mobilize SINEs in the eel through a shared 3' sequence. Cell 111: 433-444.
Kalendar, R., C. M. Vicient, O. Peleg, K. Anamthawat-Jonsson, A. Bolshoy *et al.*, 2004 Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. Genetics 166: 1437-1450.
Kallman, K. D., and S. Kazianis, 2006 The genus Xiphophorus in Mexico and central america. Zebrafish 3: 271-285.
Kamal, M., X. Xie and E. S. Lander, 2006 A large family of ancient repeat elements in the human genome is under strong selection. Proc Natl Acad Sci U S A 103: 2740-2745.
Kapitonov, V., and J. Jurka, 1996 The age of Alu subfamilies. J Mol Evol 42: 59-65.
Kapitonov, V. V., and J. Jurka, 2001 Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci U S A 98: 8714-8719.
Kapitonov, V. V., and J. Jurka, 2003 The esterase and PHD domains in CR1-like non-LTR retrotransposons. Mol Biol Evol 20: 38-46.
Kapitonov, V. V., and J. Jurka, 2005 RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLoS Biol 3: e181.
Kapitonov, V. V., and J. Jurka, 2006 Self-synthesizing DNA transposons in eukaryotes. Proc Natl Acad Sci U S A 103: 4540-4545.
Kapitonov, V. V., and J. Jurka, 2008 A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet 9: 411-412; author reply 414.
Kasahara, M., K. Naruse, S. Sasaki, Y. Nakatani, W. Qu *et al.*, 2007 The medaka draft genome and insights into vertebrate genome evolution. Nature 447: 714-719.

Kashkush, K., and V. Khasdan, 2007 Large-scale survey of cytosine methylation of retrotransposons and the impact of readout transcription from long terminal repeats on expression of adjacent rice genes. Genetics 177: 1975-1985.
Katinka, M. D., S. Duprat, E. Cornillot, G. Metenier, F. Thomarat *et al.*, 2001 Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. Nature 414: 450-453.
Katzourakis, A., R. J. Gifford, M. Tristem, M. T. Gilbert and O. G. Pybus, 2009 Macroevolution of complex retroviruses. Science 325: 1512.
Kawakami, K., 2007 Tol2: a versatile gene transfer vector in vertebrates. Genome Biol 8 Suppl 1: S7.
Kazazian, H. H., Jr., 2004 Mobile elements: drivers of genome evolution. Science 303: 1626-1632.
Kido, Y., M. Aono, T. Yamaki, K. Matsumoto, S. Murata *et al.*, 1991 Shaping and reshaping of salmonid genomes by amplification of tRNA-derived retroposons during evolution. Proc Natl Acad Sci U S A 88: 2326-2330.
Kidwell, M. G., J. F. Kidwell and J. A. Sved, 1977 Hybrid Dysgenesis in DROSOPHILA MELANOGASTER: A Syndrome of Aberrant Traits Including Mutation, Sterility and Male Recombination. Genetics 86: 813-833.
Kidwell, M. G., and D. Lisch, 1997 Transposable elements as sources of variation in animals and plants. Proc Natl Acad Sci U S A 94: 7704-7711.
Kidwell, M. G., and D. R. Lisch, 2000 Transposable elements and host genome evolution. Trends Ecol Evol 15: 95-99.
Kidwell, M. G., and J. B. Novy, 1979 Hybrid Dysgenesis in DROSOPHILA MELANOGASTER: Sterility Resulting from Gonadal Dysgenesis in the P-M System. Genetics 92: 1127-1140.
Kijima, T. E., and H. Innan, 2010 On the estimation of the insertion time of LTR retrotransposable elements. Mol Biol Evol 27: 896-904.
Kimmel, C. B., W. W. Ballard, S. R. Kimmel, B. Ullmann and T. F. Schilling, 1995 Stages of embryonic development of the zebrafish. Dev Dyn 203: 253-310.
Kimura, M., 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16: 111-120.
Kipling, D., and P. E. Warburton, 1997 Centromeres, CENP-B and Tigger too. Trends Genet 13: 141-145.
Kodama, K., S. Takagi and A. Koga, 2008 The Tol1 element of the medaka fish, a member of the hAT transposable element family, jumps in Caenorhabditis elegans. Heredity (Edinb) 101: 222-227.
Kojima, K. K., and J. Jurka, 2011 Crypton transposons: identification of new diverse families and ancient domestication events. Mob DNA 2: 12.
Kondo, N., N. Nikoh, N. Ijichi, M. Shimada and T. Fukatsu, 2002 Genome fragment of Wolbachia endosymbiont transferred to X chromosome of host insect. Proc Natl Acad Sci U S A 99: 14280-14285.
Kraaijeveld, K., 2010 Genome Size and Species Diversification. Evol Biol 37: 227-233.
Kraaijeveld, K., B. Zwanenburg, B. Hubert, C. Vieira, S. De Pater *et al.*, 2012 Transposon proliferation in an asexual parasitoid. Mol Ecol 21: 3898-3906.
Kramerov, D. A., and N. S. Vassetzky, 2005 Short retroposons in eukaryotic genomes. Int Rev Cytol 247: 165-221.
Krayev, A. S., D. A. Kramerov, K. G. Skryabin, A. P. Ryskov, A. A. Bayev *et al.*, 1980 The nucleotide sequence of the ubiquitous repetitive DNA sequence B1

complementary to the most abundant class of mouse fold-back RNA. Nucleic Acids Res 8: 1201-1215.
Kreahling, J., and B. R. Graveley, 2004 The origins and implications of Aluternative splicing. Trends Genet 20: 1-4.
Krull, M., J. Brosius and J. Schmitz, 2005 Alu-SINE exonization: en route to protein-coding function. Mol Biol Evol 22: 1702-1711.
Kumar, A., and J. L. Bennetzen, 1999 Plant retrotransposons. Annu Rev Genet 33: 479-532.
Kumar, A., and J. L. Bennetzen, 2000 Retrotransposons: central players in the structure, evolution and function of plant genomes. Trends Plant Sci 5: 509-510.
Labrador, M., M. Farre, F. Utzet and A. Fontdevila, 1999 Interspecific hybridization increases transposition rates of Osvaldo. Mol Biol Evol 16: 931-937.
Lai, J., Y. Li, J. Messing and H. K. Dooner, 2005 Gene movement by Helitron transposons contributes to the haplotype variability of maize. Proc Natl Acad Sci U S A 102: 9068-9073.
Lampert, K. P., and M. Schartl, 2008 The origin and evolution of a unisexual hybrid: Poecilia formosa. Philosophical Transactions of the Royal Society B-Biological Sciences 363: 2901-2909.
Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature 409: 860-921.
Le Rouzic, A., T. S. Boutin and P. Capy, 2007 Long-term evolution of transposable elements. Proc Natl Acad Sci U S A 104: 19375-19380.
Le Rouzic, A., and P. Capy, 2005 The first steps of transposable elements invasion: parasitic strategy vs. genetic drift. Genetics 169: 1033-1043.
Le Rouzic, A., and P. Capy, 2006 Population genetics models of competition between transposable element subfamilies. Genetics 174: 785-793.
Le Rouzic, A., and G. Deceliere, 2005 Models of the population genetics of transposable elements. Genet Res 85: 171-181.
Lecellier, C. H., and A. Saib, 2000 Foamy viruses: between retroviruses and pararetroviruses. Virology 271: 1-8.
Lerat, E., 2010 Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity (Edinb) 104: 520-533.
Levin, H. L., and J. V. Moran, 2011 Dynamic interactions between transposable elements and their hosts. Nat Rev Genet 12: 615-627.
Levy, A., S. Schwartz and G. Ast, 2010 Large-scale discovery of insertion hotspots and preferential integration sites of human transposed elements. Nucleic Acids Res 38: 1515-1530.
Li, W., and A. Godzik, 2006 Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658-1659.
Li, X., E. R. Burnight, A. L. Cooney, N. Malani, T. Brady *et al.*, 2013 piggyBac transposase tools for genome engineering. Proc Natl Acad Sci U S A 110: E2279-2287.
Lisch, D., 2013 How important are transposons for plant evolution? Nat Rev Genet 14: 49-61.
Lister, C., D. Jackson and C. Martin, 1993 Transposon-induced inversion in Antirrhinum modifies nivea gene expression to give a novel flower color pattern under the control of cycloidearadialis. Plant Cell 5: 1541-1553.

Lister, R., R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry *et al.*, 2008 Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133: 523-536.
Lister, R., M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon *et al.*, 2009 Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462: 315-322.
Litman, G. W., and J. P. Rast, 1996 The organization and structure of immunoglobulin and T-cell receptor genes in the most phylogenetically distant jawed vertebrates: evolutionary implications. Res Immunol 147: 226-233.
Litman, G. W., J. P. Rast and S. D. Fugmann, 2010 The origins of vertebrate adaptive immunity. Nat Rev Immunol 10: 543-553.
Liu, R., K. O. Koyanagi, S. Chen and Y. Kishima, 2012 Evolutionary force of AT-rich repeats to trap genomic and episomal DNAs into the rice genome: lessons from endogenous pararetrovirus. Plant J 72: 817-828.
Liu, Z., Y. Wang, Y. Shen, W. Guo, S. Hao *et al.*, 2004 Extensive alterations in DNA methylation and transcription in rice caused by introgression from Zizania latifolia. Plant Mol Biol 54: 571-582.
Llorens, C., and I. Marin, 2001 A mammalian gene evolved from the integrase domain of an LTR retrotransposon. Mol Biol Evol 18: 1597-1600.
Llorens, C., A. Munoz-Pomer, L. Bernad, H. Botella and A. Moya, 2009 Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. Biol Direct 4: 41.
Lorenzi, H. A., G. Robledo and M. J. Levin, 2006 The VIPER elements of trypanosomes constitute a novel group of tyrosine recombinase-enconding retrotransposons. Mol Biochem Parasitol 145: 184-194.
Lovsin, N., F. Gubensek and D. Kordi, 2001 Evolutionary dynamics in a novel L2 clade of non-LTR retrotransposons in Deuterostomia. Mol Biol Evol 18: 2213-2224.
Lyon, M. F., 2003 The Lyon and the LINE hypothesis. Semin Cell Dev Biol 14: 313-318.
Macfarlan, T., S. Kutney, B. Altman, R. Montross, J. Yu *et al.*, 2005 Human THAP7 is a chromatin-associated, histone tail-binding protein that represses transcription via recruitment of HDAC3 and nuclear hormone receptor corepressor. J Biol Chem 280: 7346-7358.
Makalowski, W., 2000 Genomic scrap yard: how genomes utilize all that junk. Gene 259: 61-67.
Malik, H. S., 2012 Retroviruses push the envelope for mammalian placentation. Proc Natl Acad Sci U S A 109: 2184-2185.
Malik, H. S., W. D. Burke and T. H. Eickbush, 1999 The age and evolution of non-LTR retrotransposable elements. Mol Biol Evol 16: 793-805.
Malik, H. S., and T. H. Eickbush, 1999 Modular evolution of the integrase domain in the Ty3/Gypsy class of LTR retrotransposons. J Virol 73: 5186-5190.
Malik, H. S., and T. H. Eickbush, 2001 Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. Genome Res 11: 1187-1197.
Malik, H. S., and S. Henikoff, 2005 Positive selection of Iris, a retroviral envelope-derived host gene in Drosophila melanogaster. PLoS Genet 1: e44.
Malone, C. D., and G. J. Hannon, 2009 Small RNAs as guardians of the genome. Cell 136: 656-668.
Mandal, A. K., R. Pandey, V. Jha and M. Mukerji, 2013 Transcriptome-wide expansion of non-coding regulatory switches: evidence from co-occurrence of Alu exonization, antisense and editing. Nucleic Acids Res 41: 2121-2137.

Mank, J. E., D. E. L. Promislow and J. C. Avise, 2006 Evolution of alternative sex-determining mechanisms in teleost fishes. Biological Journal of the Linnean Society 87: 83-93.
Marco, A., and I. Marin, 2009 CGIN1: a retroviral contribution to mammalian genomes. Mol Biol Evol 26: 2167-2170.
Marin, I., 2010 GIN transposons: genetic elements linking retrotransposons and genes. Mol Biol Evol 27: 1903-1911.
Mason, J. M., R. C. Frydrychova and H. Biessmann, 2008 Drosophila telomeres: an exception providing new insights. Bioessays 30: 25-37.
Matveev, V., and N. Okada, 2009 Retroposons of salmonoid fishes (Actinopterygii: Salmonoidei) and their evolution. Gene 434: 16-28.
Matzke, M. A., and A. J. Matzke, 1998a Epigenetic silencing of plant transgenes as a consequence of diverse cellular defence responses. Cell Mol Life Sci 54: 94-103.
Matzke, M. A., and A. J. Matzke, 1998b Gene silencing in plants: relevance for genome evolution and the acquisition of genomic methylation patterns. Novartis Found Symp 214: 168-180; discussion 181-166.
McClintock, B., 1951 Chromosome organization and genic expression. Cold Spring Harb Symp Quant Biol 16: 13-47.
McClintock, B., 1984 The significance of responses of the genome to challenge. Science 226: 792-801.
Medstrand, P., L. N. van de Lagemaat, C. A. Dunn, J. R. Landry, D. Svenback *et al.*, 2005 Impact of transposable elements on the evolution of mammalian gene regulation. Cytogenet Genome Res 110: 342-352.
Mejlumian, L., A. Pelisson, A. Bucheton and C. Terzian, 2002 Comparative and functional studies of Drosophila species invasion by the gypsy endogenous retrovirus. Genetics 160: 201-209.
Meng, Q., K. Chen, L. Ma, S. Hu and J. Yu, 2011 A systematic identification of Kolobok superfamily transposons in Trichomonas vaginalis and sequence analysis on related transposases. J Genet Genomics 38: 63-70.
Metcalfe, C. J., K. V. Bulazel, G. C. Ferreri, E. Schroeder-Reiter, G. Wanner *et al.*, 2007 Genomic instability within centromeres of interspecific marsupial hybrids. Genetics 177: 2507-2517.
Metcalfe, C. J., and D. Casane, 2013 Accommodating the load: The transposable element content of very large genomes. Mob Genet Elements 3: e24775.
Metcalfe, C. J., J. Filee, I. Germon, J. Joss and D. Casane, 2012 Evolution of the Australian lungfish (Neoceratodus forsteri) genome: a major role for CR1 and L2 LINE elements. Mol Biol Evol 29: 3529-3539.
Miki, Y., I. Nishisho, A. Horii, Y. Miyoshi, J. Utsunomiya *et al.*, 1992 Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. Cancer Res 52: 643-645.
Miller, W. J., S. Hagemann, E. Reiter and W. Pinsker, 1992 P-element homologous sequences are tandemly repeated in the genome of Drosophila guanche. Proc Natl Acad Sci U S A 89: 4018-4022.
Miskey, C., B. Papp, L. Mates, L. Sinzelle, H. Keller *et al.*, 2007 The ancient mariner sails again: transposition of the human Hsmar1 element by a reconstructed transposase and activities of the SETMAR protein on transposon ends. Mol Cell Biol 27: 4589-4600.
Miyao, A., K. Tanaka, K. Murata, H. Sawaki, S. Takeda *et al.*, 2003 Target site specificity of the Tos17 retrotransposon shows a preference for insertion within genes and

against insertion in retrotransposon-rich regions of the genome. Plant Cell 15: 1771-1780.
Moran, J. V., S. E. Holmes, T. P. Naas, R. J. DeBerardinis, J. D. Boeke *et al.*, 1996 High frequency retrotransposition in cultured mammalian cells. Cell 87: 917-927.
Morgan, G. T., 1995 Identification in the human genome of mobile elements spread by DNA-mediated transposition. J Mol Biol 254: 1-5.
Morgante, M., S. Brunner, G. Pea, K. Fengler, A. Zuccolo *et al.*, 2005 Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat Genet 37: 997-1002.
Murakami, M., and H. Fujitani, 1998 Characterization of repetitive DNA sequences carrying 5S rDNA of the triploid ginbuna (Japanese silver crucian carp, Carassius auratus langsdorfi). Genes Genet Syst 73: 9-20.
Muszewska, A., K. Steczkiewicz and K. Ginalski, 2013 DIRS and Ngaro Retrotransposons in Fungi. PLoS One 8: e76319.
Nakanishi, A., N. Kobayashi, A. Suzuki-Hirano, H. Nishihara, T. Sasaki *et al.*, 2012 A SINE-derived element constitutes a unique modular enhancer for mammalian diencephalic Fgf8. PLoS One 7: e43785.
Near, T. J., R. I. Eytan, A. Dornburg, K. L. Kuhn, J. A. Moore *et al.*, 2012 Resolution of ray-finned fish phylogeny and timing of diversification. Proc Natl Acad Sci U S A 109: 13698-13703.
Nekrutenko, A., and W. H. Li, 2001 Transposable elements are found in a large number of human protein-coding genes. Trends Genet 17: 619-621.
Nelson, J. S., 2006 *Fishes of the world 4th edition*. Hoboken.
Newman, J. C., A. D. Bailey, H. Y. Fan, T. Pavelitz and A. M. Weiner, 2008 An abundant evolutionarily conserved CSB-PiggyBac fusion protein expressed in Cockayne syndrome. PLoS Genet 4: e1000031.
Nikaido, M., H. Noguchi, H. Nishihara, A. Toyoda, Y. Suzuki *et al.*, 2013 Coelacanth genomes reveal signatures for evolutionary transition from water to land. Genome Res 23: 1740-1748.
Nishihara, H., Y. Terai and N. Okada, 2002 Characterization of novel Alu- and tRNA-related SINEs from the tree shrew and evolutionary implications of their origins. Mol Biol Evol 19: 1964-1972.
Novikova, O. S., and A. G. Blinov, 2009 [Origin, evolution, and distribution of different groups of non-LTR retrotransposons among eukaryotes]. Genetika 45: 149-159.
Nuzhdin, S. V., E. G. Pasyukova and T. F. Mackay, 1996 Positive association between copia transposition rate and copy number in Drosophila melanogaster. Proc Biol Sci 263: 823-831.
O'Donnell, K. A., W. An, C. T. Schrum, S. J. Wheelan and J. D. Boeke, 2013 Controlled insertional mutagenesis using a LINE-1 (ORFeus) gene-trap mouse model. Proc Natl Acad Sci U S A 110: E2706-2713.
Ogiwara, I., M. Miya, K. Ohshima and N. Okada, 2002 V-SINEs: a new superfamily of vertebrate SINEs that are widespread in vertebrate genomes and retain a strongly conserved segment within each repetitive unit. Genome Res 12: 316-324.
Ohno, S., 1972 So much "junk" DNA in our genome. Brookhaven Symp Biol 23: 366-370.
Oliver, K. R., and W. K. Greene, 2009 Transposable elements: powerful facilitators of evolution. Bioessays 31: 703-714.

Ono, R., K. Nakamura, K. Inoue, M. Naruse, T. Usami *et al.*, 2006 Deletion of Peg10, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality. Nat Genet 38: 101-106.
Orgel, L. E., and F. H. Crick, 1980 Selfish DNA: the ultimate parasite. Nature 284: 604-607.
Pallavicini, A., A. Canapa, M. Barucca, J. Alf Ldi, M. A. Biscotti *et al.*, 2013 Analysis of the transcriptome of the Indonesian coelacanth Latimeria menadoensis. BMC Genomics 14: 538.
Patel, M. R., M. Emerman and H. S. Malik, 2011 Paleovirology - ghosts and gifts of viruses past. Curr Opin Virol 1: 304-309.
Perez, J., E. Garcia-Vazquez and P. Moran, 1999 Physical distribution of SINE elements in the chromosomes of Atlantic salmon and rainbow trout. Heredity (Edinb) 83 (Pt 5): 575-579.
Petrov, D. A., 2001 Evolution of genome size: new approaches to an old problem. Trends Genet 17: 23-28.
Picard, G., J. C. Bregliano, A. Bucheton, J. M. Lavige, A. Pelisson *et al.*, 1978 Non-mendelian female sterility and hybrid dysgenesis in Drosophila melanogaster. Genet Res 32: 275-287.
Piskurek, O., and D. J. Jackson, 2011 Tracking the ancestry of a deeply conserved eumetazoan SINE domain. Mol Biol Evol 28: 2727-2730.
Piskurek, O., H. Nishihara and N. Okada, 2009 The evolution of two partner LINE/SINE families and a full-length chromodomain-containing Ty3/Gypsy LTR element in the first reptilian genome of Anolis carolinensis. Gene 441: 111-118.
Plasterk, R. H., Z. Izsvak and Z. Ivics, 1999 Resident aliens: the Tc1/mariner superfamily of transposable elements. Trends Genet 15: 326-332.
Polak, P., and E. Domany, 2006 Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. BMC Genomics 7: 133.
Poulter, R. T., and T. J. Goodwin, 2005 DIRS-1 and the other tyrosine recombinase retrotransposons. Cytogenet Genome Res 110: 575-588.
Poulter, R. T., T. J. Goodwin and M. I. Butler, 2003 Vertebrate helentrons and other novel Helitrons. Gene 313: 201-212.
Prak, E. T., and H. H. Kazazian, Jr., 2000 Mobile elements and the human genome. Nat Rev Genet 1: 134-144.
Preston, C. R., and W. R. Engels, 1996 P-element-induced male recombination and gene conversion in Drosophila. Genetics 144: 1611-1622.
Preston, C. R., J. A. Sved and W. R. Engels, 1996 Flanking duplications and deletions associated with P-induced male recombination in Drosophila. Genetics 144: 1623-1638.
Pritham, E. J., 2009 Transposable elements and factors influencing their success in eukaryotes. J Hered 100: 648-655.
Pritham, E. J., T. Putliwala and C. Feschotte, 2007 Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene 390: 3-17.
Quesneville, H., D. Nouaud and D. Anxolabehere, 2003 Detection of new transposable element families in Drosophila melanogaster and Anopheles gambiae genomes. J Mol Evol 57 Suppl 1: S50-59.
Rebollo, R., B. Horard, B. Hubert and C. Vieira, 2010 Jumping genes and epigenetics: Towards new species. Gene 454: 1-7.

Rebollo, R., M. T. Romanish and D. L. Mager, 2012 Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annu Rev Genet 46: 21-42.
Rethwilm, A., 2010 Molecular biology of foamy viruses. Med Microbiol Immunol 199: 197-207.
Rieseberg, L. H., 2001 Chromosomal rearrangements and speciation. Trends in Ecology & Evolution 16: 351-358.
Rigal, M., and O. Mathieu, 2011 A "mille-feuille" of silencing: epigenetic control of transposable elements. Biochim Biophys Acta 1809: 452-458.
Robertson, H. M., and K. L. Zumpano, 1997 Molecular evolution of an ancient mariner transposon, Hsmar1, in the human genome. Gene 205: 203-217.
Rosenfeld, M. R., J. G. Eichen, D. F. Wade, J. B. Posner and J. Dalmau, 2001 Molecular and clinical diversity in paraneoplastic immunity to Ma proteins. Ann Neurol 50: 339-348.
Rossetti, L. C., A. Goodeve, I. B. Larripa and C. D. De Brasi, 2004 Homeologous recombination between AluSx-sequences as a cause of hemophilia. Hum Mutat 24: 440.
Roulin, A., B. Piegu, R. A. Wing and O. Panaud, 2008 Evidence of multiple horizontal transfers of the long terminal repeat retrotransposon RIRE1 within the genus Oryza. Plant J 53: 950-959.
Rowold, D. J., and R. J. Herrera, 2000 Alu elements and the human genome. Genetica 108: 57-72.
Ruiz-Perez, V. L., F. J. Murillo and S. Torres-Martinez, 1996 Prt1, an unusual retrotransposon-like sequence in the fungus Phycomyces blakesleeanus. Mol Gen Genet 253: 324-333.
Sakamoto, K., and N. Okada, 1985 Rodent type 2 Alu family, rat identifier sequence, rabbit C family, and bovine or goat 73-bp repeat may have evolved from tRNA genes. J Mol Evol 22: 134-140.
Salvi, S., R. Tuberosa, E. Chiapparino, M. Maccaferri, S. Veillet *et al.*, 2002 Toward positional cloning of Vgt1, a QTL controlling the transition from the vegetative to the reproductive phase in maize. Plant Mol Biol 48: 601-613.
Santangelo, A. M., F. S. de Souza, L. F. Franchini, V. F. Bumaschny, M. J. Low *et al.*, 2007 Ancient exaptation of a CORE-SINE retroposon into a highly conserved mammalian neuronal enhancer of the proopiomelanocortin gene. PLoS Genet 3: 1813-1826.
Sarkar, A., C. Sim, Y. S. Hong, J. R. Hogan, M. J. Fraser *et al.*, 2003 Molecular evolutionary analysis of the widespread piggyBac transposon family and related "domesticated" sequences. Mol Genet Genomics 270: 173-180.
Sasaki, T., H. Nishihara, M. Hirakawa, K. Fujimura, M. Tanaka *et al.*, 2008 Possible involvement of SINEs in mammalian-specific brain formation. Proc Natl Acad Sci U S A 105: 4220-4225.
Schaack, S., C. Gilbert and C. Feschotte, 2010 Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends Ecol Evol 25: 537-546.
Schartl, M., U. Hornung, H. Gutbrod, J. N. Volff and J. Wittbrodt, 1999 Melanoma loss-of-function mutants in Xiphophorus caused by Xmrk-oncogene deletion and gene disruption by a transposable element. Genetics 153: 1385-1394.

Schartl, M., R. B. Walter, Y. Shen, T. Garcia, J. Catchen *et al.*, 2013 The genome of the platyfish, Xiphophorus maculatus, provides insights into evolutionary adaptation and several complex traits. Nat Genet 45**:** 567-572.

Schmidt, D., P. C. Schwalie, M. D. Wilson, B. Ballester, A. Goncalves *et al.*, 2012 Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell 148**:** 335-348.

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

Schonbach, C., 2004 From masking repeats to identifying functional repeats in the mouse transcriptome. Brief Bioinform 5**:** 107-117.

Schuller, M., D. Jenne and R. Voltz, 2005 The human PNMA family: novel neuronal proteins implicated in paraneoplastic neurological disease. J Neuroimmunol 169**:** 172-176.

Sekita, Y., H. Wagatsuma, K. Nakamura, R. Ono, M. Kagami *et al.*, 2008 Role of retrotransposon-derived imprinted gene, Rtl1, in the feto-maternal interface of mouse placenta. Nat Genet 40**:** 243-248.

Sela, N., B. Mersch, A. Hotz-Wagenblatt and G. Ast, 2010 Characteristics of transposable element exonization within human and mouse. PLoS One 5**:** e10907.

Selker, E. U., 2002 Repeat-induced gene silencing in fungi. Homology Effects 46**:** 439-450.

Shapiro, J. A., 1969 Mutations caused by the insertion of genetic material into the galactose operon of Escherichia coli. J Mol Biol 40**:** 93-105.

Shedlock, A. M., K. Takahashi and N. Okada, 2004 SINEs of speciation: tracking lineages with retroposons. Trends Ecol Evol 19**:** 545-553.

Shepard, S., M. McCreary and A. Fedorov, 2009 The peculiarities of large intron splicing in animals. PLoS One 4**:** e7853.

Shimoda, N., M. Chevrette, M. Ekker, Y. Kikuchi, Y. Hotta *et al.*, 1996 Mermaid: a family of short interspersed repetitive elements widespread in vertebrates. Biochem Biophys Res Commun 220**:** 226-232.

Silva, J. C., E. L. Loreto and J. B. Clark, 2004 Factors that affect the horizontal transfer of transposable elements. Curr Issues Mol Biol 6**:** 57-71.

Sinzelle, L., Z. Izsvak and Z. Ivics, 2009 Molecular domestication of transposable elements: from detrimental parasites to useful host genes. Cell Mol Life Sci 66**:** 1073-1093.

Sites, J. W., and C. Moritz, 1987 Chromosomal Evolution and Speciation Revisited. Systematic Zoology 36**:** 153-174.

Slatkin, M., 1985 Genetic differentiation of transposable elements under mutation and unbiased gene conversion. Genetics 110**:** 145-158.

Slotkin, R. K., and R. Martienssen, 2007 Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet 8**:** 272-285.

Smit, A. F., 1999 Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev 9**:** 657-663.

Smit, A. F., and A. D. Riggs, 1995 MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. Nucleic Acids Res 23**:** 98-102.

Smith, J. J., K. Sumiyama and C. T. Amemiya, 2012 A living fossil in the genome of a living fossil: Harbinger transposons in the coelacanth genome. Mol Biol Evol 29**:** 985-993.

Star, B., A. J. Nederbragt, S. Jentoft, U. Grimholt, M. Malmstrom *et al.*, 2011 The genome sequence of Atlantic cod reveals a unique immune system. Nature 477**:** 207-210.

Steinemann, M., and S. Steinemann, 1991 Preferential Y chromosomal location of TRIM, a novel transposable element of Drosophila miranda, obscura group. Chromosoma 101: 169-179.
Sun, C., J. R. Lopez Arriaza and R. L. Mueller, 2012a Slow DNA loss in the gigantic genomes of salamanders. Genome Biol Evol 4: 1340-1348.
Sun, C., D. B. Shepard, R. A. Chong, J. Lopez Arriaza, K. Hall *et al.*, 2012b LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders. Genome Biol Evol 4: 168-183.
Svoboda, Y. H., M. K. Robson and J. A. Sved, 1995 P-element-induced male recombination can be produced in Drosophila melanogaster by combining end-deficient elements in trans. Genetics 139: 1601-1610.
Syvanen, M., 2012 Evolutionary implications of horizontal gene transfer. Annu Rev Genet 46: 341-358.
Tabara, H., M. Sarkissian, W. G. Kelly, J. Fleenor, A. Grishok *et al.*, 1999 The rde-1 gene, RNA interference, and transposon silencing in C. elegans. Cell 99: 123-132.
Takahashi, K., M. Nishida, M. Yuma and N. Okada, 2001 Retroposition of the AFC family of SINEs (short interspersed repetitive elements) before and during the adaptive radiation of cichlid fishes in Lake Malawi and related inferences about phylogeny. J Mol Evol 53: 496-507.
Takasaki, N., S. Murata, M. Saitoh, T. Kobayashi, L. Park *et al.*, 1994 Species-specific amplification of tRNA-derived short interspersed repetitive elements (SINEs) by retroposition: a process of parasitization of entire genomes during the evolution of salmonids. Proc Natl Acad Sci U S A 91: 10153-10157.
Thomas, C. A., 1971 Genetic Organization of Chromosomes. Annual Review of Genetics 5: 237-&.
Thomas, J., S. Schaack and E. J. Pritham, 2010 Pervasive horizontal transfer of rolling-circle transposons among animals. Genome Biol Evol 2: 656-664.
Thomas, J., M. Sorourian, D. Ray, R. J. Baker and E. J. Pritham, 2011 The limited distribution of Helitrons to vesper bats supports horizontal transfer. Gene 474: 52-58.
Thornburg, B. G., V. Gotea and W. Makalowski, 2006 Transposable elements as a significant source of transcription regulating signals. Gene 365: 104-110.
Tian, Z., C. Rizzon, J. Du, L. Zhu, J. L. Bennetzen *et al.*, 2009 Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? Genome Res 19: 2221-2230.
Tollis, M., and S. Boissinot, 2011 The transposable element profile of the anolis genome: How a lizard can provide insights into the evolution of vertebrate genome size and structure. Mob Genet Elements 1: 107-111.
Tristem, M., P. Kabat, L. Lieberman, S. Linde, A. Karpas *et al.*, 1996 Characterization of a novel murine leukemia virus-related subgroup within mammals. J Virol 70: 8241-8246.
Tsukahara, S., A. Kobayashi, A. Kawabe, O. Mathieu, A. Miura *et al.*, 2009 Bursts of retrotransposition reproduced in Arabidopsis. Nature 461: 423-426.
Tu, Z., 1997 Three novel families of miniature inverted-repeat transposable elements are associated with genes of the yellow fever mosquito, Aedes aegypti. Proc Natl Acad Sci U S A 94: 7475-7480.
Tu, Z., J. Isoe and J. A. Guzova, 1998 Structural, genomic, and phylogenetic analysis of Lian, a novel family of non-LTR retrotransposons in the yellow fever mosquito, Aedes aegypti. Mol Biol Evol 15: 837-853.

Tudor, M., M. Lobocka, M. Goodell, J. Pettitt and K. O'Hare, 1992 The pogo transposable element family of Drosophila melanogaster. Mol Gen Genet 232: 126-134.
Ullu, E., and C. Tschudi, 1984 Alu sequences are processed 7SL RNA genes. Nature 312: 171-172.
Vela, D., A. Fontdevila, C. Vieira and M. P. Garcia Guerreiro, 2014 A genome-wide survey of genetic instability by transposition in Drosophila hybrids. PLoS One 9: e88992.
Venkatesh, B., A. P. Lee, V. Ravi, A. K. Maurya, M. M. Lian *et al.*, 2014 Elephant shark genome provides unique insights into gnathostome evolution. Nature 505: 174-179.
Vidal, F., E. Farssac, J. Tusell, L. Puig and D. Gallardo, 2002 First molecular characterization of an unequal homologous alu-mediated recombination event responsible for hemophilia. Thromb Haemost 88: 12-16.
Vieira, C., and C. Biemont, 1996 Selection against transposable elements in D. simulans and D. melanogaster. Genet Res 68: 9-15.
Villasante, A., B. de Pablos, M. Mendez-Lago and J. P. Abad, 2008 Telomere maintenance in Drosophila: rapid transposon evolution at chromosome ends. Cell Cycle 7: 2134-2138.
Vitte, C., and O. Panaud, 2003 Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice Oryza sativa L. Mol Biol Evol 20: 528-540.
Vitte, C., and O. Panaud, 2005 LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. Cytogenet Genome Res 110: 91-107.
Vitte, C., O. Panaud and H. Quesneville, 2007 LTR retrotransposons in rice (Oryza sativa, L.): recent burst amplifications followed by rapid DNA loss. BMC Genomics 8: 218.
Volff, J. N., 2006 Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays 28: 913-922.
Volff, J. N., 2009 Cellular genes derived from Gypsy/Ty3 retrotransposons in mammalian genomes. Ann N Y Acad Sci 1178: 233-243.
Volff, J. N., L. Bouneau, C. Ozouf-Costaz and C. Fischer, 2003 Diversity of retrotransposable elements in compact pufferfish genomes. Trends Genet 19: 674-678.
Volff, J. N., C. Korting, A. Froschauer, K. Sweeney and M. Schartl, 2001a Non-LTR retrotransposons encoding a restriction enzyme-like endonuclease in vertebrates. J Mol Evol 52: 351-360.
Volff, J. N., C. Korting, A. Meyer and M. Schartl, 2001b Evolution and discontinuous distribution of Rex3 retrotransposons in fish. Mol Biol Evol 18: 427-431.
Volff, J. N., C. Korting and M. Schartl, 2000 Multiple lineages of the non-LTR retrotransposon Rex1 with varying success in invading fish genomes. Mol Biol Evol 17: 1673-1684.
Voytas, D. F., and J. D. Boeke, 1992 Yeast retrotransposon revealed. Nature 358: 717.
Wallau, G. L., M. F. Ortiz and E. L. Loreto, 2012 Horizontal transposon transfer in eukarya: detection, bias, and perspectives. Genome Biol Evol 4: 689-699.
Walsh, A. M., R. D. Kortschak, M. G. Gardner, T. Bertozzi and D. L. Adelson, 2013 Widespread horizontal transfer of retrotransposons. Proc Natl Acad Sci U S A 110: 1012-1016.
Wang, Z., J. Pascual-Anaya, A. Zadissa, W. Li, Y. Niimura *et al.*, 2013 The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. Nat Genet 45: 701-706.

Webster, K. E., M. K. O'Bryan, S. Fletcher, P. E. Crewther, U. Aapola *et al.*, 2005 Meiotic and epigenetic defects in Dnmt3L-knockout mouse spermatogenesis. Proc Natl Acad Sci U S A 102**:** 4068-4073.

Weil, C. F., and S. R. Wessler, 1993 Molecular evidence that chromosome breakage by Ds elements is caused by aberrant transposition. Plant Cell 5**:** 515-522.

Weiner, A. M., P. L. Deininger and A. Efstratiadis, 1986 Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu Rev Biochem 55**:** 631-661.

Wicker, T., J. S. Robertson, S. R. Schulze, F. A. Feltus, V. Magrini *et al.*, 2005 The repetitive landscape of the chicken genome. Genome Res 15**:** 126-136.

Wicker, T., F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy *et al.*, 2007 A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8**:** 973-982.

Witte, C. P., Q. H. Le, T. Bureau and A. Kumar, 2001 Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. Proc Natl Acad Sci U S A 98**:** 13778-13783.

Wu, S. C., Y. J. Meir, C. J. Coates, A. M. Handler, P. Pelczar *et al.*, 2006 piggyBac is a flexible and highly active transposon as compared to sleeping beauty, Tol2, and Mos1 in mammalian cells. Proc Natl Acad Sci U S A 103**:** 15008-15013.

Xiao, H., N. Jiang, E. Schaffner, E. J. Stockinger and E. van der Knaap, 2008 A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. Science 319**:** 1527-1530.

Xiong, Y., and T. H. Eickbush, 1990 Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J 9**:** 3353-3362.

Xu, P., G. Widmer, Y. Wang, L. S. Ozaki, J. M. Alves *et al.*, 2004 The genome of Cryptosporidium hominis. Nature 431**:** 1107-1112.

Yang, L., and J. L. Bennetzen, 2009 Distribution, diversity, evolution, and survival of Helitrons in the maize genome. Proc Natl Acad Sci U S A 106**:** 19922-19927.

Yoder, J. A., C. P. Walsh and T. H. Bestor, 1997 Cytosine methylation and the ecology of intragenomic parasites. Trends Genet 13**:** 335-340.

Yuan, Y. W., and S. R. Wessler, 2011 The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. Proc Natl Acad Sci U S A 108**:** 7884-7889.

Zhan, X., S. Pan, J. Wang, A. Dixon, J. He *et al.*, 2013 Peregrine and saker falcon genome sequences provide insights into evolution of a predatory lifestyle. Nat Genet 45**:** 563-566.

Zhang, J., F. Zhang and T. Peterson, 2006 Transposition of reversed Ac element ends generates novel chimeric genes in maize. PLoS Genet 2**:** e164.

Zhang, X., C. Feschotte, Q. Zhang, N. Jiang, W. B. Eggleston *et al.*, 2001 P instability factor: an active maize transposon system associated with the amplification of Tourist-like MITEs and a new superfamily of transposases. Proc Natl Acad Sci U S A 98**:** 12572-12577.

Zilberman, D., and S. Henikoff, 2007 Genome-wide analysis of DNA methylation patterns. Development 134**:** 3959-3965.

ABBREVIATIONS

APE	Apurinic-apyrimidic endonuclease
CA	Capsid
CH	Chromodomain
ENV	Envelope
ERV	Endogenous Retrovirus
EVE	Endogenous Viral Elements
FV	Foamy Virus
H(G)T	Horizontal (Gene) Transfer
INT	Integrase
LARD	Large Retrotransposons derivatives
LINE	Long Interspersed Nuclear Element
LTR	Long Terminal Repeat
MA	Matrix proteins
Mb	Mega bases (bp: base pairs; kb: kilo bases; Gb: Giga bases)
MITE	Miniature Inverted Terminal Elements
My	Million years
Mya	Million years ago
NC	Nucleocapsid
nt	nucleotide
PBS	Primer Binding Sites
PPT	Polypurine Tracts
PR	Protease
REL	Restriction Enzyme-like endonuclease
RH	Ribonuclease H (RNase H)
RT	Reverse Transcriptase
RV	Retrovirus
SINE	Short Interspersed Nuclear Elements
TE	Transposable element
TIR	Terminal Inverted Repeat
TRIM	Terminal-repeat Retrotransposons in Miniature
TSD	Target Site Duplication

TABLE OF FIGURES

FIGURE 1: The central dogma **5**
FIGURE 2: Timeline showing important molecular biology and TE linked discoveries **6**
FIGURE 3: Comparison of prokaryotic and eukaryotic genomes **8**
FIGURE 4: C-value variation among the tree of life **10**
FIGURE 5: Universal classification of transposable elements **13**
FIGURE 6: Structure of the different superfamilies from the LTR retrotransposon orders **15**
FIGURE 7: Schematic representation of DIRS, Ngaro and Viper superfamilies **17**
FIGURE 8: Diversity and structure of the different LINE superfamilies **19**
FIGURE 9: Schematic organization of Penelope element **21**
FIGURE 10: Diversity and structure of DNA transposons from subclass I **23**
FIGURE 11: Helitron and Polinton structures, from subclass II **24**
FIGURE 12: Genomic rearrangements induced by transposable elements in a genic context **28**
FIGURE 13: Inversion and translocation are the two main large-scale chromosomal rearrangements **31**
FIGURE 14: Dynamics of a newly introduced mobile element from birth to death **38**
FIGURE 15: Defense mechanisms acting to restrain TE proliferation in host genomes **43**
FIGURE 16: Simplified vertebrate phylogeny that includes extinct and extant jawless lineages as well as extant jawed vertebrates **48**
FIGURE 17: Examples of genes derived from DNA transposon or LTR retrotransposon domains **50**
FIGURE 18: Simplified fish phylogeny including coelacanth and actinopterygians **56**
FIGURE 19: Protocol to annotate and analyze TEs in genomes **58**
FIGURE 20: Cover page of the Nature issue containing the coelacanth genome paper **60**
FIGURE 21: Zebrafish and cave fish repeat landscapes **62**
FIGURE 22: Transposable element superfamilies in the three *Xiphophorus* and the Amazon molly genomes **65**
FIGURE 23: Analysis of TE profiles in *X. maculatus* and *P. formosa* by crossing species-specific TE libraries **67**
FIGURE 24: Comparison of the repeat landscapes from the two African coelacanth assemblies **71**
FIGURE 25: Quantity of DNA transposons, LTR, LINE, SINE retrotransposons and unclassified elements in fish genomes **156**

FIGURE 26: Retroviridae diversity in vertebrates **158**
FIGURE 27: Phylogenetic reconstruction of retroviruses in vertebrates **161**
FIGURE 28: Expression analyses of *Gin-2* in fish using RT-qPCR experiments **166**
FIGURE 29: Expression of *Gin-2* during zebrafish embryogenesis determined by *in situ* hybridization **168**
FIGURE 30: Schematic comparison of TE diversity and content in vertebrate genomes **175**
FIGURE 31: Phylogenetic reconstruction of *Gin* genes and TE integrase domains **188**

LIST OF PUBLICATIONS

- Chalopin D, Naville M, Galiana D, Volff JN. Comparative analysis of transposable elements in fish highlights mobilome diversity and evolution in vertebrates. Submitted.

- Naville M*, Chalopin D*, Volff JN. Insertion polymorphism analysis reveals recent activity of transposable elements in the two extant species of coelacanths. Submitted (*equal contribution).

- McCaugh SE, Gross JB, Aken B, Blin M, Borowsky R, Chalopin D, Hinaux H, Jeffery W, Keene A, Ma L, et al. The cavefish genome reveals candidate genes for eye loss. Submitted to **Nature Genetics**.

- Tomaszkiewicz M, Chalopin D, Schartl M, Galiana D, Volff JN. A multicopy Y-chromosomal SGNH hydrolase gene expressed in the testis of the platyfish has been captured and mobilized by a *Helitron* transposon. **BMC genetics** in press.

- Berthelot C, Brunet F*, Chalopin D*, Juanchich A*, Bernard M, Noël B, Bento P, Dasilva C, Labadie K, Alberti A, et al. The rainbow trout genome: insights into evolution after genome duplication in vertebrates. **Nature Communication** in press (*equal contributions).

- Chen S, Zhang G, Shao C, Huang Q, Liu G, Zhang P, Song W, An N, Chalopin D, Volff JN, et al. 2014. Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. **Nature Genetics** 46:253-260.

- Forconi M, Chalopin D, Barucca M, Biscotti MA, De Moro G, Galiana D, Gerdol M, Pallavicini A, Canapa A, Olmo E, Volff JN. 2013. Transcriptional activity of transposable elements in coelacanth. **Journal of Experimental Zoology Part B** doi:10.1002/jez.b.22527 In press.

- Chalopin D, Fan S, Simakov O, Meyer A, Schartl M, Volff JN. 2013. Evolutionary active transposable elements in the genome of the coelacanth. 2013. **Journal of Experimental Zoology Part B** doi:10.1002/jez.b.22521 In Press.

- Amemiya CT, Aföldi J, Lee AP, Fan S, Brinkmann H, MacCallum I, Braasch I, Manousaki T, Schneider I, Rohner N, Organ C, Chalopin D, et al. 2013. The African coelacanth genome provides insights into tetrapod evolution. **Nature** 496:311-316.

- Schartl M, Walter RB, Shen Y, Garcia T, Catchen J, Amores A, Braasch I, Chalopin D, Volff JN, Lesch KP, et al. 2013. The genome of the platyfish, *Xiphophorus maculatus*, provides insights into evolutionary adaptation and several complex traits. **Nature Genetics** 45:567-572.

- Chalopin D, Tomaszkiewicz M, Galiana D, Volff JN. LTR retroelement-derived protein-coding genes and vertebrate evolution. G. Witzany (ed.), **Essential Agents of Life,** Springer Science+Business Media Dordrecht 2012.

- Chalopin D, Galiana D, Volff JN. 2012. Genetic innovation in vertebrates: Gypsy integrase (*GIN*) genes and other genes derived from transposable elements. **International Journal of Evolutionary Biology** 2012: 724519.

ANNEXES

1 ➔ The Southern platyfish genome (Schartl et al. 2013): Nat Genet, doi: 10.1038/ng.2604.

2 ➔ The coelacanth genome (Amemiya et al. 2013): Nature, doi: 10.1038/nature12027.

3 ➔ The tongue sole genome (Shen et al. 2014): Nat Genet, doi: 10.1038/ng.2890.

4 ➔ The rainbow trout genome (Berthelot et al. 2014): Nat commun, doi: 10.1038/ncomms4657

5 ➔ Rainbow trout analysis: TEs versus 4R ohnologs comparison using Kimura distances

6 ➔ Retroelements-derived genes (Chalopin et al. 2012): Int J Evol Biol, doi: 10.1155/2012/724519.

7 ➔ Y-chromosomal SGNH hydrolase gene (Tomaszkiewicz et al. 2014): BMC Genet.