POLYTECHNIC UNIVERSITY OF MARCHE

PhD Course in Life and Environmental Sciences

Curriculum: Biomolecular Sciences

Investigating the distribution and dynamics of transposable elements in genomes of non-model species: from molecular processes to population-level pressures

Supervisors:

Prof. Emiliano Trucchi

Prof. Marco Barucca

Ph.D. Candidate:

Lorena Ancona

XXXVI Cycle

2020-2023

Table of contents

General Introduction

Transposable elements, the dark matter of the genome

TE history timeline

Discovered in the 1940s, Barbara McClintock characterised transposable elements (TEs) as "controlling" elements due to their capacity to control gene expression.
Passionately dedicated to her work on maize, McClintock observed that chromosome breakage occurred at specific sites on maize chromosome 9, leading to variegated pigmentation in maize kernels. She identified this site as Dissociation (Ds) and demonstrated that Ds breakage was regulated by the presence of another site, called Activator (Ac) (McClintock, 1951). Consequently, the Dissociation (Ds) element emerged as the first transposable element discovered, with its transposition intricately regulated by the autonomous element "Activator" (Ac), which could also promote its own transposition.
Despite her election as a member of the National Academy of Sciences in 1944 and her historic role as the first woman President of the Genetics Society of America in 1945, McClintock encountered bewilderment and scepticism when she initially presented her pioneering results on transposition at the 1951 Cold Spring Harbor Symposium. Believing in her research, she persevered, continuing her studies by investigating another transposition system in maize, the Suppressor-Mutator (Spm) elements. Through this exploration, she uncovered the remarkable ability of certain autonomous elements to generate products with trans-regulatory activity on adjacent genes (McClintock, 1956). Recognition for her work would not come until 35 years later when she was among the first women to receive the Nobel Prize in Medicine in 1983.
In the following years, as molecular biology advanced, the investigation of transposable elements in yeast, *Drosophila*, and humans unveiled their mutagenic activity, definitively categorising them as “parasitic DNA”. It is only with the advent of genomics and whole-genome sequencing that it becomes evident that TEs are nearly ubiquitous components of eukaryotic genomes, ranging from 85% in maize, 20% in *Drosophila*, and 50% in humans. Nevertheless, the presence of numerous non-transposable fossil copies led to their designation as "junk DNA".

Early studies in functional genomics finally revealed the other side of TEs, a source of genomic novelty that can be co-opted by the host genome to perform new functions, such as regulatory activities and network rewiring, contributing to the evolution of innovations and adaptations for the host (Feschotte, 2023). Throughout eukaryotic evolution, many examples of co-option have been revealed in different species. In placental mammals, syncytin genes, essential for placental development, have been co-opted from envelope genes of diverse endogenous retrovirus (Blond et al., 2000). RAG1 and RAG2 genes in jawed vertebrates, involved in V(D)J recombination for variable antigen-binding sites, originate from the duplication of the transposase gene of an ancestral transposon. The inverted repeats (IR) of this transposon were inserted into a surface receptor gene, which underwent multiple duplications, giving rise to the V-D-J genes (Kapitonov and Jurka, 2005).

Meanwhile, transposons can be a source of adaptations in response to environmental stresses, pathogens or xenobiotic agents. There is evidence of stress-induced increases in transposable-element activity in various species.

For instance, the invasive ant species *Cardiocondyla obscurior* exhibits defined regions termed "TE islands," where TEs accumulate alongside genes likely involved in environmental adaptation. This phenomenon arises because, upon colonising new environments, the species undergoes a drastic reduction in genetic variability due to founder effects. Simultaneously, the species requires adaptive evolution to cope with the novel environmental conditions. This adaptation is facilitated by transposition bursts that generate inheritable genetic variability over a few generations, thereby facilitating the evolution of locally adapted phenotypes (Schrader et al., 2014).

Nowadays, current studies are presenting a multifaceted and intricate perspective of transposable elements, portraying them as varied and complex entities involved in a dynamic interaction with their hosts, spanning from detrimental to mutually beneficial.

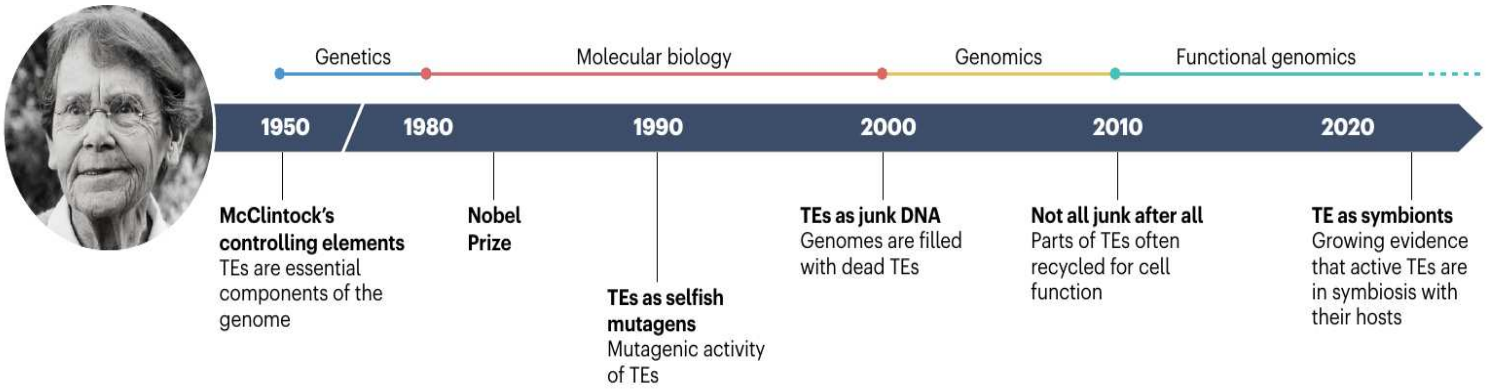


Figure 1. TE history timeline (modified from Feschotte, 2023)

Box 1. TE Classification

Transposable elements originate from different evolutionary sources and undergo continuous diversification. Consequently, they manifest in various classes, orders, and families.

They can be classified into two major classes, based on their transposition mechanism.

Class I of retrotransposons, transpose through a "*copy-and-paste*" mechanism via an RNA intermediate and a reverse transcription step. In this process, the RNA intermediate is transcribed from a genomic copy and then reverse-transcribed into a cDNA sequence that integrates into a new locus. Each transposition event produces one new copy.

The protein domains and the different integration mechanisms define the different orders of retrotransposons. **Long terminal repeat** (**LTR**) elements integrate through a cleavage and strand-transfer reaction catalysed by an integrase much like retroviruses and producing target site duplications (TSD). **DIRS** elements have a tyrosine recombinase gene instead of an integrase and therefore do not form TSDs. **PLE** elements encode a reverse transcriptase (RT) that is more closely related to telomerase than to the RT from LTR and an endonuclease.

Long and short interspersed nuclear elements (**LINEs** and **SINEs**), transpose and integrate through a process known as target-primed reverse transcription.

Class II of DNA transposons are distinguished by the number of DNA strands cut during transposition. They can mobilise via a non-replicative '*cut-and-paste*' mechanism cutting both DNA strands, with a transposase enzyme that recognizes their terminal inverted repeats (TIRs) (**TIR** elements) or with a tyrosine recombinase that involves recombination between a circular molecule and the DNA target (**Crypton** elements). Alternatively, they can mobilise through a single-strand excision using a replicative '*peel-and-paste*' mechanism (**Helitron** elements) or an extrachromosomal replication process (**Maverick** elements).

Each order is further divided into several families, the lowest level of TE taxonomy, defined by a common genetic structure according to the '80-80' rule: two elements belong to the same family if they share an 80% sequence similarity over 80% of their length.

For this reason, families are usually represented as consensus sequences, the ancestral copy reconstructed from sequence alignments of multiple copies found in a genome. Finally, TEs are also classified into autonomous and non-autonomous elements. The former can encode the enzymatic machinery necessary for their transposition, whereas the latter are mobilised through the proteins produced by their autonomous counterparts.

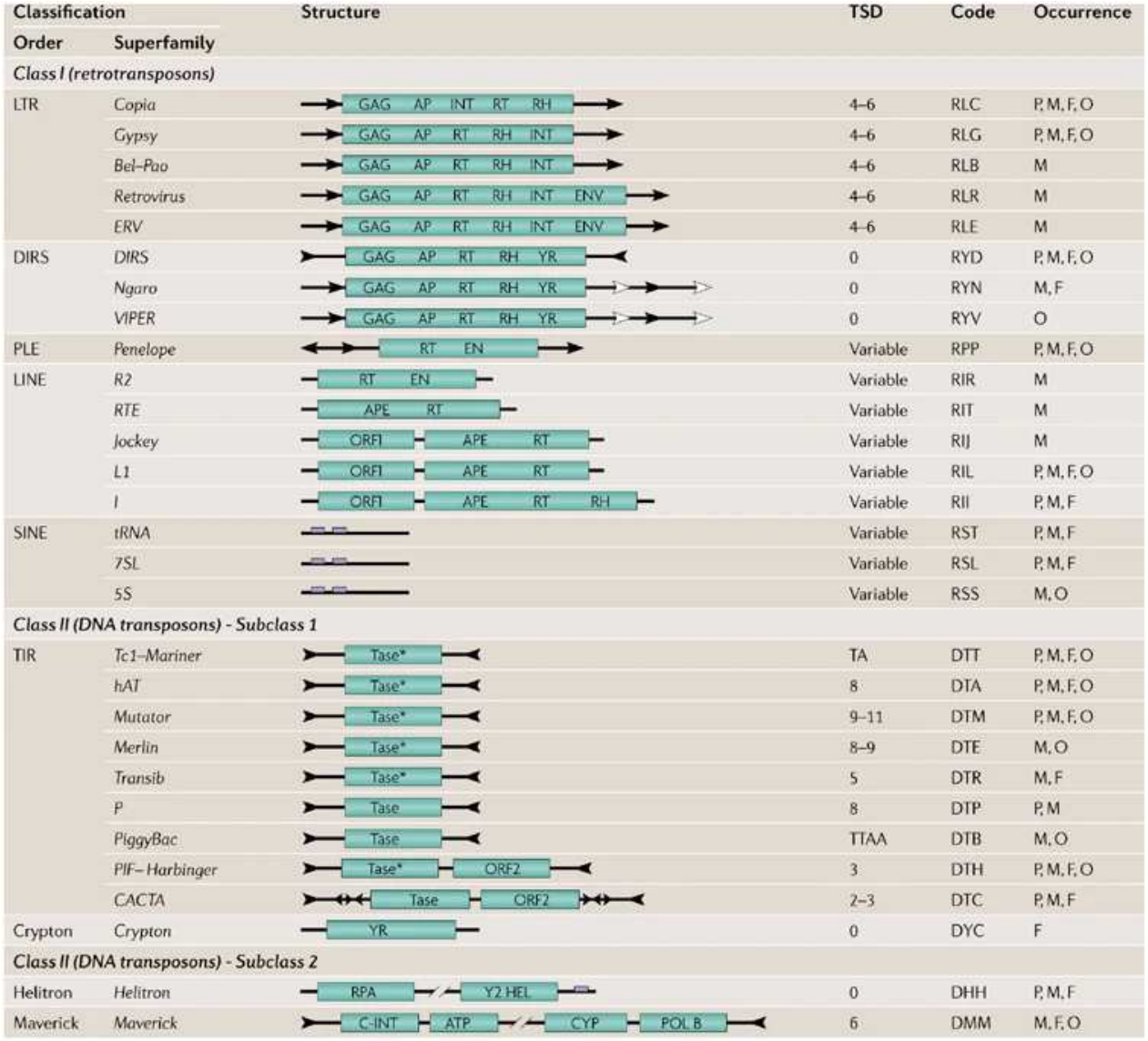


Figure 2. A classification system of transposable elements (from Wicker et al. 2007)

TE distribution and dynamics

With the advancement of new sequencing technologies, an increasing number of genomes have been characterised, revealing that transposable elements occupy a significant portion of nearly all eukaryotic genomes. The only exception lies in protists such as *Plasmodium falciparum*, which have completely purged transposable elements from their small genomes (Gardner et al., 2002).

The abundance of TEs varies considerably among different evolutionary lineages and even between closely related organisms, playing a significant role in genome size variation (Figure 3). Indeed, a positive correlation has been found between the accumulation of specific transposable element families and very large genomes.

For instance, in the Plethodontidae family of salamanders, which has undergone an extreme and independent long terminal repeat (LTR) amplification (Sun et al., 2012), or in the lungfish *Neoceratodus forsteri*, the closest living relative of tetrapods, where approximately 90% of its genome is represented by still-active transposable elements (Meyer et al., 2021). Another illustrative example is found in large plant genomes, which have also expanded in size through rapid LTR bursts over time (Baidouri et al., 2013). Conversely, certain lineages display minimal fluctuations in TE content, suggesting potential constraints on genome size. For instance, birds maintain a relatively stable genome size, potentially attributed to the metabolic expenses linked to active flight (Kapusta and Suh, 2016).

Intriguingly, the diversity of the different families of TEs present in a genome increases with their abundance only until genomes reach moderate size. Conversely, extremely large genomes exhibit lower TE diversity, primarily attributed to the proliferation of a few specific families. This observation predicts an inverse relationship between genome size and TE diversity at the largest genome sizes (Elliott and Gregory, 2015).

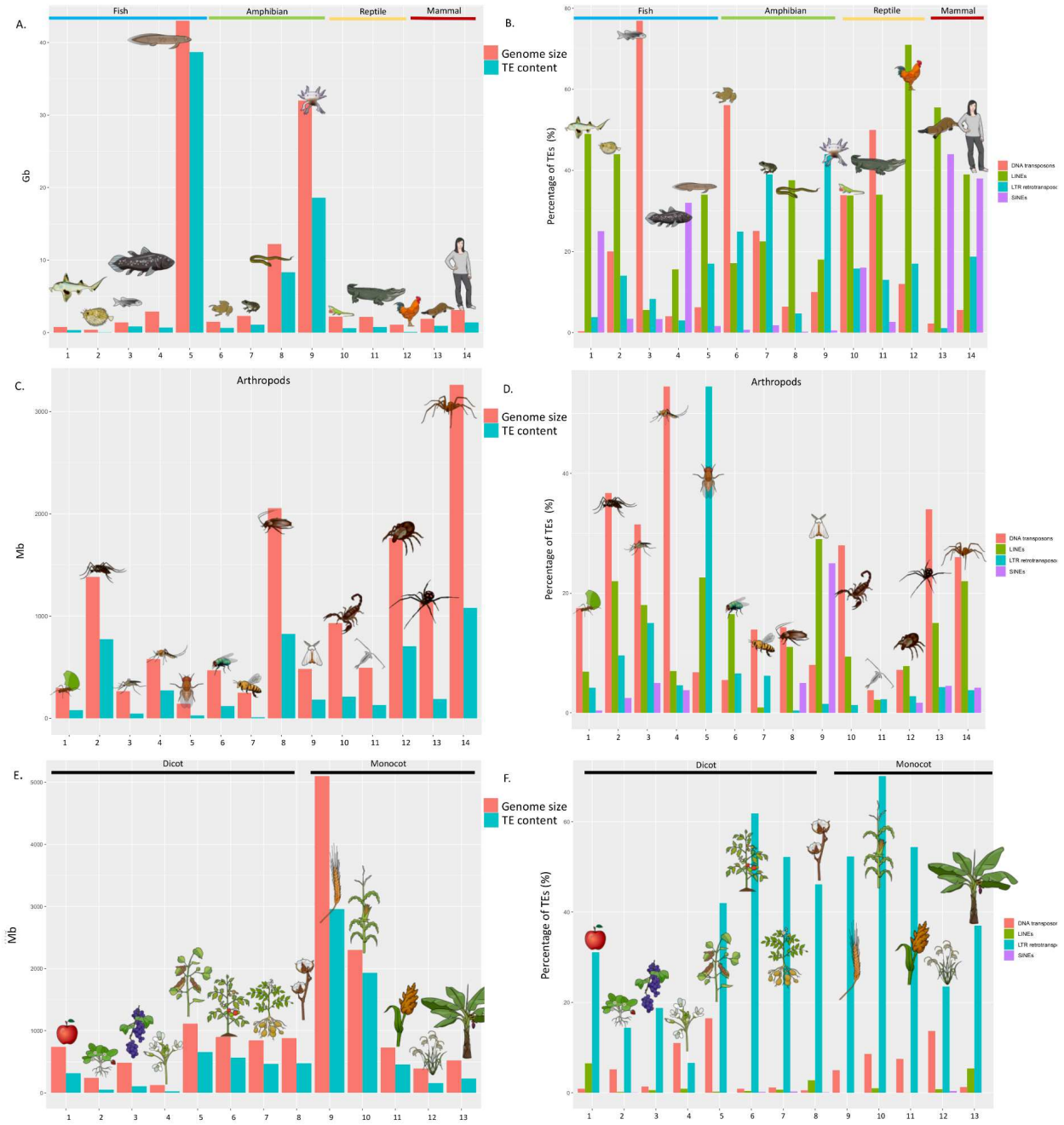


Figure 3. Distribution of TE abundance and genome size across different organisms (from Almojil et al., 2021)

The dynamics determining the expansion or contraction of a TE family in a host genome are currently subjects of controversial interpretations and ongoing investigations. Two primary categories of processes that determine and influence TE dynamics can be identified: molecular processes and population-level processes.

Molecular processes

Certainly, three factors drive the TE life cycle: (1) the rate of transposition, (2) the rate of fixation of new TE insertions, and (3) the rate of TE deletion.

In particular, the latter factor emerges as crucial in shaping TE content and genome size, as exemplified by the giant genomes of salamanders or caecilians characterised by a lower rate of DNA loss (Sun et al., 2012; Wang et al., 2021). Moreover, this pattern is also evident in bird and mammal genomes, where minimal interspecific variation in genome size is observed. This phenomenon is explained by the "accordion model" of genome size equilibrium, wherein DNA gains resulting from transposable element expansion are counteracted by DNA loss through large segmental deletions, determining the compact genomes of flying birds and bats (Kapusta et al., 2017).

Furthermore, another important process shaping TE distribution and amplification is the coevolution, alongside TE diversity and content, of different host-silencing mechanisms to control TE activity and limit their expansion (Figure 4).

One of the most widespread regulatory systems in eukaryotes is the small RNA silencing (sRNAs) system, also referred to as RNA interference (RNAi). sRNA can be differentiated by their sizes and by the Argonaute proteins with which they associate to form an RNA-induced silencing complex (RISC). This complex recognizes TE targets through base-pair complementarity and mediates the silencing of TEs at the transcriptional level through the recruitment of DNA methyltransferase (DNMT) or chromatin remodelling complexes, resulting in a repressive chromatin environment; or at the posttranscriptional level through the degradation of target TE transcripts in the cytoplasm (Rana, 2007).

In the germline, PIWI proteins, one of the Argonaute subfamilies, form specific RNA-induced silencing complexes (RISCs) with a small RNA population known as PIWI-interacting RNAs (piRNAs). piRNAs are transcribed from genomic regions called piRNA clusters in the primary processing pathway. In these clusters, mobilising

transposons can jump and be trapped, generating novel antisense piRNAs from inserted TEs. Then, the methyltransferase SETDB1 activates the transcription of the cluster by the deposition of H3K9me3. Subsequently, these piRNAs undergo processing into secondary piRNAs through the endonuclease PLD6 and the HMG protein Maelstrom (Mael), a nucleo-cytoplasmic shuttling protein capable of binding piRNA precursor transcripts and delivering them to the cytoplasm. Here, they undergo amplification through the ping-pong pathway, which simultaneously silences the target transposon sequence and amplifies the piRNA sequence (Iwasaki et al. 2015; Wang et al. 2023).

The PIWI pathway, therefore, stands as the principal safeguard system in germ cells, limiting TE proliferation at both transcriptional and post-transcriptional levels, thereby preserving normal gametogenesis and reproduction and maintaining genome integrity.

Another important system is the Krüppel-associated box (KRAB)-containing zinc finger protein (KRAB-ZFPs), the largest family of transcription factors in vertebrates. After their emergence 420 million years ago in the last common ancestor of tetrapods, KZFP genes underwent a significant expansion, coopting retrotransposon regulatory sequences. This expansion led to lineage-specific repertoires that not only control and repress TEs but also involve them in transcriptional regulatory networks (Rosspopoff and Trono, 2024).

The system binds to sequence-specific TE targets via the C-terminal array of zinc finger motifs, while the KRAB domain interacts with the KAP1/TRIM28 corepressor, which in turn serves as a scaffold protein to recruit key heterochromatin transcriptional silencing factors such as SETDB1, HP1, NuRD complex and DNA methyltransferases. The formed silencing complex implements sequence-specific transcriptional repression of transposable element activity (Ecco et al. 2017).

The correlation between transposon expansion and the host silencing mechanisms is tangled and still not fully understood. This complexity stems from the different strategies that have evolved, operating at different levels of the genome — transcriptional and post-transcriptional — and also in a tissue-specific manner.

This leaves some questions still open: whether there is a linear or non-linear correlation between TE expression and TE repression; whether there can be a balanced

dynamic between TE activity and host genome strategies; and finally, how this TE-host dynamic changes in relation to genomes of different sizes.

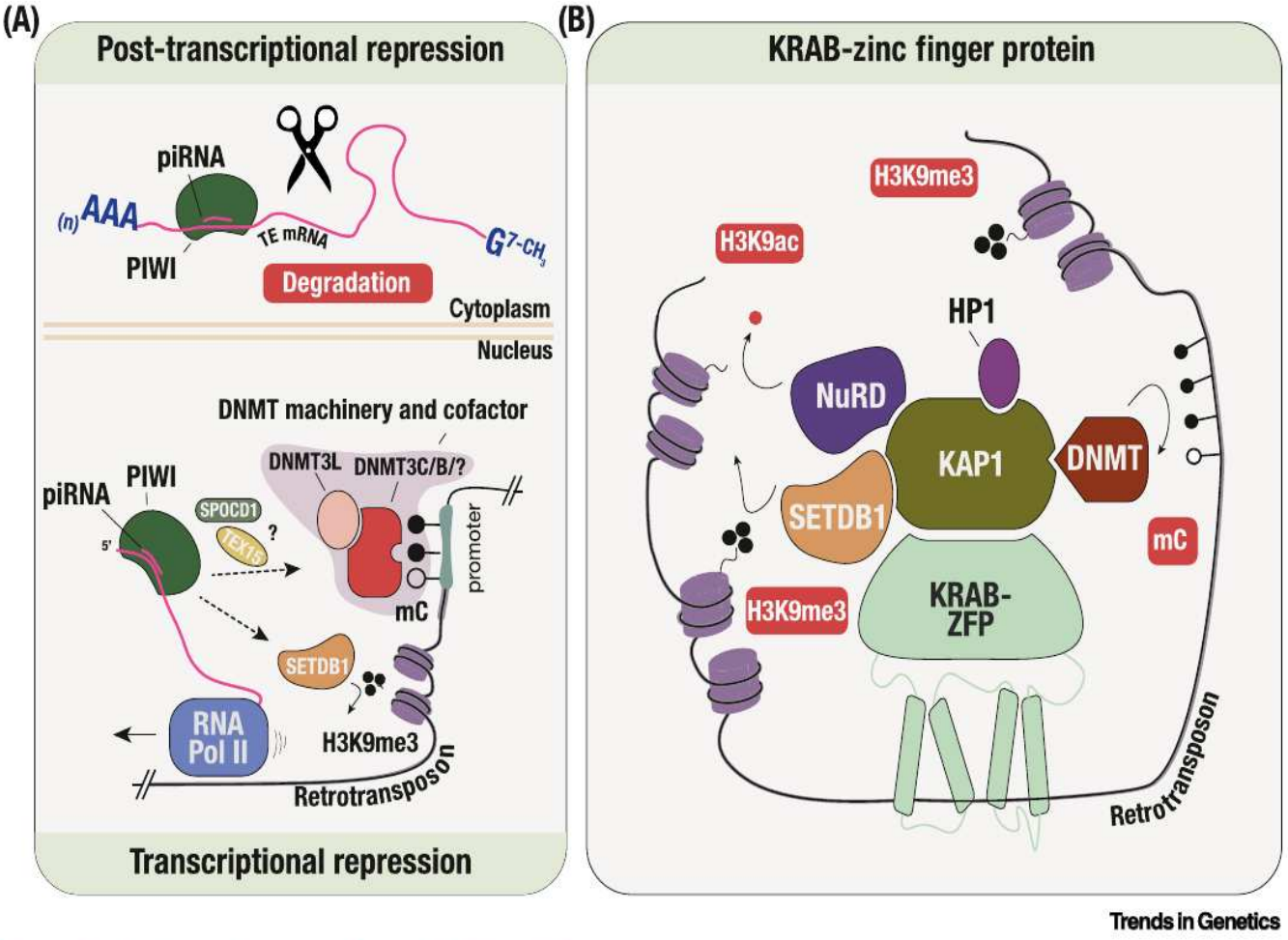


Figure 4. Host-silencing mechanisms to control TE activity (from Almeida et al., 2022)

Population-level processes

In addition to molecular processes, TE insertions are subject to population-level processes, including natural selection and genetic drift, which modulate the host organism fitness (Figure 5).

TEs can negatively affect the fitness of their host by inserting in gene or regulatory regions, or lead to large structural genomic variation (such as deletions, inversions, duplications, and translocations), or, even further, to deleterious chromosomal rearrangements through ectopic recombination between non-allelic transposon copies. Purifying selection plays a major role in preventing the fixation of deleterious TE insertions that reduce fitness in a population, thereby shaping the allele frequency spectrum of TEs. According to population genetics theory, the efficiency of selection is proportional to Ne. As a consequence, in populations with smaller effective population sizes, the fixation probability of slightly deleterious TE would be higher, due to the reduced efficiency of natural selection and a more intense influence of genetic drift. Hence, demographic changes will shape the allele frequency spectrum of TEs: reductions in Ne should lead to an excess of alleles at intermediate frequencies, while population expansions may result in an excess of rare insertions (Figure 5a). This leads to the hypothesis of a direct correlation between genome size and demographic history (Lynch and Conery, 2003).

Another important aspect is the combination of local recombination and linked selection. Due to Hill–Robertson interference (Hill and Robertson, 1966), in regions of low recombination, a deleterious TE at one site can reduce the efficacy of selection acting at neighbouring haplotypes with a different deleterious TE insertion. This interference will lead to the fixation and accumulation of more deleterious TEs, on average, than the ones inserted in regions of high recombination (Figure 5b-c) (Dolgin et al., 2008).

It is important to underline that TEs are not randomly distributed in the genome, and various models predict their heterogeneous genome-wide distribution.

The placement of a transposable element within the genome follows a two-step process. Firstly, the integration step that directs the initial allocation of insertions.

Subsequently, the action of natural selection to eliminate detrimental events and promote the fixation of advantageous insertions.

TEs can exhibit some level of insertion preference. On one hand, many elements tend to accumulate in gene-poor and low-recombining regions, where their negative effects are minimised, and they are less prone to removal by selection compared to functionally important regions, such as exons and regulatory regions.

For instance, the R1 and R2 families of LINEs preferentially target ribosomal DNA (rDNA) arrays, where they can undergo progressive purging through recombination within the array (Eickbush et al., 2015). Similarly, Ty1 and Ty3 LTR elements in the compact genome of *Saccharomyces cerevisiae* insert upstream of RNA Pol III-transcribed genes, avoiding disruption of gene expression (Sultana et al. 2017).

On the other hand, a wide variety of TEs preferentially target open regions of heterochromatin, such as 5' upstream regions of genes, where they can likely be expressed and propagated. This phenomenon is observed in some TE families in plants, such as Mu DNA transposons in maize or mPing in rice (Liu et al., 2009; Naito et al., 2009).

It is clear that TE insertion site preferences are counterbalanced by post-integration selection processes and different host regulatory mechanisms for TE activity. This intricate interplay among factors, encompassing insertion preferences, selection, recombination, and transposition, unequivocally underlies the accumulation and pervasive presence of transposable elements in the genome. Thus, the dynamic interactions within the genome shape the intricate landscape of transposable elements, highlighting their substantial role in genomic evolution.

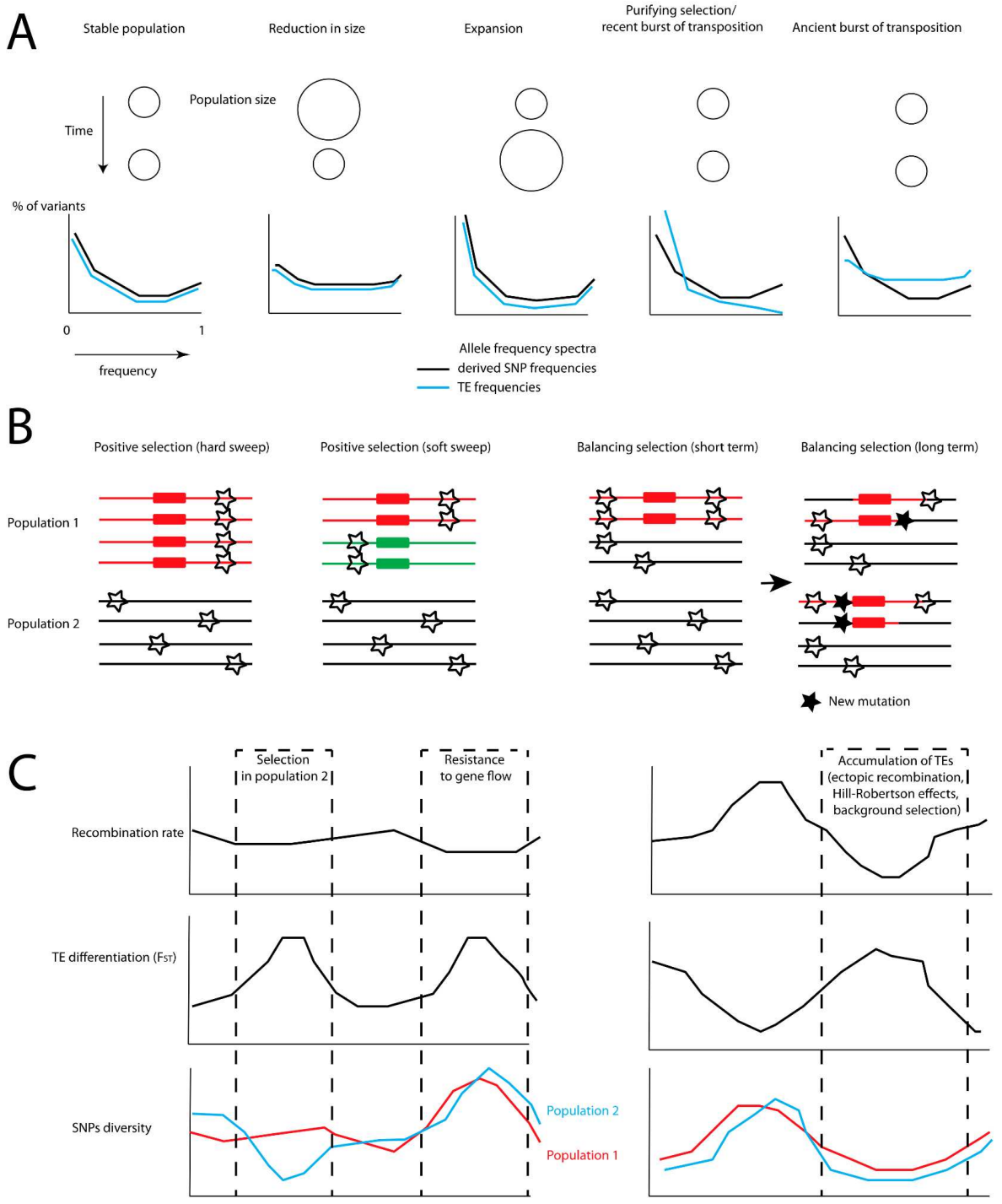


Figure 5. Population-level processes acting on TE insertions (from Bourgeois et al. 2019)

Methodological challenges in the study of TE dynamics

Despite the advent of new genome sequencing technologies, and in particular, the increasing length of sequencing reads, which has significantly facilitated the assembly and characterisation of repetitive elements, the study of transposable elements still faces significant challenges.

First and foremost, transposable elements have long been overlooked by the majority of researchers. It is worth noting that in genome-wide analyses, the common practice for standard bioinformatic pipelines is to mask out repeated regions and restrict the analysis to genes only. It follows that limiting the analysis to only a fraction of the source data actually makes so-called “whole-genome” analyses not really 'genome-wide', as well as significantly reducing the likelihood and scope of the resulting discoveries (Slotkin, 2018).

Such a disregard for TEs has led to a slow and relatively recent development of this area of research, which, on one hand, continuously enhances and enriches the tools needed to analyse the evolution of TEs, but on the other hand, it makes the approach difficult for those outside the field due to the lack of standardisation in a dedicated pipeline.

There are different methods to identify transposable elements within a genome, which can be summarised into two main strategies.

1) The first one is the homology-based approach, which detects TEs through sequence comparisons against databases of known TE consensus sequences or TE motifs. The quality and specificity of the databases used will influence the elements which are identified. Furthermore, by identifying sequences homologous to consensus ones, these methods will exclude instances that have become too divergent (Goerner-Potvin et al., 2018).

2) The second main strategy is the *de novo* approach, which identifies TEs through specific signatures, such as transposable element structure or elevated copy number. These methods mostly include several TE-order-specific tools that identify protein-coding domains, terminal repeats, or conserved sequence motifs that are specific to different TE orders.

For instance, several structure-based methods have been developed to detect LTR retrotransposons by searching for the common structural signals: the long terminal repeats, the flanking target site duplications (TSDs), the primer-binding sites (PBSs), the polypurine tracts (PPTs) and the different open reading frames (ORFs) for the gag, pol and env genes, that constitute the internal region of LTRs (Ellinghaus et al. 2008; Ou and Jiang, 2017; Valencia and Girgis, 2019).
The most common strategy for *de novo* methods is starting with the detection of similar sequences, followed by clustering methods to group related sequences into families. However, *de novo* strategies also have disadvantages, as they may fail to detect low-copy-number elements or erroneously classify TEs, leading to false positives.

A combination of the different approaches is often employed to identify the different TE families within a genome, utilising different software pipelines that exhibit varying strengths and performance on the diverse genomes under analysis.
It is important to note that the quality of the genome assembly, ranging from the type of reads used (short reads vs long reads) to the methodology employed for assembly, has an impact on TE detection and annotation. Specifically, in large genomes with a high proportion of repeated sequences, the level of assembly continuity, and thus the level of fragmentation, significantly affects TE identification.
This aspect is also crucial when conducting TE comparative analyses between different species, emphasising the importance of starting TE detection and annotation from similar genome assemblies.
Currently, the creation of a high-quality full-length transposable elements library is achievable only via manual curation, which, despite the recent development of novel bioinformatic resources (Goubert et al., 2022), remains time-consuming, especially for large genomes and even more for multiple species comparative analyses.
The TE research community is putting a lot of effort into creating open and collaborative platforms, such as the TE Hub Consortium (Elliott et al., 2021), to disseminate, share and establish a comprehensive catalogue of the different tools and protocols for wide-ranging TE analysis.
However, as mentioned above, it is only by broadening genome-wide analyses to include the repetitive and mobile DNA that we will achieve the implementation of

benchmarked and standardised methodologies. This, in turn, will make this research field more accessible to all.

References

Almojil D., Bourgeois Y., Falis M., Hariyani I., Wilcox J., Boissinot S. 2021. The Structural, Functional and Evolutionary Impact of Transposable Elements in Eukaryotes. Genes (Basel). 12(6):918. doi: 10.3390/genes12060918. PMID: 34203645; PMCID: PMC8232201.

Blond J.-L., Lavillette D., Cheynet V., Bouton O., Oriol G., Chapel-Fernandes S., Mandrand B., Mallet F., and Cosset F. L. 2000. An envelope glycoprotein of the human endogenous retrovirus HERV-W is expressed in the human placenta and fuses cells expressing the type D mammalian retrovirus receptor. J. Virol. 74, 3321–3329

Bourgeois Y., Boissinot S.. 2019. On the Population Dynamics of Junk: A Review on the Population Genomics of Transposable Elements. Genes (Basel). 10(6):419. doi: 10.3390/genes10060419. PMID: 31151307; PMCID: PMC6627506.

Dolgin E.S., Charlesworth B. 2008. The effects of recombination rate on the distribution and abundance of transposable elements. Genetics. 178(4):2169-77. doi: 10.1534/genetics.107.082743. PMID: 18430942; PMCID: PMC2323806.

Ecco G., Imbeault M., Trono D. 2017. KRAB zinc finger proteins. Development.144(15):2719-2729. doi: 10.1242/dev.132605. PMID: 28765213; PMCID: PMC7117961.

Eickbush T.H., Eickbush D.G. 2015. Integration, Regulation, and Long-Term Stability of R2 Retrotransposons. Microbiol Spectr. 3(2):MDNA3-0011-2014. doi: 10.1128/microbiolspec.MDNA3-0011-2014. PMID: 26104703; PMCID: PMC4498411.

El Baidouri M., Panaud O. 2013. Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution. Genome Biol Evol. 5(5):954-65. doi: 10.1093/gbe/evt025. PMID: 23426643; PMCID: PMC3673626.

Elliott T.A., Gregory T.R. 2015. Do larger genomes contain more diverse transposable elements?. BMC Evol Biol 15, 69. https://doi.org/10.1186/s12862-015-0339-8

Ellinghaus D., Kurtz S. & Willhoeft U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9, 18. https://doi.org/10.1186/1471-2105-9-18

Elliott T., Heitkam T., Hubley R., Quesneville H., Suh A., Wheeler T. The TE Hub Consortium. 2021. “TE Hub: a community-oriented space for sharing and connecting tools, data, resources, and methods for transposable element annotation”, Mobile DNA 12(16), doi: 10.1186/s13100-021-00244-0

Feschotte C. 2023. Transposable elements: McClintock’s legacy revisited. Nat Rev Genet 24, 797–800. https://doi.org/10.1038/s41576-023-00652-3

Gardner M.J., Hall N., Fung E., White O., Berriman M., Hyman R.W., Carlton J.M., Pain A., Nelson K.E., Bowman S., et al. 2002. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature, 419, 498–511.

Goerner-Potvin P., Bourque G. 2018. Computational tools to unmask transposable elements. Nat Rev Genet 19, 688–704. https://doi.org/10.1038/s41576-018-0050-x

Goubert C., Craig R.J., Bilat A.F. et al. 2022. A beginner’s guide to manual curation of transposable elements. Mobile DNA 13, 7. https://doi.org/10.1186/s13100-021-00259-7

Hill W.G., Robertson A. 1966. The effect of linkage on limits to artificial selection. Genet Res 8: 269–294.

Iwasaki Y.W., Siomi M.C., Siomi H. 2015. PIWI-Interacting RNA: Its Biogenesis and Functions. Annu Rev Biochem.;84:405-33. doi: 10.1146/annurev-biochem-060614-034258. Epub 2015 Mar 5. PMID: 25747396.

Kapitonov V. V., & Jurka J. 2005. RAG1 Core and V(D)J recombination signal sequences were derived from transib transposons. PLoS Biology, 3, e181. https://doi.org/10.1371/journal.pbio.0030181

Kapusta A., Suh A. 2017. Evolution of bird genomes-a transposon's-eye view. Ann N Y Acad Sci.Feb;1389(1):164-185. doi: 10.1111/nyas.13295. Epub 2016 Dec 20. PMID: 27997700.

Kapusta A., Suh A., Feschotte C. 2017. Dynamics of genome size evolution in birds and mammals. Proc Natl Acad Sci U S A. 114(8):E1460-E1469. doi: 10.1073/pnas.1616702114. Epub 2017 Feb 8. PMID: 28179571; PMCID: PMC5338432.

Liu S., Yeh C.T., Ji T., Ying K., Wu H., Tang H.M., Fu Y., Nettleton D., Schnable P.S. 2009. Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. PLoS Genet. 5(11):e1000733. doi: 10.1371/journal.pgen.1000733. Epub 2009 Nov 20. PMID: 19936291; PMCID: PMC2774946.

Lynch M., Conery J.S. 2003. The origins of genome complexity. Science. 302(5649):1401-4. doi: 10.1126/science.1089370. PMID: 14631042.

McClintock B. 1951. Mutable loci in maize. Carnegie Institution of Washington Yearbook 50, 174–181

McClintock B. 1956. Intranuclear systems controlling gene action and mutation. Brookhaven Symp. Biol.(8), 58–74

Meyer A., Schloissnig S., Franchini P., Du K., Woltering J.M., Irisarri I., Wong W.Y., Nowoshilow S., Kneitz S., Kawaguchi A. et al. 2021. Giant lungfish genome elucidates the conquest of land by vertebrates. Nature, 590, 284–289

Naito K., Zhang F., Tsukiyama T. et al. 2009. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature 461, 1130–1134. https://doi.org/10.1038/nature08479

Ou S., Jiang N. 2018. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. Plant Physiol.176(2):1410-1422. doi: 10.1104/pp.17.01310. Epub 2017 Dec 12. PMID: 29233850; PMCID: PMC5813529.

Rana T. 2007. Illuminating the silence: understanding the structure and function of small RNAs. Nat Rev Mol Cell Biol 8, 23–36. https://doi.org/10.1038/nrm2085

Rosspopoff O., Trono D. 2024. Take a walk on the KRAB side: (Trends in Genetics, 39:11 p:844-857, 2023). Trends Genet;40(2):203-205. doi: 10.1016/j.tig.2023.12.007. Epub 2023 Dec 30. Erratum for: Trends Genet. 2023 Nov;39(11):844-857. PMID: 38160062.

Schrader L., Kim J., Ence D. et al. 2014. Transposable element islands facilitate adaptation to novel environments in an invasive species. Nat Commun 5, 5495. https://doi.org/10.1038/ncomms6495

Slotkin R.K. 2018. The case for not masking away repetitive DNA. Mobile DNA 9, 15. https://doi.org/10.1186/s13100-018-0120-9

Sultana T., Zamborlini A., Cristofari G. et al. 2017. Integration site selection by retroviruses and transposable elements in eukaryotes. Nat Rev Genet 18, 292–308 . https://doi.org/10.1038/nrg.2017.7

Sun C., Shepard D.B., Chong R.A., Lopez Arriaza J., Hall K., Castoe T.A., Feschotte C., Pollock D.D., Mueller R.L. 2012. LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders. Genome Biol. Evol., 4, 168–183

Sun C., López Arriaza J.R., Mueller RL. 2012. Slow DNA loss in the gigantic genomes of salamanders. Genome Biol Evol. 4(12):1340-8. doi: 10.1093/gbe/evs103. PMID: 23175715; PMCID: PMC3542557.

Valencia J.D., Girgis H.Z. 2019. LtrDetector: A tool-suite for detecting long terminal repeat retrotransposons de-novo. BMC Genomics 20, 450. https://doi.org/10.1186/s12864-019-5796-9

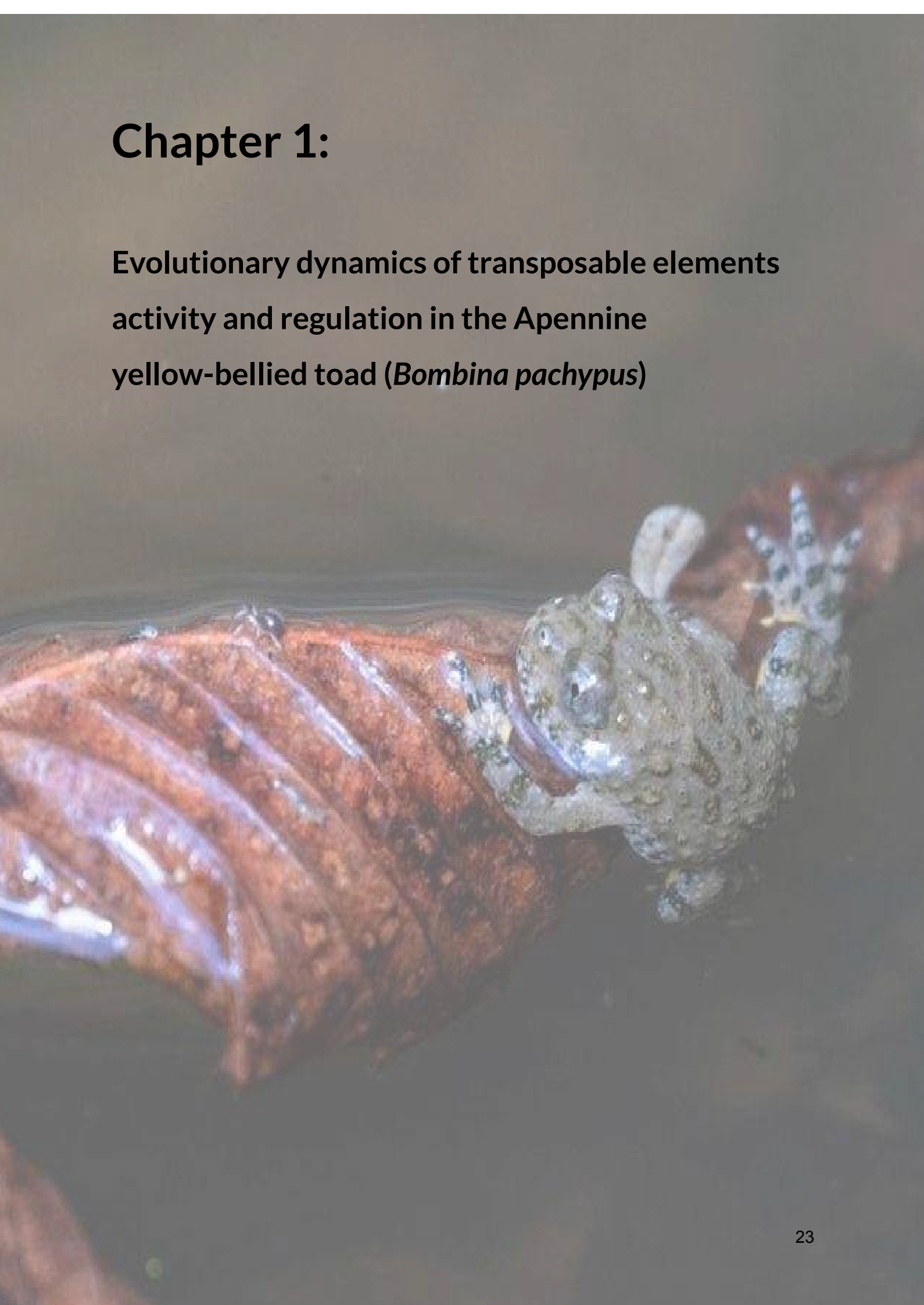
Wang X., Ramat A., Simonelig M. et al. 2023. Emerging roles and functional mechanisms of PIWI-interacting RNAs. Nat Rev Mol Cell Biol 24, 123–141. https://doi.org/10.1038/s41580-022-00528-0

Wang J., Itgen M.W., Wang H., Gong Y., Jiang J., Li J., Sun C., Sessions S.K., Mueller RL. 2021. Gigantic Genomes Provide Empirical Tests of Transposable Element Dynamics Models. Genomics Proteomics Bioinformatics.19(1):123-139.doi: 10.1016/j.gpb.2020.11.005. Epub 2021 Mar 4. PMID: 33677107; PMCID: PMC8498967.

Wicker T., Sabot F., Hua-Van A. et al. 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8, 973–982. https://doi.org/10.1038/nrg2165

Chapter 1:

Evolutionary dynamics of transposable elements activity and regulation in the Apennine yellow-bellied toad (*Bombina pachypus*)



Chapter 1

Evolutionary dynamics of transposable elements activity and regulation in the Apennine yellow-bellied toad (*Bombina pachypus*)

Authors:

Lorena Ancona[1], Flávia A. Nitta Fernandes[1], Roberto Biello[2], Andrea Chiocchio[3], Tiziana Castrignanò[3], Marco Barucca[1], Daniele Canestrelli[3], Emiliano Trucchi[1]

Affiliations:

1 Department of Life and Environmental Sciences, Polytechnic University of Marche, Italy

2 Department of Life Sciences and Biotechnology, University of Ferrara, Italy

3 Department of Ecological and Biological Sciences, Tuscia University, Italy

Introduction

Transposable elements (TEs) are DNA sequences that replicate and mobilise within a host genome, influencing genome architecture and evolution. Long considered junk DNA, their high evolutionary potential is increasingly being discovered (Frank et al. 2022; Choudhary et al. 2023). Transposons can have a double impact on the host genome: from the disruption of coding regions and genomic rearrangements via ectopic recombination, to being an important source of genomic novelty by acting as gene regulatory elements and generating new genes through domestication (Chuong et al. 2017; Schrader and Schmitz 2019).

Intriguingly, TEs can be the driver of genome expansion, playing a major role in genome size variation (Chalopin et al. 2015; Sotero-Caio et al. 2017). In particular, a positive correlation was detected between the accumulation of specific TE families and species with very large genomes, with notable examples among vertebrates (Sun et al. 2012; Meyer et al. 2021). However, merely considering the genomic abundance of TEs provides little insight into their past and recent amplification history, thereby neglecting the evolutionary dynamics between these elements and their host genome. In addition, while the characterisation of TE diversity and abundance is now an important step in genome annotation, the expression of TEs is poorly studied in large genomes (Rogers et al. 2018; Carducci et al. 2021; Wang et al. 2021(a);2023). It is therefore necessary to study TE activity and its relationship to genome size to understand the dynamics of transposon expansion.

The concept of TE activity is often misunderstood and directly associated with TE mobilisation. However, it is important to recognise that TE activity encompasses various interconnected levels. The first one is TE transcription, which can be initiated with an internal TE promoter or as co-transcription within a hosting gene (i.e., the gene with the TE insertion), serving as a prerequisite for TE mobilisation (Lanciano et al. 2020). Subsequently, translation of the different TE protein domains becomes necessary to enable their enzymatic activity and successive mobilisation. Thus, assessing TE expression constitutes the primary step in investigating TE activity.

An inverse relationship between genome size and the proportion of transcriptionally active TE copies has been detected in previous simulation studies, suggesting that

genomes with more TE copies have a lower number of active TE families, due to the competition between different elements (Kijima and Innan 2013; Boissinot et al. 2016). Moreover, the more the TE copies increase, the more the host genome is activated to control and limit their expansion, for instance by degradation of TE transcripts through RNA silencing, thus resulting in fewer active sequences (Roessler et al. 2018). In this genomic arms race, the host genome has evolved different mechanisms to repress TE mobilisation.

In the germline, transposons find a breeding ground for expansion and transmission to the next generation, facilitated by the deletion of global methylation patterns that determine cellular potency during germ-cell development. The PIWI-interacting RNAs pathway is the specific safeguard in germ cells to maintain genome integrity, which acts both at the transcriptional level through methylation of target TEs, and at the posttranscriptional level through degradation of target TE transcripts (Iwasaki et al. 2015). In addition to the small RNA silencing pathways, the Krüppel-associated box domain zinc finger proteins (KRAB-ZFPs), the largest family of transcription factors in vertebrates, together with the nucleosome remodelling and deacetylase (NuRD) complex, play an important role in transcriptional repression (Ecco et al. 2017). KZFP genes, after they first emerged in the last common ancestor of tetrapods, seem to have co-opted retrotransposon regulatory sequences to control and involve them in transcriptional regulatory networks (Bruno et al. 2019; Playfoot et al. 2021).

Nevertheless, the relationship between transposon expansion and TE silencing among vertebrates is complex, with studies often revealing different patterns and dynamics among species (Carducci et al. 2021; Liu et al. 2022; Wang et al. 2023). Particularly in large genomes, it remains unclear whether the heightened activity and amplification of TEs result from a lower efficiency of silencing mechanisms, and whether distinct mechanisms and dynamics operate in different tissues.

Amphibians represent one of the groups with the largest genomes in the animal kingdom, exhibiting highly variable genome sizes among the three orders (ranging from 1 Gb in *Platyplectrum ornatum* to 120 Gb in *Necturus lewisi*), and an exceptionally high abundance of repetitive elements. This makes them an excellent model for studying transposon dynamics (Wang et al. 2021(a); Haley and Mueller, 2022).

Here, we investigate the expression and regulation patterns of TEs in the Apennine yellow-bellied toad (*Bombina pachypus*), an anuran species endemic to the Italian peninsula, showing one of the largest genome sizes among the Anura order (approximately 10 Gb).

Using transcriptomic data from somatic and germline tissues, we investigate the complex dynamics acting between transposons and the host genome, aiming to answer the following questions:

1) What are the transcriptional activity patterns of TEs in the large genome of *B. pachypus*, and which TE families exhibit the highest expression levels? 2) Are distinct patterns of TE expression between somatic and germline tissues determining a preferential pathway for TE expression and, likely, propagation? 3) What is the contribution of TE silencing gene pathways in the different tissues? Are there tissue-specific control systems to counteract TE expansion?

Materials and Methods

Study species and limitations

Bombina pachypus is an anuran species endemic to the Italian peninsula, which has been listed as Endangered due to habitat loss, climate change, and vulnerability to chytridiomycosis infection (Barbieri et al. 2004; Andreone et al. 2009; Canestrelli et al. 2013). The species is distributed south of the Po Valley, through the Apennines region, and south to the Aspromonte massif in Calabria. Genetically, it is subdivided into two main genetic clusters: a southern cluster, in the putative glacial refugia for the species, which is the hotspot of genetic diversity for this species; and a northern cluster, resulting from a post-glacial range expansion, with lower genetic variability than the southern counterpart (Canestrelli et al. 2006). *B. pachypus* is known to have suffered a drastic population decline of more than 50% across its range, with the exception of Calabrian populations that showed the highest levels of intrapopulation genetic variation, although a more recent demographic decline has also been observed in these populations (Zampiglia et al. 2019; Martino et al. 2022).

The large genome size of this toad makes it a suitable model for studying the contribution of TEs to genome expansion.

We analysed six adult yellow-bellied toad individuals (three females and three males) that were collected during a population monitoring program in the spring of 2020 and 2023, from the Aspromonte massif in Calabria, southern Italy. Target sampling and lab breeding (two strategies potentially useful to increase sample size) were not feasible, as this species is threatened and strictly protected.

Sampling procedures were approved by the Italian Ministry of Ecological Transition and ISPRA (permit number: 20824, 18-03-2020).

RNA extraction, library preparation, and sequencing

Individuals were dissected to obtain brain and gonad tissues, and samples were stored in RNAprotect Tissue Reagent (Qiagen) until laboratory processing.

RNA extraction was performed using the RNeasy Plus Kit (Qiagen), according to the manufacturer's protocol, followed by RNA quality and quantification procedures with a

spectrophotometer and a Bioanalyzer (Agilent Cary60 UV-vis and Agilent 2100, respectively - Agilent Technologies, Santa Clara, USA).

Library preparation was performed separately for each sample and sequencing was performed by NOVOGENE (UK) COMPANY LIMITED, using the NEBNext® Ultra ™ RNA Library Prep Kit for Illumina® and Illumina NovaSeq platforms respectively, as described in Chiocchio et al. 2022. We obtained on average 52.8 million reads for each brain sample library and 58.3 million reads for each gonad sample library.

Transcriptome assemblies

To explore the dynamics of TE activity and investigate different expression patterns in both somatic and germline tissues of *B. pachypus*, we concatenated two different transcriptome assemblies of the brain and gonads tissues.

The brain transcriptome was already assembled (Chiocchio et al. 2022). We then assembled another transcriptome version starting from gonad RNA samples and concatenated it with the "after CD-HIT-est" version of the brain assembly as described below.

First, raw read quality was examined using FastQC 0.11.5 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and MultiQC v.1.9 (Ewels et al. 2016), and trimming was performed with Trimmomatic v.0.39 (Bolger et al. 2014) (setting the option SLIDINGWINDOW: 4: 15, MINLEN: 36, and HEADCROP: 13) to remove low-quality bases and adapter sequences. After the cleaning step and removal of low-quality reads, 395,999,482 clean reads (*i.e.*, 96% of raw reads) were maintained for building the *de novo* transcriptome assembly.

Brain and gonad transcriptomes were *de novo* assembled separately, using the bioinformatic protocol described in Chiocchio et al. 2022 (Figure 2) with modifications. Briefly, we used rnaSPAdes v.3.14.1 (Bushmanova et al. 2019) to construct two optimised *de novo* transcriptomes. Transcriptome quality was validated using BUSCO v.4.1.4 (Simão et al. 2015), DETONATE v.1.11 (Li et al. 2014) and TransRate v.1.0.3 (Smith-Unna et al. 2016). We removed potential redundancy with CD-HIT-est v.4.8.1 (Fu et al. 2012) using the default parameters, corresponding to a similarity of 95%. We then concatenated the two transcriptome assemblies and ran another step of

CD-HIT-est to remove redundant transcripts. Finally, we filtered out contigs with less than 1 transcript per million (TPM) to remove extremely low-expressed transcripts.

Transcriptome annotation

All the unique transcripts were converted to peptide sequence using TransDecoder v.5.5.0 (Haas, BJ. https://github.com/TransDecoder/TransDecoder). Sequences were searched against the nonredundant NCBI protein database using DIAMOND v.0.9.10 (Buchfink et al. 2015) with an E-value cut-off of ≤1 × 10−5. BLAST2GO v.5.0 (Conesa et al. 2005) and INTERPROSCAN v.2.5.0 (Quevillon et al. 2005) were used to assign Gene Ontology (GO) terms. Protein domains were annotated by searching against the InterPro v.32.0 (Hunter et al. 2012) and Pfam v.27.0 (Punta et al. 2012) databases, using INTERPROSCAN v.5.52 (Quevillon et al. 2005) and HMMER v.3.3 (Finn et al. 2011), respectively.

TE detection and expression

With the aim of discovering how transposons can expand and mobilise in a large genome and to explore the dynamics of their activity in different tissues, we first identified which TE families are present in the *B. pachypus* transcriptome, and then investigated which of these families are active elements.

We first detected and annotated TEs in the transcriptome assembly using the Extensive *de novo* TE Annotator (EDTA) v1.9.9 (Ou et al. 2019), a *de novo* pipeline that combines a suite of best-performing packages and includes a final step with RepeatModeler (Flynn et al. 2020) to generate a library of high-quality non-redundant TE sequences. Afterwards, we refined the library with DeepTE (Yan et al. 2020), which classifies unknown elements at order and superfamily level, based on convolutional neural networks.

Subsequentially, we estimated individual expression levels of the different TE families in the different tissues, mapping each individual sample of brain and male and female gonads to the TE library (TE families consensus sequences found in the transcriptome) with SalmonTE v.0.4 (Jeong et al. 2018). TE counts were normalised with edgeR (Robinson et al. 2010) using the TMM method, which is recommended for

between-sample comparisons, as it takes into account sequencing depth, RNA composition and gene length. Differential expression analyses were conducted between different tissues and different sexes with edgeR, using the negative binomial Generalised Linear Model (GLM) (glmQLFit and glmQLFTest functions in edgeR). First, we tested for sex-based expression differences by conducting pairwise comparisons between male and female samples of brain and gonads separately (male brain versus female brain, male gonad versus female gonad). Subsequently, we explored differentially expressed TEs between brain and gonads by contrasting one tissue against the other (male brain + female brain versus male gonad + female gonad). TE families with an adjusted P-value < 0.05 and log2 fold change ≥2 were considered as differentially expressed.

Characterisation and expression of TE-silencing gene pathways

To investigate if there are tissue-specific strategies to control TE mobilisation, we searched the literature (Biscotti et al. 2017; Almeida et al. 2022) to compile a list of genes involved in TE host silencing mechanisms.

We selected and analysed the activity of 32 target genes, including: germline-specific repressors that prevent the vertical inheritance of TEs, transcriptional repressors that act through methylation, and post-transcriptional repressors that promote TE transcript degradation, all of which are further described below, to gain a broad view of the different silencing dynamics.

We first functionally annotated our concatenated transcriptome and then used tBLASTn (Gertz et al. 2006) to search for transcripts which could be orthologous to the following set of TE regulatory genes: 1) members of Ago (AGO1, AGO2, AGO4) and Piwi (PIWIL1, PIWIL2, PIWIL4) pathways; 2) genes involved in small RNA biogenesis (DICER, DROSHA, DGCR8, PLD6, MAEL, SETDB1); 3) genes participating in KRAB-ZFPs repression complex and chromatin-related corepressors (SETDB1, Trim28/KAP1, HP1a, HP1b, HP1g, DNMT1, DNMT3A, PRMT5); 4) genes related to the NuRD complex (CHD3, CHD4, CHD5, HDAC1, HDAC2, MBD2, MBD3, MTA1, MTA2, p66aplha, p66beta, RBBP4, RBBP7). Then we translated the 32 target sequences into their protein sequences, looking for the most complete CDS region. In the case of a

sequence fragmented into several frames, we manually curated the sequence in order to reconstruct the CDS region and identify the 5' and 3' UTR regions.

To estimate the individual expression levels of the 32 target genes in the different tissues, we mapped each individual sample of brain and gonad tissues to the transcriptome with Salmon v1.4.0 (Patro et al. 2017), followed by normalisation of all transcripts with edgeR (using the TMM method). Afterwards, we extracted the expression values of the 32 genes of interest.

Differential expression analyses were conducted between different tissues and different sexes with edgeR, using the negative binomial Generalised Linear Model (GLM) (glmQLFit and glmQLFTest functions in edgeR) with the same contrasts as described for TE differential analysis. Genes with adjusted P-value < 0.05 and log2 fold change ≥2 were considered as differentially expressed.

Results

TE expression

We obtained a concatenated transcriptome with 1,738,562 contigs. BUSCO assessment showed a completeness score of 93.9%, with 38.6% represented as single-copy and 55.3% as duplicated (S:38.6%, D:55.3%). The level of duplication, likely due to the high number of repeated sequences, does not affect our analysis as the creation of the TE library permits searches for consensus sequences among the duplicates.

Our TE detection and expression pipeline identified 22 active superfamilies in the three different tissues of *B. pachypus* (brain, testes and ovaries). When estimating individual TE expression in the three tissues, we found on average higher TE expression in ovaries and testes compared to the brain (Figure 1; Supplementary Table S1-S3). In particular, retrotransposons were the most active class in both ovaries and testes (average normalised counts: ovaries 706,236; testes 701,833; brain 602,183), with the LTR/Gypsy family showing the highest expression, followed by the L1 family. On the other hand, DNA transposons were the most active class in ovaries (average normalised counts: ovaries 563,452; testes 325,878; brain 270,719), with the highest expression described by DNA/TcMar and DNA/hAT families.

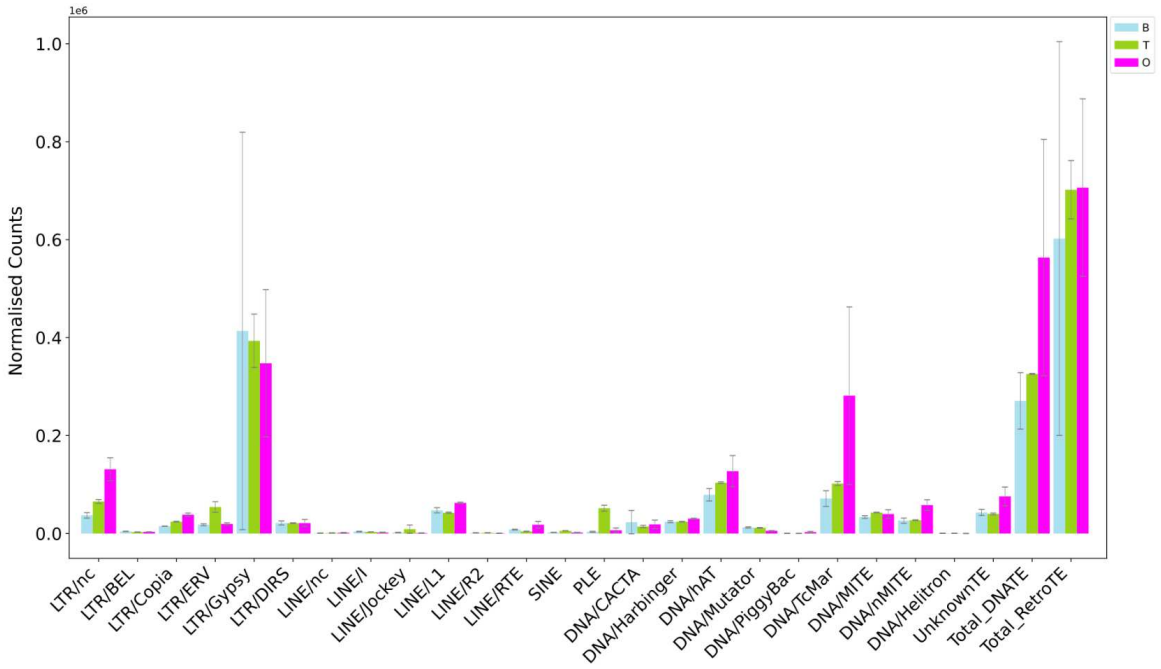


Figure 1. Expression levels of TEs in *B. pachypus*:

Total_DNATE: total of all DNA transposons; Total_RetroTE: total of all Retrotransposons (from LTR/nc to PLE families); B: brain (blue), T: testes (green), O: ovaries (pink). Nc: unclassified elements at the family level.

The female individual BP402 showed very high peaks in some TE families, particularly for the brain tissue, which was found to be an outlier in the MDS plot (Supplementary Figure 1). Different normalisation methods were tested, resulting in the same pattern. In addition, we characterised TE expression without BP402 and obtained the same global expression pattern (Supplementary Figure 2). In comparison to the other samples, BP402 was sampled during a different breeding season, which may have had an effect on the observed expression patterns.

By analysing patterns of expression between testes and ovaries, we identified a total of 400 differentially expressed TEs (logFC ≥2): 231 were more highly expressed in testes, while 169 were more highly expressed in ovaries. In addition to the higher number of overexpressed TEs, the testes also showed a much higher expression level than the ovaries, with a maximum logFC of 22 compared to 8 in the ovaries (Figure 2; Supplementary Table S4). Of the overexpressed transposons in the testes: 53% were retrotransposons, 42% were DNA transposons and the remaining 5% were unknown. Of those which were overexpressed in the ovaries: 53% were DNA transposons, 38% were retrotransposons and the remaining 10% were unknown.

Analyses of sex-based expression differences of TEs in the brain yielded no significant differentially expressed TEs (DE-TEs).

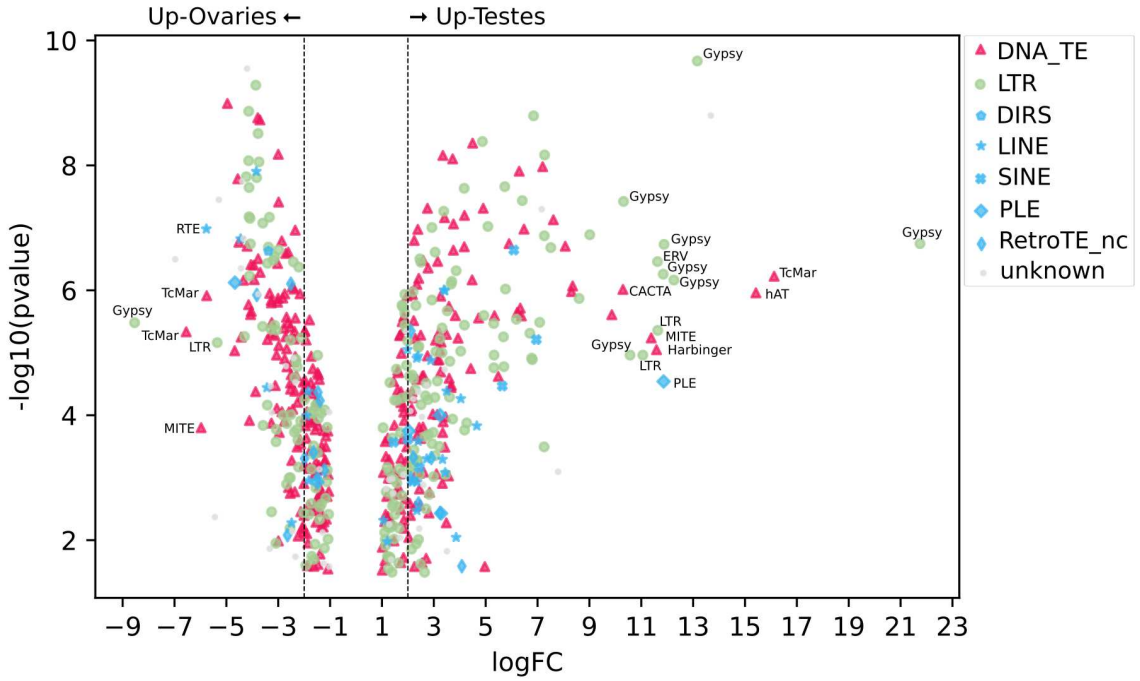


Figure 2. Volcano plot showing the overexpressed DE-TEs between ovaries and testes:
Overexpressed TEs in ovaries (left) and in testes (right). Dotted lines indicate logFC ±2

TE silencing gene pathways expression and dynamics

Firstly, we investigated whether there were different expression dynamics of TE-silencing gene pathways between the brain and gonads in *B. pachypus*.

We observed tissue-specific strategies in the gonads, with heightened activity observed in most of the TE-silencing genes that we analysed, with the exception of the AGO genes pathways (AGO1,2,4, DICER and DROSHA), which exhibited low expression across all tissues (Figure 3; Supplementary Table S5-S6).

Regarding the other Argonaute proteins, which are essential components of the RNA-induced silencing complex (RISC) (Peters and Meister 2007), the PIWI family (which is specifically expressed in germ cells), showed the highest expression in both the ovaries and the testes. PIWIL1 and PIWIL2 were active in both gonad tissues, while PIWIL4 displayed higher expression levels, primarily in males. Moreover, PIWIL4 showed the greatest fold change (5.6 logFC) among the overexpressed silencing genes between testes and ovaries (Figure 4; Supplementary Table S7).

In addition, genes involved in the primary and secondary biogenesis of piRNAs (the endonuclease PLD6, the transposon silencer MAEL, and the histone methyltransferase SETDB1) exhibited greater expression levels in gonad tissues compared to the brain. Specifically, all three of these genes together with PIWI genes, showed elevated expression and significant fold changes when examining differentially expressed genes (DEGs) between brain and gonad tissues (Figure 5; Supplementary Table S8-S9).

Considering the sequence-specific TE targeting through the KRAB-ZFPs complex, we detected robust expression in the ovaries for all the genes involved in this complex: the scaffold protein Trim28/KAP1, the histone methyltransferase SETDB1, the heterochromatin proteins HP1(a, b, g), and the maintenance DNA methyltransferases DNMT1. All these genes, with the exception of HP1g, were significantly overexpressed in the ovaries compared to the testes. Furthermore, the *de novo* methyltransferase DNMT3A, although showing low expression levels, was significantly overexpressed in the testes compared to the ovaries (Figure 4; Supplementary Table S7).

Furthermore, in relation to the NuRD complex, which is likewise linked with the KRAB-ZFPs complex, we noted increased expression in gonad tissues, especially in females. The majority of NuRD genes exhibited significant overexpression in ovaries,

with only three of them (CHD3, CHD5, and MBD2) displaying overexpression in testes when examining DEGs between testes and ovaries (Figure 4; Supplementary Table S7). CHD5 and p66beta transcripts made an exception, showing a brain-specific higher expression compared to the gonads.

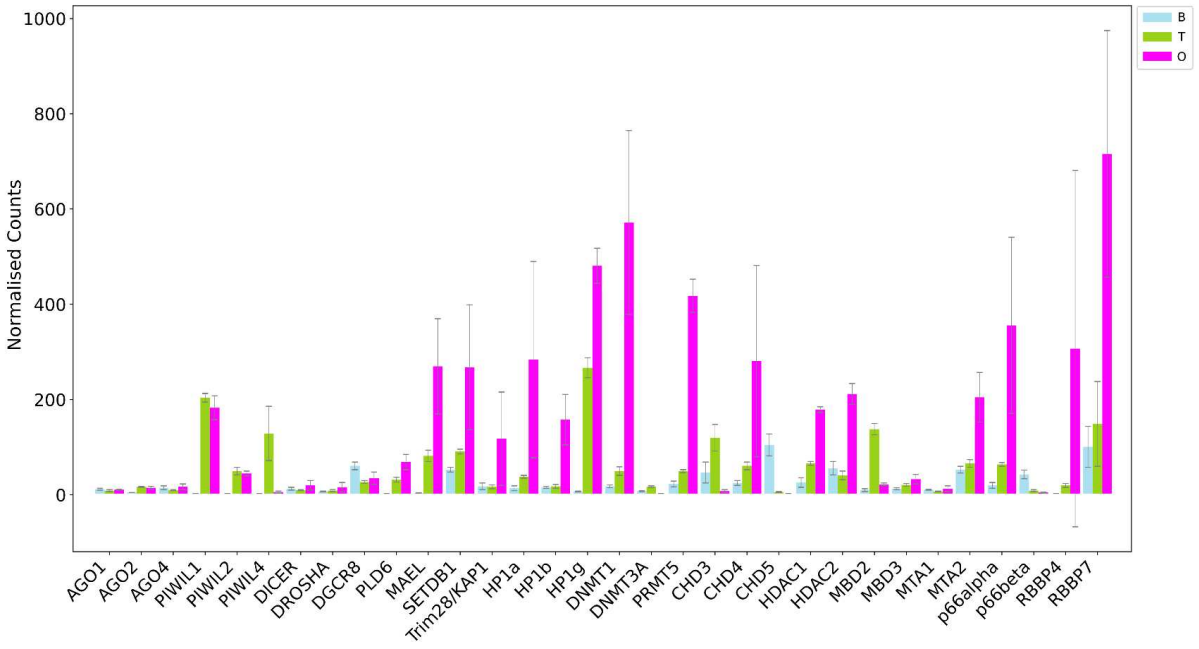


Figure 3. Expression levels of key genes involved in negative regulation of TE activity in *B. pachypus*:
B: brain (blue), T: testes (green), O: ovaries (pink).

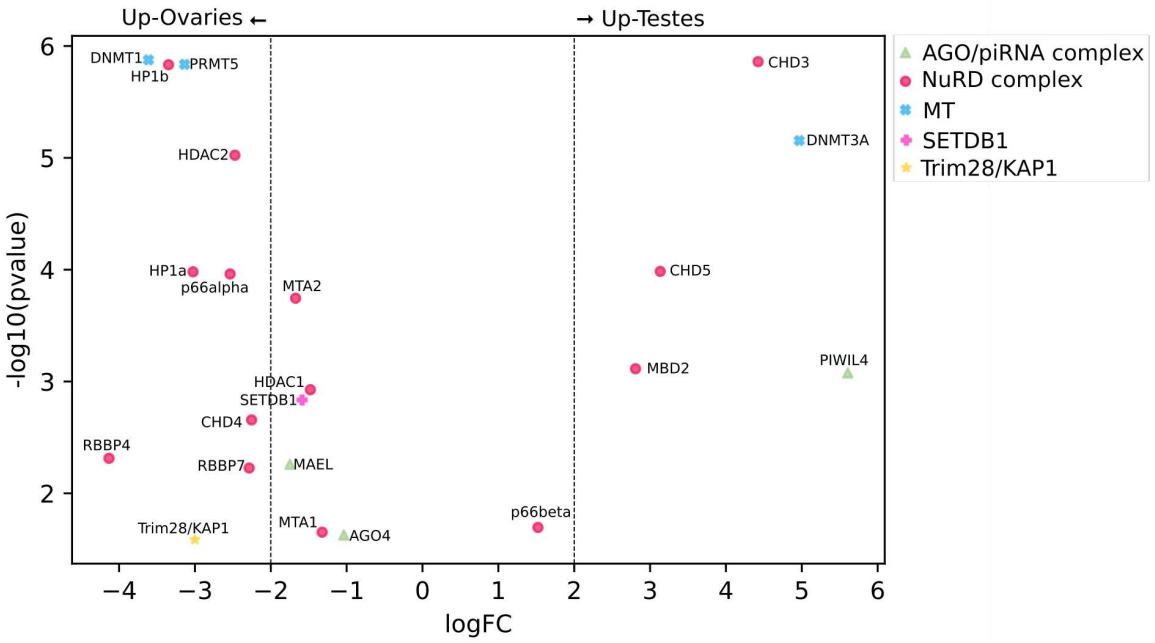


Figure 4. Volcano plots showing the overexpressed DEGs (TE-silencing gene pathways) between ovaries and testes:

Overexpressed DEGs in ovaries (left) and testes (right). Dotted lines indicate logFC ±2

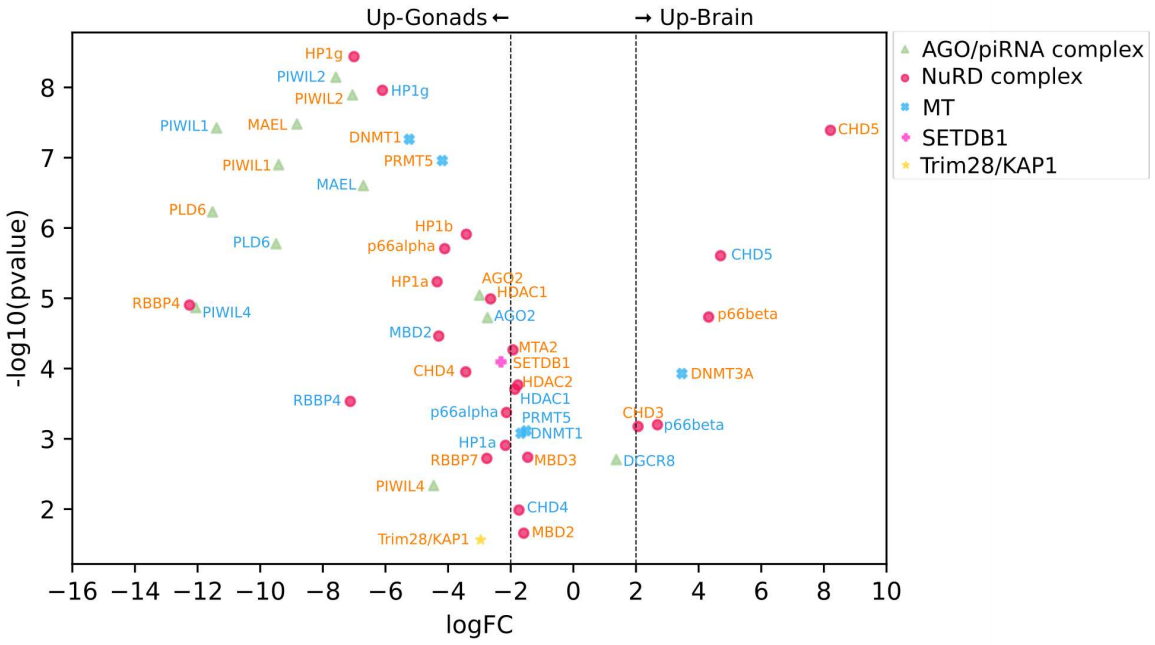


Figure 5. Volcano plots showing the overexpressed DEGs (TE-silencing gene pathways) between gonads and brain:

Overexpressed DEGs in gonads (left; orange: ovaries, blue: testes) and brain (right; orange: brain female, blue: brain male). Dotted lines indicate logFC ±2

Discussion

While it is now well-established that one of the drivers of genomic gigantism is the expansion and accumulation of transposons, the genomic fraction of TE copies alone does not capture the complex dynamics and conflicts occurring between the host genome and their mobile elements. This is because a significant portion of transposable elements in large genomes consists of fossil, truncated and degenerated copies that can no longer be actively mobilised. This is common not only in plants but also in animal genomes (Novák et al. 2020; Wang et al. 2021(b)).

Hence, within the scope of this study, we explored the activity of transposable elements and the host genome safeguard system in both somatic and germline tissues of *B. pachypus*, to deepen our understanding of TE expansion in large genomes.

Gonad-specific TE activity

Transposable elements exhibited higher gonad-specific activity in *B. pachypus* compared to the brain. In addition, TEs displayed a slight sex-biased expression pattern, with the most active retrotransposon families primarily expressed in the testes (LTR/Gypsy), while the most active DNA-TE families being prevalent in the ovaries (DNA/hAT and DNA/TcMar) (Figure 1). Nevertheless, in further exploring the TE activity between the two sexes, we found that the testes exhibited a significantly higher number of differentially expressed TEs and substantially larger fold changes compared to the ovaries (Figure 2). These findings highlight the male gonad as a notably permissive environment for TE expression, a pattern consistent with prior findings in the salamander *Ranodon sibiricus* (Wang et al. 2023), in the two grasshopper species *Locusta migratoria manilensis* and *Angaracris rhodopa* (Liu et al. 2022), and also in *Drosophila melanogaster* (Lawlor et al. 2021). In particular, these previous studies revealed how different dynamics operate and sometimes cooperate, contributing to increased TE expression in the male gonad.

In *D. melanogaster*, Lawlor et al. (2021) demonstrated that higher TE expression in males results from distinct TE dynamics on the sex-limited chromosome. Specifically, they reported increased TE expression in primary spermatocytes, which was co-expressed with Y-linked fertility factors. This was further confirmed by substantially

higher TE copy numbers in males compared to females, as predicted by TE insertions located on the Y-chromosome.

This enhanced TE expression in the testes may be associated with the accumulation of TEs on the Y-chromosome, as a consequence of the suppression of recombination (Muller's ratchet process) (Peona et al. 2021). Such a scenario is possible for *B. pachypus*, which has heteromorphic sex chromosomes with a male heterogamety system (XY). However, to explore this hypothesis, further analyses of the sex chromosome in *B. pachypus* using genomic data will be required.

On the other hand, in the salamander *R. sibiricus*, which has homomorphic sex chromosomes, Wang et al. (2023) similarly observed a twofold higher TE expression in the testes compared to the ovaries. Furthermore, this increased male expression appeared to correlate with specific higher expression of piRNA pathway genes, similar to what we have found here in our analysis of *B. pachypus*.

The observed patterns invoke the Red Queen hypothesis (McLaughlin and Malik 2017) to explain the intricate dynamic interplay between TE expansion and host genome repression strategies, characterised by an ongoing arms race between bursts of TE activity and subsequent enhancements in TE silencing mechanisms. Consequently, an evolutionary feedback loop ensues wherein TEs evolve to propagate themselves, and the host genome evolves to restore TE suppression.

Also supporting this hypothesis, a comparative study across 12 vertebrate species emphasised a positive relationship between expression levels of recent TEs and different TE silencing pathways in the male germline (Pasquesi et al. 2020).

Together, this suggests that the male gonads allow a more permissive genomic arena for TE expression.

Multifaceted dynamics between TE expansion and host genome regulation

The tangled interaction and evolution between transposon expansion and the host genome defence system remain poorly understood.

B. pachypus exhibited remarkable gonad-specific expression of distinct gene pathways involved in TE regulation within its large genome (Figure 3). Considering that the

germline is the principal route for the propagation and inheritance of TEs in subsequent generations, enhanced silencing defence systems are expected, especially in genomes with a large amount of TEs.

The piRNA pathway is widely recognised as the primary safeguarding system in the germline. While traditionally associated with testis-specific TE silencing, in *B. pachypus*, it has instead shown heightened expression in both gonads. Notably, the elevated expression of PIWI genes, coupled with substantial fold changes in the endonuclease PLD6 — responsible for cleaving long transcripts produced from piRNA clusters — along with MAEL, the nucleo-cytoplasmic shuttling protein, and SETDB1, activator of piRNA clusters (Soper et al. 2008; Biscotti et al. 2017; Wang et al. 2023), collectively suggest an intense activity associated with the secondary biosynthesis of piRNAs through the Ping-Pong cycle.

The Ping-Pong cycle requires ongoing expression of the piRNA clusters as well as target transposons. Essentially, the higher the transposon activity in the genome, the greater the probability of TEs leaping into a piRNA cluster region, thereby triggering the generation of novel antisense piRNAs. This amplification loop directs piRNA production toward transcriptionally active and highly mobilised transposons. From this perspective, the increased expression of TE mRNAs and TE-derived piRNA abundance, as we have found in *B. pachypus*, provides support for the rapid evolution of the piRNA system in counteracting TE propagation in this biological system.

Regarding TE-sequence-specific transcriptional repression, recent evidence indicates that TEs have been instrumental in shaping the evolution and diversification of zinc finger genes, which characterise the KZFP repressor system across metazoans (Wells et al. 2023). Wang et al. (2021) found a surprisingly high number of KRAB domains in amphibian genomes with an average of 675 domains per genome, despite the low number of species analysed.

In *B. pachypus*, we observed greater ovary-specific expression across the entire silencing complex (Figure 3). The elevated expression included TRIM28/KAP1, the pivotal scaffold protein responsible for recruiting the other heterochromatin transcriptional silencing factors: SETDB1, HP1, the NuRD complex, and DNA methyltransferases. Moreover, the strong directionality of the complex towards the

female gonad is corroborated by its overexpression in the ovaries when analysing differentially expressed pathways between the testes and the ovaries (Figure 4).

The same female-specific pattern was observed in the African lungfish (Wang et al. 2021(b)) and two large salamander genomes: *Cynops orientalis* and *Ranodon sibiricus* (Carducci et al. 2021; Wang et al. 2023).

This enhanced control in female gonads could also be linked to the involvement of these transcriptional gene pathways in oogenesis, as maternal effect genes that regulate oocyte gene expression. Several studies have also demonstrated the crucial role of SETDB1 and other KRAB and NuRD-related genes in oocyte meiotic progression and embryonic development in different species (Clough et al. 2014; Brici et al. 2017; Stäubli et al. 2021)

Interestingly, a recent study prompted a reevaluation of the primary role of metazoan KZFP as transcriptional silencers (Wells et al. 2023). In particular, this study encouraged the consideration of KZFP as potential genome stabilisers, preventing ectopic recombination through heterochromatin formation in the repetitive regions. This hypothesis is further supported by the presence of KZF genes in genomic regions marked by H3K9me3, suggesting their role in stabilising repeats rather than suppressing them transcriptionally. This would explain the high expression of TEs in the gonads of *B. pachypus* and the high number of copies expected to be retained in the genome. Moreover, it is now well established that TEs play a role in gene regulatory networks, and KZFPs have also been implicated as crucial contributors to their domestication (Chuong et al. 2017; Rosspopoff et al. 2023).

The related expansion of TEs and KZFPs in *B. pachypus* suggests a broader spectrum of dynamics that may have acted in the past and may still be active between TEs and the host genome, extending beyond a simple arms race dynamic. The growing evidence of the functional role of TEs in host adaptation and evolution delineates a more complex and multifaceted perspective on TE activity. TEs are no longer viewed as merely competing with the host genome cellular functions, but rather as integral components of a dynamic interaction, encompassing both detrimental and mutually beneficial effects. Further genomic analyses will be necessary to explore the number of KZF genes and their localisation in the genome of *B. pachypus*, contributing to a deeper understanding of the different evolutionary dynamics involved.

Conclusions

Our study contributes to understanding the complex dynamics acting between TE expansion and host genome silencing mechanisms in the large genome of *B. pachypus*.
Anurans show remarkable variation in genome size, ranging from 1 Gb in *Platyplectrum ornatum* to approximately 10 Gb in the *Bombina* genus. Nevertheless, they are characterised by less extreme genome sizes as those observed in Urodels or lungfish, suggesting that Anurans could serve as a valuable model system to investigate the complex evolutionary dynamics, either as an arms race or as a mutualistic advantage, between TEs and host genomes.
We found consistent differences in TE expression patterns between somatic and germline tissues, with male gonads exhibiting significantly higher transcriptional activity emerging as a more permissive environment for TE expression. Our analysis also revealed heightened activity of TE silencing mechanisms in both male and female gonads, offering new insights into the potential interacting dynamics between TE elements and the host genome. On one front, heightened TE activity is associated with increased activation of TE repressive mechanisms in the host. This is exemplified by the feedback amplification loop observed in the high activity of the piRNA Ping-Pong cycle, initiated by transcriptionally active transposons (as also observed by Wang et al. 2023). On the other hand, the higher expression of the KZFP system in the female gonad suggests its potential role as genome stabiliser rather than mere transcriptional silencers, potentially acting to safeguard DNA from ectopic recombination events (as hypothesised by Wells et al. 2023). Moreover, given the coevolutionary relationship between TEs and KZFPs during the evolutionary transition underlying the emergence of tetrapods, it becomes essential to explore a potential role of TEs in this process. TEs may have played pivotal roles in driving tetrapod-specific adaptations and innovations, with KZFPs emerging as crucial contributors to their domestication.

References

Almeida MV, Vernaz G, Putman ALK, Miska EA. 2022. Taming transposable elements in vertebrates: from epigenetic silencing to domestication. Trends Genet. 38:529–553.

Andreone F, Corti C, Sindaco R, Romano A, Giachi F, Vanni S & Delfino G. 2009. *Bombina pachypus*. The IUCN Red List of Threatened Species 2009: e.T54450A86629977.http://dx.doi.org/10.2305/IUCN.UK.2009.RLTS.T54450A11147957.en

Barbieri F, Bernini F, Guarino F, and Venchi A. 2004. Distribution and conservation status of *Bombina variegata* in Italy (Amphibia, Bombinatoridae). Ital. J. Zool. 71, 83–90. doi: 10.1080/11250003.2004.9525541

Biscotti MA et al. 2017. The small non-coding RNA processing machinery of two living fossil species, lungfish and coelacanth, gives new insights into the evolution of the Argonaute protein family. Genome Biol. Evol. 9:438–453.

Boissinot S, Sookdeo A. 2016. The Evolution of LINE-1 in Vertebrates. Genome Biol. Evol. 8:3485–3507.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30:2114–2120.

Brici D et al. 2017. Setd1b, encoding a histone 3 lysine 4 methyltransferase, is a maternal effect gene required for the oogenic gene expression program. Development. 144:2606–2617.

Bruno M, Mahgoub M, Macfarlan TS. 2019. The Arms Race Between KRAB-Zinc Finger Proteins and Endogenous Retroelements and Its Impact on Mammals. Annu. Rev. Genet. 53:393–416.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat. Methods. 12:59–60.

Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. 2019. rnaSPAdes: a de novo

transcriptome assembler and its application to RNA-Seq data. Gigascience. 8. doi: 10.1093/gigascience/giz100.

Canestrelli D, Cimmaruta R, Costantini V, Nascetti G. 2006. Genetic diversity and phylogeography of the Apennine yellow-bellied toad *Bombina pachypus*, with implications for conservation. Mol. Ecol. 15:3741–3754.

Canestrelli D, Zampiglia M, and Nascetti G. 2013. Widespread occurrence of *Batrachochytrium dendrobatidis* in contemporary and historical samples of the endangered *Bombina pachypus* along the Italian Peninsula. PLoS ONE8:e63349. Doi:10.1371/journal.pone.0063349

Carducci F et al. 2021. Investigation of the activity of transposable elements and genes involved in their silencing in the newt *Cynops orientalis*, a species with a giant genome. Sci. Rep. 11:14743.

Chalopin D, Naville M, Plard F, Galiana D, Volff J-N. 2015. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. Genome Biol. Evol. 7:567–580.

Chiocchio A et al. 2022. Brain de novo transcriptome assembly of a toad species showing polymorphic anti-predatory behavior. Sci Data. 9:619.

Choudhary MNK, Quaid K, Xing X, Schmidt H, Wang T. 2023. Widespread contribution of transposable elements to the rewiring of mammalian 3D genomes. Nat. Commun. 14:634.

Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. Nat. Rev. Genet. 18:71–86.

Clough E, Tedeschi T, Hazelrigg T. 2014. Epigenetic regulation of oogenesis and germ stem cell maintenance by the Drosophila histone methyltransferase Eggless/dSetDB1. Dev. Biol. 388:181–191.

Conesa A et al. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 21:3674–3676.

Ecco G, Imbeault M, Trono D. 2017. KRAB zinc finger proteins. Development. 144:2719–2729.

Ewels P, Magnusson M, Lundin S, Käller M. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 32:3047–3048.

Falcon F, Tanaka EM, Rodriguez-Terrones D. 2023. Transposon waves at the water-to-land transition. Curr. Opin. Genet. Dev. 81:102059.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39:W29–37.

Flynn JM et al. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proc. Natl. Acad. Sci. U. S. A. 117:9451–9457.

Frank JA et al. 2022. Evolution and antiviral activity of a human protein of retroviral origin. Science. 378:422–428.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 28:3150–3152.

Gertz EM, Yu Y-K, Agarwala R, Schäffer AA, Altschul SF. 2006. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. BMC Biol. 4:41.

Haley AL, Mueller RL. 2022. Transposable Element Diversity Remains High in Gigantic Genomes. J. Mol. Evol. 90:332–341.

Iwasaki YW, Siomi MC, Siomi H. 2015. PIWI-Interacting RNA: Its Biogenesis and Functions. Annu. Rev. Biochem. 84:405–433.

Jeong H-H, Yalamanchili HK, Guo C, Shulman JM, Liu Z. 2017. An ultra-fast and scalable quantification pipeline for transposable elements from next generation sequencing data. In: Biocomputing 2018. WORLD SCIENTIFIC pp. 168–179.

Kijima TE, Innan H. 2013. Population genetics and molecular evolution of DNA sequences in transposable elements. I. A simulation framework. Genetics.

195:957–967.

Lanciano S, Cristofari G. 2020. Measuring and interpreting transposable element expression. Nat. Rev. Genet. 21:721–736.

Lawlor MA, Cao W, Ellison CE. 2021. A transposon expression burst accompanies the activation of Y-chromosome fertility genes during Drosophila spermatogenesis. Nat. Commun. 12:6854.

Li B et al. 2014. Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biol. 15:553.

Liu X et al. 2022. Transposable element expansion and low-level piRNA silencing in grasshoppers may cause genome gigantism. BMC Biol. 20:243.

Martino G, Chiocchio A, Siclari A, Canestrelli D. 2022. Distribution and conservation status of threatened endemic amphibians within the Aspromonte mountain region, a hotspot of Mediterranean biodiversity. NC. 50:1–22.

McLaughlin RN Jr, Malik HS. 2017. Genetic conflicts: the usual suspects and beyond. J. Exp. Biol. 220:6–17.

Meyer A et al. 2021. Giant lungfish genome elucidates the conquest of land by vertebrates. Nature. 590:284–289.

Novák P et al. 2020. Repeat-sequence turnover shifts fundamentally in species with large genomes. Nat Plants. 6:1325–1329.

Ou S et al. 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 20:275.

Pasquesi GIM et al. 2020. Vertebrate Lineages Exhibit Diverse Patterns of Transposable Element Regulation and Expression across Tissues. Genome Biol. Evol. 12:506–521.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods. 14:417–419.

Peona V et al. 2021. The avian W chromosome is a refugium for endogenous retroviruses with likely effects on female-biased mutational load and genetic incompatibilities. Philos. Trans. R. Soc. Lond. B Biol. Sci. 376:20200186.

Peters L, Meister G. 2007. Argonaute proteins: mediators of RNA silencing. Mol. Cell. 26:611–623.

Playfoot CJ et al. 2021. Transposable elements and their KZFP controllers are drivers of transcriptional innovation in the developing human brain. Genome Res. 31:1531–1545.

Punta M et al. 2012. The Pfam protein families database. Nucleic Acids Res. 40:D290–301.

Quevillon E et al. 2005. InterProScan: protein domains identifier. Nucleic Acids Res. 33:W116–20.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 26:139–140.

Roessler K, Bousios A, Meca E, Gaut BS. 2018. Modeling Interactions between Transposable Elements and the Plant Epigenetic Response: A Surprising Reliance on Element Retention. Genome Biol. Evol. 10:803–815.

Rogers RL et al. 2018. Genomic Takeover by Transposable Elements in the Strawberry Poison Frog. Mol. Biol. Evol. 35:2913–2927.

Rosspopoff O, Trono D. 2023. Take a walk on the KRAB side. Trends Genet. 39:844–857.

Schrader L, Schmitz J. 2019. The impact of transposable elements in adaptive evolution. Mol. Ecol. 28:1537–1549.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31:3210–3212.

Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. 2016. TransRate: reference-free quality assessment of de novo transcriptome assemblies. Genome Res. 26:1134–1144.

Soper SFC et al. 2008. Mouse maelstrom, a component of nuage, is essential for spermatogenesis and transposon repression in meiosis. Dev. Cell. 15:285–297.

Sotero-Caio CG, Platt RN II, Suh A, Ray DA. 2017. Evolution and Diversity of Transposable Elements in Vertebrate Genomes. Genome Biol. Evol. 9:161–177.

Stäubli A, Peters AH. 2021. Mechanisms of maternal intergenerational epigenetic inheritance. Curr. Opin. Genet. Dev. 67:151–162.

Sun C et al. 2012. LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders. Genome Biol. Evol. 4:168–183.

Wang J et al. 2021 (a). Gigantic Genomes Provide Empirical Tests of Transposable Element Dynamics Models. Genomics Proteomics Bioinformatics. 19:123–139.

Wang J et al. 2023. Transposable element and host silencing activity in gigantic genomes. Front Cell Dev Biol. 11:1124374.

Wang K et al. 2021 (b). African lungfish genome sheds light on the vertebrate water-to-land transition. Cell. 184:1362–1376.e18.

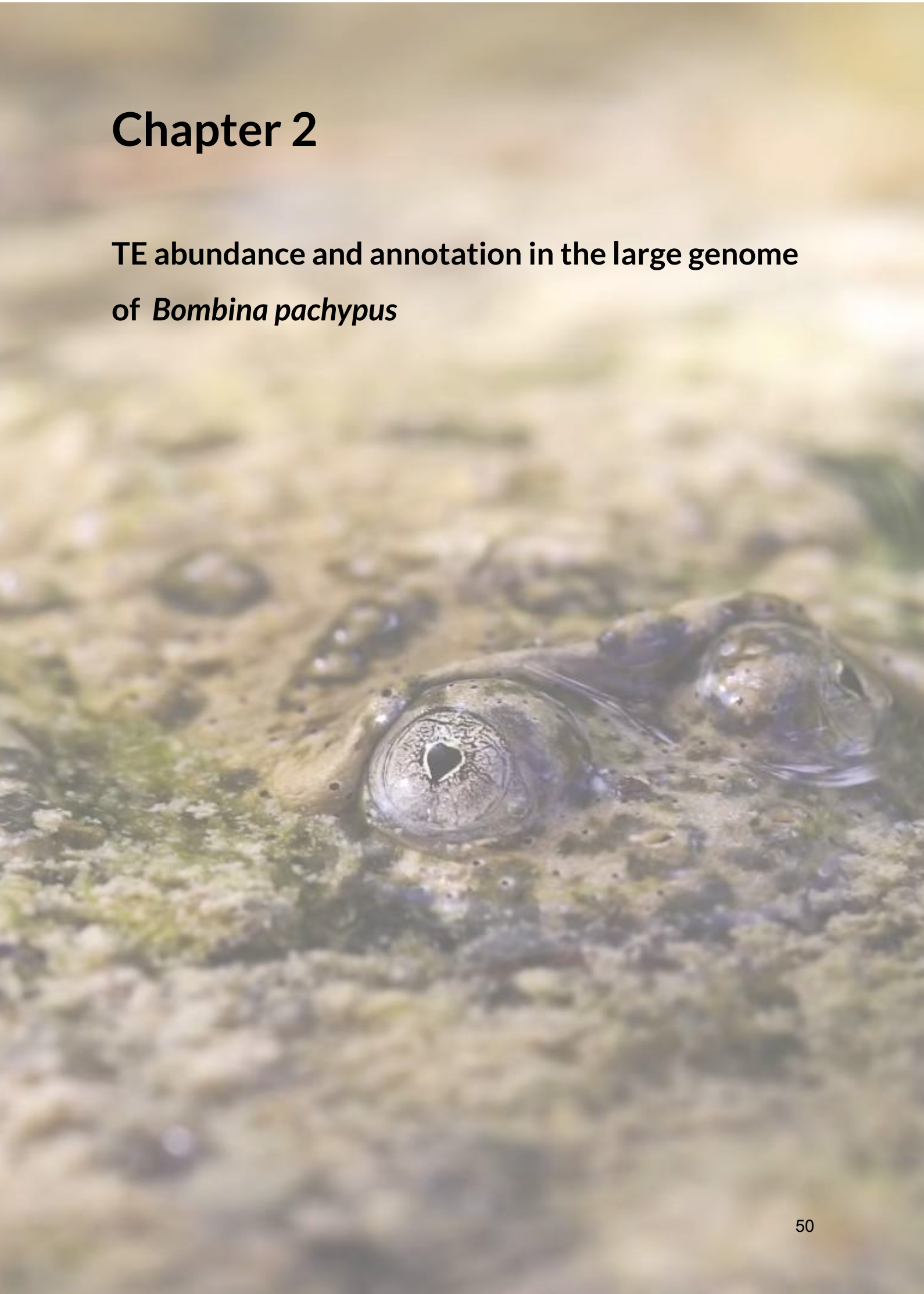
Wells JN et al. 2023. Transposable elements drive the evolution of metazoan zinc finger genes. Genome Res. 33:1325–1339.

Yan H, Bombarely A, Li S. 2020. DeepTE: a computational method for de novo classification of transposons with convolutional neural network. Bioinformatics. 36:4269–4275.

Zampiglia M et al. 2019. Drilling Down Hotspots of Intraspecific Diversity to Bring Them Into On-Ground Conservation of Threatened Species. Frontiers in Ecology and Evolution. 7. doi: 10.3389/fevo.2019.00205.

Chapter 2

TE abundance and annotation in the large genome of *Bombina pachypus*



Chapter 2

TE abundance and annotation in the large genome of *Bombina pachypus*

Authors:

Lorena Ancona[1], Roberto Biello[2], Alessio Iannucci[3], Andrea Benazzo[2], Claudio Ciofi[3], Daniele Canestrelli[4], Marco Gerdol[5], Samuele Greco[5], Francesca Raffini[6], Giorgio Bertorelle[2], Emiliano Trucchi[1]

Affiliations:

1 Department of Life and Environmental Sciences, Polytechnic University of Marche, Italy

2 Department of Life Sciences and Biotechnology, University of Ferrara, Italy

3 Department of Biology, University of Florence, Italy

4 Department of Ecological and Biological Sciences, Tuscia University, Italy

5 Department of Life Sciences, University of Trieste, Italy

6 Department of Biology and Evolution of Marine Organisms, Stazione Zoologica Anton Dohrn, Napoli, Italy

Author contribution

This chapter is based on the analyses of the *B. pachypus* reference genome which have been assembled by collaborators from the project Endemixit (P.I. Prof Bertorelle, University of Ferrara). Details of Materials and Methods and Results concerning the genome assembly are only summarised here.

Introduction

Genome sizes exhibit significant variation among different species, with differences exceeding 200,000-fold among eukaryotes (Gregory, 2001). Interestingly, this wide diversity appears to have no correlation with either organismal complexity or the number of genes—a phenomenon often referred to as the 'C-value enigma'.

In eukaryotes, genome size variation is primarily influenced by changes in the non-coding regions of the genome, particularly through variations in the proportion of repetitive and mobile elements, as well as the amount and length of introns (Lynch et al. 2011). Both adaptive and non-adaptive hypotheses have been proposed to elucidate this extreme variability in genome size evolution, although this still remains a subject of controversial debate.

Concerning the non-adaptive or nearly neutral models, Lynch and Conery (2003), proposed the Mutational Hazard Hypothesis (MHH) to account for the variation in the complex architecture of genomes. The underlying assumption of MHH is that all forms of DNA insertions may have slightly deleterious effects on organismal fitness, increasing the risk of accumulating deleterious mutations (mutational hazard). Consequently, the susceptibility of a genome to accumulate non-coding DNA is driven by the ratio between the mutation rate and the genetic drift. In populations with smaller effective population size (Ne), where the efficiency of purifying selection is reduced, slightly deleterious insertions may be more likely to spread and fix due to genetic drift, thereby leading to larger genomes.

On the contrary, adaptive hypotheses suggest that genome size variation is a trait under selection due to its direct phenotypic effect on organismal fitness (Cavalier-Smith, 1978,2005; Gregory and Hebert 1999). In particular, such hypotheses are based on the correlation of genome size with various organismal traits, including cell and nucleus size (Gregory, 2001), developmental time and life cycle complexity (Arnqvist et al. 2015).

However, several studies failed to prove the adaptive hypotheses. For instance, a comparative analysis on large amphibian genomes with different reproductive strategies and life cycles found no significant correlation between genome size and life history complexity in any of the three amphibian orders (Liedtke et al. 2018).

On the other hand, among vertebrates, some groups, like birds, have consistently tight genome size across genera and species (Kapusta et al. 2017), whereas other groups, such as lungfish and amphibians, exhibit significant variation in genome size, with examples of exceedingly gigantic genomes. If the non-adaptive mutational hazard hypothesis was correct, why do we not observe gigantic genomes more randomly across the tree of Life? Or are there specific molecular features, for example affecting transposition or deletion rates, making some classes of organisms more susceptible to genome size runaways in case of high drift - low selection conditions?

As already mentioned, amphibians exhibit considerable interspecific variation in genome size across their three orders, ranging from 1 Gb to 10 Gb in the anurans *Platyplectrum ornatum* and *Bombina bombina*, respectively, to the maximum size known for caecilians (*Siphonops annulatus*, 13.7 Gb) and urodels (*Necturus lewisi*, 120 Gb). Importantly, the proliferation and accumulation of transposable elements have been identified as the main cause of their large and variable genomes, with genomes being predominantly composed of a substantial amount of TEs. For instance, the strawberry poison frog *Oophaga pumilio* has a genome consisting of 70% TEs, with phylogenetic evidence suggesting horizontal transfer for some elements (Rogers et al. 2018). Similarly, the caecilian *Ichthyophis bannanicus* exhibits 78% of its 12 Gb genome consisting of repeated sequences, primarily dominated by DIRS (Wang et al. 2021).

Among the three orders, anurans show relatively moderate genome sizes compared to urodels and gymnophions, but the range of variation from small-compact genomes in the order of 1 Gb to rather large up to 10Gb makes them ideal models for investigating genome size evolution and genomic gigantism. However, the evolutionary mechanisms underlying anuran genome size variation and their genomic characterisation remain poorly understood, primarily due to the computational challenges associated with assembling their large and repetitive genomes (Sun et al. 2020; Zuo et al. 2023).

By investigating the genomic distribution patterns of TEs and their ancient and recent expansion dynamics, here, we study the mechanisms underlying genome gigantism in the Italian endemic Apennine yellow-bellied toad, *Bombina pachypus*, featuring one of the largest genomes among anurans.

After assembling a chromosome-level genome for this species, we first describe the

abundance, diversity and genome localisation of the different TE families in *B. pachypus* in comparison with its closest species, *B. bombina*, and eight other anuran species for a range of genome sizes from 1 to 10 Gb. Then, we use TE expression data for *B. pachypus* and *de novo* sequenced genomes from different populations of this species to investigate the most recent dynamics of TE activity.

In fact, *B. pachypus* presents a clear genetic pattern of "southern richness and northern purity" (Hewitt, 2000), characterised by smaller and less genetically diverse populations in the Northern areas of the Italian Apennines compared to their southern counterparts in Aspromonte massif in Calabria region. The species is then structured into two main genetic clusters: the southern cluster, which has been the putative glacial refugia for the species in the past, representing the hotspot of genetic diversity (i.e., population core); and the northern cluster, generated by post-glacial range expansion (i.e., population edge), exhibiting lower genetic variability compared to the southern cluster (Canestrelli et al. 2006). Core and edge populations are expected to have experienced different demographic trajectories during the range expansion, with the edge characterised by lower population size, higher genetic drift and lower selection efficiency than the more stable core population. Genomic data from the two extremes of the species range provide therefore a unique opportunity to investigate the very recent dynamics of TE expansion and to understand the impact of selection on TE accumulation and genome expansion.

With this study, we seek to shed light on the following questions:

1) What role do transposable elements play in the evolution of genome size in anuran species with large genomes? 2) Is there a direct correlation between genome size and TE abundance? 3) What is the distribution and diversity of TEs in the large genome of *B. pachypus*? 4) What are the recent dynamics of transposition in populations characterised by different demographic histories?

Materials and Methods

De Novo Genome Assembly

One male individual of *B. pachypus* was sampled in the South of Italy (Aspromonte) (permit number 20824, 18-03-2020). The individual was immediately frozen in liquid nitrogen to preserve the integrity of nucleic acids. High molecular weight DNA was isolated from phalanx tissue using the Nanobind Tissue big DNA kit (Circulomics Inc., Baltimore, USA). DNA quality and fragment length were checked in a pulse field gel electrophoresis and DNA concentration was measured with fluorometric and spectrophotometric assays using a Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, California, US) and a TECAN Nanoquant Infinite 200 Pro (Tecan Mannedorf, Switzerland), respectively. Fragments of 35,000 bp length were selected using a Blue Pippin device (Sage Science; Beverly, MA, USA). Isolated fragments were used to prepare the DNA library with a SMRTbell express template prep kit 2.0 (Pacific Biosciences, Menlo Park, California, US) according to manufacturer's protocols. The library was run on eight PacBio SMRT Cells 1M in continuous long-read sequencing (CLR) mode on a PacBio Sequel platform.

To confirm the chromosomal structure of our assembly, a karyotype for the Apennine yellow-bellied toad was generated using a cultured cell protocol. A gingival tissue biopsy was obtained from the male individual. Cells were cultured in a medium composed of 50% RPMI1640 and 50% Iscove's Modified Dulbecco's Medium, supplemented with 10% fetal bovine serum, 1% penicillin (10,000 units/mL) - streptomycin (10 mg/mL), 1% gentamycin sulfate (10 mg/mL), 0.5% amphotericin B (250 mg/mL) and 1% L-glutamine (200 mM) and incubated at 37 °C with 5% CO2. Chromosome preparations were made following standard procedures (Stanyon and Galleni 1991). In brief, after 4 h of treatment in 0.01 ng/mL colcemid, the cells are removed by standard trypsination and placed in a 15 mL tube. Cells are then centrifuged at 10,000 g, surnatant is removed and substituted with a 1:1 mixture of 0.075M KCl and 0.4% sodium citrate (hypotonic treatment). After a 20 min exposure at 37 °C the cells are pelleted by centrifugation and fixed in methanol:acetic acid fixative

(at a ratio of 3:1). Slides are then prepared by dropping metaphases with a Pasteur pipette onto a clean glass microscope slide. Diploid number and chromosome morphology were determined from the analyses of 20 mitotic cells stained with DAPI. Karyotype was arranged according to the standard ursid karyotype set (CIT).

Part of the cultured cells were harvested and sent to Dovetail Genomics (Scott's Valley, CA) to construct chromatin conformation capture libraries using the Omni-C kit from Dovetail Genomics. Omni-C libraries were sequenced paired-end on an Illumina NovaSeq 6000 System using a 300-cycle Reagent Kit v1.5.

Wtdbg2 v2.5 (Ruan and Li, 2020) was used to assemble the reads with parameters "-p 21 -S 4 -s 0.05 -L 5000 -g 10g". The initial draft assemblies were polished by first mapping PacBio subreads to the genome using pbmm2 v1.4.0 (https://github.com/PacificBiosciences/pbmm2) and then error correction was performed using gcpp v2.0.0 (https://github.com/PacificBiosciences/GenomicConsensus).
Additionally, purge_dups v1.2.3 (Guan et al. 2020) was used to remove haplotigs and contig overlaps in the resulting assembly. A first round of scaffolding was performed, mapping HiC data to the contigs with Chromap v0.2.4 (Zhang et al. 2021) and then using YaHS v1.2a.1 (Zhou et al. 2023) for proximity-ligation-based scaffolding. To improve the assembly, we applied the reference-guided software RagTag v2.0.1 (Alonge et al. 2022) using the genome of *B. bombina* (GCF_027579735.1) to scaffold the contigs. Finally, we performed another round of scaffolding, mapping again the HiC data to the scaffolded assembly with Chromap v0.2.4 and then using YaHS v1.2a.1. Hi-C contact maps were generated with JuicerTools v.1.9.9 (Durand et al. 2016) and used for manual curation of the scaffolded assembly in Juicebox v.1.11.08 (Durand et al. 2016).
The completeness of the assembly was assessed with Compleasm v0.2.4 (Huang and Li 2023), using the tetrapoda_odb10 database. The assembly base QV was calculated with Merqury v1.3 (Rhie et al. 2020), with a k-mer database constructed from short reads using Meryl v1.4 (Rhie et al. 2020).

Detection and annotation of TEs

To identify and annotate TEs in the genome assembly of *B. pachypus*, we initially constructed a *de novo* repeat library using the Extensive *de novo* TE Annotator (EDTA) v1.9.9 (Ou et al. 2019), a *de novo* pipeline that combines a suite of best-performing packages and includes a final step with RepeatModeler (Flynn et al. 2020), to generate a library of high-quality non-redundant TE sequences. Subsequently, we refined the library with DeepTE (Yan et al. 2020), which employs convolutional neural networks to classify unknown elements at the order and superfamily levels. Following this step, we used RepeatMasker v4.1.2 (Smit et al. 2013-2015) with the final TE library to annotate and mask the assembled genome.

Then, we parsed the RepeatMasker output file using the RM_TRIPS script (https://github.com/clbutler/RM_TRIPS) to exclude repeats not classified as TEs, merge closely localised TE fragments with matching identity, and eliminate fragments shorter than 80 bp (default script settings). Then, we filtered the RM_TRIPS output for TE orders and families, calculating their respective abundances and standardising them by determining the percentage of the total genome length represented by TEs.

Transposable element family diversity

To test whether TEs diversity was inversely correlated with genome size (Elliott and Gregory, 2015), diversity of the overall genomic TE community was measured, using both the Simpson's and Shannon diversity indices (Simpson, 1949; Shannon 1948).

We considered the different TE families as "species'' and the total numbers of base pairs for each TE family as individuals per "species''. All the unknown repeats (i.e. Unknown TEs; LTR/nc) were excluded from the analysis, as were TEs that could only be annotated down to the level of Class. Simpson's diversity index is expressed as the variable D, calculated by: $D = \frac{\sum n\,(n-1)}{N\,(N-1)}$. D is the probability that two individuals at random pulled from a community will be from the same species. We reported the more intuitive Gini-Simpson's index, expressed as 1—D. The Shannon's diversity index is

represented by the variable H, which is calculated by $H = -\sum_{i=1}^{s} pi\ lnpi$. The higher the value of H, the greater the diversity.

Genomic localisation of TEs

In order to detect the genomic localisation of TEs, we first generated the TE annotation gtf file from the RepeatMasker (.out) output file, using the makeTEgtf.pl script (https://labshare.cshl.edu/shares/mhammelllab/www-data/TEtranscripts/TE_GTF/).
After, Bedtools intersect (Quinlan and Hall, 2010) was employed to detect TEs located in intragenic and intergenic regions of the genome, using TE and genome annotation files. The relative abundances of the different TE families in intra- and intergenic regions were filtered and compared to the proportion of genic and intergenic regions of the genome (Supplementary Table S1). To this end, we assumed that the average TE length for each family was the same in both intra- and intergenic regions.
In detail, we first calculated the ratio between the abundance of TEs (i.e. number of TEs) in intra- and intergenic regions for each different family (R1). Secondly, we calculated the ratio between the number of bases covered by intra- and intergenic regions of the genome (R2). Finally, we used the formula $1 - (R1 \div R2)$ to obtain the enrichment of TE families in intra- and intergenic regions of the genome, relative to the background content of the genome.

Amplification history of TEs

To summarise the overall amplification history of TEs, we utilised the RepeatMasker script calcDivergenceFromAlign.pl with the RepeatMasker (.align) output file. This script calculates the Kimura distances between TE genome copies and their respective TE consensus sequences from the library, providing a measure of the historical dynamics of TE expansion. Histograms (so-called *TE landscapes*) were plotted for each family.
Additionally, to investigate the correlation between TE family abundances and expressions, we compared the genomic abundances with transcriptome expressions. In particular, we used TE abundances annotated on the genome of *B. pachypus* (as

previously detailed) along with available TE expression data for three different tissues of *B. pachypus:* brain, ovary, and testis (Ancona et al. (submitted); Chapter 1 of this thesis).

To compare the two distributions, we calculated the percentages of abundance and expression of each TE family relative to the total number of TEs present in the genome and transcriptome, respectively. Specifically regarding the transcriptome, we computed the average TE expression across all three tissues.

Recent transposition of TEs in two different populations of *Bombina pachypus*

To investigate the recent dynamic of transposition and the impact of selection on TE expansion, we characterised the abundance of TEs in two populations of B. *pachypus* with markedly different effective population size (Ne): one from the southern refugium and one from the margin of the northern expansion range.

Specifically, we collected a total of 20 individuals, 10 individuals from the South of Italy (Massetti Pollino; F. Argentino Pollino; Aspromonte) and 10 individuals from the North of Italy (Bagno di Romagna). The samples were preserved in 96% ethanol at -20 °C and the total DNA was extracted using the PureLink Genomic DNA Mini Kit (Invitrogen). DNA integrity was assessed by 1.5% agarose gel electrophoresis and DNA concentration was measured using a Qubit 4 fluorometer Broad Range Assay (Invitrogen). Short-read genomic libraries were constructed using a Illumina DNA PCR-Free Prep Kit (Illumina) according to the manufacturer's protocol. Target coverage was 10-15× for all samples. Libraries were sequenced paired-end on an Illumina NovaSeq 6000 System using a 300-cycle S2 Reagent Kit v1.5.

To the scope of TE quantification per individual, we subsampled 10 million reads for each sample and performed trimming with bbduk (ftr=149; minlength=150; http://jgi.doe.gov/data-and-tools/bb-tools/) to ensure uniform read length. RepeatMasker v4.1.2 was used with the *B. pachypus* genome TE library (as described before) to annotate TEs in each individual. RepeatMasker output files were then parsed and filtered using the RM_TRIPS script as described above.

For each population, we calculated the average percentage of TE families across 10 individuals and plotted the differences in abundance between the two populations for each TE family. We then performed the Mann-Whitney U test (Mann and Whitney, 1947) and the Kolmogorov-Smirnov test (Massey 1951) to statistically examine the differences in abundance between the two populations.

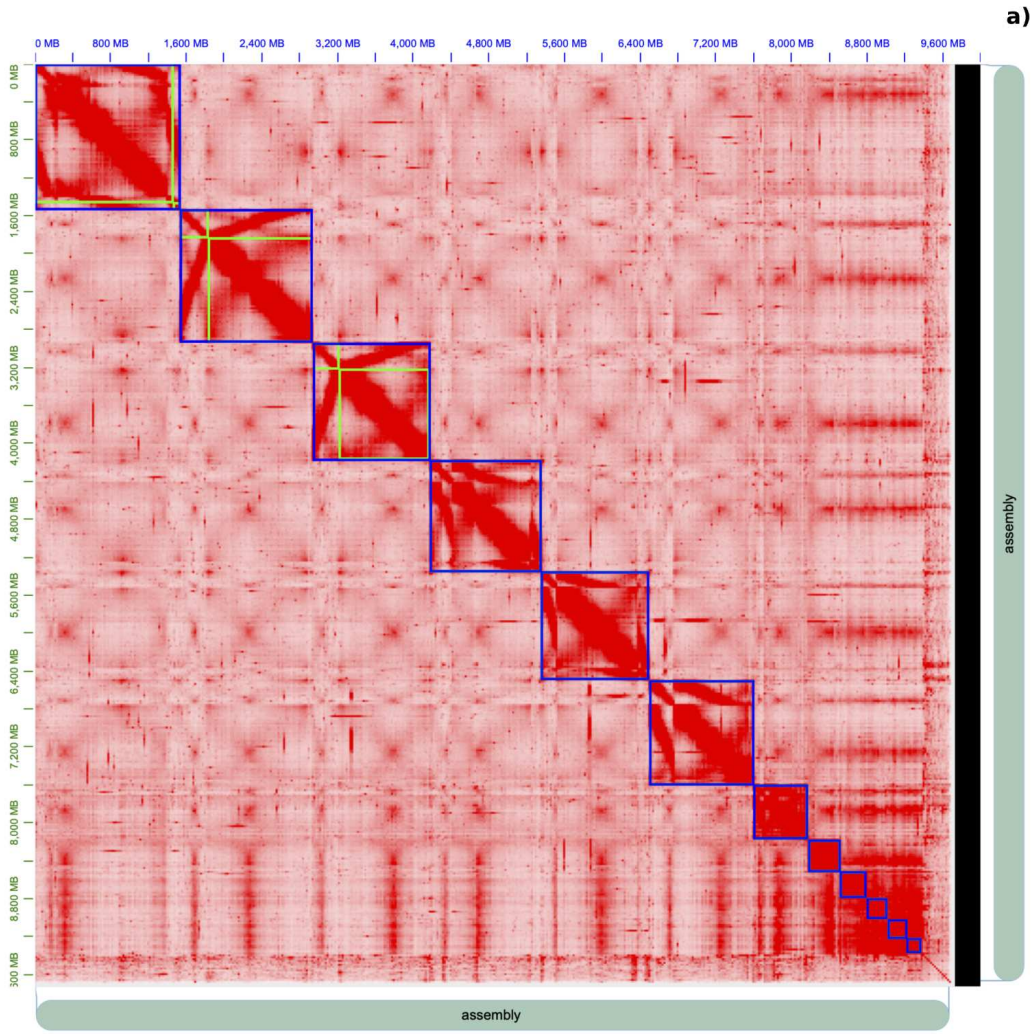
TE abundance among anuran genomes

With the aim of investigating the contribution of TEs to genome size evolution in Anura, we selected whole-genome sequences of ten anuran species (*Platyplectrum ornatum*, *Xenopus tropicalis*, *Dendropsophus ebraccatus*, *Xenopus laevis*, *Leptobrachium leishanense*, *Discoglossus pictus*, *Gastrophryne carolinensis*, *Bufo bufo*, *Bombina pachypus* and *Bombina bombina*) ranging in size from 1 to 10 Gb. Additionally, we included a caecilian species, (*Rhinatrema bivittatum*) with a genome size of 5 Gb, as an outgroup.
Samples information, including accession number, sequencing technology and assembly level are available in (Supplementary Table S6).
Considering the significant impact of genome assembly quality on TE detection and annotation, we ensured that all selected genome assemblies were generated using long-read sequencing technology and had similar assembly levels. Nine out of eleven assemblies were at the chromosome level, while the remaining two were at the scaffold level (*Platyplectrum ornatum* and *Dendropsophus ebraccatus*).
We identified and annotated TEs in each genome using the same pipeline described above, and tested for a correlation between genome size and TE abundance using linear regression. Finally, we measured the diversity of the overall genomic TE community in each genome, using both the Simpson's and Shannon diversity indices.

Results

TE diversity and distribution in the genome of *Bombina pachypus*

We obtained a chromosome-level genome assembly comprising 12 chromosomes and 16,500 scaffolds (Figure 1). The *B. pachypus* genome assembly has a total genome size of 9.69 Gb, with a scaffold N50 of 1,179 Mb and a scaffold L50 of 4 (Table 1).
Compleasm assessment of universal tetrapoda genes showed a completeness value of 84.29%. Among these, 71.26% were identified as single-copy genes, while 1.11% were identified as duplicated genes and 11.92% as fragmented genes.



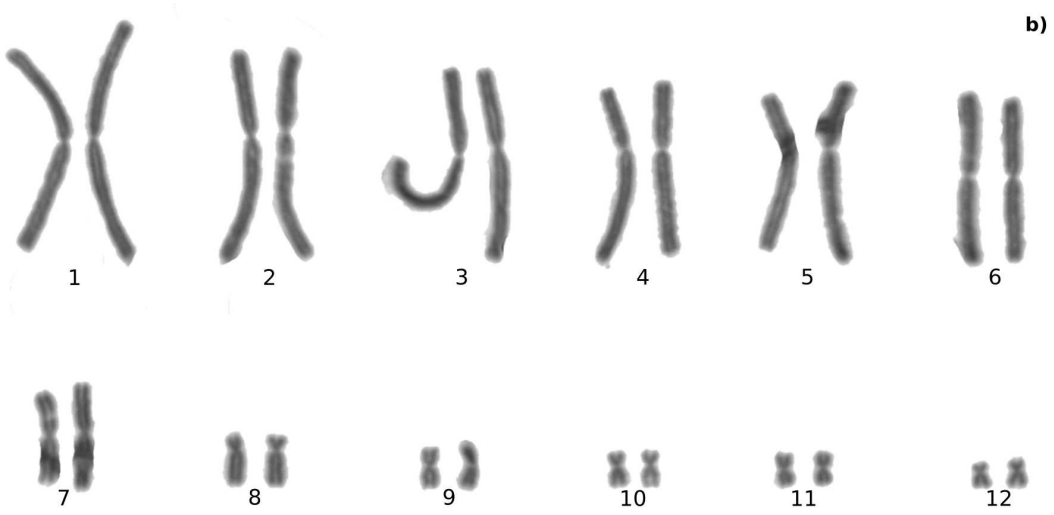


Figure 1. Hi-C contact map and karyotype of *B. pachypus* genome:

a) The genomic Hi-C contact map illustrates interactions between distinct regions of the genome. Grid cells represent the intensity of interactions between genomic regions, with darker colours indicating more frequent interactions. Highlighted squares represent the 12 chromosomes. b) Karyotype of *B.pachypus*.

Table 1. Genome assembly summary statistics

GENOME STATISTICS	VALUES
ASSEMBLY SIZE (MB)	9,689
NUMBER OF SCAFFOLDS	16,500
LARGEST SCAFFOLD (MB)	1,524
SCAFFOLD N50 (MB)	1,179
SCAFFOLD L50	4
SCAFFOLD N90 (MB)	279
SCAFFOLD L90	9

Table 2. TE classification and abundance

ORDER	FAMILY	TE LENGTH (BP)	PERCENTAGE OF THE GENOME
LTR	BEL	11032269	0,11
	COPIA	95082345	0,98
	ERV	320841455	3,31
	GYPSY	941938099	9,72
	LTR/nc	166612762	1,72
DIRS	DIRS	251804648	2,60
LINE	LINE/I	20584054	0,21
	LINE/Jockey	13999233	0,14
	LINE/L1	458382512	4,73
	LINE/R2	41040306	0,42
	LINE/RTE	45524912	0,47
	LINE/nc	75874428	0,78
SINE	SINE	1641285	0,02
PLE	PLE	251716771	2,60
	nLTR/nc	71017466	0,73
	ClassI/nc	207875225	2,15
TIR	DNA/CACTA	116822737	1,21
	DNA/Harbinger	112686771	1,16
	DNA/hAT	1957493356	20,20
	DNA/Mutator	195086565	2,01
	DNA/PiggyBac	19472228	0,20
	DNA/P	197389	0,00
	DNA/TcMar	467595527	4,83
MITE	DNA/MITE	144452390	1,49
nMITE	DNA/nMITE	709670250	7,32
HELITRON	DNA/Helitron	36666921	0,38
	ClassII/nc	4225290	0,04
	Unknown TEs	497131273	5,13
	Total_RetroTE (ClassI)	2974967770	30,70
	Total_DNATE (ClassII)	3764369424	38,85
	Total_TEs	7236468467	74,69

Transposable elements constitute 74.69% of the *B. pachypus* genome assembly. Class II of DNA transposons is the most abundant class, representing 38.85% of the genome, followed by Class I of Retrotransposons, which account for 30.70% of the genome (Figure 2, Table 2).

Among the ClassII of DNA TEs, the hAT family shows the highest abundance (20.20% of the genome), followed by nMITE (7.32%) and TcMar (4.83%) families. Among the ClassI of retrotransposons, the Gypsy family has the second highest abundance in the genome (9.72%), followed by L1 (4.73%), ERV (3.31%), DIRS (2.60%) and PLE (2.60%) families (Figure 2, Table 2).

The diversity of the overall genomic TE community was measured using both Simpson's and Shannon's diversity indices, resulting in a Gini-Simpson Index (1-D) value of 0.85 and a Shannon Index (H) value of 2.27.

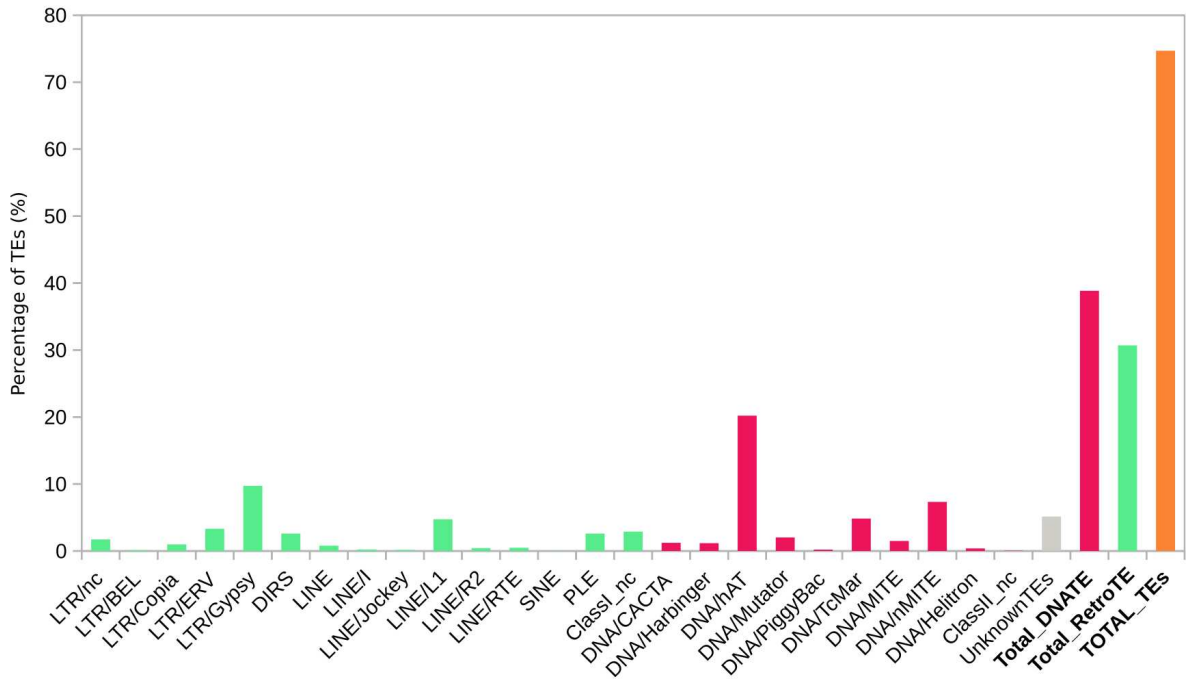


Figure 2. TE abundance in *B. pachypus* genome:

Total amount and relative proportions of TE families in the genome of *B. pachypus*. Total_DNATE: total of all DNA transposons represented by red bars (from DNA/CACTA to ClassII_nc families); Total_RetroTE: total of all Retrotransposons represented by green bars (from LTR/nc to ClassI_nc families). Nc: unclassified elements at the family level.

After identifying and annotating TEs in the *B. pachypus* genome, we further investigated their genomic localisation in intragenic and intergenic regions using TE and genome annotation data. More specifically, our aim was to identify TE families that are enriched in intra- and intergenic regions, relative to the background content of the genome.

From a total of more than 20 million TEs (20,680,244) annotated in the genome, we identified 7,924,404 TEs (38.32%) in intragenic regions and 12,799,101 TEs (61.89%) in intergenic regions (Supplementary Table S1). It is important to note that a single element may be inserted in multiple regions, which explains why the number of TEs identified in both intragenic and intergenic regions exceeds the total number of TEs annotated in the genome. The proportion of total TEs in intragenic and intergenic regions (0.62) is in line with the overall proportion of genic and intergenic region lengths in the genome (0.64). Regarding the enrichment of TEs in intra- and intergenic regions, we found that most of TE families were slightly more present in intergenic regions. However, the DIRS elements exhibited greater enrichment in intragenic regions.

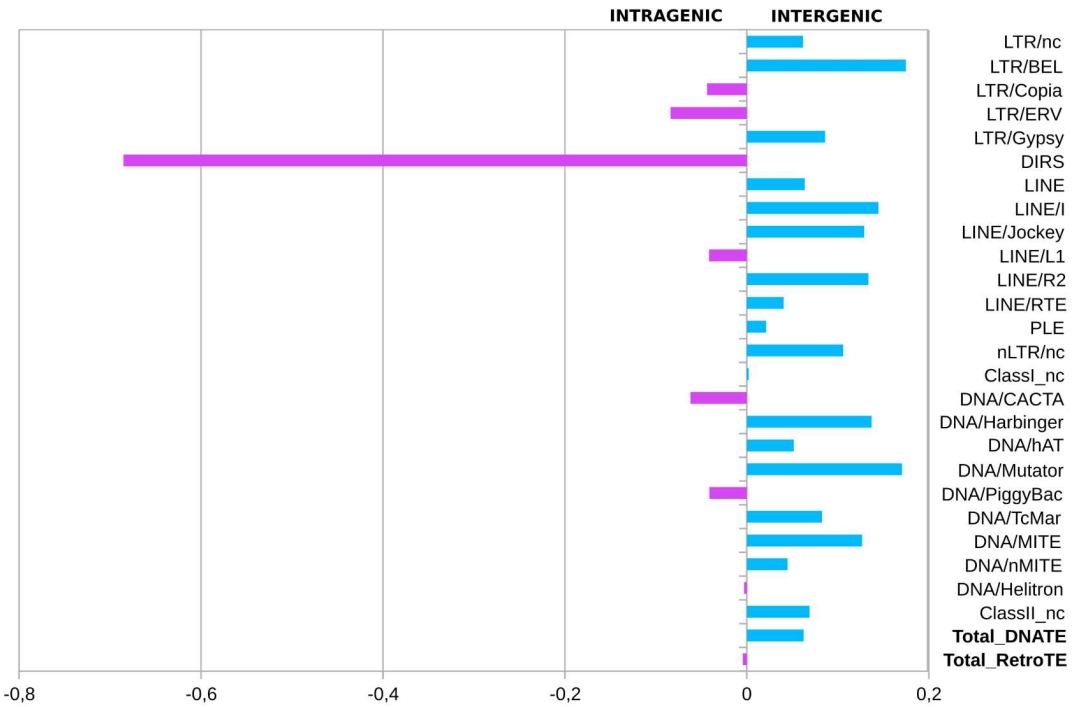


Figure 3. TE enrichment:

Enrichment of the different TE families in intragenic (left) and intergenic (right) regions of the genome. The X-axis shows the ratio between number of TEs and the length in base pairs of intragenic and intergenic regions (1 - R1/R2; see Methods).

TE amplification history

We studied the amplification history of the most abundant families of TEs from both Class I and Class II. Their landscapes of abundance throughout different levels of genetic differentiation were plotted to represent the genetic distances between TE genome copies and their respective ancestral TE sequences.

All class I families (LTR/Gypsy; LTR/ERV; DIRS; LINE/L1; PLE) show bimodal distributions, characterised by an ancient burst of expansion (i.e. 40 substitutions, Kimura distance), followed by further, relatively recent, expansions. In contrast, Class II families (DNA/hAT; DNA/nMITE; DNA/TcMar) exhibit unimodal distributions with a single wave of expansion, probably indicative of a more constant transpositional history (Figure 4). This single wave of transposition of Class II families appears to overlap the time frame of the second wave of transposition of the Class I families.

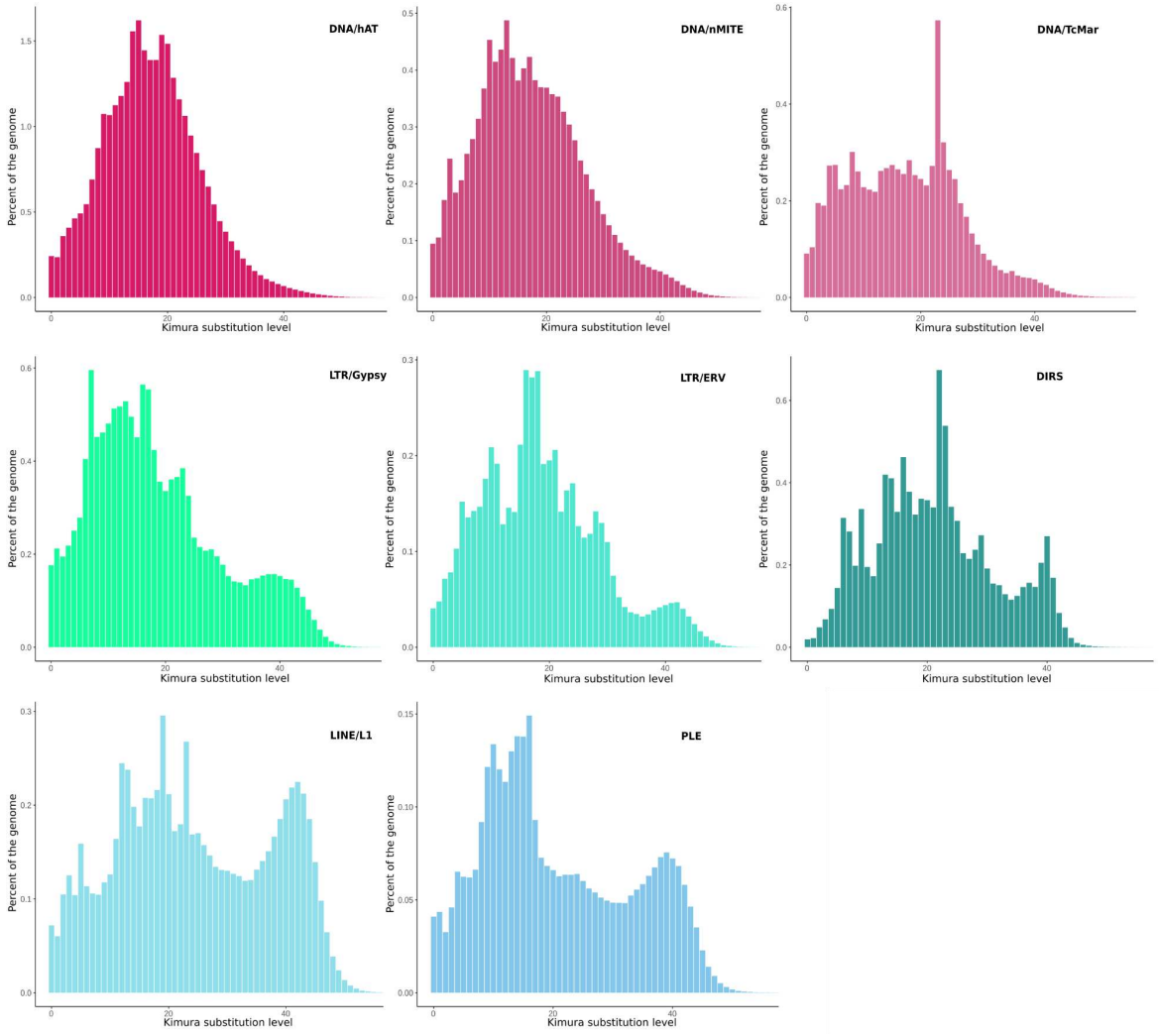


Figure 4. TE landscapes:

TE amplification history plots of the most abundant TE families in *B. pachypus*; Class II (first row), Class I (second and third row) The Y-axis shows the genomic coverage of TE families, the X-axis shows the number of substitutions (Kimura distances).

To further investigate the dynamics of TE expansion in the large genome of *B. pachypus*, we compared the genomic abundances with the transcriptome expressions to determine whether there was a linear or non-linear correlation. Specifically, we aimed to ascertain whether the most abundant TE families were also the most expressed.

The linear regression showed a R^2 value of 0.31 (p-value = 0.004), indicating that approximately 31% of the variance in transcriptome expression can be explained by genome abundance. Interestingly, the most abundant families are not the most expressed. For instance, the DNA/hAT family, which constitutes the most abundant family in the genome (27% of the total abundant TEs), is not among the most expressed families (9% of the total expressed TEs). Conversely, the LTR/Gypsy family, although not the most abundant (13% of the total abundant TEs), exhibits the highest expression levels (35% of the total expressed TEs) (Figure 5; Supplementary Table S2-S3). This suggests that while DNA transposons may possess a higher number of copies in the genome, they are not necessarily the most active or highly expressed.

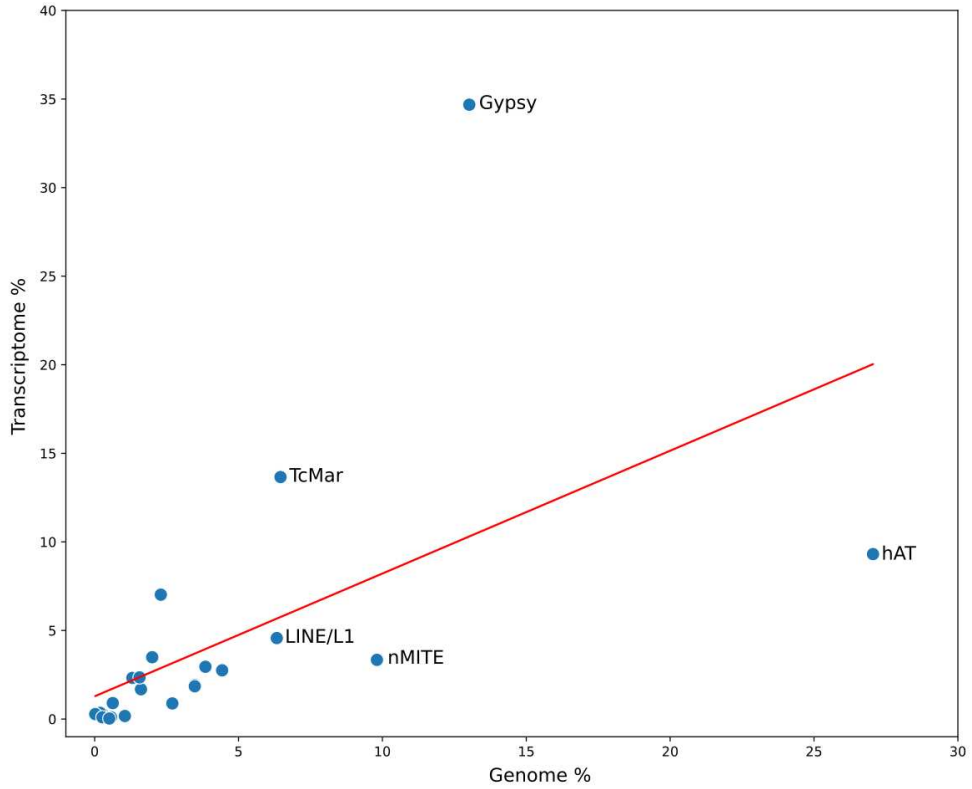


Figure 5. Correlation between TE abundance and expression:

Scatterplot showing the correlation between TE abundance in the genome and TE expression in the transcriptome. Each point represents a TE family, with the x-axis indicating the percentage of TE elements in the genome and the y-axis indicating the percentage of TE elements expressed in the transcriptome. The red line represents the linear regression fit to the data.

Recent dynamic of transposition in two populations of *Bombina pachypus*

In order to explore the recent dynamics of transposition in *B. pachypus*, we evaluated the abundance of TEs in two populations with markedly different effective population size (Ne). One population originated from the southern genetic cluster, characterised by higher genetic diversity, while the other originated from the northern genetic cluster, the result of post-glacial re-colonization of the Italian peninsula from the southern cluster, and characterised by lower genetic diversity. This approach also enabled us to evaluate the impact of drift and selection on TE expansion dynamics.

We found slightly higher TE abundance in the Northern population, which is statistically significant despite the low number of individuals analysed (MWU test p value=0.0009; K–S test p value=0.002) (Figure 6; Supplementary Table S5).

Elements from both DNA transposons and Retrotransposons classes appear to have had recent transposition activity. DNA/hAT, LTR/Gypsy and DIRS elements displayed a more significant difference in abundance in the northern population. Conversely, LTR/Copia and LINE/RTE elements showed a slightly higher abundance in the southern population.

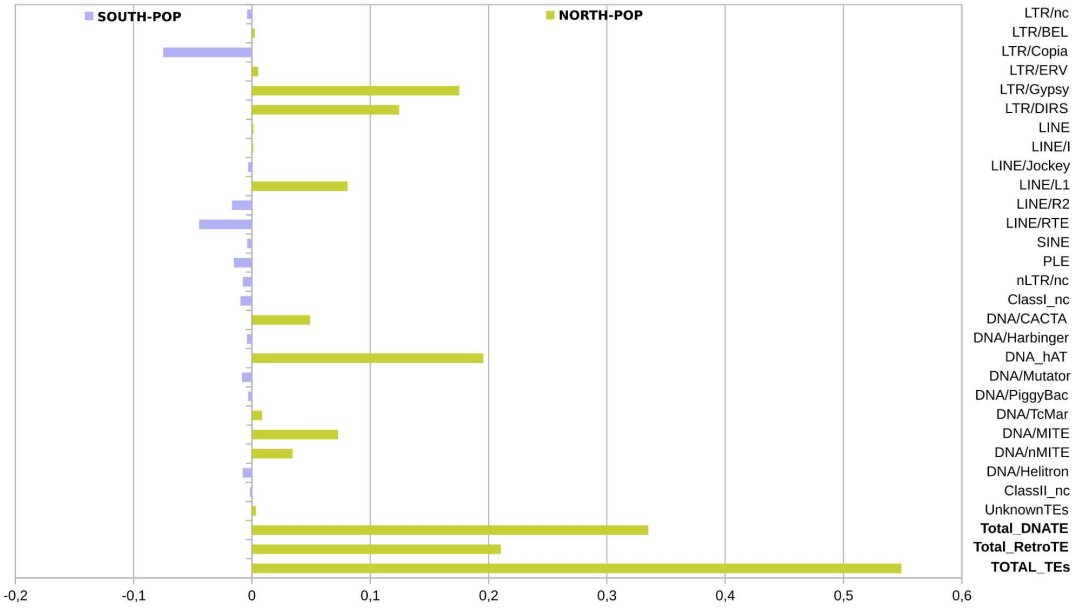


Figure 6. Differences in TE abundance in two different populations of *B. pachypus*:

Differences in TE percentage abundance for each family between the northern and southern populations.

Comparative analysis of TEs among anurans

With the objective of exploring the contribution of TEs to genome size evolution in Anura, we conducted a comparative analysis of TE abundance across ten anuran species with genome sizes ranging from 1 to 10 Gb. Additionally, we included a caecilian species (*R. bivittatum*) with a 5 Gb genome as an outgroup.

Our analysis unveiled a significant correlation between genome size and TE abundance (R^2 = 0.99; p-value = $1.2600228013683217e^{-10}$), highlighting a positive association between larger genome sizes and heightened TE content (Figure 7).

We observed substantial variations in TE abundance between the two extremes of the studied species. Specifically, we detected a TE abundance of 20% in *P. ornatum*, characterised by a genome size of 1.1 Gb, contrasting with a notably higher TE abundance of 77% in *B. bombina*, within its 10 Gb genome (Figure 8). The drastic reduction in TE content observed in *P. ornatum* appears to primarily involve LINE and LTR elements, compared to the other species (Figure 8).

It is noteworthy that all analysed anuran genomes, except for *B. bombina*, exhibit a higher content of Class II of DNA transposons compared to retrotransposons, with the former being twice as abundant in six of the ten species (*P. ornatum*, *D. ebraccatus*, *X. laevis*, *L. leishanense*, *D. pictus*, *B. bufo*) (Figure 8; Supplementary Table S7).

Across all anuran genomes, the hAT family represents the most abundant family among Class II of DNA transposons, followed by nMITE and TcMar families. Within ClassI of retrotransposons, the Gypsy family shows the highest content, followed by ERV and L1 elements.

B. pachypus exhibits TE composition patterns similar to the other smaller anura genomes. In contrast, the closest species *B. bombina*, while displaying a comparable total TE content, has a genome more enriched with retrotransposons than DNA transposons (Figure 8). Specifically, the total abundance of LINE elements in *B. bombina* is nearly double that observed in *B. pachypus*. This heightened prevalence of retrotransposons is a characteristic trait shared by the genomes of urodeles and caecilians, as evidenced by the two-lined caecilian *R. bivittatum*, which has 42% of its genome represented by retrotransposons and 20% by DNA TEs.

Despite the distinct content of TE families, the two closest species present comparable patterns of TE history amplification across both Class I and Class II major abundant families (Figure 9, columns eight and nine).

Notably, upon analysing the amplification history of different TE families across species, a consistent unimodal distribution was observed for the DNA/hAT family in all analysed anuran genomes. Conversely, DNA/nMITE and DNA/TcMar families exhibit more recent bursts of expansion (0-5 Kimura substitution level) in almost all species. Regarding Class I families, LTR/Gypsy, LTR/ERV and LINE/L1 families show multimodal distributions in all the species except one (the smallest *P. ornatum* genome), characterised by an ancient burst of expansion (i.e. 40 Kimura substitution level), followed by more recent expansion bursts.

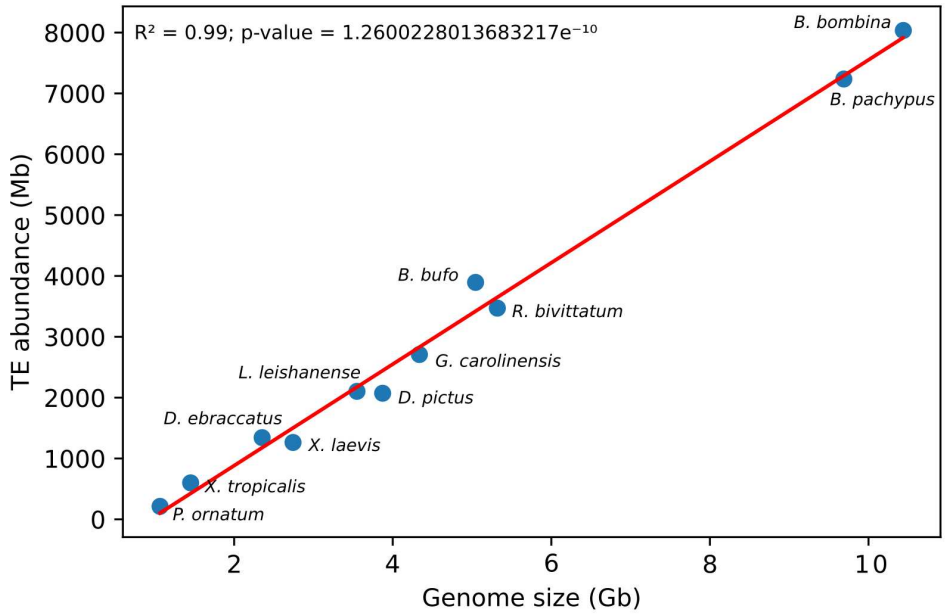


Figure 7. Linear regression between genome size and TE abundance:

The X-axis shows the genome size in Gb, the Y-axis shows the length of TEs in Mb.

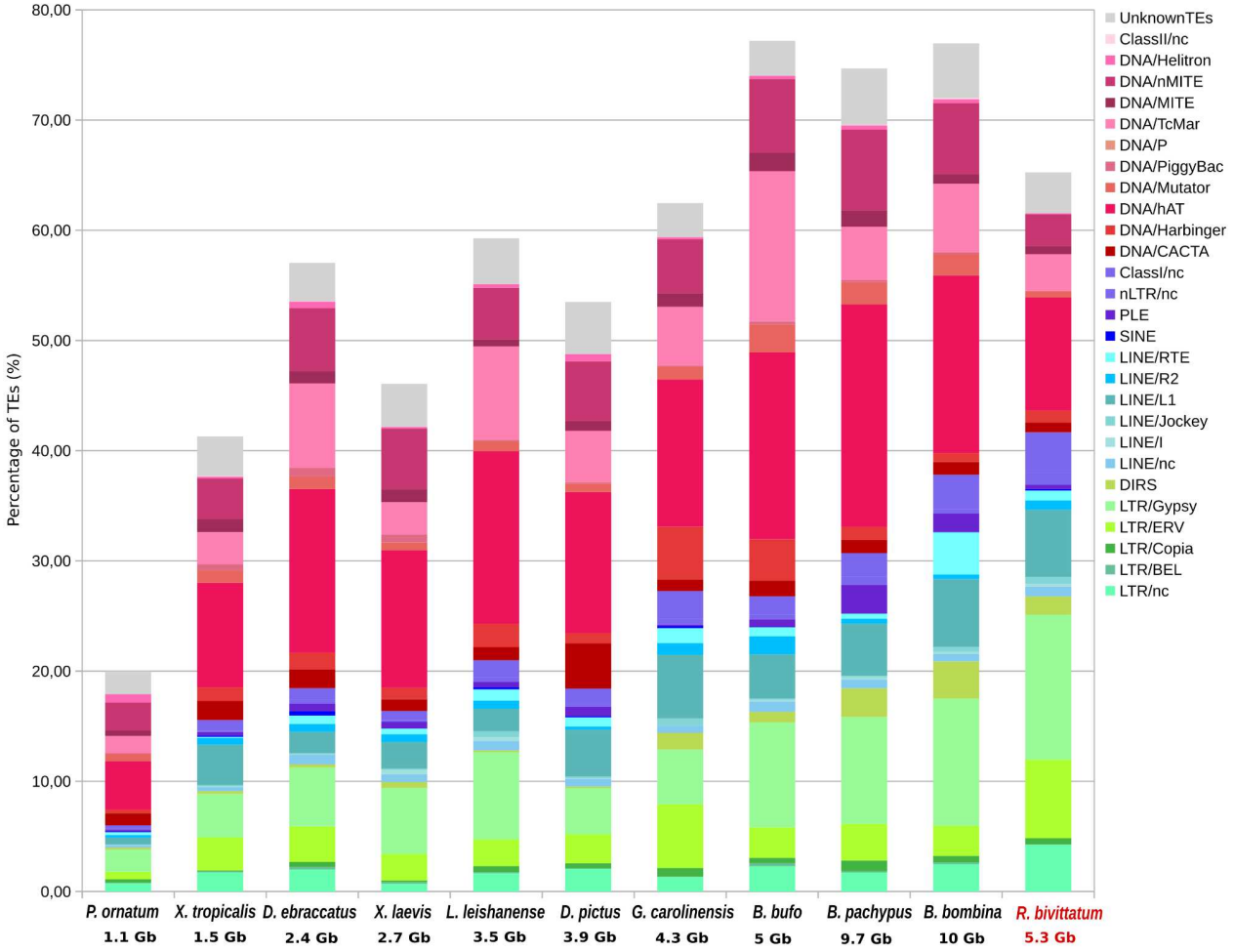


Figure 8. TE Abundance and Distribution across anuran species:

The different anuran species analysed are in order of genome size on the x-axis, *R. bivittatum* is a caecilian species included as an outgroup. The y-axis represents the percentages of TE abundance. Different colours correspond to distinct TE families, as indicated in the legend.

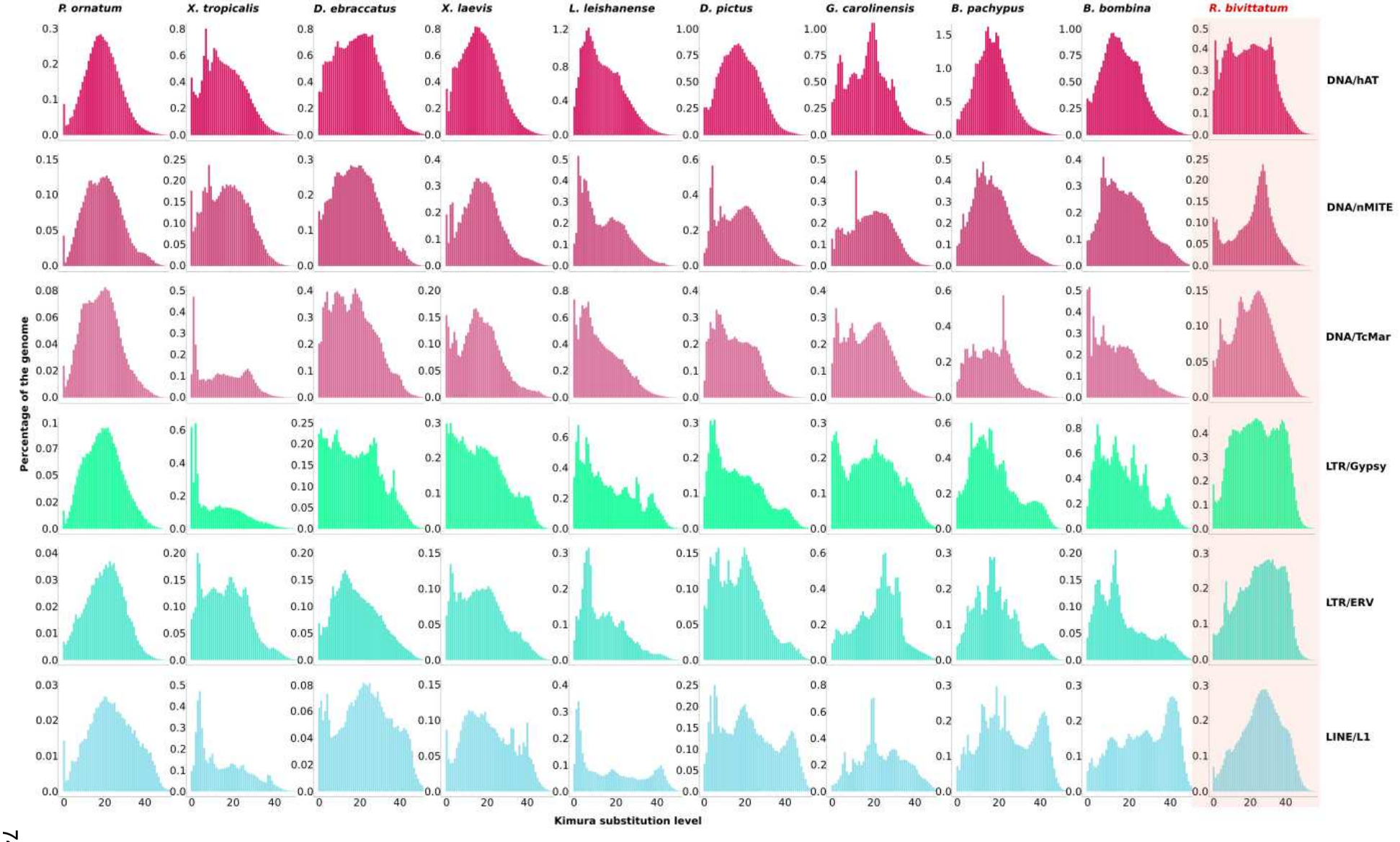

P. ornatum
X. tropicalis
D. ebraccatus
X. laevis
L. leishanense
D. pictus
G. carolinensis
B. pachypus
B. bombina
R. bivittatum
DNA/hAT
DNA/nMITE
DNA/TcMar
LTR/Gypsy
LTR/ERV
LINE/L1
Percentage of the genome
Kimura substitution level

Figure 9. TE landscapes:

TE amplification history plots of the most abundant TE families across the different anuran species (the caecilian outgroup is highlighted in the last column). The Y-axes show the genomic coverage of TE families, varying in scale, the X-axes show the number of substitutions (Kimura distances).

Ultimately, we explored the relationship between genome size and TE family diversity by initially performing a PCA, and subsequently measuring diversity indices for each species. The PCA in Figure 10 shows a correlation between TE diversity composition and genome size. Intermediate-sized genomes tend to cluster together along PC1, while the smallest genome of *P. ornatum* and the two largest genomes of *B. pachypus* and *B. bombina* occupy opposite extremes. The two-lined caecilian *R. bivittatum* appears as an outlier. However, examination of the diversity indices, measured using both Simpson's and Shannon's diversity indices, reveals similar values across all genomes (Table 3). This suggests that variation in genome size is not associated with substantial changes in TE family diversity.

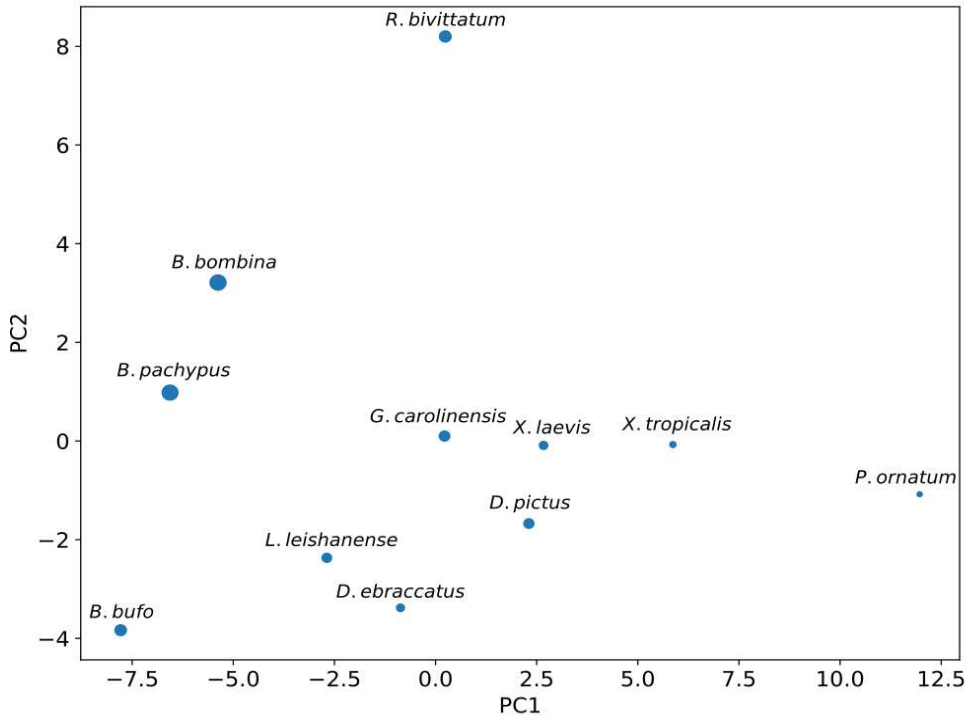


Figure 10. Principal Component Analysis (PCA) of TE diversity composition across anuran species:

Dots dimensions are scaled with the size of the genome for each species.

Table 3. TE diversity indices

	Genome size	Shannon Index (H)	Gini-Simpson Index (1-D)
P. ornatum	1.1 Gb	2.39	0.87
X. tropicalis	1.5 Gb	2.34	0.87
D. ebraccatus	2.4 Gb	2.30	0.85
X. laevis	2.7 Gb	2.27	0.84
L. leishanense	3.5 Gb	2.21	0.84
D. pictus	3.9 Gb	2.30	0.86
G. carolinensis	4.3 Gb	2.40	0.88
B. bufo	5 Gb	2.29	0.86
B. pachypus	9.7	2.27	0.85
B. bombina	10 Gb	2.35	0.87
R. bivittatum	5.3 Gb	2.24	0.85

Discussion

One of the long-standing debates in evolutionary biology concerns the underlying processes that have led to the evolution of genome size variation and its complex architecture among eukaryotes.

In this study, we delve into the analysis of large genomes in anuran species, aiming to elucidate the contribution of transposable elements to genome size evolution. Our investigation encompasses both ancient and recent dynamics of TEs that could have contributed to increase in genome size, as well as the influence of effective population size on TE amplification.

TE amplification dynamics in the large genome of *Bombina pachypus*

Three-quarters of the Apennine yellow-bellied toad's large genome are represented by transposable elements, highlighting their role in shaping the intricate architecture of this genome. Class I of DNA transposons predominantly populated the genome through a single expansion wave, contrasting with Class II of retrotransposons, which appeared to have undergone two distinct waves of expansion over the evolution of the *B. pachypus* genome (Figure 4).

Moreover, our analysis revealed a non-linear relationship between the abundance of TE and their expression levels (Figure 5). Specifically, while DNA transposons were the most abundant class within the genome, retrotransposons emerged as the most highly expressed. This suggests complex dynamics in the expansion of TEs in the genome of *B. pachypus*, where the most abundant TE families may not necessarily be the most highly expressed. We speculate that this may be attributed to a higher abundance of degenerating copies of DNA transposons that are no longer active, a scenario commonly observed in large genomes. Indeed, genomes with a substantial amount of TEs often exhibit extensive regions derived from older waves of transposition, followed by periods of degeneration, eventually forming what are commonly referred to as "cemeteries of TEs" (Sirijovski et al. 2005; Wang et al. 2021).

Another plausible scenario involves a differential efficiency or greater specificity of host silencing mechanisms (Ancona et al. (submitted); Chapter 1 of this thesis) towards

different families of TEs. Different studies on transposon dynamics have revealed a very heterogeneous and intricate picture of the relationship between TE families expansion and host silencing strategies. For instance, in *Drosophila*, distinct somatic and germline piRNA clusters have evolved to target different families of TEs, with the former mainly targeting elements from the Gypsy family, while the latter suppressing a broad range of elements (Malone et al. 2009).

Furthermore, although the genomic localisation of TEs in intra- and intergenic regions has revealed an enrichment of TEs in intergenic regions, it should be noted that this insertion enrichment is modest (with an enrichment in intergenic regions of 0.2 greater than in intragenic ones) (Figure 3). This observation could indicate a lower efficiency in the action of silencing mechanisms aimed at preventing TE insertions within genes, as we would typically expect a higher proportion of TEs in intergenic regions compared to intragenic regions under opposite conditions. The only exception is the DIRS family, which shows a significant insertion preference in intragenic regions. It is also possible that TE insertions in gene introns have no more fitness consequences than intergenic insertions, as long as the coding sequence and the splicing sites are not affected. This could explain the gigantic introns observed in different eukaryotic lineages (Gozashti et al. 2022) and also in *B. pachypus*.

To gain deeper insights into the recent dynamics of TEs within the host genome, population-level genomic data offer unprecedented opportunities to explore their recent evolutionary history. Firstly, they allow for the analysis of TE abundance and distribution across different populations of a given species, which may have experienced distinct population processes and dynamics, including different demographic histories. Secondly, these data afford the investigation of the recent impact of selection on TE dynamics, providing a new perspective for understanding the underlying processes driving variations in genome size.

Upon comparing population genomic data from two different populations of *B. pachypus* — one from the southern refugium and the other from the margin of the northern expansion range — we observe a slightly higher abundance of TEs in the northern population. This finding supports the hypothesis of relaxation of purifying selection and consequent greater impact of genetic drift during the northern

expansion, which would have allowed for a greater expansion of TEs in this population. This is consistent with the non-adaptive hypothesis of genome size variation, which predicts that under condition of reduced efficiency of selection and major impact of random genetic drift, reduced population size provides a permissive environment for the proliferation of mobile genetic elements, consequently resulting in genome size expansion (Lynch and Conery, 2003).

TE expansion and distribution among anuran species with different genome size

To further explore the contribution of TEs to genome size evolution, we compared the diversity and distribution of transposable elements in *B. pachypus* with those in its closest relative, *B. bombina*, as well as with eight other anuran species with genome sizes ranging from 1 to 10 Gb.

Our results unveiled a positive correlation between genome size and TE abundance, showing a progressive accumulation of transposable elements that increases linearly with the species' genome size (Figure 7). This underscores the significant role of transposable element amplification dynamics in driving genome size variation, a trend supported by recent comparative analyses (Cong et al. 2022; Zuo et al. 2023).

Moreover, different models have attempted to explain the relationship between genome size and TE family diversity. Early models predicted an inverse relationship, suggesting that smaller genomes would harbour more diverse TE communities, while larger genomes would only allow the amplification of a subset of TE families, resulting in reduced TE diversity (Petrov et al. 2003; Furano et al. 2004). However, our comparative analysis revealed similar estimates for the diversity indexes across all genomes, indicating that variation in genome size is not associated with substantial changes in TE family diversity. Large genomes maintain high levels of TE diversity, as also emphasised by a recent study on salamander genomes (Haley and Mueller, 2022).

Detailed analysis of TE composition showed a consistently higher abundance of DNA transposons compared to retrotransposons across all anuran species, except for *B. bombina*, which instead exhibits a genome predominantly dominated by

retrotransposons. As previously discussed, the expansion of specific classes or families of TEs is influenced by various processes and dynamics, including the number of active families and the cross-reactivity of different silencing mechanisms targeting different families, which in turn leads to competition among TEs to evade host silencing (Abrusan et al. 2006; Roessler et al. 2018). All these processes significantly impact the expansion and contraction of the different TE families. Additionally, due to their relatively simple structure, DNA transposons undergo more frequent horizontal transfer (Schaack et al. 2010), a phenomenon that may also have occurred in anuran genomes.

The two closest species, *B. pachypus* and *B. bombina*, share a comparable overall TE content, with percentages of 75% and 77%, respectively. However, they show distinct patterns in the families predominantly associated with genome expansion, despite exhibiting similar amplification histories (Figure 9).

Remarkably, *B. bombina* stands out with a unique genomic profile among the analysed anuran species. It displays a higher proportion of retrotransposons relative to DNA transposons, indicating a distinct genomic landscape in this species. This greater expansion of retrotransposons aligns it more closely with the TE expansion dynamics observed in the genomes of urodeles and caecilians (Sun et al. 2012; Sun and Mueller, 2014; Wang et al. 2021). However, consistent differences in TE abundance and diversity have been observed among other closely related species (Vieira and Biémont, 2004; Hollister et al. 2011; Kawahara et al. 2023). Further investigation into TE expression dynamics and localisation will be necessary in *B. bombina* to gain a deeper understanding of the underlying dynamics driving changes in its genomic architecture.

Conclusions

Our comprehensive analysis of TE abundance and dynamics across the genomes of *B. pachypus* and nine other anuran species provided valuable insights into the contribution of transposable elements to genome size evolution. Our findings highlight the predominant role of TE amplification in driving genome size expansion, as evidenced by a significant positive correlation between genome size and TE copy numbers across species.

Interestingly, the diversity of TE families remains remarkably consistent across the different genomes, challenging prior expectations of decreased diversity with genome expansion. This suggests that variations in genome size do not necessarily entail substantial alterations in TE family diversity, but rather an increase in the copy number of TEs in multiple families.

Additionally, our investigation into TE recent dynamics within distinct populations of *B. pachypus*, characterised by different effective population sizes, has shed light on the selective pressures governing TE accumulation and genome expansion. This underscores the critical need for integrating population-level genomic data to glean deeper insights into the evolutionary dynamics between TEs and host genomes.

References

Abrusán G., Krambeck H.J. 2006. Competition may determine the diversity of transposable elements. Theor Popul Biol. 70(3):364-75. doi: 10.1016/j.tpb.2006.05.001.

Alonge M., Lebeigle L., Kirsche M., Jenike K., Ou S., Aganezov S., Wang X., Lippman Z.B., Schatz M.C., Soyk S. 2022. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. Genome Biol. 23(1):258. doi: 10.1186/s13059-022-02823-7. PMID: 36522651; PMCID: PMC9753292.

Ancona L., Nitta Fernandes F.A., Biello R., Chiocchio A., Castrignanò T., Barucca M., Canestrelli D., Trucchi E. (Submitted). Evolutionary dynamics of transposable elements activity and regulation in the Apennine yellow-bellied toad (*Bombina pachypus*).

Arnqvist G., Sayadi A., Immonen E., Hotzy C., Rankin D., Tuda M., Hjelmen C.E. and Johnston J. S. 2015. Genome size correlates with reproductive fitness in seed beetlesProc.R.Soc.B.2822015142120151421. http://doi.org/10.1098/rspb.2015.1421

Canestrelli D., Cimmaruta R., Costantini V., Nascetti G. 2006. Genetic diversity and phylogeography of the Apennine yellow-bellied toad Bombina pachypus, with implications for conservation. Mol. Ecol. 15:3741–3754.

Cavalier-Smith T. 1978. Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. J Cell Sci. Dec;34:247-78. doi: 10.1242/jcs.34.1.247. PMID: 372199.

Cavalier-Smith T. 2005. Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. Ann Bot. 95(1):147-75. doi: 10.1093/aob/mci010. PMID: 15596464; PMCID: PMC4246715.

Cong Y., Ye X., Mei Y., He K., Li F. 2022. Transposons and non-coding regions drive the intrafamily differences of genome size in insects. iScience. 25(9):104873. doi: 10.1016/j.isci.2022.104873. PMID: 36039293; PMCID: PMC9418806.

Durand N.C., Shamim M.S., Machol I., Rao S.S., Huntley M.H., Lander E.S., Aiden E.L. 2016. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst. 3(1):95-8. doi: 10.1016/j.cels.2016.07.002. PMID: 27467249; PMCID: PMC5846465.

Elliott T.A., Gregory T.R. 2015. Do larger genomes contain more diverse transposable elements?. BMC Evol Biol 15, 69. https://doi.org/10.1186/s12862-015-0339-8

Flynn J.M., Hubley R., Goubert C., Rosen J., Clark A.G., Feschotte C., Smit A.F. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 117(17):9451-9457. doi: 10.1073/pnas.1921046117. Epub 2020 Apr 16. PMID: 32300014; PMCID: PMC7196820.

Furano A.V., Duvernell D.D., Boissinot S. 2004. L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish. Trends Genet. 20(1):9-14. doi: 10.1016/j.tig.2003.11.006. PMID: 14698614.

Gozashti L., Roy S.W., Thornlow B., Kramer A, Ares M Jr, Corbett-Detig R. 2022. Transposable elements drive intron gain in diverse eukaryotes. Proc Natl Acad Sci U S A. 119(48):e2209766119. doi: 10.1073/pnas.2209766119.

Gregory T.R., Hebert PD. 1999. The modulation of DNA content: proximate causes and ultimate consequences. Genome Res. 9(4):317-24. PMID: 10207154.

Gregory TR. 2001. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol Rev Camb Philos Soc. 76(1):65-101. Doi: 10.1017/s1464793100005595. PMID: 11325054.

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(5):830-43. doi: 10.1006/bcmd.2001.0457. PMID: 11783946.

Guan D., McCarthy S.A., Wood J., Howe K., Wang Y., Durbin R. 2020. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics. 36(9):2896-2898. doi: 10.1093/bioinformatics/btaa025. PMID: 31971576; PMCID: PMC7203741.

Haley A.L., Mueller R.L. 2022. Transposable Element Diversity Remains High in Gigantic Genomes. J Mol Evol. 90(5):332-341. doi: 10.1007/s00239-022-10063-3. Epub 2022 Jun 25. PMID: 35751655.

Hewitt G. 2000. The genetic legacy of the Quaternary ice ages. Nature 405, 907–913. https://doi.org/10.1038/35016000

Hollister J.D., Smith L.M., Guo Y.L., Ott F., Weigel D., Gaut B.S. 2011. Transposable elements and small RNAs contribute to gene expression divergence between Arabidopsis thaliana and Arabidopsis lyrata. Proc Natl Acad Sci U S A. 108(6):2322-7. doi: 10.1073/pnas.1018222108.

Huang N., Li H. 2023. compleasm: a faster and more accurate reimplementation of BUSCO. Bioinformatics. 39(10):btad595. doi: 10.1093/bioinformatics/btad595. PMID: 37758247; PMCID: PMC10558035.

Kapusta A., Suh A., Feschotte C. 2017. Dynamics of genome size evolution in birds and mammals. Proc Natl Acad Sci U S A. 114(8):E1460-E1469. doi: 10.1073/pnas.1616702114. Epub 2017 Feb 8. PMID: 28179571; PMCID: PMC5338432.

Kawahara K., Inada T., Tanaka R., Dayi M., Makino T., Maruyama S., Kikuchi T., Sugimoto A., Kawata M. 2023. Differentially Expressed Genes Associated with Body Size Changes and Transposable Element Insertions between Caenorhabditis elegans and Its Sister Species, Caenorhabditis inopinata. Genome Biol Evol. 15(4):evad063. doi: 10.1093/gbe/evad063. PMID: 37071793; PMCID: PMC10139442.

Liedtke H.C., Gower D.J., Wilkinson M. et al. 2018. Macroevolutionary shift in the size of amphibian genomes and the role of life history and climate. Nat Ecol Evol 2, 1792–1799. https://doi.org/10.1038/s41559-018-0674-4

Lynch M., and Conery J.S. 2003. The origins of genome complexity. Science 302:1401–1404.

Lynch M., Bobay L.M., Catania F., Gout J.F., Rho M. 2011. The repatterning of eukaryotic genomes by random genetic drift. Annu Rev Genomics Hum Genet.12:347-66. doi: 10.1146/annurev-genom-082410-101412. PMID: 21756106; PMCID: PMC4519033.

Malone C.D., Brennecke J., Dus M., Stark A., McCombie W.R., Sachidanandam R., Hannon G.J. 2009. Specialized piRNA pathways act in germline and somatic tissues of the Drosophila ovary. Cell. 137(3):522-35. doi: 10.1016/j.cell.2009.03.040. Epub 2009 Apr 23. PMID: 19395010; PMCID: PMC2882632.

Mann H.B. and Whitney D.R. 1947. On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other. Annals of Mathematical Statistics, 18, 50-60. http://dx.doi.org/10.1214/aoms/1177730491

Massey, Frank J. 1951. The Kolmogorov-Smirnov Test for Goodness of Fit." Journal of the American Statistical Association 46, no. 253 (1951): 68–78. https://doi.org/10.2307/2280095.

Ou S., Su W., Liao Y., Chougule K., Agda J. R. A., Hellinga A. J., Lugo C. S. B., Elliott T. A., Ware D., Peterson T., Jiang N., Hirsch C. N. and Hufford M. B. 2019. Benchmarking Transposable Element Annotation Methods for Creation of a Streamlined, Comprehensive Pipeline. Genome Biol. 20(1): 275.

Petrov D.A., Aminetzach Y.T., Davis J.C., Bensasson D., Hirsh A.E. 2003. Size matters: non-LTR retrotransposable elements and ectopic recombination in Drosophila. Mol Biol Evol. 20(6):880-92. doi: 10.1093/molbev/msg102.

Quinlan A.R. and Hall I.M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26, 6, pp. 841–842

Rhie A., Walenz B.P., Koren S. et al. 2020. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol 21, 245. https://doi.org/10.1186/s13059-020-02134-9

Roessler K., Bousios A., Meca E., Gaut B.S. 2018. Modeling Interactions between Transposable Elements and the Plant Epigenetic Response: A Surprising Reliance on

Element Retention. Genome Biol Evol. 10(3):803-815. doi: 10.1093/gbe/evy043. PMID: 29608716; PMCID: PMC5841382.

Rogers R.L. et al. 2018. Genomic Takeover by Transposable Elements in the Strawberry Poison Frog. Mol. Biol. Evol. 35:2913–2927.

Ruan J., Li H. 2020. Fast and accurate long-read assembly with wtdbg2. Nat Methods 17, 155–158. https://doi.org/10.1038/s41592-019-0669-3

Schaack S., Gilbert C., Feschotte C. 2010. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends Ecol Evol. 25(9):537-46. doi: 10.1016/j.tree.2010.06.001.

Shannon C.E. 1948. A mathematical theory of communication. Bell Syst Tech J 27:379–423

Simpson E.H. 1949. Measurement of diversity. Nature 163:688–688

Sirijovski N., Woolnough C., Rock J., Joss J.M. 2005. NfCR1, the first non-LTR retrotransposon characterized in the Australian lungfish genome, Neoceratodus forsteri, shows similarities to CR1-like elements. J Exp Zool B Mol Dev Evol. 304(1):40-9. doi: 10.1002/jez.b.21022. PMID: 15593278.

Smit A.F.A., Hubley R. & Green P. 2013-2015. RepeatMasker Open-4.0. <http://www.repeatmasker.org>

Sun C., Shepard D.B., Chong R.A., López Arriaza J., Hall K., Castoe T.A., Feschotte C., Pollock D.D., Mueller RL. 2012. LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders. Genome Biol Evol. 4(2):168-83. doi: 10.1093/gbe/evr139.

Sun C., Mueller R.L. 2014. Hellbender genome sequences shed light on genomic expansion at the base of crown salamanders. Genome Biol Evol. 6(7):1818-29. doi: 10.1093/gbe/evu143. PMID: 25115007; PMCID: PMC4122941.

Sun Y.B., Zhang Y., Wang K. 2020. Perspectives on studying molecular adaptations of amphibians in the genomic era. Zool Res. 41(4):351-364. doi: 10.24272/j.issn.2095-8137.2020.046. PMID: 32390371; PMCID: PMC7340517.

Vieira C., Biémont C. 2004. Transposable element dynamics in two sibling species: Drosophila melanogaster and Drosophila simulans. Genetica. 120(1-3):115-23. doi: 10.1023/b:gene.0000017635.34955.b5. PMID: 15088652.

Wang J., Itgen M.W., Wang H., Gong Y., Jiang J., Li J., Sun C., Sessions S.K., Mueller R.L. 2021. Gigantic Genomes Provide Empirical Tests of Transposable Element Dynamics Models. Genomics Proteomics Bioinformatics. 19(1):123-139. doi: 10.1016/j.gpb.2020.11.005. Epub 2021 Mar 4. PMID: 33677107; PMCID: PMC8498967.

Wang K., Wang J., Zhu C., Yang L., Ren Y., Ruan J., Fan G., Hu J., Xu W., Bi X., Zhu Y., Song Y., et al. 2021. African lungfish genome sheds light on the vertebrate water-to-land transition. Cell. 184(5):1362-1376.e18. doi: 10.1016/j.cell.2021.01.047. Epub 2021 Feb 4. PMID: 33545087.

Yan H., Bombarely A., Li S. 2020. DeepTE: a computational method for de novo classification of transposons with convolutional neural network. Bioinformatics, Volume 36, Issue 15, 1 August 2020, Pages 4269–4275.

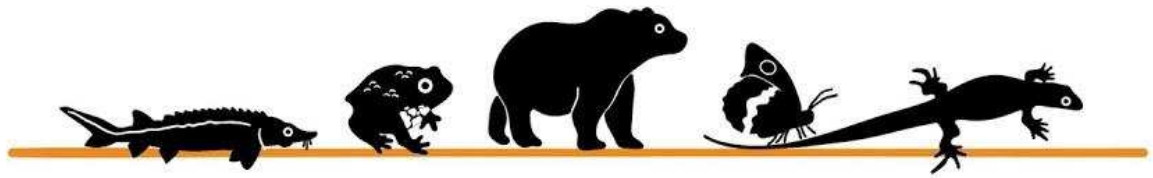
Zhang H., Song L., Wang X. et al. 2021. Fast alignment and preprocessing of chromatin profiles with Chromap. Nat Commun 12, 6566. https://doi.org/10.1038/s41467-021-26865-w

Zhou C., McCarthy S.A., Durbin R. 2023. YaHS: yet another Hi-C scaffolding tool. Bioinformatics. 39(1):btac808. doi: 10.1093/bioinformatics/btac808. PMID: 36525368; PMCID: PMC9848053.

Zuo B., Nneji L.M. & Sun YB. 2023. Comparative genomics reveals insights into anuran genome size evolution. BMC Genomics 24, 379. https://doi.org/10.1186/s12864-023-09499-8

Chapter 3

Characterization of TE abundance and distribution in endangered Italian endemic species within the Endemixit project



ENDEMIXIT

Population Genomics of Italian Endemics

Project introduction

Italy is a biodiversity hotspot, but several endemic species, representing a unique biological heritage, are endangered. Main threats are related to human activities causing fragmentation and decline in their population size. Extinction risks can be reduced by improving knowledge of genetic variation and developing conservation strategies aimed at preventing genetic erosion. Small and declining populations are routed to radical changes in their genetic diversity as natural selection is less efficient and random genetic drift becomes the major player. Genetic drift can lead to the accumulation of deleterious mutations, i.e., the mutation load, affecting individual and population fitness and, in turn, further reducing population size even to extinction. Understanding the genome-wide dynamics of this process can reduce a species extinction risk.

The enormous improvement of next generation sequencing techniques, computing resources, and statistical methods now allow the study of complete genomes from several individuals virtually in any non-model species. Genomes can be screened to predict the deleterious effects of different mutation types and, ultimately, to estimate the mutation load in single individuals and in populations, and to predict its impact on fitness. Using five Italian iconic endangered endemics as model species (a mammal, a reptile, an amphibian, a fish, and an insect) and an unprecedented effort of massive sequencing, bioinformatics and population genomics analyses, the ENDEMIXIT project (P.I.: Giorgio Bertorelle, University of Ferrara, www.endemixit.com) proposes a comprehensive conservation genomics action with three major goals:

i) understand the dynamics of the accumulation of deleterious mutations in small populations, and its impact on individual fitness and extinction risks;

ii) estimate the genomic susceptibility to extinction due to the mutation load, predict the consequences of a strategy of genetic rescue, and propose conservation actions;

iii) boost the interaction between research and practice in conservation genomics, and increase public awareness about biodiversity erosion and innovative molecular tools to prevent it.

For each endemic species, ENDEMIXIT produced five *de novo* genomes and resequenced twenty individuals from two populations with different estimated population size. Computational approaches were used to estimate demographic histories and different measures of mutation load and to quantify the genomic susceptibility to extinction due to load (GSEm). Computer simulations were also used to predict how the accumulated load might affect the outcome of a genetic rescue strategy before possibly implementing it, leading to the genomic susceptibility to extinction due to rescue (GSMr). In addition, functional assays were also carried out to study i) in vitro the correlation between bioenergetic and cellular functions and the negative effects predicted in silico for fixed deleterious mutations, and ii) in vivo the segregation pattern of the load and its correlation with individual fitness in controlled inbred and outbred crosses. Finally, ENDEMIXIT worked to reduce the existing gap between researchers, practitioners, and citizens by establishing a Conservation Genomic Consortium, organising a public exhibition on genetics, biodiversity, and conservation, and activating a series of dissemination activities (https://youtu.be/mL_JzgOqk7c) with the final aim of favouring a novel attitude about genetic studies and how these can help developing species and environment protection plans.

Author contribution

As part of the ENDEMIXIT project, I have been involved in characterising the abundance and distribution of transposable elements in the target species of the project: *Podarcis raffonei*, *Hipparchia sbordoni* and *Ursus arctos marsicanus* (still in early stage of preparation, not presented below). This chapter includes the following papers on *Podarcis raffonei* and *Hipparchia sbordonii*:

- A high-quality reference genome for the critically endangered Aeolian wall lizard, *Podarcis raffonei*. J Hered. 2023 May 25;114(3):279-285. doi: 10.1093/jhered/esad014. PMID: 36866448.

- Chromosome-level reference genome of the Ponza grayling (*Hipparchia sbordonii*), an Italian endemic and endangered butterfly. (Under revision)

A high-quality reference genome for the critically endangered Aeolian wall lizard, *Podarcis raffonei*

Authors

Maëva Gabrielli[1*], Andrea Benazzo[1*], Roberto Biello[1], **Lorena Ancona[2]**, Silvia Fuselli[1], Alessio Iannucci[3], Jennifer Balacco[4], Jacqueline Mountcastle[4], Alan Tracey[5], Gentile Francesco Ficetola[6,7], Daniele Salvi[8], Marco Sollitto[9], Olivier Fedrigo[4], Giulio Formenti[4,10], Erich D. Jarvis[4], Marco Gerdol[9], Claudio Ciofi[3], Emiliano Trucchi[2], Giorgio Bertorelle[1]

Affiliations

1 - Department of Life Sciences and Biotechnology, University of Ferrara, Ferrara, Italy

2 - Department of Life and Environmental Sciences, Marche Polytechnic University, Ancona, Italy

3 - Department of Biology, University of Florence, Florence, Italy

4 - Vertebrate Genome Laboratory, The Rockefeller University, New York, NY, USA

5 – Tree of Life, Wellcome Sanger Institute, Cambridge, United Kingdom

6 - Department of Environmental Sciences and Policy, University of Milan, Milan, Italy

7 - Laboratoire d'Ecologie Alpine (LECA), CNRS, Université Grenoble Alpes and Université Savoie Mont Blanc, Grenoble, France

8 - Department of Health, Life & Environmental Sciences - University of L'Aquila, L'Aquila, Italy

9 - Department of Life Sciences, University of Trieste, Trieste, Italy

10 - Howard Hughes Medical Institute, Chevy Chase, MD, USA

Abstract

The Aeolian wall lizard, *Podarcis raffonei*, is an endangered species endemic to the Aeolian archipelago, Italy, where it is present only in 3 tiny islets and a narrow promontory of a larger island. Because of the extremely limited area of occupancy, severe population fragmentation and observed decline, it has been classified as Critically Endangered by the International Union for the Conservation of Nature (IUCN). Using Pacific Biosciences (PacBio) High Fidelity (HiFi) long-read sequencing,

Bionano optical mapping and Arima chromatin conformation capture sequencing (Hi-C), we produced a high-quality, chromosome-scale reference genome for the Aeolian wall lizard, including Z and W sexual chromosomes. The final assembly spans 1.51 Gb across 28 scaffolds with a contig N50 of 61.4 Mb, a scaffold N50 of 93.6 Mb, and a BUSCO completeness score of 97.3%. This genome constitutes a valuable resource for the species to guide potential conservation efforts and more generally for the squamate reptiles that are underrepresented in terms of available high-quality genomic resources.

Key words:

conservation genetics, de novo assembly, Endemixit, Hi-C, Lacertids, PacBio HiFi

Introduction

The Aeolian wall lizard *Podarcis raffonei* (Fig. 1A) is one of the most endangered vertebrate in Europe (Gippoliti et al. 2017). It is endemic to 4 islands of the Aeolian archipelago, located North-East of Sicily, with an extremely restricted distribution range including 3 islets less than 0.01 km2 (La Canna, Scoglio Faraglione, Strombolicchio) and a larger island (Vulcano, 21.2 km2) where it currently occupies nonetheless a very limited area (Bonardi et al. 2022). The total area of occupancy has been estimated to be as small as 5,000 m2 (Ficetola et al. 2021) and the total population size is estimated at about 2,000 individuals (Capula et al. 2002; Lo Cascio et al. 2014; Gippoliti et al. 2017; Ficetola et al. 2018, 2021). As a result, the Aeolian wall lizard has been listed as Critically Endangered in the Red List of Endangered Species of the IUCN (2009). The main threats to its survival include interactions with the invasive Italian wall lizard *Podarcis siculus*, combined with habitat degradation (Capula et al. 2002). This is particularly visible on the island Vulcano, where the intense habitat change that occurred in the last 50 yr may have favored the spread of P. siculus, leading to a sharp decline in the *P. raffonei* population (Capula et al. 2002). The production of highly contiguous genomes has greatly accelerated in the last decade, refining our understanding of the genomic basis of organismal traits, the chromosome evolution, and allowing the detection of natural selection through genomic scans (Geneva et al. 2022). Furthermore, reference genomes can be key for conservation genomics as they may permit, in combination with whole-genome resequencing data, to assess genetic diversity, investigate inbreeding depression, or characterize deleterious mutations (Formenti et al. 2022b). High-quality reference genomes are unevenly distributed across the tree of life, and some clades, such as the squamate reptiles, are underrepresented (Pinto et al. 2022; Card et al. 2023). Here, we present a high-quality chromosome-scale reference genome for the Aeolian wall lizard, produced as part of the Endemixit project (www.endemixit.com). Our final genome assembly spans 1.51 Gb across 28 scaffolds, with a scaffold N50 of 93.6 Mb and a BUSCO completeness score of 97.3%. This high-quality reference genome is a valuable resource to assess the genetic diversity in the 4 extant populations of the Aeolian wall lizard and better develop the conservation strategy for this species.

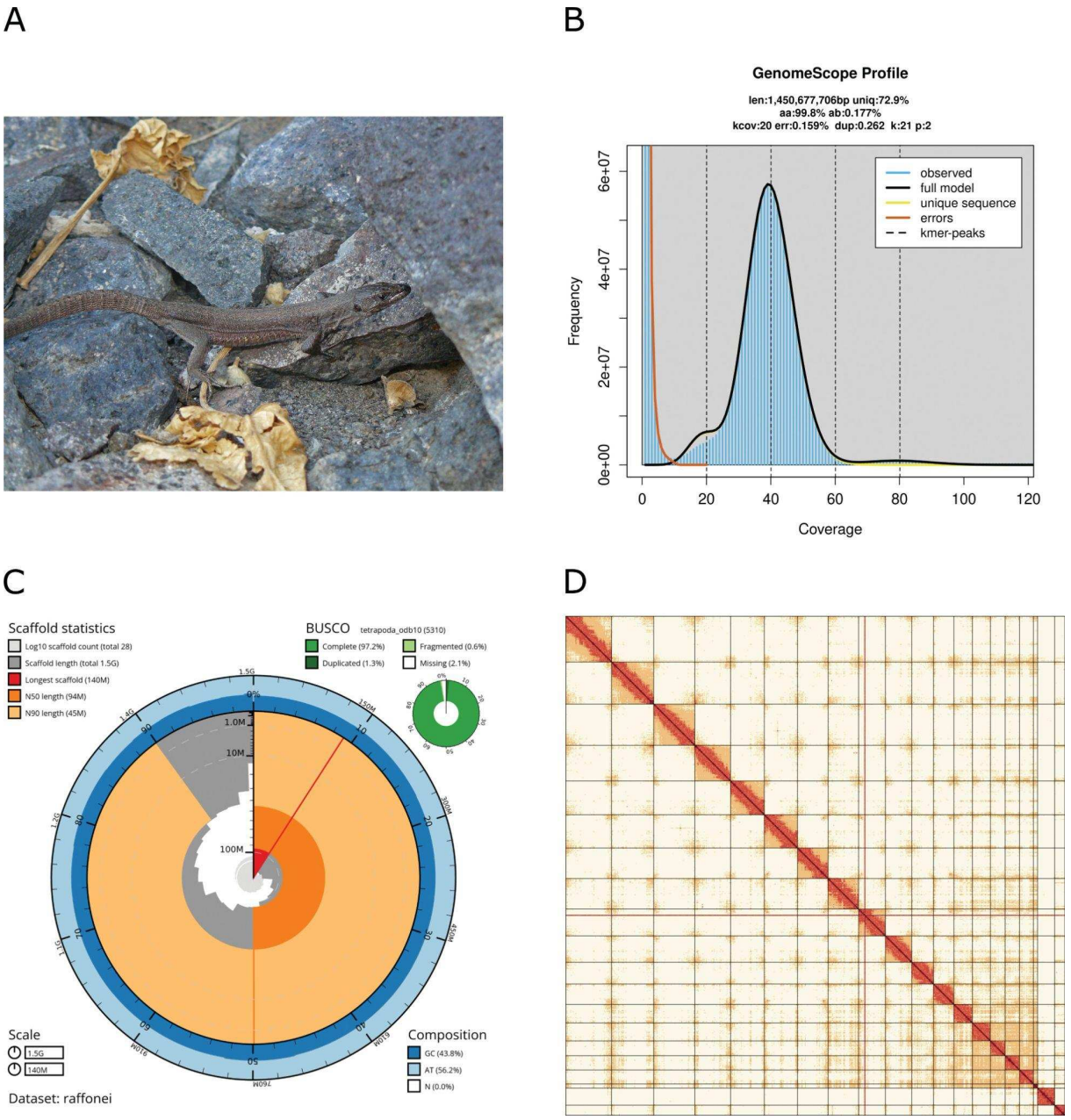


Fig. 1. A) Photography of an individual of *Podarcis raffonei*, on La Canna stack (Photo credit: Daniele Salvi), and visual overview of genome assembly metrics. B) K-mer spectra output and corresponding genome size and heterozygosity estimated with GenomeScope 2.0. C) BlobToolKit Snail plot showing a graphical representation of the quality metrics presented in Table 2 for the *Podarcis raffonei* primary assembly (rPodRaf1.pri). D) Hi-C contact map for the 20 scaffolds of the primary genome assembly generated with PretextSnapshot.

Methods

Biological materials

An adult female was collected on the 31st of July 2020 by D. Salvi on the stack of La Canna (38°34'56.13″N to 14°31'16.61″E; see Supplementary Fig. 1), in the Aeolian archipelago, in a small terrace at 50 m a.s.l. on the eastern slope of the stack, reached by climbing with the technical assistance of the mountain guide Lorenzo Inzigneri. A piece of tail was cut and immediately frozen in liquid nitrogen until the final storage at −80 °C.

Nucleic acid extraction, library preparation, and sequencing

All the following steps were carried out at the Vertebrate Genomes Project (VGP, https://vertebrategenomesproject. org/) lab. High molecular weight (HMW) DNA was extracted from muscle with the Circulomics HMW DNA extraction standard TissueRuptor protocol with the Nanobind Tissue Big DNA Kit (PN NB-900-701-01). DNA absorbance was checked as quality and purity control with Nanodrop and average fragment length was verified with a Pulsed Field Gel Electrophoresis (PFGE). Genomic data from 3 different sequencing technologies were used for the assembly: Pacific Biosciences (PacBio) High Fidelity (HiFi) reads, Bionano optical maps, and Hi-C reads from Arima Genomics. PacBio HiFi libraries were prepared using the Pacific Biosciences Express Template Prep Kit 2.0. The library was then size selected (>10 kb) using the Circulomics Short Read Eliminator. The PacBio library was sequenced on 2 PacBio 8M v3 SMRT Cells on a PacBio Sequel II and 1 PacBio 8M SMRT Cell on a PacBio Sequel IIe using the sequencing kit 2.0 and a 30-h movie. An aliquot of the HMW DNA was labeled for Bionano Genomics optical mapping using the Bionano Prep Direct Label and Stain (DLS) Protocol and run on 1 Saphyr instrument chip flowcell. Hi-C libraries were generated by Arima Genomics (https:// arimagenomics.com/) using muscle in vivo cross-linking with the Arima-HiC kit with 2-enzyme proximity ligation. Proximally ligated DNA was subjected to shearing, size selection (~200 to 600 bp) with SPRI beads, and enrichment with streptavidin beads for the biotin-labeled DNA. KAPA Hyper Prep kit was employed to generate libraries compatible with Illumina

technologies. Libraries were amplified through PCR, purified with SPRI beads and sequenced on an Illumina HiSeq X (~60× coverage) after a quality check with Bioanalyzer and qPCR.

Nuclear genome assembly

The genome of the Aeolian wall lizard was assembled following the VGP assembly pipeline v2.0 (Rhie et al. 2021), as outlined in Table 1. Briefly, PacBio HiFi long reads were processed using hifiasm (Cheng et al. 2021, 2022) producing a set of primary contigs representing the initial haploid assembly and separating alternative haplotypic variants. Primary contigs were then processed with purge_dups (Guan et al. 2020) to identify residual haplotype duplication in the assembly. Such duplicated sequences were moved to the alternate assembly that was then exposed to a second round of purge_dups to obtain the final set of nonredundant haplotypic variants. Primary contigs were anchored to scaffolds using Bionano optical maps, adjusting the gap size according to the observed optical distance with the bionano_solve pipeline v3.6.1 (Chan et al. 2018). A second round of scaffolding was performed using Hi-C data. Paired-end reads were aligned to the primary assembly using the Arima genomics' pipeline (https://github.com/ArimaGenomics/mapping_pipeline) and the obtained contact data were used to guide the scaffolding procedure using salsa2 (Ghurye et al. 2017, 2019). Hi-C contact maps were generated and visually inspected using PretextSuite (https://github.com/wtsi-hpag/PretextView; https://github.com/wtsi-hpag/PretextMap; https://github.com/wtsi-hpag/ PretextSnapshot) before and after the last scaffolding step. The resulting primary and alternate assemblies were screened for residual contaminations (Howe et al. 2021) and manual curation was performed on the primary assembly using the gEVAL browser release 73 (Howe et al. 2021), PretextView and HiGlass (Kerpedjiev et al. 2018) to anchor scaffolds to chromosomes and check their coherence.

Genome size estimation and quality assessment

We estimated the genome size from the PacBio HiFi reads using a k-mer-based approach. The distribution of k-mers of length 21 was generated using meryl v1.3

(Miller et al. 2008) and GenomeScope 2.0 (Ranallo-Benavidez et al. 2020) was subsequently used to infer the genome length, genome-wide heterozygosity, and error rate. We assessed the quality of our genome assembly using 2 independent methods. First, we used the BUSCO quality control tool to check for genome completeness using a set of conserved single-copy orthologous genes. We ran BUSCO v5.3.2 (Manni et al. 2021) in the genome mode with default parameters on the tetrapod dataset (tetrapoda_odb10) that contains 5,310 orthologous genes. Second, we used Mercury v1.3 (Rhie et al. 2020) to estimate the base level accuracy (QV) and the assembly completeness comparing the k-mers in the assembly and those observed in the HiFi reads. All assembly metrics were computed using gfastats v1.2.3 (Formenti et al. 2022a).

Identification of repetitive elements and gene annotation

To identify repetitive elements, we first generated a *de novo* repeat library using the Extensive *de novo* TE Annotator (EDTA) v1.9.9 (Ou et al. 2019) and DeepTE (Yan et al. 2020) to refine classifications within this library. We then used the final library to mask the genome with RepeatMasker v4.1.2 (Smit et al. n.d.). We used the same pipeline to identify repeats in the genome of *Podarcis muralis* (assembly PodMur_1.0; Andrade et al. 2019). For gene prediction, we first downloaded RNA-seq reads available on NCBI from various tissues of closely related species (4 species of the genus *Podarcis*; see Supplementary Table 1). Quality control and trimming for adapters and low-quality bases (quality score <20) of the raw reads were performed using fastqc v0.11.8 (Andrews 2010) and TrimGalore v0.5.0 (https://github.com/FelixKrueger/TrimGalore), respectively. High-quality reads were then mapped to the soft-masked assembly with hisat2 v2.1.0 (Kim et al. 2015), and sorted with samtools v1.10 (Li et al. 2009). All the BAM files were filtered to remove invalid splice junctions with Portcullis v1.1.2 (Mapleson et al. 2018). Filtered RNA-seq alignments were passed to Braker v2.1.6 (Hoff et al. 2016, 2019), together with amino acid sequences of the whole exome of 22 closely related species from the order Squamata belonging to 11 families including 3 Lacertidae (P. muralis, Lacerta agilis, and Zootoca vivipara; see Supplementary Table 2). The Braker gene prediction pipeline was run with the options "--softmasking --prg=gth --gth2traingenes." The resulting gene set was further filtered by evidence, keeping only

gene predictions supported by RNA-seq or protein evidence using a BRAKER2 script (selectSupportedSubsets.py). The completeness of the final gene set was checked with BUSCO v5.3.2 (Manni et al. 2021) using the longest transcript of each gene as the representative transcript.

Mitochondrial genome sequencing and assembly

To characterize the entire sequence of the mitochondrial DNA via Sanger sequencing, we designed 4 different, and partially overlapping, amplicons of expected length between 4 and 7.3 kb. Primers were designed based on mitochondrial DNA sequences of congeneric species (P. siculus NC_011609.1, P. muralis NC_011607 and NC_011609). Amplifications were carried out starting from 50 ng of extracted DNA, in a 50 µL reaction with 0.2 µM primers and 1.25 u of PrimeSTAR GXL DNA Polymerase. Amplification primers and additional internal primers were used for Sanger sequencing reactions (see Supplementary Table 3). Fragments were visually inspected and manually assembled to reconstruct the mitochondrial sequence.

Comparative analyses with *P. muralis*

We performed a synteny comparison with the *P. muralis* assembly (PodMur_1.0; Andrade et al. 2019), the only chromosome scale assembly presently available for the *Podarcis* genus. Phylogenetic reconstructions based on whole-genome data suggest that the 2 species diverged ~18 Mya during Miocene (Yang et al. 2021). We used minimap2 (Li 2018) to map the genome assembly of *P. raffonei* to the genome reference of *P. muralis* allowing a maximum sequence divergence of 5% (parameter -x asm20). We then filtered the alignment by mapping quality (>60) and length of the mapped fragments (>1 Mb) and plotted the alignment between the 18 autosomes and Z sexual chromosome (the W chromosome being absent from the *P. muralis* assembly) using Circos v0.69-8 (Krzywinski et al. 2009). Synteny between the 2 species was finally used to annotate the scaffolds of the *P. raffonei* assembly as chromosomes.

Table 1: Pipeline and software used for the genome assembly

Assembly	Software	Version
K-mer counting	Meryl	1.3
Estimation of genome size and heterozygosity	GenomeScope2	2.0
De novo assembly (contigging)	HiFiasm	0.16.1-r375
Remove low-coverage, duplicated contigs	purge_dups	1.2.5
Scaffolding		
Bionano Scaffolding	bionano_solve	3.6.1
Hi-C mapping for SALSA	Arima Genomics mapping pipeline	Commit 2e74ea4
Hi-C Scaffolding	salsa2	2.3
Hi-C Contact map generation		
Short-read alignment	bwa	0.7.17
SAM/BAM processing	samtools	1.10
Pairs processing	bedtools	2.30
Contact map visualization	PretextView	0.2.2
	PretextMap	0.1.8
	PretextSnapshot	0.0.4
Genome assembly refinement		

Manual curation and contamination screening	gEVAL	release 73
Genome quality assessment		
Basic assembly metrics	gfastats	1.2.3
Assembly completeness	BUSCO	5.3.2
	merqury	1.3
Repeat element identification		
Repeat identification	EDTA	1.9.9
	DeepTE	Commit babd65e
Repeat annotation	RepeatMasker	4.1.2
Gene annotation		
RNA-seq read quality control	fastqc	0.11.8
	TrimGalore	0.5.0
Mapping RNA-seq reads-genome	hisat2	2.1.0
Filtering splice junctions	Portcullis	1.1.2
Gene prediction	Braker	2.1.6
Comparison to *P. muralis*		
Genome–genome alignment	minimap2	2.22
Synteny visualisation	Circos	0.69-8

Results

The final genome size (1.51 Gb) is in agreement with the size estimated from the k-mer analysis with GenomeScope 2.0 (Fig. 1B) and very close to the genome size of *P. muralis* (1.51 Gb, Andrade et al. 2019). The k-mer spectrum shows a bimodal distribution with 2 major peaks, at ~20- and ~40-fold coverage, corresponding to heterozygous and homozygous states, respectively. Based on PacBio HiFi reads, we estimated a 0.159% sequencing error rate and a 0.177% nucleotide heterozygosity rate (Fig. 1B). The mitochondrial genome size is 17,038 bp, in agreement with the mitochondrial genome size of other species of Podarcis (17,311 bp for *P. muralis* and 17,297 bp for *P. siculus*; Podnar et al. 2009). The primary assembly contains 28 scaffolds for a total length of 1.51 Gb, with a contig N50 of 61.4 Mb, a scaffold N50 of 93.6 Mb, a longest contig size of 104.8 Mb, and a longest scaffold size of 139.1 Mb (Table 2; Fig. 1C). The alternate assembly contains 4,811 scaffolds spanning 182 Mb, having a N50 of 38.4 kb. This assembly is highly contiguous, as shown in the Hi-C contact map (Fig. 1D), with the 20 first scaffolds being of chromosome length and corresponding to the 18 autosomes and the 2 sexual chromosomes Z and W (see Supplementary Table 4). The sequencing depth of the HiFi reads along chromosomes is approximately uniform and does not reveal discrepancies in the assembly (see Supplementary Fig. 2). The completeness of the assembly is very high, with a BUSCO completeness score of 97.3% ([Single copy: 96.0%, Duplicated: 1.3%], Fragmented: 0.6%, Missing: 2.1%) using the tetrapod gene set and a k-mer completeness of 99.5%. Per base quality (QV) as estimated by Merqury is 62, corresponding to less than 1 incorrect nucleotide per megabase. In total, 22,463 protein-coding genes were predicted. The BUSCO completeness of the gene annotation using the same tetrapod gene set was 92.1% ([Single copy: 91.1%, Duplicated: 1.0%], Fragmented: 3.9%, Missing: 4.0%). The identification of repetitive elements resulted in a 48.2% repeat content, falling within the range of repeat contents for other squamate species (24.4% to 73.0%; Pasquesi et al. 2018). In Lacertidae and Teiidae, the repeat content was estimated to be 45.1% and 44.5%for *P. muralis* and Salvator merianae (Roscito et al. 2018), respectively (see Supplementary Tables 5 and 6). The major class of repetitive elements was constituted by LTR elements and DNA transposons (see Supplementary Table 5). The alignment of the genomes of *P. muralis*

and *P. raffonei* revealed a very high congruency in the chromosomal organization (Fig. 2). The only chromosomal segment that did not map to the homologous chromosome from the other species was a 1.5 Mb segment of the chromosome 2 of *P. raffonei* that mapped to the chromosome 18 of *P. muralis*. We analyzed the depth of coverage profile and the reads mapping in the edges of this segment of chromosome 2 in *P. raffonei* and did not find any discrepancies in the assembly (see Supplementary Fig. 3). The 2 species have a similar number of genes (24,656 protein-coding genes were predicted in *P. muralis*; Andrade et al. 2019).

Table 2: Genome assembly statistics.

Measure	rPodRaf1
Total length	1.513 Gb
Number of scaffolds	28
Scaffold L50/N50	7 scaffolds ; 93.6 Mb
Longest scaffold	139.1 Mb
Number of contigs	53
Contig L50/N50	10 contigs ; 61.4 Mb
Longest contig	104.8 Mb
BUSCO completeness:	97.3%
Single copy	5,095
Duplicated	67
Fragmented	34
Missing	114
Total	5,310

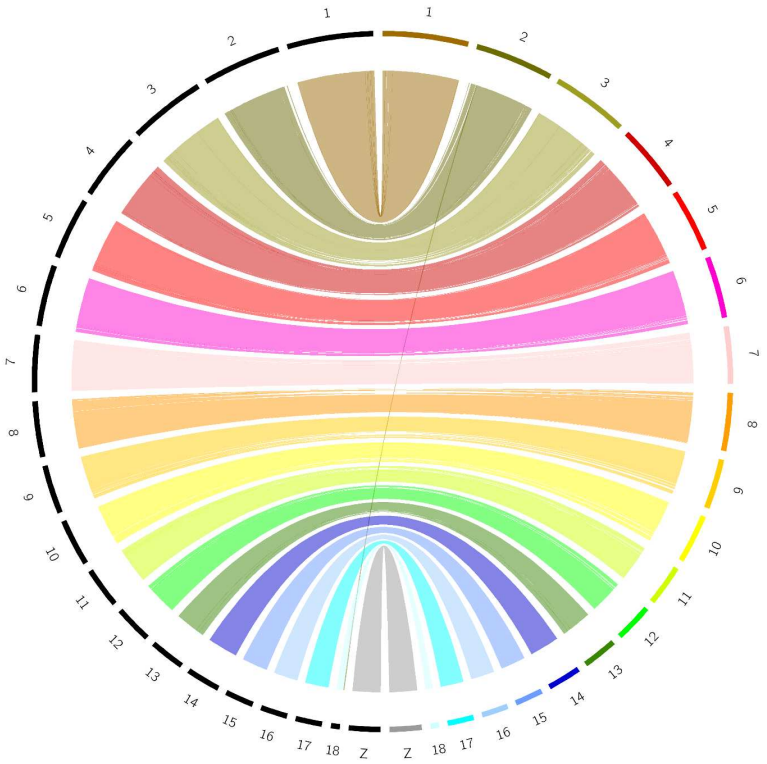


Fig. 2. Comparison of the chromosomal structure between the 18 autosomal chromosomes and Z chromosome between *P. raffonei* (right) and *P. muralis* (left). The different colors correspond to the different chromosomes of *P. raffonei*. The chromosomes were aligned using minimap2 and the resulting alignment between fragments longer than 1 Mb is represented with a ribbon plot using Circos.

Discussion

We present here the first chromosome-scale genome assembly for the Aeolian wall lizard (scaffold N50 of 93.6 Mb). Several metrics indicate that our genome assembly possesses a very high quality being chromosome-scale, accurate and complete. It constitutes a useful resource for squamates, a group composed of ~11,000 species for which only 29 high-quality genome assemblies are currently available (Card et al. 2023). In comparison to the other squamates, the *P. raffonei* assembly has a high scaffold N50 and the highest BUSCO completeness score (see Supplementary Table 6). The alignment between the genomes of *P. raffonei* and *P. muralis* showed a very high synteny, suggesting that both assemblies are structurally accurate and that the 2 species share a very similar chromosomal organization. Only 1 segment of the chromosome 2 of *P. raffonei* mapped to the chromosome 18 of *P. muralis*. This finding could be a biological chromosomal rearrangement between these 2 species (that belong to distinct clades of the genus *Podarcis;* Salvi et al. 2021; Yang et al. 2021) or a disjunction in the genome assembly of *P. muralis*.

The genome assembly of the Aeolian wall lizard, one of the most endangered vertebrate species in Europe, is a useful resource to better plan conservation efforts. Previous studies have highlighted that the Aeolian wall lizard exhibits low levels of genetic diversity and that the populations inhabiting different islands show a very reduced gene flow, constituting additional threats to this species (Capula 2004). Accordingly, our genome assembly suggests a very low heterozygosity (0.177% as estimated by GenomeScope), the lowest value documented among 7 species belonging to distinct squamate families (see Supplementary Table 6). The genome resequencing of several individuals from different islands is in progress to comprehensively characterize the genetic diversity of this species and evaluate its extinction risk.

References

Andrade P, Pinho C, Pérez i de Lanuza G, Afonso S, Brejcha J, Rubin C-J, Wallerman O, Pereira P, Sabatino SJ, Bellati A, et al. Regulatory changes in pterin and carotenoid genes underlie balanced color polymorphisms in the wall lizard. Proc Natl Acad Sci USA. 2019;116(12):5633–5642.

Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. http://www.bioinformatics.babraham.ac.uk/ projects/fastqc/. [accessed 2022 January 8].

Bonardi A, Francesco Ficetola G, Razzetti E, Canedoli C, Falaschi M, Parrino EL, Rota N, Padoa-Schioppa E, Roberto S. ReptIslands: Mediterranean islands and the distribution of their reptile fauna. Glob Ecol Biogeogr. 2022;31(5):840–847.

Capula M. Low genetic variation in a critically endangered Mediterranean lizard: conservation concerns for Podarcis raffonei (Reptilia, Lacertidae). Ital J Zool. 2004;71(suppl 1):161–166.

Capula M, Luca L, Marco AB, Arianna C. The decline of the Aeolian wall lizard, Podarcis raffonei: causes and conservation proposals. Oryx. 2002;36(1):66–72.

Card DC, Bryan Jennings W, Scott VE. Genome evolution and the future of phylogenomics of non-avian reptiles. Animals. 2023;13(3):471.

Chan S, Lam E, Saghbini M, Bocklandt S, Hastie A, Cao H, Holmlin E, Borodkin M. Structural variation detection and analysis using Bionano optical mapping. In: Bickhart DM, editor. Copy number variants: methods and protocols. Methods in Molecular Biology. New York (NY): Springer; 2018. p. 193–203.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18(2):170–175.

Cheng H, Jarvis ED, Fedrigo O, Koepfli K-P, Urban L, Gemmell NJ, Li H. Haplotype-resolved assembly of diploid genomes without parental data. Nat Biotechnol. 2022;40(9):1332–1335.

Ficetola GF, Barzaghi B, Melotto A, Muraro M, Lunghi E, Canedoli C, Lo Parrino E, Nanni V, Silva-Rocha I, Urso A, et al. N-mixture models reliably estimate the abundance of small vertebrates. Sci Rep. 2018;8(1):10357.

Ficetola GF, Silva-Rocha I, Carretero MA, Vignoli L, Sacchi R, Melotto A, Scali S, Salvi D. Status of the largest extant population of the critically endangered Aeolian lizard Podarcis raffonei (Capo Grosso, Vulcano Island). PLoS One. 2021;16(6):e0253631.

Formenti G, Abueg L, Brajuka A, Brajuka N, Gallardo-Alba C, Giani A, Fedrigo O, Jarvis ED. Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs. Bioinformatics. 2022a:38(17):4214–4216.

Formenti G, Theissinger K, Fernandes C, Bista I, Bombarely A, Bleidorn C, Ciofi C, Crottini A, Godoy JA, Höglund J, et al.; European Reference Genome Atlas (ERGA) Consortium. The era of reference genomes in conservation genomics. Trends Ecol Evol. 2022b:37(3):197–202.

Geneva AJ, Park S, Bock DG, de Mello PLH, Sarigol F, Tollis M, Donihue CM, Reynolds RG, Feiner N, Rasys AM, et al. Chromosome-scale genome assembly of the brown Anole (Anolis sagrei), an emerging model species. Commun Biol. 2022;5(1):1–13.

Ghurye J, Pop M, Koren S, Bickhart D, Chin C-S. Scaffolding of long read assemblies using long range contact information. BMC Genomics. 2017;18(1):527.

Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, Phillippy AM, Koren S. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLoS Comput Biol. 2019;15(8):e1007273.

Gippoliti S, Capula M, Ficetola GF, Salvi D, Andreone F. Threatened by legislative conservationism? The case of the critically endangered Aeolian lizard. Front Ecol Evol. 2017;5:130. doi:10.3389/ fevo.2017.00130

Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics. 2020;36(9):2896–2898.

Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMarkET and AUGUSTUS. Bioinformatics. 2016;32(5):767–769.

Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-genome annotation with BRAKER. Methods Mol Biol. 2019;1962:65–95.

Howe K, Chow W, Collins J, Pelan S, Pointon D-L, Sims Y, Torrance J, Tracey A, Wood J. Significantly improving the quality of genome assemblies through curation. GigaScience. 2021;10(1):giaa153.

IUCN. Podarcis raffonei. The IUCN Red List of Threatened Species; 2009;e.T61552A12514822.

Kerpedjiev P, Abdennur N, Lekschas F, McCallum C, Dinkla K, Strobelt H, Luber JM, Ouellette SB, Azhir A, Kumar N, et al. HiGlass: webbased visual exploration and analysis of genome interaction maps. Genome Biol. 2018;19(1):125.

Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–360.

Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19(9):1639–1645. doi:10.1101/ gr.092759.109

Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–3100.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–2079.

Lo Cascio P, Grita F, Guarino L, Speciale C. A little is better than none: new insights into the natural history of the Aeolian wall lizard Podarcis raffonei from La Canna stack (Squamata Sauria). Naturalista sicil. 2014;38(2):355–366.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for

scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2021;38(10):4647–4654.

Mapleson D, Venturini L, Kaithakottil G, Swarbreck D. Efficient and accurate detection of splice junctions from RNA-Seq with Portcullis. GigaScience. 2018;7(12):giy131. doi:10.1093/gigascience/ giy131

Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, Johnson J, Li K, Mobarry C, Sutton G. Aggressive assembly of pyrosequencing reads with mates. Bioinformatics. 2008;24(24):2818–2824.

Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 2019;20(1):275.

Pasquesi GIM, Adams RH, Card DC, Schield DR, Corbin AB, Perry BW, Reyes-Velasco J, Ruggiero RP, Vandewege MW, Shortt JA, et al. Squamate reptiles challenge paradigms of genomic repeat element evolution set by birds and mammals. Nat Commun. 2018;9(1):2774.

Pinto BJ, Keating SE, Nielsen SV, Scantlebury DP, Daza JD, Gamble T. Chromosome-level genome assembly reveals dynamic sex chromosomes in neotropical leaf-litter geckos (Sphaerodactylidae: Sphaerodactylus). J Hered. 2022;113(3):272–287.

Podnar M, Pinsker W, Mayer W. Complete mitochondrial genomes of three lizard species and the systematic position of the Lacertidae (Squamata). J Zool Syst Evol Res. 2009;47(1):35–41.

Ranallo-Benavidez TR, Kamil SJ, Michael CS. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. 2020;11(1):1432.

Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021;592(7856):737–746.

Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: referencefree quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 2020;21(1):245.

Roscito JG, Sameith K, Pippel M, Francoijs K-J, Winkler S, Dahl A, Papoutsoglou G, Myers G, Hiller M. The Genome of the tegu lizard Salvator merianae: combining Illumina, PacBio, and optical mapping data to generate a highly contiguous assembly. GigaScience. 2018;7(12):giy141.

Salvi D, Pinho C, Mendes J, James Harris D. Fossil-calibrated time tree of Podarcis wall lizards provides limited support for biogeographic calibration models. Mol Phylogenet Evol. 2021;161: 107169.

Smit AFA, Hubley R, Green P. RepeatMasker. n.d. http://repeatmasker. org. [accessed 2022 March 1].

Yan H, Bombarely A, Li S. DeepTE: a computational method for de novo classification of transposons with convolutional neural network. Bioinformatics. 2020;36(15):4269–4275.

Yang W, Feiner N, Pinho C, While GM, Kaliontzopoulou A, James Harris D, Salvi D, Uller T. Extensive introgression and mosaic genomes of Mediterranean endemic lizards. Nat Commun. 2021;12(1):2762

Chromosome-level reference genome of the Ponza grayling (*Hipparchia sbordonii*), an Italian endemic and endangered butterfly

Authors

Sebastiano Fava[1], Marco Sollitto[2], Mbarsid Racaku[2], Alessio Iannucci[3], Andrea Benazzo[4], **Lorena Ancona[1]**, Paolo Gratton[5], Fiorella Florian[2], Alberto Pallavicini[2], Claudio Ciofi[3], Donatella Cesaroni[5], Marco Gerdol[2], Valerio Sbordoni[5], Giorgio Bertorelle[4], Emiliano Trucchi[1]

Affiliations

1 Department of Life and Environmental Sciences, Marche Polytechnic University, Ancona, Italy

2 Department of Life Sciences, University of Trieste, Trieste, Italy

3 Department of Biology, University of Florence, Florence, Italy

4 Department of Life Sciences and Biotechnology, University of Ferrara, Ferrara, Italy

5 Department of Biology, University of Rome "Tor Vergata", Rome, Italy

* Corresponding author:

Abstract

Islands are crucial evolutionary centers, providing unique opportunities for differentiation of novel biodiversity and thriving endemic species. Islands are also fragile ecosystems where the distinctive biodiversity they host is more exposed to environmental and anthropogenic pressures than on continents.

The Ponza grayling, *Hipparchia sbordonii*, is an endemic butterfly species which is currently supposed to be present in two tiny islands of the Pontine archipelago, Italy, occupying an area which is smaller than 10 km^2. It has been classified as Endangered by the IUCN because of the extremely limited area of occupancy, the severe population fragmentation and the recent demographic decline.

Using a combination of long and short read sequencing, bulk transcriptome RNA sequencing and synteny analysis with phylogenetically close butterflies, we produced a highly contiguous chromosome-scale annotated reference genome for the Ponza

grayling, including 28 autosomes and the Z sexual chromosomes. The final assembly spans 388.61 Gb with a contig N50 of 14.5 Mb and a BUSCO completeness score of 98.5%.

Such high-quality genomic resource for the *Hipparchia sbordonii* opens up new opportunities for detailed estimates of genetic diversity and genetic load to better assess its conservation status, and for the investigations of the genomic novelties characterizing the evolutionary path of this endemic island species.

Keywords

Conservation genomics, Island biogeography, Endemic species, *Endemixit*, Nymphalidae

Introduction

Although islands contribute only 6.7% of land surface area, they harbor ~20% of the Earth's biodiversity (Sayre et al., 2018)(Kier et al., 2009). Unfortunately, they also account for ~50% of the threatened species and 75% of the known extinctions since European expansion around the globe (Russell and Kueffer, 2019). Due to their geological and geographical history and characteristics, islands act simultaneously as cradles of evolutionary diversity and museums of formerly widespread lineages, achieving outstanding endemicity (Cronk, 1997). Nevertheless, the majority of these endemic species are inherently vulnerable due to genetic and demographic factors linked with the way islands are colonized (Fernández-Palacios et al. 2021). Additionally, island populations can be small in size and, by definition, they can not easily move to track their habitat, therefore, are often more at risk of extinction (Frankham, R. 1997). Small populations are characterized by reduced genetic diversity, higher effects of genetic drift and, hence, higher realized genetic load (Bertorelle et al 2022). Genomics is emerging as an effective tool for conservation, providing more detailed estimates of different types of genetic diversity, like, for example, of adaptive genetic variation (Stange et al 2021), which could be used to inform targeted and effective strategies to protect endangered species (Segelbacher et al 2022). Reference genomes are the necessary first step in conservation genomics as they constitute the backbone for whole-genome population-level investigations to properly assess annotated genetic diversity, also as structural variants (Pokrovac & Pezer 2022). The production of highly contiguous genomes has greatly accelerated in the last decade, refining our understanding of the genomic basis of organismal traits, chromosome evolution and allowing the detection of recent selection (Zhang et al 2021). Here we present the high-quality chromosome-level genome of the endangered island endemic Ponza grayling, *Hipparchia sbordonii (*Nymphalidae Satyrinae*)*. The Ponza grayling is found only in the Pontine archipelago, located East of Naples, with an extremely restricted range of occurrence (**Figure 1A**). The historical total area of occupancy is limited to the three islands of Ponza, Palmarola and Zannone and has been estimated to be as small as 16 km^2 (https://www.iucnredlist.org/species/173231/64640021). Ponza, the main island of the Ponzian archipelago, is about 70 km from the Aurunci mountains, the likely

source range of the ancestral population that could have colonized the archipelago from the continent. The second-largest island in the archipelago, Ventotene, is about 40 Km South-East of Ponza, but it has no records of the presence of this butterfly. (**Figure 1B**). Moreover, the butterfly has not been found in Palmarola and Zannone in recent years (Bonelli et al. 2018), thus further reducing its distribution. As a result, the status of the Ponza grayling as Endangered in the Red List of Endangered Species of the IUCN (2009) could be reviewed for the worse. The main threat to its survival appears to be improper land management with the implementation of new agricultural practices in spite of the traditional ones, which were more favorable to the survival of *H. sbordonii* populations (Bonelli et al. 2018). In addition, reduced hunting activity and poaching of sparrows has led to an increase in the number of birds such as *Muscicapa striata*, an insectivorous bird specialized in preying on insects in flight, like *Hipparchia sbordonii*, likely leading to higher predation pressure on the grayling (Sbordoni 2018). The high-quality reference genome of the Ponza grayling is a valuable resource to investigate the genomic consequences of thriving at small (and further reducing) population size in order to propose accurate conservation strategies. It also constitutes the basis to explore the genomic features characterizing the unique evolutionary pathways of this butterfly, a natural experiment of island biodiversity.

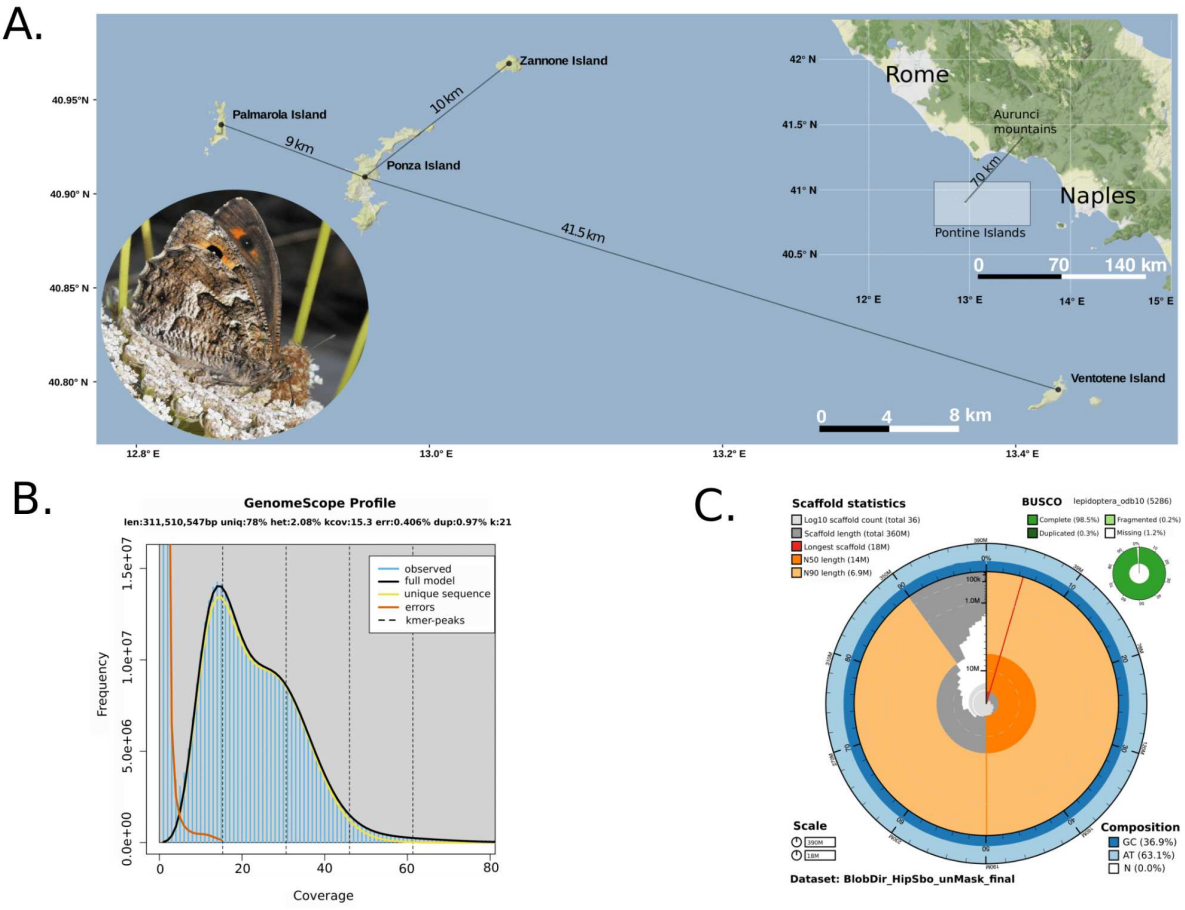


Figure 1: A) An individual of *Hipparchia sbordinii*, in Ponza Island (Photo credit: Valerio Sbordoni) and the geographic area of *H. sbordonii* population included in this study. B) K-mer spectra output and corresponding genome size and heterozygosity estimated with GenomeScope 2.0. C) BlobToolKit Snail plot showing a graphical representation of the quality metrics for the Hipparchia sbordonii primary assembly (HipSbo_unMask_final).

Materials and Methods

Sampling, genomic DNA extraction and sequencing

One specimen of Hipparchia sbordonii was sampled in Ponza Island in June 2019. The individual was immediately frozen in liquid nitrogen to preserve the integrity of nucleic acids. High molecular weight DNA was isolated from head and thorax using the Nanobind Tissue big DNA kit (Circulomics Inc., Baltimore, USA). DNA quality and fragment length were checked by pulse field gel electrophoresis and DNA concentration was measured with fluorometric and spectrophotometric assays using a Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, California, US) and a TECAN Nanoquant Infinite 200 Pro (Tecan Mannedorf, Switzerland), respectively. Fragments of 30 Kbp length were selected using a Blue Pippin device (Sage Science; Beverly, MA, USA). Isolated fragments were used to prepare the DNA library with a SMRTbell express template prep kit 2.0 (Pacific Biosciences, Menlo Park, California, US) according to manufacturer's protocols. The library was run on four PacBio SMRT Cells 1M in continuous long-read sequencing (CLR) mode on a PacBio Sequel platform. Extracted DNA was also used to construct a short-read genomic library using a Illumina DNA PCR-Free Prep Kit (Illumina) according to the manufacturer's protocol. Target coverage was 10-15X. The library was sequenced paired-end on an Illumina NovaSeq 6000 System using a 300-cycle Reagent Kit v1.5.

RNA extraction and sequencing

Upon sagittal dissection, approximately half of the body of an adult male individual was placed in a plastic tube with 1 ml RNA-Solv reagent (Omega Bio-tek, Norcross, GA, USA) and five paramagnetic beads. RNA extraction was performed following the manufacturer’s instructions, after grinding the tissues for 1 min with a bead-beater homogenizer. RNA was further purified using a Direct-zol™ RNA Miniprep kit (Zymo Research, Irvine, CA, USA), with an additional DNaseI treatment to remove residual genomic DNA contamination. Total extracted RNA was used as an input for the preparation of a poly(A)-selected library with a TruSeq library preparation kit (Illumina, San Diego, CA, USA), which was subjected to RNA-sequencing on an Illumina NovaSeq

6000 platform at the Genomic Core Facility of AREA Science Park (Trieste, Italy), using a 2 × 150 bp paired-end sequencing strategy. Raw reads were trimmed with fastp (Chen at al. 2018), removing sequencing adapters and nucleotides characterized by poor quality scores. After trimming, reads shorter than 75 nucleotides were discarded.

Primary genome assembly

In this study, we applied a multi-assembler approach to reconstruct chromosomal-scale genome without using Hi-C data (**Figure 2**). Firstly, PacBio CLR reads and Illumina reads were filtered to remove remnant adapter sequences. After trimming with Trimmomatic (Bolger et al. 2014), Illumina short reads were used to estimate the genome size using a k-mer based approach with Jellyfish (Marçais et al. 2011). The distribution of k-mers (*k* = 31) was calculated and GenomeScope 2.0 (Ranallo-Benavidez et al. 2020) was then used to infer the genome size, repeat content, and genome-wide heterozygosity. The genome of *Hipparchia sbordonii* was assembled with CANU (Koren et al. 2017) using PacBio CLR long reads, as a set of primary contigs, representing the initial haploid assembly, and separating alternative haplotypic variants. Following the assembly, PacBio subreads were mapped to assembly and sorted using pbmm2 package (https://github.com/PacificBiosciences/pbmm2/) from SMRT analysis software (https://github.com/PacificBiosciences/pbbioconda). The mapped PacBio subreads, were then used for polishing of the contigs with GCpp v 2.0.2 (https://github.com/PacificBiosciences/gcpp) that use Arrow algorithm (https://github.com/PacificBiosciences/GenomicConsensus). Following polishing with GCpp, we carried out two rounds of polishing using the Illumina reads to further fix the indel errors in the contigs with POLCA (POLishing by Calling Alternatives) (Zimin et al. 2020). Primary contigs were then processed with purge_dups (Guan et al. 2020) to identify residual haplotype duplication in the assembly. Such duplicated sequences were moved to the alternate assembly to remain with a final set of non-redundant haplotypic variants. The contigs included in the primary assembly were anchored to scaffolds, exploiting long read information, using LRscaf (Qin et al. 2019). TGS-GapCloser was then used to fill the gaps originated during the scaffolding step due to the presence of repeats or regions characterized by low coverages (Xu et al. 2020). After gap filling, a final polishing step was performed with POLCA, using Illumina reads,

to correct any errors made in the scaffolding and gap filling steps. The assembly was subjected to a second round of analysis with Purge_dups to remove any duplicates that may have been added during the last scaffolding and gap filling steps, obtaining the haploid primary genome assembly. The two haplotypes fasta files generated by the purging process, representing duplicated genomic sequences most likely ascribable to the alternative haplotype, were merged. The GetOrganelle and MITOS toolkits were then used to assemble and annotate the mitochondrial genome sequence, respectively (Jin et al. 2020) (Bankevich et a.l 2012) (Bernt et al. 2013).

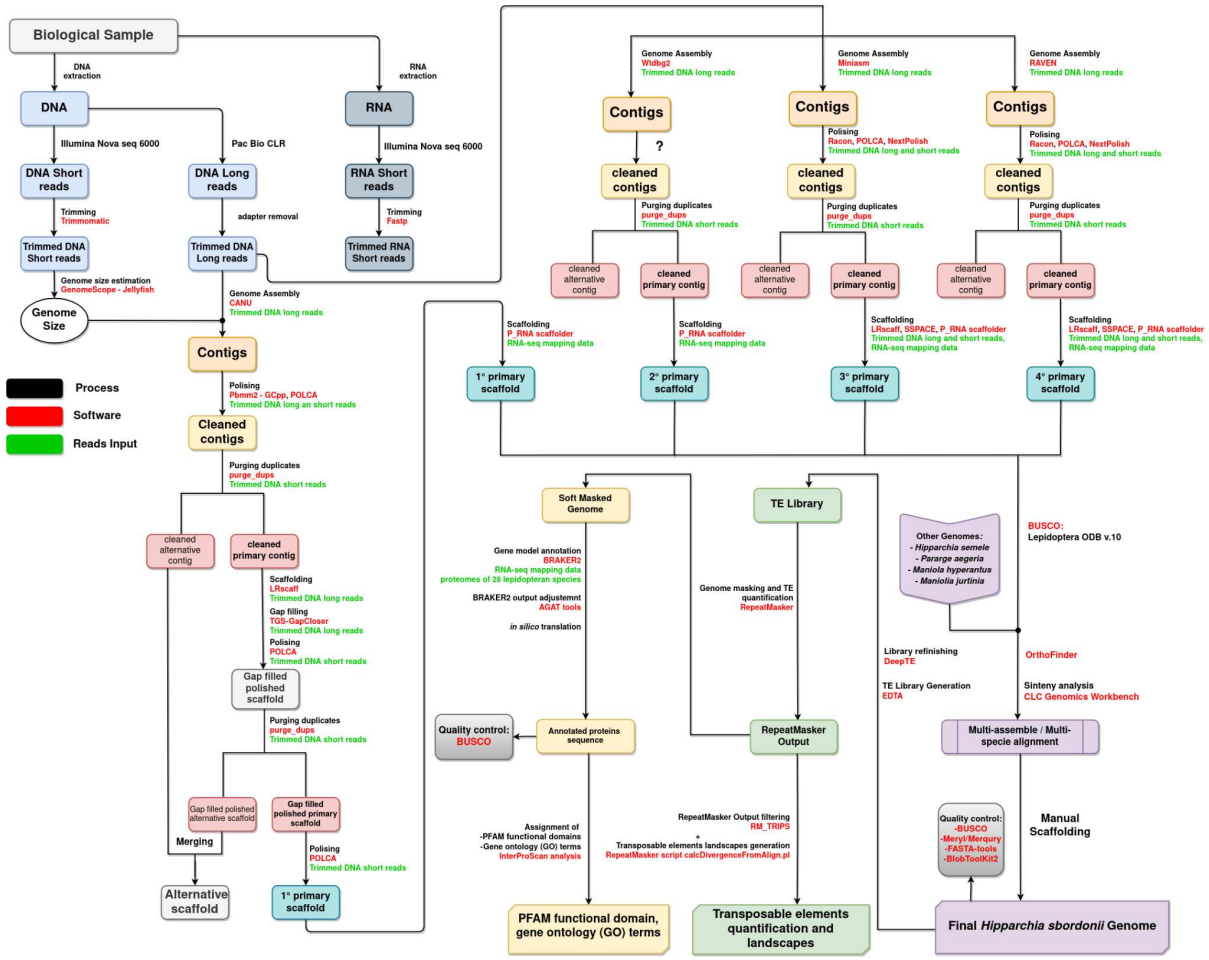


Figure 2: General bioinformatic pipeline to assemble, compare and annotate the *Hipparchia sbordonii* chromosomal-scale genome utilizing a multi-assembler approach, using short and long reads as well as hybrid approaches.

Auxiliary genome assemblies

The genome was also assembled using PacBio long reads with other three distinct assemblers, Wtdbg2 (Ruan and Li, 2020), miniasm (Li, 2016), and Raven (Vaser and Šikić, 2021), to explore whether employing a different assembly algorithm could result in the assembly of challenging regions and help with the manual curation described in the following paragraph. Following the assembly, three distinct polishing tools were sequentially employed to correct the errors associated with the assembled contigs. Namely, Racon (Vaser et al., 2017), POLCA (Zimin et al. 2020) and NextPolish (Hu et al., 2020) were utilized, exploiting PacBio long reads, genomic Illumina short reads and a combination of both read types, respectively. Upon polishing, the primary contigs generated by the three assembly algorithms underwent further processing with purge_dups (Guan et al., 2020) to detect and remove haplotype duplications, leading to non-redundant draft assemblies. Three distinct scaffolders, namely LRScaf (Qin et al., 2019), SSPACE (Boetzer et al., 2011), and P_RNA scaffolder (Zhu et al., 2018) were employed to improve the contiguity of the assembly, by exploiting genomic PacBio long reads, genomic Illumina short reads, and RNA-seq mapping data, respectively.

Manual curation and synteny analysis

The primary genome assembly obtained with Canu was selected as the reference genome for further manual curation due to its superior quality compared to the three other assemblies. BUSCO (Manni et al. 2021) analysis using the Lepidoptera ODB v.10 database for both the reference and the three alternative assemblies allowed to assign coordinates to single-copy orthologous genes conserved across all lepidopteran insects. These annotated genes served as anchors for filling the gaps present in the reference assembly, by utilizing alternative assemblies whenever those gaps were resolved. However, since the BUSCO Lepidoptera database had a relatively small number of orthologous genes (5286 genes), we used OrthoFinder (Emms et al. 2019) to increase the density of detected 1:1 orthologous genes, thereby enhancing the efficiency of the gap-filling and scaffold reordering process. OrthoFinder was further used to identify 1:1 orthologous genes between *H. sbordonii* and the available genomes of other species belonging to Satyrinae. We selected species based on their

phylogenetic relatedness and the availability of a chromosome-scale genome assembly. The species that fulfilled these criteria were *Hipparchia semele* (GCA_933228805.1), *Pararge aegeria* (GCF_905163445.1), *Maniola jurtina* (GCF_905333055.1), and *Maniola hyperantus* (GCF_902806685.1). The approach involved leveraging synteny and orthologous gene information from closely related species to enhance the genome assembly of *H. sbordonii*, with the aim to achieve a near-chromosome level of assembly. The Whole Genome Alignment plugin in CLC Genomics Workbench21® (Qiagen, Hilden, Germany) was used to align the genome assemblies, facilitating the visualization of large-scale events like inversions and translocations. To this end, following preliminary tests, whole-genome alignments were computed by setting the seed value to 115 and by not allowing mismatches, which reduced background noise and ensured a stringent alignment. Whenever a reliable match between *H. sbordonii* and the other Satyrinae species was found, the sequence information from the alternate assemblies was used to close the gaps between the scaffolds of the reference genome. This strategy was strictly applied only when the hit was unequivocal, ensuring accurate merging. In summary, we merged distinct scaffolds into super-scaffolds only if the following criteria were met: (1) the joining of the two neighboring scaffolds was supported by at least one of the three available auxiliary assemblies; (2) the relative placement of the two joined scaffolds was corroborated by synteny data from at least one chromosome-scale genome assembly of a species belonging to the Satyrinae subfamily.

Genome assemblies quality assessment

Prior to gene annotation, we assessed the quality of the genome assembly using two independent methods. First, we used BUSCO quality control tool to check for genome completeness using a set of conserved single-copy orthologous genes. We ran BUSCO v5.2.2 in the genome mode with default parameters against the lepidopteran dataset included in ODB v.10 (lepidoptera_odb10). Second, we used Merqury v1.3 (Rhie et al. 2020) to estimate the base level accuracy (QV) and the assembly completeness, by comparing the k-mers represented in the assembly and those observed in the Illumina reads. All assembly metrics were computed using FASTA-tools (https://github.com/b-brankovics/fasta_tools). To have a summarized graphical

representation of the quality of the genome assembly we employed BlobToolKit2 (Challis et al. 2020).

Repetitive elements, gene models and ncRNA annotation

To identify and annotate repetitive elements, we first generated a *de novo* repeat library using the Extensive *denovo* TE Annotator (EDTA) v1.9.9 (Ou et al. 2019). Subsequently, we refined the library using DeepTE (Yan et al. 2020), which employs convolutional neural networks to classify unknown elements at the order and superfamily levels. Then we used RepeatMasker v4.1.2 (Smit et al. 2013-2015) with the final library to mask the genome and we parsed the RepeatMasker output file with RM_TRIPS script (https://github.com/clbutler/RM_TRIPS). Transposable elements landscapes were generated using the RepeatMasker script calcDivergenceFromAlign.pl. The final version of the *Hipparchia sbordonii* genome underwent gene model annotation using BRAKER2 (Hoff et al., 2019), a comprehensive pipeline that combines *ab initio* gene prediction tools, sequence homology information and transcriptomic evidence. For this purpose, the proteomes of 28 lepidopteran species were obtained from NCBI and used to create a custom reference protein database for homology detection. Moreover, Illumina paired-end RNAseq data generated from the whole body of a single *H. sbordonii* individual was supplied to provide transcriptomic evidence. AGAT tools (Dainat et al., 2022) was employed to adjust the output of BRAKER. The final set of annotated proteins was evaluated using BUSCO. The protein sequences generated from the *in silico* translation of annotated gene models were subjected to InterProScan analysis (Jones et al., 2014), to assign PFAM functional domain (Mistry et al., 2021), and gene ontology (GO) terms (Ashburner et al., 2000). Additionally, INFERNAL (Kalvari et al., 2018) with cmscan was used to annotate the most conserved classes of non-coding RNAs (ncRNAs), and tRNAscan-SE was used to predict transfer RNA (tRNA) genes.

Table 1. Pipeline and software used for the genome assembly.

Assembly	Software	Version
K-mer counting	Meryl	1.3
Estimation of genome size and heterozygosity	GenomeScope2	2.0
De novo assembly (contigging)	Canu	2.1.1
Remove low-coverage, duplicated contigs	purge_dups	1.2.5
polishing with short reads Illumina	Polca	
polishing with long reads PacBio	Arrow	2.0.2
Scaffolding	LRScaf	1.1.9
Short-read alignment	Bwa	2.2.1
SAM/BAM processing	Samtools	1.12
Gap filling	TGS-GapCloser	1.1.1
Genome assembly refinement		
Manual curation	OrthoFinder 1	2.5
	CLC Genomics Workbench®	21
Genome quality assessment	fasta_tools	
Basic assembly metrics	BUSCO	5.5.0
Assembly completeness	Merqury	1.3
General contamination screening	BlobToolKit	2.3.3
Repeat identification and annotation	EDTA	1.9.9

	DeepTE	Commit babd65e
	RepeatMasker	4.1.2
Gene annotation	BRAKER2	2.1.6
Mapping RNA-seq reads genome	HISAT2	2.2.1
Transcriptome assembly	Oyster River Protocol	2.3.3
Comparison to *H. semele*	Orthofinder	2.5
Genome–genome alignment	CLC Genomics Workbench®	21
Synteny visualization	RIdeogram	
Mitochondrial genome assembly	GetOrganelle	
Mitochondrial genome annotation	MITOS	

Results

The final genome size of *H. sbordonii* (388.61 Mb) is consistent with the size predicted by the k-mer spectra with Genomescope2.0 (**Figure 1C**) and closely resembles the genome size of *H. semele* for which a reference genome is available (403 Mb; NCBI Accession: GCA_933228805.2). The k-mer spectrum shows a bimodal distribution with two major peaks, at ~15 and ~30-fold coverage, corresponding to heterozygous and homozygous states, respectively. Based on Illumina reads, we estimated a 0.406% sequencing error rate and 2.08% nucleotide heterozygosity rate (**Figure 1B**). The mitochondrial genome size is 15,321 bp, in agreement with the mitochondrial genome size of its sister species *H. semele* (15,223 bp; OW121739.2)(**Table S4**; **Figure S2**). The primary assembly of *H. sbordonii* contains 36 scaffolds with a N50 of 14.5 Mb and the longest scaffold of 17.7 Mb (**Table S1**; **Figure 1C**). The alternative assembly contains 1606 scaffolds spanning 352.9 Mb, having a N50 of 409.7 Kb. The completeness of the primary assembly is very high, with a BUSCO completeness score of 98.5% ([S:98.2%, D:0.3%], F:0.2%, M:1.3%) using the lepidoptera gene set, a k-mer completeness of 83,01 %, and a per-base quality value (QV) of 40.87. In total, 16,346 protein-coding genes were predicted. The BUSCO completeness of the gene annotation using the lepidoptera gene set was 97% ([S:96.2%, D:0.8%], F:1.2%, M:2.3%). Following the annotation of non-coding RNA, we identified 6,068 putative tRNAs, of which 24 as potential suppressor tRNAs, 614 unknown isotypes, and 5430 standard tRNAs . Results are consistent with those obtained on *H. semele* (**Table S5**).

The alignment of the genomes of *H. sbordonii* and *H. semele*, revealed that despite the presence of high synteny in the genomes of all Satyrinae, we could detect the presence of 10 intra-chromosomal inversions exceeding a size of 10 Kb predicted to have occurred after the split between *H. sbordonii* and *H. semele*. The main features of these inversions are summarized in **Table S2** and their chromosomal locations are highlighted in the dot plot shown in **Figure 3B**. Although most of these inversions were relatively small, four of them exceeded 100Kb. In detail, one 104 Kb inverted block was detected in *H. semele* chromosome 6, a 169 Kb inversion was observed in chromosome 13 and a 604 Kb terminal inversion was found in chromosome 18. Nevertheless, the most significant inversion event involved chromosome 25, with over 2,7 Mb of genomic

sequence (including 26 protein-coding genes) being found in a reverse order between *H. semele* and *H. sbordonii*. In addition to the aforementioned inversions, the interspecies comparison also allowed the identification of a 220 Kb-long intrachromosomal inversion in chromosome 27.

Also comparing our assembly to another phylogenetically-close member of the family Satyrinae, i.e. *Maniola jurtina*, which displayed the same karyotype (Wiemers et al. 2020), we confirmed that 23 out of the 28 autosomes expected to be present *H. sbordonii* were correctly assembled to their full-length **(Figure S1**). In detail, chromosomes 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 17, 18, 19, 20, 21, 22, 24, 25 and 28 (numbers refer to the karyotype of *H. semele*) were complete. On the other hand, the sequences of chromosomes 15, 18, 23 and 26 were split between two scaffolds in the *H. sbordonii* assembly. Chromosome 27 displayed the highest level of fragmentation, corresponding to three contigs in the Ponza grayling. About the two chromosomes involved in sex determination, W was not present in the *H. sbordonii* reference genome due to the fact that the sequenced individual was a male. On the other hand, the Z chromosome was present and matched two contigs in *H. sbordonii* (**Table S1**).

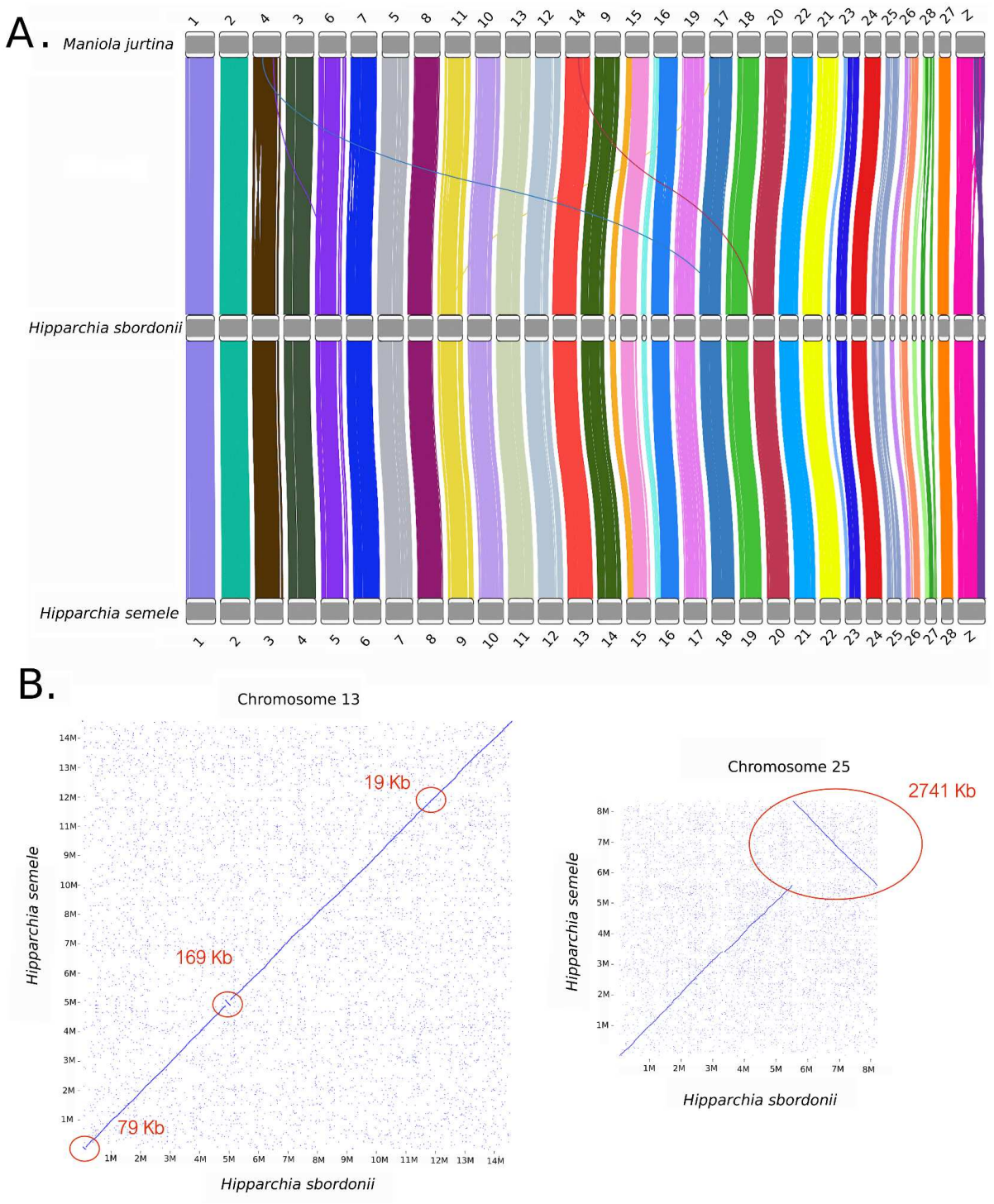


Figure 3. A) Conserved synteny between the chromosome scale assemblies of *Maniola jurtina* (version ilManJurt1.1, top), *Hipparchia semele* (ilHipSeme1.2, bottom) and the genome assembly of *Hipparchia sbordonii* reported in this study (middle). Synteny blocks are highlighted by lines connecting the orthologous genes identified in the three species, coloured based on their placement on each of the 36 scaffolds obtained in *H. sbordonii*. Note that the W chromosome is not reported in this plot due to its absence in the *H. sbordonii* reference genome, obtained from a male individual. B) Dot plot depicting two synteny relationships between chromosome 13 and 25 of *H. semele* and the corresponding scaffolds of *H. sbordonii* (inversions are highlighted by red circles).

Discussion

Island endemics have a greater susceptibility to anthropogenic changes due to small range size, geographic isolation and a peculiar evolutionary history characterized by potentially low initial founding size and long-term maintenance of small populations; island endemics are therefore predisposed to lower genetic diversity and higher rates of inbreeding (Frankham et al. 1997). We presented the high-quality chromosome-scale genome assembly for the Ponza grayling (36 scaffolds, N50: 14.5 Mb), a beneficial asset to investigate the genomic peculiarities of this endangered endemic island butterfly. It also constitutes a useful resource for studying butterfly evolution in general, a group composed of ~157,000 species for which only 766 genome assemblies are currently available (searched in NCBI database with keywords "butterfly assembly", 21/02/2024). In comparison to the other butterfly genomes, *H. sbordonii* assembly has a high scaffold N50 and one of the highest BUSCO completeness scores (**Figure 1C**). The alignment between the genomes of *H. sbordonii* and *H. semele* showed a very high synteny, suggesting that both assemblies are structurally accurate and that the two species share a very similar chromosomal organization. The creation of a high-density synteny map between *H. sbordonii* and *H. semele* also offered the opportunity to investigate the genomic architecture of the two species, highlighting the presence of structural variants. As recently reported in other species, including butterflies, chromosomal rearrangements, such as inversions and translocations, even if characterized by recent occurrence, may provide a strong contribution to reproductive isolation, establishing barriers to gene flow (Le Moan et al., 2023; Mackintosh et al., 2023). Such a preliminary overview of the structural variants present in the *Hipparchia* genus might be highly relevant for the upcoming population genomics analyses planned on both *H. sbordonii* and mainland graylings. In this study, we successfully reconstructed a chromosomal-scale genome utilizing a multi-assembler approach, bulk transcriptome RNA sequencing, and synteny analysis with phylogenetically close butterflies (**Figure 3**). Despite missing Hi-C data, our genomic reconstruction exhibits notable integrity, as evidenced by N50 of 14.5 Mb and BUSCO completeness score of 98.5%. While Hi-C data is a powerful tool for refining genome structures, our strategy showcases the efficacy of alternative methodologies in

achieving chromosomal-scale assemblies, especially in cases where Hi-C might be impractical to get or present technical challenges.

References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ... & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. Nature genetics, 25(1), 25-29.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., ... & Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology, 19(5), 455-477.

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., ... & Stadler, P. F. (2013). MITOS: improved de novo metazoan mitochondrial genome annotation. Molecular phylogenetics and evolution, 69(2), 313-319.

Bertorelle, G., Raffini, F., Bosse, M., Bortoluzzi, C., Iannucci, A., Trucchi, E., ... & Van Oosterhout, C. (2022). Genetic load: genomic estimates and applications in non-model animals. Nature Reviews Genetics, 23(8), 492-503.

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., & Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics, 27(4), 578-579.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114-2120.

Bonelli, S., Casacci, L. P., Barbero, F., Cerrato, C., Dapporto, L., Sbordoni, V., ... & Balletto, E. (2018). The first red list of Italian butterflies. Insect Conservation and Diversity, 11(5), 506-521.

Challis, R., Richards, E., Rajan, J., Cochrane, G., & Blaxter, M. (2020). BlobToolKit–interactive quality assessment of genome assemblies. G3: Genes, Genomes, Genetics, 10(4), 1361-1374.

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics, 34(17), i884-i890.

Cronk, Q. C. (1997). Islands: stability, diversity, conservation. Biodiversity & Conservation, 6, 477-493.

Dainat J, Hereñú D, LucileSol, pascal-git . 2022. NBISweden/AGAT: AGAT-v0.8.1. Zenodo Available from: https://zenodo.org/record/5834795.

Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome biology, 20, 1-14.

Fernández-Palacios, J. M., Kreft, H., Irl, S. D., Norder, S., Ah-Peng, C., Borges, P. A., ... & Drake, D. R. (2021). Scientists' warning–The outstanding biodiversity of islands is in peril. Global Ecology and Conservation, 31, e01847.

Frankham, R. (1997). Do island populations have less genetic variation than mainland populations?. Heredity, 78(3), 311-327.

Guan, Dengfeng, Shane A. McCarthy, Jonathan Wood, Kerstin Howe, Yadong Wang, and Richard Durbin. 2020. 'Identifying and Removing Haplotypic Duplication in Primary Genome Assemblies'. Bioinformatics (Oxford, England) 36 (9): 2896–98. https://doi.org/10.1093/bioinformatics/btaa025.

Hoff, K. J., Lomsadze, A., Borodovsky, M., & Stanke, M. (2019). Whole-genome annotation with BRAKER. Gene prediction: methods and protocols, 65-95.

Hu, J., Fan, J., Sun, Z., & Liu, S. (2020). NextPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics, 36(7), 2253-2255.

Jin, J. J., Yu, W. B., Yang, J. B., Song, Y., DePamphilis, C. W., Yi, T. S., & Li, D. Z. (2020). GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome biology, 21, 1-31.

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., ... & Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics, 30(9), 1236-1240.

Kalvari, I., Nawrocki, E. P., Argasinska, J., Quinones-Olvera, N., Finn, R. D., Bateman, A., & Petrov, A. I. (2018). Non-coding RNA analysis using the Rfam database. Current protocols in bioinformatics, 62(1), e51.

Kier, G., Kreft, H., Lee, T. M., Jetz, W., Ibisch, P. L., Nowicki, C., ... & Barthlott, W. (2009). A global assessment of endemism and species richness across island and mainland regions. Proceedings of the National Academy of Sciences, 106(23), 9322-9327.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome research, 27(5), 722-736.

Le Moan, A., Stankowski, S., Rafajlovic, M., Martinez Ortega, O., Faria, R., Butlin, R., & Johannesson, K. (2023). Coupling of 12 chromosomal inversions maintains a strong barrier to gene flow between ecotypes. bioRxiv, 2023-09.

Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics, 32(14), 2103-2110.

Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22(13), 1658-1659.

Mackintosh, A., Vila, R., Laetsch, D. R., Hayward, A., Martin, S. H., & Lohse, K. (2023). Chromosome fissions and fusions act as barriers to gene flow between Brenthis fritillary butterflies. Molecular biology and evolution, 40(3), msad043.

Manni, M., Berkeley, M. R., Seppey, M., & Zdobnov, E. M. (2021). BUSCO: assessing genomic data quality and beyond. Current Protocols, 1(12), e323.

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics, 27(6), 764-770.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L., ... & Bateman, A. (2021). Pfam: The protein families database in 2021. Nucleic acids research, 49(D1), D412-D419.

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R., Hellinga, A. J., ... & Hufford, M. B. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome biology, 20(1), 1-18.

Pokrovac, I., & Pezer, Ž. (2022). Recent advances and current challenges in population genomics of structural variation in animals and plants. Frontiers in Genetics, 13, 1060898.

Qin, M., Wu, S., Li, A., Zhao, F., Feng, H., Ding, L., & Ruan, J. (2019). LRScaf: improving draft genomes using long noisy reads. BMC genomics, 20(1), 1-12.

Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nature Communications, 11(1), 1-10.

Rhie, A., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome biology, 21(1), 1-27.

Ruan, J., & Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. Nature methods, 17(2), 155-158.

Russell, J. C., & Kueffer, C. (2019). Island biodiversity in the Anthropocene. Annual Review of Environment and Resources, 44, 31-60.

Sayre, R., Noble, S., Hamann, S., Smith, R., Wright, D., Breyer, S., ... & Reed, A. (2019). A new 30 meter resolution global shoreline vector and associated global islands database for the development of standardized ecological coastal units. Journal of Operational Oceanography, 12(sup2), S47-S56.

Sbordoni, V., Aspetti genetici ed ecologici del declino di popolazioni di farfalle e altri insetti, in Atti Accademia Nazionale Italiana di Entomologia, LXVI, 2018, pp. 159-168.

Segelbacher, G., Bosse, M., Burger, P., Galbusera, P., Godoy, J. A., Helsen, P., ... & Buzan, E. (2022). New developments in the field of genomic technologies and their relevance to conservation management. Conservation Genetics, 23(2), 217-242.

Smit A, Hubley R, Green P. 2013–2015. RepeatMasker Open-4.0. http://www.repeatmasker.org.

Stange, M., Barrett, R.D.H. & Hendry, A.P. The importance of genomic variation for biodiversity, ecosystems and people. Nat Rev Genet 22, 89–105 (2021). https://doi.org/10.1038/s41576-020-00288-7

Vaser, R., & Šikić, M. (2021). Time-and memory-efficient genome assembly with Raven. Nature Computational Science, 1(5), 332-336.

Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. Genome research, 27(5), 737-746.

Wiemers, M., Chazot, N., Wheat, C. W., Schweiger, O., & Wahlberg, N. (2020). A complete time-calibrated multi-gene phylogeny of the European butterflies. ZooKeys, 938, 97.

Xu, M., Guo, L., Gu, S., Wang, O., Zhang, R., Peters, B. A., ... & Zhang, Y. (2020). TGS-GapCloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. GigaScience, 9(9), giaa094.

Yan, H., Bombarely, A., & Li, S. (2020). DeepTE: a computational method for de novo classification of transposons with convolutional neural network. Bioinformatics, 36(15), 4269-4275.

Zhang, L., Steward, R. A., Wheat, C. W., & Reed, R. D. (2021). High-quality genome assembly and comprehensive transcriptome of the painted lady butterfly Vanessa cardui. Genome Biology and Evolution, 13(7), evab145.

Zhu, B. H., Xiao, J., Xue, W., Xu, G. C., Sun, M. Y., & Li, J. T. (2018). P_RNA_scaffolder: a fast and accurate genome scaffolder using paired-end RNA-sequencing reads. BMC genomics, 19, 1-13.

Zimin, A. V., & Salzberg, S. L. (2020). The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. PLoS computational biology, 16(6), e1007981.

Supplementary Materials

Chapter 1: Evolutionary dynamics of transposable elements activity and regulation in the Apennine yellow-bellied toad (*Bombina pachypus*)

Supplementary Table S1:
Values of normalised TE counts for the brain tissue

TE order/family	BR100	BR103	BR95	BR87	BR92	BR402	Average Expression	Standard Deviation
LTR/BEL	4061	3987	4024	3687	4037	5138	4156	501
LTR/Copia	14374	14971	14740	14969	15281	14820	14859	301
LTR/ERV	17898	18571	18474	19206	19403	14138	17948	1943
LTR/Gypsy	323971	213951	276423	224246	206381	1236876	413641	405781
LTR/nc	36440	33985	35715	34345	33984	48154	37104	5505
DIRS	21339	22983	22715	24351	23690	13444	21420	4037
LINE/I	3862	4207	4033	4392	4251	2295	3840	779
LINE/Jockey	2173	1942	2196	1829	2157	887	1864	501
LINE/L1	46482	49368	48915	51473	50698	37077	47335	5311
LINE/R2	1269	1297	1361	1412	1372	801	1252	227
LINE/RTE	9339	7016	8075	8323	8084	7339	8029	814
LINE/nc	920	959	1011	899	952	668	902	121
SINE	2028	2017	2294	2402	2218	1646	2101	269
PLE	2999	3309	2795	2816	2845	5646	3401	1116
nLTR/nc	7541	7289	7313	7433	7226	10931	7956	1462
ClassI/nc	15249	16196	15883	16290	16159	18463	16373	1091

DNA/CACTA	17814	12018	15338	11590	10929	71465	23193	23794
DNA/Harbinger	23944	24929	24900	24319	25566	20518	24029	1808
DNA/hAT	76288	72836	72118	74025	74409	104869	79091	12710
DNA/Mutator	12782	12833	13182	12179	12979	10229	12364	1099
DNA/PiggyBac	230	157	131	199	181	810	285	260
DNA/TcMar	68163	62680	66625	64430	62875	103837	71435	16017
DNA/MITE	33419	33391	32856	31298	32205	38665	33639	2589
DNA/nMITE	24386	24018	23229	24973	23560	37099	26211	5369
DNA/Helitron	703	470	345	658	461	197	472	189
Unknown	41445	39867	39926	40622	40795	55435	43015	6113
Total RetroTE	509945	402049	465967	418073	398738	1418324	602183	402116
Total DNATE	257730	243332	248725	243671	243165	387690	270719	57577
Total TE	809120	685249	754618	702365	682697	1861448	915916	465759

Supplementary Table S2:

Values of normalised TE counts for the male gonad tissue

TE order/family	MG100	MG103	MG95	Average Expression	Standard Deviation
LTR/BEL	2463	2906	3092	2820	323
LTR/Copia	23903	23614	24912	24143	681
LTR/ERV	54135	64946	43131	54071	10908
LTR/Gypsy	455969	368656	355402	393342	54640
LTR/nc	64235	62005	69769	65337	3998
DIRS	21332	20445	21834	21204	703
LINE/I	2968	3137	3267	3124	150
LINE/Jockey	18365	2446	5946	8919	8366

LINE/L1	43162	40847	43188	42399	1344
LINE/R2	1586	1156	1347	1363	215
LINE/RTE	4642	3477	4463	4194	627
LINE/nc	1075	1061	1181	1106	66
SINE	5197	5473	5741	5471	272
PLE	47599	48843	58638	51693	6046
nLTR/nc	4188	4164	4296	4216	70
ClassI/nc	19810	17236	18249	18432	1297
DNA/CACTA	11651	16397	14470	14173	2387
DNA/Harbinger	23644	23862	24740	24082	580
DNA/hAT	102906	102519	105598	103675	1677
DNA/Mutator	10869	11462	12194	11508	664
DNA/PiggyBac	150	110	113	124	22
DNA/TcMar	106250	98742	100938	101976	3860
DNA/MITE	42624	43645	42166	42812	757
DNA/nMITE	27559	27586	26209	27118	788
DNA/Helitron	400	534	298	411	118
Unknown	38870	42190	38818	39959	1932
Total RetroTE	770632	670412	664455	701833	59656
Total DNATE	326052	324857	326726	325878	946
Total TE	1135555	1037459	1029999	1067671	58907

Supplementary Table S3:

Values of normalised TE counts for the female gonad tissue

TE order/family	FG87	FG92	FG402	Average Expression	Standard Deviation
LTR/BEL	3207	3045	3431	3228	194
LTR/Copia	35581	37087	42250	38306	3498
LTR/ERV	21383	20308	17115	19602	2220
LTR/Gypsy	252739	269198	521047	347662	150382
LTR/nc	120760	114667	157881	131103	23390
DIRS	25498	25439	12840	21259	7291
LINE/I	2229	2433	1229	1964	644
LINE/Jockey	1179	1134	567	960	341
LINE/L1	61945	61333	64153	62477	1484
LINE/R2	729	684	433	615	159
LINE/RTE	23467	19333	10651	17817	6541
LINE/nc	2034	1640	821	1498	619
SINE	1601	1703	2424	1909	449
PLE	4644	3059	11988	6564	4764
nLTR/nc	7425	6856	11799	8693	2704
ClassI/nc	38492	32563	56681	42579	12568
DNA/CACTA	14179	12983	28678	18613	8737
DNA/Harbinger	30185	31254	29243	30227	1006
DNA/hAT	113616	104467	163702	127262	31888
DNA/Mutator	5401	5953	5141	5499	415
DNA/PiggyBac	1995	1932	4417	2781	1417
DNA/TcMar	178576	174861	490808	281415	181349
DNA/MITE	37125	32449	49306	39627	8703
DNA/nMITE	53804	49922	70310	58012	10826

DNA/Helitron	21	6	21	16	8
Unknown	66180	63526	97591	75766	18948
Total RetroTE	602913	600483	915312	706236	181069
Total DNATE	434903	413828	841627	563452	241136
Total TE	1103996	1077836	1854530	1345454	441067

Supplementary Table S4:

Differentially expressed TEs between male and female gonad

TE	TE order/family	logFC	logCPM	F	PValue	FDR
TE_00000471	LTR/Gypsy	21,75	14,95	288,67	1,79E-07	3,11E-06
TE_00000710	DNA/TcMar	16,12	9,33	210,59	5,96E-07	6,47E-06
TE_00000523	DNA/hAT	15,42	8,72	111,15	1,10E-06	9,82E-06
TE_00000604	unknown	13,68	10,48	270,37	1,60E-09	2,15E-07
TE_00000671	LTR/Gypsy	13,16	12,22	384,16	2,13E-10	1,46E-07
TE_00000278	LTR/Gypsy	12,26	5,69	90,22	6,86E-07	7,13E-06
TE_00000685	LTR/Gypsy	11,88	5,17	163,90	1,84E-07	3,11E-06
TE_00000874	LINE/Penelope	11,86	5,31	53,02	2,88E-05	0,000110024594561
TE_00000560	LTR/Gypsy	11,84	12,91	94,00	5,52E-07	6,13E-06
TE_00000483	LTR/nc	11,64	5,00	81,85	4,38E-06	2,60E-05
TE_00000380	LTR/ERV	11,63	4,94	142,90	3,47E-07	4,41E-06
TE_00000727	DNA/Harbinger	11,59	4,91	69,61	8,94E-06	4,35E-05
TE_00000428	DNA/MITE	11,38	8,30	59,67	5,79E-06	3,08E-05
TE_00000478	LTR/nc	11,06	4,44	66,52	1,09E-05	5,08E-05
TE_00000927	LTR/Gypsy	10,57	3,95	96,77	1,09E-05	5,08E-05
TE_00000737	LTR/Gypsy	10,32	7,14	154,03	3,78E-08	1,25E-06
TE_00000769	DNA/CACTA	10,29	10,12	84,51	9,69E-07	9,09E-06
TE_00000482	DNA/MITE	9,86	3,63	93,15	2,46E-06	1,75E-05

TE_00000249	LTR/Copia	9,01	5,92	123,22	1,29E-07	2,60E-06
TE_00000755	LTR/ERV	8,60	6,39	79,30	1,35E-06	1,08E-05
TE_00000425	DNA/MITE	8,35	6,30	86,59	8,52E-07	8,30E-06
TE_00000705	DNA/Harbinger	8,31	8,49	83,27	1,05E-06	9,48E-06
TE_00000720	DNA/TcMar	8,07	6,83	113,92	1,97E-07	3,22E-06
TE_00000444	unknown	7,79	2,55	22,64	0,000802505158901	0,001820187525243
TE_00000662	DNA/MITE	7,61	7,73	136,20	7,44E-08	1,83E-06
TE_00000664	LTR/Gypsy	7,51	7,47	112,85	2,07E-07	3,24E-06
TE_00000474	LTR/nc	7,27	6,41	209,34	6,81E-09	4,77E-07
TE_00000387	LTR/Gypsy	7,26	7,36	122,23	1,34E-07	2,67E-06
TE_00000869	LTR/Gypsy	7,25	4,26	29,07	0,000320908623787	0,000811710048402
TE_00000824	DNA/nMITE	7,20	8,33	193,95	1,05E-08	5,68E-07
TE_00000247	unknown	7,16	9,52	146,26	5,03E-08	1,48E-06
TE_00000823	LTR/ERV	7,08	6,97	66,85	3,26E-06	2,13E-05
TE_00000159	SINE/tRNA	6,95	8,44	58,94	6,16E-06	3,18E-05
TE_00000208	LTR/Gypsy	6,85	11,54	270,11	1,61E-09	2,15E-07
TE_00000382	LTR/Gypsy	6,80	7,57	50,78	1,29E-05	5,69E-05
TE_00000495	LTR/BEL	6,77	6,95	51,30	1,22E-05	5,49E-05
TE_00000026	LTR/Gypsy	6,70	5,16	79,91	4,87E-06	2,75E-05
TE_00000623	DNA/TcMar	6,48	7,73	128,01	1,05E-07	2,23E-06
TE_00000568	LTR/ERV	6,41	12,47	154,93	3,66E-08	1,25E-06
TE_00000443	DNA/TcMar	6,38	9,99	70,01	2,57E-06	1,78E-05
TE_00000287	DNA/MITE	6,33	10,56	74,05	1,93E-06	1,43E-05
TE_00000475	DNA/TcMar	6,29	6,46	188,09	1,24E-08	6,14E-07
TE_00000379	DNA/TcMar	6,22	9,42	71,79	2,26E-06	1,64E-05
TE_00000535	LTR/nc	6,18	8,23	68,35	2,91E-06	1,95E-05
TE_00000598	SINE/tRNA	6,08	7,78	111,00	2,27E-07	3,40E-06
TE_00000011	DNA/hAT	5,90	9,91	115,94	1,79E-07	3,11E-06

TE_00000788	LTR/ERV	5,76	6,84	84,73	9,55E-07	9,05E-06
TE_00000452	LTR/Copia	5,74	6,97	169,90	2,19E-08	8,60E-07
TE_00000694	LTR/nc	5,70	2,85	48,15	1,67E-05	7,10E-05
TE_00000959	LTR/Gypsy	5,69	5,24	60,14	5,56E-06	3,01E-05
TE_00000193	SINE/tRNA	5,64	7,41	41,43	3,41E-05	0,000125818499875
TE_00000451	DNA/MITE	5,49	4,11	44,66	2,39E-05	9,45E-05
TE_00000749	DNA/TcMar	5,34	9,97	70,12	2,55E-06	1,78E-05
TE_00000225	LTR/nc	5,32	11,46	52,49	1,09E-05	5,08E-05
TE_00001024	LTR/nc	5,32	3,91	47,72	1,74E-05	7,33E-05
TE_00000240	LTR/Gypsy	5,31	4,92	66,31	3,40E-06	2,16E-05
TE_00000082	LTR/Gypsy	5,09	6,48	130,27	9,50E-08	2,18E-06
TE_00001035	DNA/hAT	4,97	1,36	6,79	0,026518206453036	0,040331538410609
TE_00000539	LTR/ERV	4,93	5,75	70,23	2,53E-06	1,78E-05
TE_00000750	DNA/nMITE	4,91	7,15	147,18	4,86E-08	1,48E-06
TE_00000761	LTR/Copia	4,87	6,36	228,50	4,15E-09	3,79E-07
TE_00001050	DNA/Helitron	4,73	8,43	68,96	2,78E-06	1,88E-05
TE_00000647	LINE/nc	4,66	3,56	29,99	0,000147888130527	0,000418262393655
TE_00000415	DNA/Mutator	4,50	7,42	226,07	4,41E-09	3,79E-07
TE_00000610	DNA/hAT	4,48	9,47	90,33	6,82E-07	7,13E-06
TE_00000553	DNA/nMITE	4,43	10,35	47,46	1,79E-05	7,50E-05
TE_00000808	LTR/Gypsy	4,27	3,06	30,74	0,00013269308372	0,000382511906143
TE_00000640	LTR/Copia	4,19	4,47	28,93	0,0001728789568	0,000477034982401
TE_00000773	DNA/nMITE	4,18	6,59	140,29	6,33E-08	1,76E-06
TE_00000239	LTR/nc	4,18	12,66	65,51	3,61E-06	2,27E-05
TE_00000740	DNA/MITE	4,17	7,95	113,50	2,01E-07	3,22E-06
TE_00000466	LTR/ERV	4,17	6,84	168,03	2,33E-08	8,60E-07
TE_00000630	ClassI/nc	4,08	2,98	6,46	0,026105119198674	0,039971043046041
TE_00000805	LINE/RTE	4,04	6,56	37,48	5,44E-05	0,000183384996068

TE_00000246	LTR/Copia	4,04	4,89	54,09	9,43E-06	4,53E-05
TE_00000763	DNA/hAT	3,93	9,65	59,57	5,84E-06	3,09E-05
TE_00001033	LTR/Gypsy	3,86	7,15	96,28	4,86E-07	5,51E-06
TE_00000722	LINE/nc	3,86	3,55	9,72	0,009023957213464	0,015469640937366
TE_00001068	DNA/TcMar	3,83	5,79	70,13	2,55E-06	1,78E-05
TE_00000053	DNA/hAT	3,76	7,07	132,45	8,67E-08	2,03E-06
TE_00000074	DNA/MITE	3,75	7,15	110,92	2,28E-07	3,40E-06
TE_00000544	DNA/MITE	3,73	6,99	204,01	7,87E-09	5,04E-07
TE_00000488	LTR/nc	3,73	4,05	31,75	0,000115004306916	0,000341047254991
TE_00000282	LTR/ERV	3,69	11,42	89,44	7,19E-07	7,34E-06
TE_00000262	DNA/hAT	3,68	3,83	41,70	3,31E-05	0,000123361251367
TE_00000497	LTR/Copia	3,67	11,10	87,74	7,95E-07	7,89E-06
TE_00000266	DNA/Harbinger	3,67	4,98	40,96	3,60E-05	0,000131768007754
TE_00000457	LTR/ERV	3,62	8,41	50,16	1,37E-05	5,95E-05
TE_00000035	LTR/nc	3,60	8,29	37,99	5,11E-05	0,000174561685696
TE_00000429	DNA/TcMar	3,59	5,88	43,98	2,57E-05	0,000100905858304
TE_00000467	DNA/hAT	3,58	11,98	74,93	1,81E-06	1,37E-05
TE_00000421	DNA/hAT	3,57	4,00	19,09	0,000940180590305	0,002073218737595
TE_00000526	LTR/Gypsy	3,52	8,73	72,60	2,13E-06	1,57E-05
TE_00000996	unknown	3,50	4,55	8,06	0,015060570261539	0,024409879378941
TE_00001011	LINE/RTE	3,50	9,39	39,82	4,11E-05	0,000146766048885
TE_00000549	DNA/hAT	3,49	12,36	61,21	5,09E-06	2,86E-05
TE_00000697	DNA/hAT	3,48	7,44	11,61	0,005288283555396	0,009474841370085
TE_00000220	LTR/Copia	3,46	5,15	19,37	0,00088846294325	0,001970230821931
TE_00000327	DNA/MITE	3,45	6,57	32,34	0,000105990590266	0,00031747302092
TE_00000551	LINE/Jockey	3,44	11,17	19,73	0,000828595132606	0,00186705278788
TE_00000790	DNA/MITE	3,43	3,91	24,93	0,000324942683817	0,000817904511462
TE_00000977	LINE/RTE	3,42	5,26	83,67	1,02E-06	9,40E-06

TE_00000952	unknown	3,42	4,32	47,13	1,85E-05	7,65E-05
TE_00001074	DNA/hAT	3,41	6,02	138,15	6,88E-08	1,78E-06
TE_00000896	LINE/I	3,39	5,10	84,18	9,88E-07	9,19E-06
TE_00000782	LTR/Gypsy	3,38	7,19	144,26	5,43E-08	1,56E-06
TE_00000216	DNA/TcMar	3,36	6,57	31,01	0,000127617780327	0,000369948172184
TE_00000010	DNA/MITE	3,35	9,77	208,66	6,94E-09	4,77E-07
TE_00000367	DNA/hAT	3,34	10,88	28,33	0,000189070912569	0,00050973116571
TE_00000842	LINE/R2	3,34	5,72	22,33	0,000508942111927	0,001235831198843
TE_00000331	DNA/hAT	3,33	3,35	17,76	0,001232858907709	0,002607193427778
TE_00000456	LTR/Gypsy	3,32	6,67	64,19	4,00E-06	2,46E-05
TE_00000829	LTR/ERV	3,32	7,76	53,42	1,00E-05	4,77E-05
TE_00000585	DNA/TcMar	3,28	6,51	57,88	6,74E-06	3,39E-05
TE_00000527	LINE/Penelope	3,28	3,56	12,94	0,003728605885971	0,006983523183888
TE_00000340	DNA/hAT	3,27	6,33	80,67	1,24E-06	1,01E-05
TE_00000210	DNA/MITE	3,26	8,44	46,19	2,03E-05	8,37E-05
TE_00000212	LINE/Penelope	3,24	13,92	32,77	0,000100016932046	0,000301805479156
TE_00000607	ClassI/nc	3,24	2,85	13,03	0,003649742443247	0,006860718035394
TE_00000536	LTR/Gypsy	3,24	10,41	59,29	5,97E-06	3,13E-05
TE_00000774	DNA/CACTA	3,24	5,38	54,43	9,14E-06	4,43E-05
TE_00000334	LTR/Gypsy	3,21	8,72	25,11	0,00031534835475	0,000803554326178
TE_00000626	DNA/hAT	3,20	11,06	50,64	1,30E-05	5,73E-05
TE_00000486	DNA/TcMar	3,20	7,52	61,95	4,79E-06	2,73E-05
TE_00000151	DNA/Harbinger	3,15	7,79	102,64	3,45E-07	4,41E-06
TE_00000033	DNA/MITE	3,14	6,44	50,67	1,30E-05	5,73E-05
TE_00000810	DNA/TcMar	3,14	4,22	32,95	9,75E-05	0,000296960672394
TE_00000350	LTR/ERV	3,13	8,92	43,11	2,83E-05	0,000108951588016
TE_00000229	DNA/Harbinger	3,10	2,47	19,48	0,000870362260869	0,001939986724011
TE_00000360	LTR/Gypsy	3,08	7,76	53,85	9,64E-06	4,61E-05

TE_00000322	DNA/TcMar	3,07	12,75	58,69	6,29E-06	3,20E-05
TE_00000541	LTR/nc	3,05	9,02	23,10	0,00044392767012	0,00109864114044
TE_00000803	LTR/Gypsy	3,05	11,03	109,43	2,45E-07	3,52E-06
TE_00000753	LTR/Gypsy	3,04	5,77	37,86	5,19E-05	0,000176867650121
TE_00000812	LTR/ERV	2,99	4,32	18,98	0,000960879456316	0,002114344560592
TE_00000641	LTR/nc	2,98	3,27	28,35	0,000188479033909	0,00050973116571
TE_00000673	DNA/TcMar	2,96	9,30	33,07	9,60E-05	0,000292968388168
TE_00001004	LTR/Gypsy	2,93	6,72	25,70	0,000286021307748	0,00073793497399
TE_00000300	LTR/Gypsy	2,93	7,07	129,34	9,88E-08	2,22E-06
TE_00001008	LTR/Gypsy	2,93	7,67	78,39	1,43E-06	1,14E-05
TE_00000508	nLTR/nc	2,91	5,94	22,61	0,000483715943737	0,001188559176039
TE_00000574	LINE/L1	2,89	9,50	50,49	1,32E-05	5,78E-05
TE_00000427	DNA/TcMar	2,83	4,03	16,23	0,00171237498808	0,003513262400991
TE_00000885	LTR/nc	2,83	4,05	15,88	0,001853534773003	0,00374334224215
TE_00000122	DNA/MITE	2,81	4,78	13,00	0,003677343924765	0,006900034418831
TE_00000283	LTR/nc	2,79	11,85	27,87	0,000203075486033	0,000545765368713
TE_00000762	DNA/nMITE	2,78	9,04	98,06	4,41E-07	5,11E-06
TE_00000286	LTR/Gypsy	2,78	4,44	37,21	5,62E-05	0,000187158297268
TE_00000328	DNA/TcMar	2,75	9,08	147,11	4,87E-08	1,48E-06
TE_00001062	DNA/hAT	2,75	10,98	52,22	1,12E-05	5,17E-05
TE_00000146	unknown	2,73	6,01	41,54	3,37E-05	0,000124586329047
TE_00001005	LINE/L1	2,73	7,02	22,45	0,000497846215249	0,001214603532238
TE_00000433	LTR/Gypsy	2,71	6,47	42,37	3,07E-05	0,000114763589454
TE_00000624	DNA/hAT	2,70	7,09	7,28	0,01956456736418	0,030825394686769
TE_00000301	LTR/Gypsy	2,70	9,43	13,61	0,00315813665484	0,006080591469767
TE_00000185	LTR/nc	2,68	4,61	38,35	4,89E-05	0,000168763965991
TE_00000578	LTR/Gypsy	2,64	5,00	27,33	0,000220525453946	0,000588067877188
TE_00000477	LTR/Copia	2,64	4,98	5,84	0,032702114801907	0,048840206187508

TE_00000409	DNA/MITE	2,61	6,29	19,82	0,000813666924401	0,001841456723644
TE_00000571	DNA/hAT	2,60	5,47	44,95	2,32E-05	9,27E-05
TE_00000924	unknown	2,60	6,39	17,47	0,00130924639226	0,002751817264384
TE_00000529	DNA/hAT	2,59	9,98	103,40	3,32E-07	4,34E-06
TE_00000706	DNA/hAT	2,58	8,62	29,69	0,000154535019199	0,000435738086922
TE_00000548	DNA/Harbinger	2,56	4,73	44,73	2,37E-05	9,44E-05
TE_00000934	DNA/TcMar	2,55	2,89	6,80	0,023078302879383	0,035922788192343
TE_00000175	unknown	2,55	5,54	32,31	0,000106476195657	0,000317582178953
TE_00000436	LTR/Gypsy	2,51	4,56	35,73	6,77E-05	0,000220385143349
TE_00000680	LTR/Gypsy	2,50	6,55	102,05	3,56E-07	4,43E-06
TE_00000112	DNA/TcMar	2,50	8,10	26,91	0,000235805180303	0,000622380936247
TE_00000819	DNA/MITE	2,49	6,77	60,88	5,23E-06	2,90E-05
TE_00000703	LTR/Gypsy	2,49	2,93	6,68	0,024083465479601	0,037206790980461
TE_00000670	DNA/MITE	2,47	7,21	14,17	0,002756793075975	0,00538827737577
TE_00000754	LTR/Gypsy	2,46	8,55	104,47	3,14E-07	4,26E-06
TE_00000616	LTR/Gypsy	2,46	7,45	19,37	0,000889658491298	0,001970230821931
TE_00000760	DNA/MITE	2,45	4,96	16,78	0,001518609504942	0,003153329998189
TE_00000489	SINE/tRNA	2,44	5,92	20,70	0,000687369355482	0,001608650074135
TE_00000564	DNA/TcMar	2,44	3,66	19,60	0,000849778204863	0,00190645892917
TE_00000793	unknown	2,44	4,24	10,84	0,006534156524342	0,011409897687175
TE_00000395	DNA/hAT	2,44	9,17	29,58	0,000157080992086	0,000440509738676
TE_00000336	LTR/Gypsy	2,42	8,20	56,14	7,84E-06	3,89E-05
TE_00000388	DNA/Mutator	2,42	7,65	91,29	6,45E-07	6,93E-06
TE_00000613	unknown	2,41	5,74	39,58	4,23E-05	0,000147823461438
TE_00000091	ClassI/nc	2,41	3,81	14,44	0,002583413148191	0,005100608257976
TE_00000606	LTR/nc	2,41	1,73	9,45	0,009767189350458	0,016578518765909
TE_00000689	LTR/nc	2,40	9,17	55,58	8,25E-06	4,07E-05
TE_00000632	LINE/RTE	2,40	8,86	26,30	0,000259659568124	0,000676688571474

TE_00000417	LINE/R2	2,39	9,33	52,12	1,13E-05	5,19E-05
TE_00000384	DNA/hAT	2,38	8,83	127,74	1,06E-07	2,23E-06
TE_00000802	LTR/Gypsy	2,37	4,49	21,98	0,000542180862969	0,001301234071127
TE_00000231	LTR/Gypsy	2,36	8,66	22,20	0,000520958059621	0,00125908364761
TE_00000343	DNA/MITE	2,36	5,75	27,64	0,000210412300755	0,000564014271114
TE_00001037	LINE/RTE	2,35	5,35	18,00	0,001173849822432	0,002497758797421
TE_00000743	LINE/L1	2,35	6,47	13,37	0,003347360977197	0,006397178756422
TE_00000089	DNA/TcMar	2,35	7,51	43,60	2,68E-05	0,000103965509096
TE_00000154	DNA/hAT	2,35	8,29	87,56	8,04E-07	7,90E-06
TE_00000540	LINE/Jockey	2,34	10,11	51,37	1,22E-05	5,49E-05
TE_00000258	LTR/Gypsy	2,32	9,90	8,93	0,011459663349984	0,019136525205799
TE_00000310	LTR/Gypsy	2,31	6,78	75,42	1,75E-06	1,33E-05
TE_00000929	DNA/hAT	2,30	6,09	14,02	0,002859022227979	0,005556517776411
TE_00000461	DNA/hAT	2,28	6,19	59,39	5,93E-06	3,12E-05
TE_00000422	LTR/Gypsy	2,28	5,83	40,26	3,90E-05	0,000140765297721
TE_00001026	DNA/Harbinger	2,25	4,58	6,42	0,02646570107412	0,040331538410609
TE_00000184	DNA/Harbinger	2,25	5,69	18,94	0,000968760131274	0,002122633663429
TE_00000517	DNA/MITE	2,24	8,16	118,56	1,59E-07	2,98E-06
TE_00000695	LTR/Gypsy	2,23	8,97	24,34	0,000358654083365	0,000900562077939
TE_00000993	DNA/TcMar	2,22	9,81	67,66	3,06E-06	2,03E-05
TE_00000615	LTR/Copia	2,22	10,43	22,41	0,000501922248457	0,001221659812281
TE_00000371	unknown	2,22	5,81	37,39	5,50E-05	0,000184783232898
TE_00000480	LTR/nc	2,22	5,73	43,81	2,62E-05	0,000102423237844
TE_00000621	ClassI/nc	2,21	9,56	22,71	0,000475429336647	0,001170985860191
TE_00000583	LTR/Gypsy	2,20	10,81	83,12	1,06E-06	9,48E-06
TE_00000046	DNA/MITE	2,19	7,04	37,76	5,26E-05	0,00017840094661
TE_00000207	DNA/hAT	2,17	7,71	81,52	1,17E-06	9,89E-06
TE_00000215	LTR/Gypsy	2,17	8,63	45,16	2,27E-05	9,15E-05

TE_00000236	SINE/tRNA	2,17	4,18	18,15	0,001137177352483	0,002429745399093
TE_00000779	DNA/TcMar	2,16	6,01	39,60	4,22E-05	0,000147823461438
TE_00000307	DNA/TcMar	2,15	7,99	34,58	7,85E-05	0,000249976687963
TE_00000338	LTR/nc	2,14	6,08	8,30	0,013936093876448	0,022792470492067
TE_00000042	DNA/hAT	2,14	8,74	83,13	1,06E-06	9,48E-06
TE_00000856	LTR/ERV	2,13	7,88	20,70	0,000687417328191	0,001608650074135
TE_00000521	LTR/ERV	2,13	3,61	20,74	0,000681849911491	0,001606550476391
TE_00000780	ClassI/nc	2,12	8,28	62,71	4,51E-06	2,63E-05
TE_00000407	ClassI/nc	2,11	6,54	18,66	0,001024493387941	0,002216513996551
TE_00000361	DNA/MITE	2,10	7,15	33,95	8,53E-05	0,000269169930543
TE_00000383	unknown	2,10	8,37	33,93	8,56E-05	0,000269169930543
TE_00000441	DNA/nMITE	2,10	9,64	12,65	0,004022271688632	0,007439040112309
TE_00000596	DNA/nMITE	2,07	9,51	14,46	0,002570819215882	0,005100608257976
TE_00000281	LTR/Gypsy	2,06	7,96	44,71	2,38E-05	9,44E-05
TE_00001066	DNA/hAT	2,04	6,18	9,71	0,009042252006477	0,015475296966309
TE_00000902	LINE/Penelope	2,03	6,07	28,71	0,000178704795767	0,000489490289692
TE_00000180	SINE/tRNA	2,00	7,22	26,69	0,000243880878488	0,000640420016794
TE_00000699	DNA/TcMar	-2,00	4,81	12,66	0,004004772825215	0,007419974067544
TE_00000744	DNA/Mutator	-2,01	4,38	17,73	0,001240070031278	0,002617080311409
TE_00000244	DNA/TcMar	-2,02	11,25	10,37	0,007454356054364	0,012929236047232
TE_00000620	LTR/nc	-2,02	6,68	13,05	0,003629747554152	0,006845695203837
TE_00000448	DNA/MITE	-2,04	7,01	42,75	2,94E-05	0,000112094084339
TE_00000173	DNA/MITE	-2,05	11,25	11,16	0,005976973272301	0,010598344359131
TE_00000890	DNA/hAT	-2,07	7,20	38,14	5,02E-05	0,000171947001184
TE_00000373	DNA/TcMar	-2,07	6,18	10,85	0,006518684214909	0,011402173067434
TE_00000206	LTR/Gypsy	-2,10	10,21	59,69	5,78E-06	3,08E-05
TE_00001040	DNA/hAT	-2,13	7,47	13,19	0,003500942191509	0,006629307048877
TE_00000007	DNA/Harbinger	-2,13	6,54	25,63	0,000289095258566	0,000744005752718

TE_00000352	LTR/Gypsy	-2,14	7,88	31,25	0,000123405924201	0,000360778792565
TE_00000601	DNA/TcMar	-2,16	5,89	26,60	0,000247411711457	0,000648042858436
TE_00000921	DNA/hAT	-2,16	7,64	81,95	1,14E-06	9,87E-06
TE_00000728	LTR/nc	-2,21	7,71	98,76	4,24E-07	4,98E-06
TE_00000311	LTR/nc	-2,21	7,27	20,78	0,000676434912234	0,001597438968937
TE_00001000	DNA/hAT	-2,23	2,21	9,87	0,008617138446479	0,014796816766665
TE_00000899	LTR/Gypsy	-2,24	5,83	26,58	0,00024842132422	0,000649040016696
TE_00000157	DNA/hAT	-2,25	9,61	36,40	6,22E-05	0,00020374198067
TE_00000954	LTR/Gypsy	-2,27	6,34	48,05	1,68E-05	7,12E-05
TE_00000795	DNA/hAT	-2,27	10,04	24,97	0,000322772075168	0,00081442733881
TE_00000224	LTR/Gypsy	-2,28	8,89	29,09	0,000168794001882	0,000468267231029
TE_00000936	DNA/hAT	-2,30	11,38	45,00	2,31E-05	9,26E-05
TE_00000414	LTR/Copia	-2,32	6,61	50,90	1,27E-05	5,65E-05
TE_00000492	LTR/ERV	-2,32	5,09	49,97	1,39E-05	6,01E-05
TE_00000798	unknown	-2,34	1,36	7,46	0,018427193069804	0,029256712689289
TE_00000844	DNA/nMITE	-2,34	9,45	68,90	2,79E-06	1,88E-05
TE_00000661	unknown	-2,34	7,06	42,69	2,96E-05	0,000112358835702
TE_00000741	DNA/TcMar	-2,34	5,92	51,98	1,15E-05	5,22E-05
TE_00000312	DNA/CACTA	-2,35	8,79	126,97	1,09E-07	2,26E-06
TE_00000867	DNA/MITE	-2,36	2,09	16,31	0,001682636857896	0,003465948330124
TE_00000134	DNA/TcMar	-2,36	8,13	39,78	4,13E-05	0,000146958755748
TE_00000323	LTR/nc	-2,37	5,90	44,23	2,50E-05	9,87E-05
TE_00000251	DNA/TcMar	-2,37	7,37	62,64	4,53E-06	2,63E-05
TE_00000094	LTR/Copia	-2,38	8,74	28,44	0,000186032865683	0,000507599295011
TE_00000674	DNA/hAT	-2,39	7,56	59,17	6,04E-06	3,15E-05
TE_00000922	DNA/nMITE	-2,39	6,54	84,72	9,56E-07	9,05E-06
TE_00001049	unknown	-2,40	9,62	51,30	1,22E-05	5,49E-05
TE_00000385	LTR/Gypsy	-2,42	8,37	102,36	3,50E-07	4,41E-06

TE_00000005	DNA/Harbinger	-2,42	9,77	50,03	1,38E-05	6,00E-05
TE_00000877	DNA/hAT	-2,44	6,18	34,42	8,02E-05	0,000254728068101
TE_00000439	unknown	-2,46	6,64	10,46	0,007264836299953	0,012639703880342
TE_00000377	DNA/TcMar	-2,46	8,03	86,03	8,82E-07	8,50E-06
TE_00000892	DNA/MITE	-2,48	3,08	22,10	0,000530933058048	0,001277209594186
TE_00000972	LINE/L1	-2,48	11,93	11,61	0,005286673218893	0,009474841370085
TE_00000971	DNA/nMITE	-2,50	11,61	45,29	2,24E-05	9,07E-05
TE_00000512	LTR/Gypsy	-2,51	7,18	32,92	9,80E-05	0,000297309527978
TE_00000263	ClassI/nc	-2,51	5,93	87,91	7,87E-07	7,89E-06
TE_00000859	DNA/nMITE	-2,53	9,08	40,43	3,83E-05	0,000139023560607
TE_00000969	DNA/hAT	-2,53	7,48	64,62	3,87E-06	2,39E-05
TE_00001030	DNA/Harbinger	-2,53	9,39	42,40	3,06E-05	0,000114763589454
TE_00000573	LTR/Copia	-2,53	4,81	58,71	6,28E-06	3,20E-05
TE_00000939	LTR/Gypsy	-2,53	2,65	18,78	0,0010000479927	0,002186545611159
TE_00001053	DNA/hAT	-2,54	4,80	16,05	0,001782192793795	0,003627658704529
TE_00000271	LTR/BEL	-2,55	2,25	10,87	0,006475295005194	0,011345508396198
TE_00001054	LTR/Copia	-2,58	7,76	18,73	0,001010258382119	0,002199549895246
TE_00000970	DNA/hAT	-2,58	11,28	33,69	8,83E-05	0,000274537664589
TE_00000797	DNA/hAT	-2,60	2,22	17,04	0,001437618173051	0,003003283308884
TE_00000677	DNA/TcMar	-2,61	8,14	109,38	2,45E-07	3,52E-06
TE_00000065	DNA/MITE	-2,63	7,74	66,57	3,33E-06	2,13E-05
TE_00000228	nLTR/nc	-2,64	7,31	9,99	0,008316514314219	0,014304404620457
TE_00001006	DNA/hAT	-2,66	7,38	48,05	1,68E-05	7,12E-05
TE_00000051	LTR/nc	-2,67	5,95	30,58	0,000135829080947	0,000389376698715
TE_00001072	DNA/hAT	-2,67	5,69	55,05	8,65E-06	4,25E-05
TE_00000784	LTR/nc	-2,68	3,06	17,63	0,001267163325868	0,002668801127133
TE_00000857	DNA/TcMar	-2,68	10,30	69,35	2,70E-06	1,85E-05
TE_00000716	unknown	-2,70	9,86	48,34	1,63E-05	7,00E-05

TE_00000935	DNA/hAT	-2,71	5,46	108,41	2,57E-07	3,64E-06
TE_00000925	DNA/hAT	-2,72	10,27	60,18	5,55E-06	3,01E-05
TE_00000879	DNA/hAT	-2,75	7,23	81,91	1,14E-06	9,87E-06
TE_00001012	DNA/MITE	-2,75	5,04	51,06	1,25E-05	5,60E-05
TE_00000794	DNA/MITE	-2,76	9,02	31,14	0,000125408663726	0,00036456828441
TE_00000245	DNA/MITE	-2,79	7,31	64,05	4,05E-06	2,47E-05
TE_00000302	LTR/nc	-2,82	9,46	32,80	9,96E-05	0,000301369311221
TE_00000378	DNA/MITE	-2,83	5,41	79,58	1,33E-06	1,08E-05
TE_00000534	DNA/TcMar	-2,86	4,73	33,72	8,80E-05	0,000274537664589
TE_00000342	DNA/TcMar	-2,86	10,06	118,09	1,62E-07	2,99E-06
TE_00000937	DNA/MITE	-2,96	4,29	28,34	0,000188989793245	0,00050973116571
TE_00000910	DNA/Harbinger	-2,98	9,24	56,35	7,70E-06	3,84E-05
TE_00000303	DNA/TcMar	-2,99	8,67	153,34	3,87E-08	1,25E-06
TE_00000462	LTR/Gypsy	-2,99	10,51	111,13	2,25E-07	3,40E-06
TE_00000257	DNA/Harbinger	-2,99	9,69	9,27	0,010298002831116	0,017365259676
TE_00000405	DNA/hAT	-3,00	8,24	210,22	6,65E-09	4,77E-07
TE_00000658	DNA/nMITE	-3,00	9,49	79,39	1,34E-06	1,08E-05
TE_00000707	DNA/hAT	-3,02	9,91	101,17	3,73E-07	4,53E-06
TE_00000912	DNA/hAT	-3,03	5,98	77,27	1,55E-06	1,21E-05
TE_00000944	DNA/TcMar	-3,05	3,93	34,64	7,79E-05	0,000249000857568
TE_00000978	DNA/TcMar	-3,07	6,96	52,50	1,09E-05	5,08E-05
TE_00000786	LTR/Gypsy	-3,07	9,86	63,14	4,35E-06	2,60E-05
TE_00000901	LTR/nc	-3,07	9,52	29,10	0,000168412699507	0,000468267231029
TE_00000289	LTR/Gypsy	-3,08	4,47	26,16	0,000265662117633	0,000688852526124
TE_00000516	LTR/Gypsy	-3,09	0,27	10,86	0,011205754781387	0,018742850785075
TE_00000988	DNA/hAT	-3,11	10,85	41,58	3,36E-05	0,000124563611545
TE_00000845	DNA/nMITE	-3,13	8,59	60,77	5,28E-06	2,91E-05
TE_00000955	LTR/nc	-3,13	7,90	106,14	2,88E-07	4,02E-06

TE_00000450	LTR/nc	-3,15	4,38	65,48	3,62E-06	2,27E-05
TE_00000666	LTR/ERV	-3,19	6,59	103,57	3,29E-07	4,34E-06
TE_00000868	unknown	-3,25	10,84	33,51	9,05E-05	0,000278642478643
TE_00000943	DNA/TcMar	-3,25	6,05	66,74	3,29E-06	2,13E-05
TE_00000818	LTR/Copia	-3,26	1,62	16,90	0,003521699516103	0,006656399085382
TE_00000205	DNA/hAT	-3,26	9,70	30,67	0,000134062804871	0,000385383884753
TE_00000458	DNA/hAT	-3,26	4,71	76,87	1,59E-06	1,23E-05
TE_00000979	DNA/hAT	-3,26	7,06	61,81	4,85E-06	2,75E-05
TE_00000739	DNA/hAT	-3,27	11,37	78,23	1,45E-06	1,14E-05
TE_00000410	LTR/nc	-3,28	7,93	63,53	4,22E-06	2,53E-05
TE_00000735	unknown	-3,29	7,74	41,38	3,43E-05	0,000126077567866
TE_00000479	unknown	-3,33	9,05	8,36	0,013697548650109	0,022437889217322
TE_00000359	DNA/TcMar	-3,33	13,46	82,17	1,12E-06	9,87E-06
TE_00000532	LTR/Copia	-3,34	9,36	138,59	6,77E-08	1,78E-06
TE_00000849	DIRS	-3,38	9,43	110,27	2,35E-07	3,46E-06
TE_00000498	LTR/nc	-3,38	13,59	113,33	2,03E-07	3,22E-06
TE_00000470	LTR/Gypsy	-3,41	12,16	101,31	3,70E-07	4,53E-06
TE_00000832	DNA/TcMar	-3,42	6,65	58,95	6,15E-06	3,18E-05
TE_00000940	LTR/nc	-3,42	8,15	35,60	6,88E-05	0,000221908150583
TE_00000951	LINE/RTE	-3,43	9,61	40,91	3,62E-05	0,000132042400601
TE_00000966	LTR/Gypsy	-3,58	6,96	30,11	0,000145377052519	0,000414445077901
TE_00000965	LTR/Gypsy	-3,60	6,14	133,23	8,40E-08	2,02E-06
TE_00000299	LTR/Copia	-3,60	7,21	64,90	3,79E-06	2,35E-05
TE_00000957	DNA/TcMar	-3,70	8,56	95,15	5,17E-07	5,80E-06
TE_00000198	DNA/TcMar	-3,71	8,49	262,92	1,87E-09	2,15E-07
TE_00000503	LTR/nc	-3,74	13,41	200,05	8,79E-09	5,04E-07
TE_00000900	unknown	-3,77	2,86	81,72	1,15E-06	9,87E-06
TE_00000852	DNA/hAT	-3,77	10,10	104,56	3,13E-07	4,26E-06

TE_00000142	LTR/Copia	-3,78	10,29	240,69	3,09E-09	3,19E-07
TE_00000648	DNA/Harbinger	-3,79	10,70	266,52	1,73E-09	2,15E-07
TE_00000725	ClassI/nc	-3,83	12,05	81,29	1,19E-06	9,96E-06
TE_00000009	LTR/Gypsy	-3,83	9,80	180,47	1,57E-08	7,02E-07
TE_00000052	LINE/RTE	-3,85	11,85	187,91	1,25E-08	6,14E-07
TE_00000177	LTR/Copia	-3,86	9,11	329,09	5,20E-10	1,79E-07
TE_00000587	DNA/MITE	-3,87	8,01	39,65	4,19E-05	0,000147823461438
TE_00000411	DNA/TcMar	-3,92	9,69	90,48	6,76E-07	7,13E-06
TE_00000886	DNA/nMITE	-3,95	9,93	100,06	3,96E-07	4,69E-06
TE_00001039	DNA/MITE	-4,04	6,85	72,29	2,18E-06	1,60E-05
TE_00000865	DNA/Harbinger	-4,05	6,12	100,45	3,88E-07	4,65E-06
TE_00000888	DNA/nMITE	-4,06	8,86	70,69	2,45E-06	1,75E-05
TE_00000161	LTR/Gypsy	-4,07	7,23	115,66	1,81E-07	3,11E-06
TE_00000455	LTR/nc	-4,09	9,49	137,54	7,05E-08	1,78E-06
TE_00000883	DNA/nMITE	-4,11	6,60	31,36	0,000121438567871	0,000356560044151
TE_00001069	LTR/Gypsy	-4,11	7,81	92,88	5,88E-07	6,46E-06
TE_00000866	LTR/nc	-4,12	10,10	139,02	6,65E-08	1,78E-06
TE_00000182	LTR/Gypsy	-4,12	6,65	168,91	2,26E-08	8,60E-07
TE_00000058	LTR/Gypsy	-4,14	10,87	278,26	1,36E-09	2,15E-07
TE_00000166	LTR/Gypsy	-4,14	12,40	201,60	8,42E-09	5,04E-07
TE_00000791	DNA/TcMar	-4,14	4,88	75,87	1,70E-06	1,31E-05
TE_00000933	DNA/MITE	-4,20	5,88	113,95	1,97E-07	3,22E-06
TE_00000502	unknown	-4,20	10,52	365,71	2,83E-10	1,46E-07
TE_00000001	LTR/Copia	-4,23	11,70	181,40	1,52E-08	7,02E-07
TE_00000928	LTR/nc	-4,30	7,75	60,30	5,49E-06	3,00E-05
TE_00000967	DNA/TcMar	-4,34	4,41	90,11	6,91E-07	7,13E-06
TE_00001052	unknown	-4,34	8,02	119,88	1,49E-07	2,87E-06
TE_00000814	unknown	-4,39	11,76	176,16	1,79E-08	7,40E-07

TE_00000806	DNA/Harbinger	-4,43	2,72	60,01	5,62E-06	3,02E-05
TE_00000335	unknown	-4,43	5,53	97,82	4,46E-07	5,12E-06
TE_00000873	LINE/L1	-4,45	10,54	119,76	1,50E-07	2,87E-06
TE_00000756	DNA/hAT	-4,52	11,80	117,12	1,70E-07	3,07E-06
TE_00001009	DNA/TcMar	-4,56	8,45	178,95	1,64E-08	7,06E-07
TE_00001063	LINE/Penelope	-4,67	8,51	88,65	7,53E-07	7,62E-06
TE_00001045	DNA/PiggyBac	-4,68	9,72	54,27	9,28E-06	4,48E-05
TE_00000853	DNA/nMITE	-4,96	9,88	292,64	1,02E-09	2,15E-07
TE_00000905	unknown	-5,29	7,44	155,80	3,55E-08	1,25E-06
TE_00000406	LTR/nc	-5,35	12,31	57,65	6,88E-06	3,44E-05
TE_00000438	unknown	-5,44	7,36	12,42	0,004266056765068	0,007847719396703
TE_00000296	DNA/TcMar	-5,76	12,83	80,91	1,22E-06	1,01E-05
TE_00000884	LINE/RTE	-5,77	8,69	128,20	1,04E-07	2,23E-06
TE_00000528	DNA/MITE	-5,97	7,47	29,52	0,000158322019921	0,000442238489553
TE_00000400	DNA/TcMar	-6,55	14,49	62,39	4,62E-06	2,65E-05
TE_00000496	unknown	-6,97	10,81	104,24	3,18E-07	4,26E-06
TE_00000945	LTR/Gypsy	-8,54	8,63	66,69	3,30E-06	2,13E-05

Supplementary Table S5:

Normalised count values of TE silencing gene pathways in the brain tissue

Gene	BR100	BR103	BR95	BR87	BR92	BR402	Average Expression	Standard Deviation
AGO1	10	11	9	12	9	5	9	2
AGO2	2	3	2	2	2	1	2	0
AGO4	8	11	10	11	11	19	12	4
PIWIL1	0	0	0	0	0	1	0	0
PIWIL2	0	0	0	0	0	0	0	0

PIWIL4	0	0	0	0	0	0	0	0
DICER	9	12	8	13	12	6	10	3
DROSHA	3	3	5	6	6	4	4	1
DGCR8	67	61	61	61	57	42	58	8
PLD6	0	0	0	0	0	0	0	0
MAEL	1	0	1	0	1	1	1	0
SETDB1	46	49	46	52	47	61	50	5
Trim28/KAP 1	11	17	16	23	20	0	15	7
HP1a	7	6	10	12	8	21	11	5
HP1b	12	12	11	13	12	18	13	2
HP1g	6	3	3	4	2	5	4	1
DNMT1	10	19	14	18	16	11	15	3
DNMT3A	5	6	4	7	6	2	5	1
PRMT5	16	19	15	20	15	34	20	6
CHD3	55	70	72	25	22	22	44	22
CHD4	17	18	17	30	24	22	22	5
CHD5	84	100	82	145	117	83	102	23
HDAC1	17	20	15	20	20	44	23	10
HDAC2	43	48	41	54	47	82	53	14
MBD2	6	7	7	9	9	1	7	3
MBD3	11	10	8	11	8	13	10	2
MTA1	7	8	7	10	8	7	8	1
MTA2	46	48	46	47	46	66	50	7
p66alpha	13	15	14	17	15	29	17	6
p66beta	40	43	33	56	43	28	40	9

RBBP4	0	0	0	0	0	0	0	0
RBBP7	122	115	37	118	43	154	98	43

Supplementary Table S6:

Normalised count values of TE silencing gene pathways in male and female gonad tissues

Gene	MG100	MG103	MG95	Average Expression	Standard Deviation	FG87	FG92	FG402	Average Expression	Standard Deviation
AGO1	8	6	5	6	2	10	9	7	8	1
AGO2	14	15	13	14	1	11	10	16	12	3
AGO4	6	7	7	7	1	10	11	21	14	6
PIWIL1	201	210	192	201	9	196	150	192	180	25
PIWIL2	54	49	38	47	8	43	36	47	42	5
PIWIL4	111	189	79	126	57	2	1	4	3	2
DICER	9	7	6	7	1	22	24	5	17	10
DROSHA	8	5	5	6	2	18	19	2	13	10
DGCR8	26	26	21	24	3	44	34	19	32	13
PLD6	30	33	24	29	5	52	64	83	66	16
MAEL	88	84	66	79	12	204	214	382	267	100
SETDB1	94	85	85	88	5	375	299	120	265	131
Trim28/KAP1	18	11	14	14	4	194	144	6	115	98
HP1a	31	34	38	35	3	143	182	518	281	206
HP1b	15	11	19	15	4	203	164	98	155	53
HP1g	277	239	276	264	21	466	449	520	478	37
DNMT1	55	46	38	47	9	766	560	381	569	193
DNMT3A	14	15	12	14	2	1	0	0	0	0

PRMT5	47	50	44	47	3	405	386	454	415	35
CHD3	148	111	93	117	28	8	6	2	5	3
CHD4	66	59	49	58	8	434	349	52	278	201
CHD5	4	4	3	3	1	0	0	1	0	0
HDAC1	67	62	60	63	4	183	170	175	176	6
HDAC2	44	42	27	38	9	216	226	184	209	22
MBD2	148	131	125	135	12	22	20	15	19	3
MBD3	20	20	14	18	3	39	31	19	30	10
MTA1	5	3	4	4	1	11	16	3	10	6
MTA2	69	55	66	63	8	240	222	142	202	52
p66alpha	66	59	58	61	4	514	393	151	353	185
p66beta	6	8	4	6	2	3	2	0	2	1
RBBP4	17	13	22	17	4	118	59	734	304	374
RBBP7	183	211	44	146	89	892	415	831	713	259

Supplementary Table S7:

Differentially expressed TE silencing gene pathways between male and female gonad

Gene	logFC	logCPM	F	PValue	FDR
PIWIL4	5,61	5,01	19,53	0,000837767404548	0,002789439373773
DNMT3A	4,97	2,60	56,63	7,01E-06	0,000126969640543
CHD3	4,42	5,72	77,68	1,38E-06	5,57E-05
CHD5	3,13	5,70	32,18	0,000103647579519	0,000648477445728
MBD2	2,81	5,39	19,96	0,000768733960486	0,002619109944095

p66beta	1,52	4,48	7,16	0,020186304289968	0,034859504385645
AGO4	-1,04	3,46	6,72	0,02359140508699	0,039691625061363
MTA1	-1,32	2,91	6,88	0,022258634654854	0,037788153390704
HDAC1	-1,48	6,15	17,83	0,00118393568929	0,00360799106324
SETDB1	-1,59	6,82	16,85	0,001461973072038	0,004228253629811
MTA2	-1,67	6,51	28,39	0,000179956130458	0,000940911047638
MAEL	-1,75	6,44	11,40	0,005503164705093	0,011971326857067
CHD4	-2,25	6,57	15,02	0,002204728069807	0,005797498811675
RBBP7	-2,28	8,04	11,13	0,005937593284885	0,012739847044089
HDAC2	-2,47	6,46	53,32	9,47E-06	0,00015035984194
p66alpha	-2,54	6,81	31,80	0,000109263557375	0,000671060203525
Trim28/KAP1	-3,00	5,31	6,45	0,025943158341325	0,042970489968416
HP1a	-3,03	6,40	32,12	0,000104433614972	0,000651958055066
PRMT5	-3,14	6,97	76,84	1,46E-06	5,74E-05
HP1b	-3,35	5,62	76,73	1,47E-06	5,75E-05
DNMT1	-3,61	7,33	78,19	1,33E-06	5,48E-05
RBBP4	-4,13	6,33	11,86	0,004860329132571	0,010824241021597

Supplementary Table S8:

Differentially expressed TE silencing gene pathways between brain and male gonad

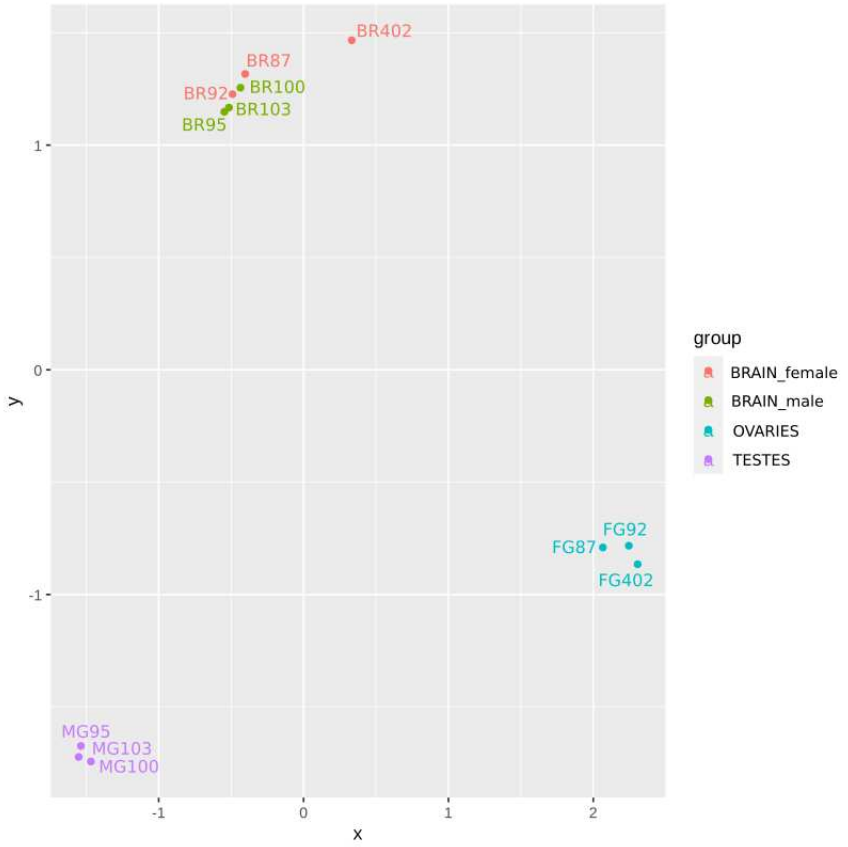
Gene	logFC	logCPM	F	PValue	FDR
CHD5	4,70	5,70	69,44	2,48E-06	9,68E-05
p66beta	2,69	4,48	21,00	0,000630497497095	0,003404836246187
DGCR8	1,37	5,43	15,50	0,001973446996459	0,008133156290217
PRMT5	-1,51	6,97	19,96	0,000769319213224	0,003946615732878
DNMT1	-1,67	7,33	19,56	0,000833204961407	0,004188729858989
CHD4	-1,74	6,57	9,23	0,010306888650887	0,029839407708797
HDAC1	-1,87	6,15	27,81	0,000197047002244	0,001469091071228
p66alpha	-2,14	6,81	23,20	0,000421765870416	0,002536140024681
HP1a	-2,17	6,40	17,62	0,001237229125272	0,005653616348686
AGO2	-2,74	2,89	46,31	1,89E-05	0,000311814303864
MBD2	-4,30	5,39	40,86	3,45E-05	0,000450801862546
HP1g	-6,10	7,55	187,45	1,10E-08	9,42E-06
MAEL	-6,71	6,44	106,93	2,50E-07	3,07E-05
RBBP4	-7,12	6,33	25,32	0,000293391563854	0,001949696913199
PIWIL2	-7,58	4,49	202,22	7,18E-09	8,95E-06
PLD6	-9,49	4,58	74,82	1,68E-06	7,89E-05
PIWIL1	-11,39	6,57	150,58	3,78E-08	1,39E-05
PIWIL4	-12,06	5,01	49,58	1,36E-05	0,000254881414044

Supplementary Table S9:

Differentially expressed TE silencing gene pathways between brain and female gonad

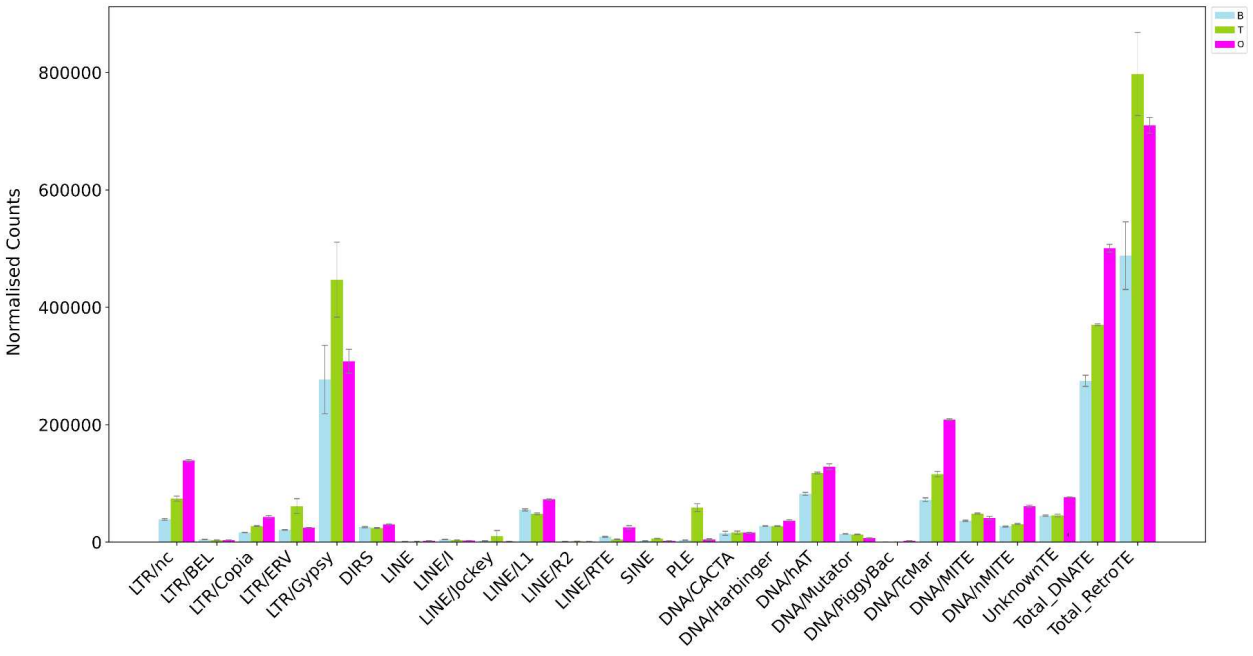
Gene	logFC	logCPM	F	PValue	FDR
CHD5	8,21	5,70	148,54	4,08E-08	3,21E-06

p66beta	4,32	4,48	46,50	1,85E-05	0,000130286936562
DNMT3A	3,47	2,60	31,27	0,000117862290903	0,000511157751824
CHD3	2,07	5,72	20,70	0,000667674035823	0,001963200541503
MBD3	-1,46	4,10	15,84	0,001828953598165	0,004429081584729
MBD2	-1,58	5,39	6,91	0,022001301888668	0,03461377556209
HDAC2	-1,77	6,46	28,72	0,000171172941327	0,000682071096018
MTA2	-1,93	6,51	37,08	5,43E-05	0,000285216072625
SETDB1	-2,31	6,82	34,04	8,04E-05	0,000383866500301
HDAC1	-2,64	6,15	52,58	1,02E-05	8,48E-05
RBBP7	-2,76	8,04	15,68	0,001897051030174	0,004561549991646
Trim28/KAP1	-2,96	5,31	6,31	0,027352659497943	0,041687202557848
AGO2	-2,99	2,89	53,81	9,05E-06	7,83E-05
HP1b	-3,42	5,62	79,43	1,23E-06	2,05E-05
CHD4	-3,44	6,57	31,65	0,000111545955942	0,000489630371855
p66alpha	-4,11	6,81	72,53	1,97E-06	2,79E-05
PRMT5	-4,18	6,97	124,13	1,10E-07	5,20E-06
HP1a	-4,35	6,40	58,76	5,82E-06	5,75E-05
PIWIL4	-4,46	5,01	12,05	0,004619446379909	0,009435290942545
DNMT1	-5,24	7,33	140,95	5,47E-08	3,68E-06
HP1g	-7,00	7,55	227,64	3,65E-09	1,18E-06
PIWIL2	-7,05	4,49	182,60	1,28E-08	1,87E-06
MAEL	-8,83	6,44	154,18	3,32E-08	2,94E-06
PIWIL1	-9,41	6,57	121,13	1,26E-07	5,58E-06
PLD6	-11,52	4,58	91,19	5,90E-07	1,31E-05
RBBP4	-12,26	6,33	50,39	1,25E-05	9,86E-05



Supplementary Figure 1:

MDS plot of the dataset



Supplementary Figure 2:

Expression levels of TEs in *B. pachypus* without the female individual BP402

Chapter 2: TE abundance and annotation in the large genome of *Bombina pachypus*

Supplementary Table S1:

Genomic localisation of TEs in *B. pachypus* genome

	TEs in INTRAGENIC REGIONS		TEs in INTERGENIC REGIONS			
TE order/family	**N of TEs**	**TE Percentage**	**N of TEs**	**TE Percentage**	**R1-TE**	**TE Enrichment Ratio**
LTR/BEL	14656	0,07	27576	0,13	0,53	0,17
LTR/Copia	165890	0,80	246738	1,19	0,67	-0,04
LTR/ERV	312955	1,51	448249	2,17	0,70	-0,08
LTR/Gypsy	837957	4,05	1423324	6,88	0,59	0,09
LTR/nc	209890	1,01	347331	1,68	0,60	0,06
DIRS	222680	1,08	205127	0,99	1,09	-0,69
LINE/I	18109	0,09	32865	0,16	0,55	0,14
LINE/Jockey	13897	0,07	24771	0,12	0,56	0,13
LINE/L1	445801	2,16	664533	3,21	0,67	-0,04
LINE/R2	41905	0,20	75094	0,36	0,56	0,13
LINE/RTE	42949	0,21	69488	0,34	0,62	0,04
LINE/nc	69806	0,34	115733	0,56	0,60	0,06
SINE	3	0,00	13	0,00	0,23	0,64
PLE	229819	1,11	364521	1,76	0,63	0,02
nLTR/nc	62422	0,30	108387	0,52	0,58	0,11
ClassI/nc	182222	0,88	283484	1,37	0,64	0,00
DNA/CACTA	164956	0,80	241136	1,17	0,68	-0,06

DNA/Harbinger	136656	0,66	245877	1,19	0,56	0,14
DNA/hAT	2225795	10,76	3643686	17,62	0,61	0,05
DNA/Mutator	192156	0,93	359650	1,74	0,53	0,17
DNA/PiggyBac	28124	0,14	41935	0,20	0,67	-0,04
DNA/P	243	0,00	386	0,00	0,63	0,02
DNA/TcMar	570118	2,76	964915	4,67	0,59	0,08
DNA/MITE	222251	1,07	395091	1,91	0,56	0,13
DNA/nMITE	825277	3,99	1341242	6,49	0,62	0,04
DNA/Helitron	48215	0,23	74621	0,36	0,65	0,00
ClassII/nc	8081	0,04	13474	0,07	0,60	0,07
Unknown	631571	3,05	1039854	5,03	0,61	0,06
Total RetroTE	2870961	13,88	4437234	21,46	0,65	0,00
Total DNATE	4421872	21,38	7322013	35,41	0,60	0,06
Total TE	7924404	38,32	12799101	61,89	0,62	0,04

Supplementary Table S2:

Abundance percentages of the different TE families in relation to the total amount of TEs in the genome of *B. pachypus*

TE order/family	TE Length (bp)	TE Percentage
LTR/BEL	11032269	0,15
LTR/Copia	95082345	1,31
LTR/ERV	320841455	4,43
LTR/Gypsy	941938099	13,02
LTR/nc	166612762	2,30

DIRS	251804648	3,48
LINE/I	20584054	0,28
LINE/Jockey	13999233	0,19
LINE/L1	458382512	6,33
LINE/R2	41040306	0,57
LINE/RTE	45524912	0,63
LINE/nc	75874428	1,05
SINE	1641285	0,02
PLE	251716771	3,48
nLTR/nc	71017466	0,98
ClassI/nc	207875225	2,87
DNA/CACTA	116822737	1,61
DNA/Harbinger	112686771	1,56
DNA/hAT	1957493356	27,05
DNA/Mutator	195086565	2,70
DNA/PiggyBac	19472228	0,27
DNA/P	197389	0,00
DNA/TcMar	467595527	6,46
DNA/MITE	144452390	2,00
DNA/nMITE	709670250	9,81
DNA/Helitron	36666921	0,51
ClassII/nc	4225290	0,06
Unknown	497131273	6,87
Total RetroTE	2974967770	41,11
Total DNATE	3764369424	52,02
Total TE	7236468467	100

Supplementary Table S3:
Expression percentages of the different TE families in relation to the total amount of TEs expressed in the transcriptome of *B. pachypus*

TE order/family	Average Expression (counts)	TE Percentage
LTR/BEL	3401	0,31
LTR/Copia	25769	2,32
LTR/ERV	30540	2,75
LTR/Gypsy	384882	34,68
LTR/nc	77848	7,02
DIRS	21294	1,92
LINE/I	2976	0,27
LINE/Jockey	3914	0,35
LINE/L1	50737	4,57
LINE/R2	1077	0,10
LINE/RTE	10013	0,90
LINE/nc	1169	0,17
SINE	3160	0,28
PLE	20553	1,85
ClassI/nc	32750	2,95
DNA/CACTA	18660	1,68
DNA/Harbinger	26113	2,35
DNA/hAT	103342	9,31
DNA/Mutator	9790	0,88
DNA/PiggyBac	1064	0,10
DNA/TcMar	151609	13,66
DNA/MITE	38692	3,49

DNA/nMITE	37114	3,34
DNA/Helitron	300	0,03
Unknown	52913	4,77
Total DNATE	386683	34,85
Total RetroTE	670084	60,39
Total TE	1109680	100

Supplementary Table S4:

***B. pachypus* samples information**

Sample ID	Population origin (Large, Small)	Location	Sex
n_68	L	Masseti Pollino - Sud	Putative M
n_69	L	Masseti Pollino - Sud	Putative M
n_73	L	Masseti Pollino - Sud	Putative M
n_78	L	F. Argentino Pollino - Sud	Putative M
n_82	L	F. Argentino Pollino - Sud	Putative M
n_87	L	Aspromonte - Sud	Putative M
n_91	L	Aspromonte - Sud	Putative M
n_92	L	Aspromonte - Sud	Putative M
n_95	L	Aspromonte - Sud	Putative M
n_100	L	Aspromonte - Sud	Putative M
n_201	S	Bagno di Romagna (Nord)	Putative M
n_203	S	Bagno di Romagna (Nord)	Putative M

n_205	S	Bagno di Romagna (Nord)	Putative M
n_206	S	Bagno di Romagna (Nord)	Putative M
n_207	S	Bagno di Romagna (Nord)	Putative M
n_208	S	Bagno di Romagna (Nord)	Putative M
n_210	S	Bagno di Romagna (Nord)	Putative M
n_211	S	Bagno di Romagna (Nord)	Putative M
n_212	S	Bagno di Romagna (Nord)	Putative M
n_213	S	Bagno di Romagna (Nord)	Putative M

Supplementary Table S5:

Abundance percentages of the different TE families among the two different populations of *B. pachypus*

	SOUTH-POP		NORTH-POP	
TE order/family	**Average TE Length (bp)**	**TE Percentage**	**Average TE Length (bp)**	**TE Percentage**
LTR/BEL	2302162	0,15	2337045	0,16
LTR/Copia	17945408	1,20	16820963	1,12
LTR/ERV	44889434	2,99	44968130	3,00
LTR/Gypsy	125118436	8,34	127747910	8,52
LTR/nc	29223561	1,95	29161380	1,94
DIRS	68530657	4,57	70397909	4,69
LINE/I	3852622	0,26	3868735	0,26
LINE/Jockey	2979403	0,20	2931025	0,20

LINE/L1	71894076	4,79	73107507	4,87
LINE/R2	6605859	0,44	6353182	0,42
LINE/RTE	16368687	1,09	15700189	1,05
LINE/nc	11306868	0,75	11327775	0,76
SINE	348596	0,02	289332	0,02
PLE	25225690	1,68	24997510	1,67
nLTR/nc	6868399	0,46	6753040	0,45
ClassI/nc	43564550	2,90	43418451	2,89
DNA/CACTA	22160825	1,48	22897056	1,53
DNA/Harbinger	18434660	1,23	18371717	1,22
DNA/hAT	227195289	15,15	230129567	15,34
DNA/Mutator	24793700	1,65	24667207	1,64
DNA/PiggyBac	2109315	0,14	2062196	0,14
DNA/P	51304	0,00	50678	0,00
DNA/TcMar	91485078	6,10	91612240	6,11
DNA/MITE	26914546	1,79	28006615	1,87
DNA/nMITE	83457204	5,56	83973800	5,60
DNA/Helitron	4213010	0,28	4096981	0,27
ClassII/nc	815860	0,05	789674	0,05
Unknown	89196454	5,95	89248296	5,95
Total RetroTE	477024408	31,80	480180083	32,01
Total DNATE	501630790	33,44	506657730	33,78
Total TE	1067851652	71,19	1076086108	71,74

Supplementary Table S6:
Samples information

Species	Order	Genome Size	Assembly ID	Sequencing technology	Assembly level
Platyplectrum ornatum	Anura	1,1 Gb	GCA_016617825.1	Illumina HiSeq, Oxford Nanopore MinION	Scaffold-level 148.035 scaffolds
Xenopus tropicalis	Anura	1,5 Gb	GCA_000004195.4	PacBio Sequel, Illumina HiSeq	Chromosome-level 10 chromosomes, 166 scaffolds
Dendropsophus ebraccatus	Anura	2,4 Gb	GCA_027789725.1	PacBio Sequel I CLR; 10X Gemonics linked reads; Bionano Genomics DLS; Arima Genomics Hi-C v1; Ilumina WGS	Scaffold-level 2.569 scaffolds
Xenopus laevis	Anura	2,7 Gb	GCA_017654675.1	PacBio RSII	Chromosome-level 18 chromosomes, 54 scaffolds
Leptobrachium leishanense	Anura	3,5 Gb	GCA_009667805.1	PacBio RSII	Chromosome-level 13 chromosomes, 5.302 scaffolds
Discoglossus pictus	Anura	3,9 Gb	GCA_027410445.1	PacBio Sequel I CLR; 10X Gemonics linked reads; Bionano Genomics DLS; Arima Genomics Hi-C v1	Chromosome-level 14 chromosomes, 1.317 scaffolds
Gastrophryne carolinensis	Anura	4,3 Gb	GCA_027917425.1	PacBio Sequel II HiFi; Arima Hi-C v2	Chromosome-level 11 chromosomes, 1.002 scaffolds
Bufo bufo	Anura	5 Gb	GCF_905171765.1	PacBio data, 10X Genomics Chromium data, BioNano data, Arima Hi-C data	Chromosome-level 11 chromosomes, 1.306 scaffolds

Bombina pachypus	Anura	9,6 Gb		PacBio Sequel I CLR, Hi-C data	Chromosome-level 12 chromosomes, 16.500 scaffolds
Bombina bombina	Anura	10 Gb	*GCF_027579735.1	PacBio Sequel I CLR	Chromosome-level 12 chromosomes, 2.962 scaffolds
Rhinatrema bivittatum	Gymno phiona	5,3 Gb	GCF_901001135.1	PacBio data, 10X Genomics Chromium data, BioNano data, Hi-C data	Chromosome-level 19 chromosomes, 1.329 scaffolds

Supplementary Table S7:

Abundance percentages of the different TE families across the amphibian species

	1,1Gb	**1,5Gb**	**2,4Gb**	**2,7Gb**	**3,5Gb**	**3,9Gb**	**4,3Gb**	**5Gb**	**9,7Gb**	**10Gb**	**5,3Gb**
TE order/family	***P. ornatum***	***X. tropicalis***	***D. ebraccatus***	***X. laevis***	***L. leishanense***	***D. pictus***	***G. carolinensis***	***B. bufo***	***B. pachypus***	***B. bombina***	***R. bivittatum***
LTR/BEL	0,03	0,10	0,24	0,16	0,09	0,05	0,04	0,29	0,11	0,17	0,03
LTR/Copia	0,33	0,08	0,45	0,14	0,55	0,45	0,78	0,46	0,98	0,57	0,57
LTR/ERV	0,67	3,00	3,21	2,42	2,43	2,65	5,78	2,77	3,31	2,74	7,12
LTR/Gypsy	2,00	3,96	5,37	6,00	7,92	4,18	4,96	9,53	9,72	11,55	13,17
LTR/nc	0,75	1,73	2,00	0,69	1,66	2,05	1,32	2,29	1,72	2,48	4,24
DIRS	0,19	0,26	0,27	0,52	0,16	0,15	1,52	0,97	2,60	3,37	1,65
LINE/I	0,05	0,05	0,04	0,42	0,34	0,04	0,02	0,15	0,21	0,22	0,26
LINE/Jockey	0,05	0,12	0,12	0,08	0,55	0,17	0,62	0,14	0,14	0,43	0,59
LINE/L1	0,64	3,67	1,93	2,43	2,05	4,25	5,76	4,02	4,73	6,15	6,11
LINE/R2	0,24	0,61	0,72	0,71	0,74	0,27	1,07	1,63	0,42	0,42	0,85
LINE/RTE	0,26	0,10	0,75	0,51	1,01	0,84	1,35	0,81	0,47	3,83	0,89
LINE/nc	0,18	0,36	0,87	0,70	0,84	0,70	0,67	0,91	0,78	0,66	0,90
SINE	0,06	0,14	0,39	0,03	0,20	0,12	0,20	0,06	0,02	0,03	0,14
PLE	0,16	0,30	0,69	0,59	0,49	0,87	0,14	0,66	2,60	1,69	0,38

nLTR/nc	0,12	0,09	0,38	0,15	0,40	0,25	0,50	0,41	0,73	0,37	1,07
ClassI/nc	0,27	1,00	1,03	0,83	1,56	1,37	2,53	1,68	2,15	3,14	3,69
DNA/CACTA	1,11	1,73	1,71	1,06	1,20	4,13	1,04	1,41	1,21	1,15	0,91
DNA/Harbinger	0,32	1,23	1,49	1,04	2,10	0,91	4,78	3,78	1,16	0,80	1,07
DNA/hAT	4,39	9,48	14,89	12,49	15,67	12,80	13,38	16,96	20,20	16,13	10,28
DNA/Mutator	0,66	1,13	1,14	0,73	0,90	0,75	1,16	2,52	2,01	1,95	0,57
DNA/PiggyBac	0,07	0,56	0,75	0,71	0,12	0,11	0,10	0,27	0,20	0,15	0,02
DNA/P	np	np	0,02	0,01	np	0,03	0,00	0,00	0,00	0,01	0,00
DNA/TcMar	1,57	2,93	7,65	2,91	8,48	4,66	5,33	13,64	4,83	6,21	3,34
DNA/MITE	0,49	1,18	1,10	1,18	0,66	0,91	1,25	1,67	1,49	0,87	0,70
DNA/nMITE	2,54	3,69	5,72	5,51	4,67	5,39	4,89	6,69	7,32	6,47	2,90
DNA/Helitron	0,76	0,16	0,59	0,14	0,31	0,65	0,18	0,29	0,38	0,33	0,09
ClassII/nc	0,00	0,06	0,04	0,01	0,02	0,00	0,02	0,03	0,04	0,12	0,02
Unknown	2,07	3,57	3,47	3,90	4,14	4,75	3,08	3,15	5,13	4,95	3,68
Total RetroTE	6,00	15,56	18,45	16,39	20,99	18,41	27,27	26,78	30,70	37,82	41,66
Total DNATE	11,91	22,16	35,12	25,78	34,14	30,35	32,12	47,26	38,85	34,19	19,91
Total TE	19,98	41,29	57,04	46,06	59,27	53,51	62,47	77,19	74,69	76,96	65,25

Chapter 3:

A high-quality reference genome for the critically endangered Aeolian wall lizard, Podarcis raffonei

Supplementary Table 1: RNA-seq data from NCBI used for genome annotation.

SRA ID	Number of bases (bp)	Species	Tissue	Sex	Age	Reference
SRR3201591	732,444,173	Podarcis cretensis	myoskeletal tissue	-	adult	Heidelberg Institute for Theoretical Studies
SRR3201796	2,785,453,744	Podarcis cretensis	myoskeletal tissue	-	adult	Heidelberg Institute for Theoretical Studies
SRR3479613	3,332,793,152	Podarcis siculus	brain	male	adult	Trapanese et al. 2017
SRR3479614	4,386,464,340	Podarcis siculus	testis	male	adult	Trapanese et al. 2017
SRR3479616	3,853,684,290	Podarcis siculus	brain	male	adult	Trapanese et al. 2017
SRR3479618	4,258,304,834	Podarcis siculus	testis	male	adult	Trapanese et al. 2017
SRR3479621	5,419,530,720	Podarcis siculus	brain	male	adult	Trapanese et al. 2017
SRR3479624	4,143,993,640	Podarcis siculus	testis	male	adult	Trapanese et al. 2017
SRR5859153	8,293,892,000	Podarcis muralis	embryonic tissue	-	embryo	Feiner et al. 2018
SRR5859154	8,913,923,200	Podarcis muralis	embryonic tissue	-	embryo	Feiner et al. 2018
SRR5859155	8,472,098,200	Podarcis muralis	embryonic tissue	-	embryo	Feiner et al. 2018
SRR5859156	8,154,765,000	Podarcis muralis	embryonic tissue	-	embryo	Feiner et al. 2018

SRR7152529	3,217,297,430	Podarcis siculus	brain	-	-	University of Naples Federico II
SRR7152530	2,591,554,556	Podarcis siculus	brain	-	-	University of Naples Federico II
SRR8468518	90,033,600	Podarcis muralis	skin	male	adult	Andrade et al. 2019
SRR8468521	15,275,708,727	Podarcis muralis	muscle	male	adult	Andrade et al. 2019
SRR8468522	15,066,658,193	Podarcis muralis	skin	male	adult	Andrade et al. 2019
SRR8468523	23,699,932,512	Podarcis muralis	testis	male	adult	Andrade et al. 2019
SRR8468525	25,523,402,259	Podarcis muralis	brain	male	adult	Andrade et al. 2019
SRR8468526	14,166,746,290	Podarcis muralis	duodenum	male	adult	Andrade et al. 2019
SRR8468527	6,582,380,402	Podarcis muralis	whole	-	embryo	Andrade et al. 2019
SRR8468528	5,999,697,073	Podarcis muralis	whole	-	embryo	Andrade et al. 2019
SRR9090247	5,749,012,355	Podarcis liolepis	mix of various organs	-	adult	Braunschweig University of Technology
SRR9090248	3,520,632,005	Podarcis muralis	mix of various organs	-	adult	Braunschweig University of Technology

Supplementary Table 2: Protein data from NCBI used for genome annotation.

GenBank code	Species	Family	Reference
GCA_900067755.1	*Pogona vitticeps*	Agamidae	Georges et al. 2015

GCA_001185365.2	*Pantherophis guttatus*	Colubridae	Ullate-Agote et al. 2020
GCA_009769535.1	*Thamnophis elegans*	Colubridae	Vertebrate Genomes Project
GCA_001077635.2	*Thamnophis sirtalis*	Colubridae	Wilson RK, the McDonnell Genome Institute, Washington University School of Medicine
GCA_000090745.2	*Anolis carolinensis*	Dactyloidae	Alföldi et al. 2011
GCA_009733165.1	*Naja naja*	Elapidae	Suryamohan et al. 2020
GCA_900518725.1	*Notechis scutatus*	Elapidae	BABS Genome project
GCA_000516915.1	*Ophiophagus hannah*	Elapidae	Vonk et al. 2013
GCA_900518735.1	*Pseudonaja textilis*	Elapidae	BABS Genome project
GCA_009819535.1	*Lacerta agilis*	Lacertidae	Vertebrate Genomes Project
GCA_004329235.1	*Podarcis muralis*	Lacertidae	Andrade et al. 2019
GCA_011800845.1	*Zootoca vivipara*	Lacertidae	Yurchenko, Recknagel, et Elmer 2020
GCA_020142125.1	*Phrynosoma platyrhinos*	Phrynosomatidae	Koochekian et al. 2022
GCA_019175285.1	*Sceloporus undulatus*	Phrynosomatidae	Westfall et al. 2021
GCA_000186305.2	*Python bivittatus*	Pythonidae	Castoe et al. 2013
GCA_004798865.1	*Varanus komodoensis*	Varanidae	Lind et al. 2019
GCA_018340635.1	*Bothrops jararaca*	Viperidae	Almeida et al. 2021
GCA_018446365.1	*Crotalus adamanteus*	Viperidae	Hogan et al. 2021

GCA_016545835.1	*Crotalus tigris*	Viperidae	Margres et al. 2021
GCA_001527695.3	*Protobothrops mucrosquamatus*	Viperidae	Aird et al. 2017
GCA_001447785.1	*Gekko japonicus*	Gekkonidae	Liu et al. 2015
GCA_021028975.2	*Sphaerodactylus townsendi*	Sphaerodactylidae	Pinto et al. 2022

Supplementary Table 3: Long_PCR and sequencing primers for mitochondrial DNA sequencing

Primer name	Reaction type	Primer sequence
tSer_12682	Amplicon 1	GCTGCTAACTCTAATAACTAAGAAT
Pod_seq_16695	Amplicon 1 + Sanger	TTTTAGGGTTGCGTTCGTGG
Pod_seq_3332	Amplicon 2 + Sanger	TGGTGCTCGGTTTGTTTCTG
Pod_seq_15306	Amplicon 2 + Sanger	TGATAACCCCGTCCTAGTAGC
srRNA_691	Amplicon 3 + Sanger	TCAGCCTATATACCGCCGTC
ATP8_7955	Amplicon 3	AGGGCCATGGTCAGGTTCA
ND6_13608	Amplicon 4	GTCTTCGTGCAGTTAGGTTC
Pod_seq_6260	Amplicon 4 +Sanger	GAGCTTACTTCACCTCAGCT
Pod_seq_11002	Sanger	TGCCTACGACAAACAGACCTA
Pod_seq_1278	Sanger	CCCTGTACCTCCTGCATCAT

Pod_seq_12819	Sanger	AGGTTATGGATGATTGCGCC
Pod_seq_13357	Sanger	CCCAACACTTCATCGCATCA
Pod_seq_14115	Sanger	CACCAAAACCTGCGACTTGA
Pod_seq_15120	Sanger	ATACCGCCCACTATCTCAGC
Pod_seq_17248	Sanger	CAGGACTGAACAACAAAGCCT
Pod_seq_1987	Sanger	CCTGCCCAGTGACTCTTTA
Pod_seq_3277	Sanger	TGGCTAAGGGTCATGTTGGT
Pod_seq_3873	Sanger	AAAGCTCTTGGGCCCATACC
Pod_seq_3998	Sanger	GATAACTGGCGGCCGTAATG
Pod_seq_40	Sanger	CATCTTCAGTGCCGTGCTTT
Pod_seq_4450	Sanger	TCAATCGGACACCTAGGCTG
Pod_seq_4621	Sanger	TGTTGGGGAGGCTGTTCATA
Pod_seq_5116	Sanger	GCCTCGATCCTGCAAAACTT
Pod_seq_5483	Sanger	AACCCGGAACCCTTCTTGG
Pod_seq_7787	Sanger	GCCTCAACTTAATCCTGCCC
Pod_seq_782	Sanger	GCTACACCTTGACCTGACGT

Supplementary Table 4: Length of the scaffolds of the nuclear and mitochondrial genome assemblies.

Scaffold name	Length (bp)
SUPER_1	139,138,986
SUPER_2	127,263,675
SUPER_3	124,660,641
SUPER_4	108,497,525
SUPER_5	102,317,954
SUPER_6	100,096,178
SUPER_7	92,543,381
SUPER_8	93,623,253
SUPER_9	80,976,960
SUPER_10	79,254,795
SUPER_11	66,434,231
SUPER_12	61,439,401
SUPER_13	56,248,303
SUPER_14	53,986,714
SUPER_15	45,100,313
SUPER_16	42,970,794

SUPER_17	42,074,563
SUPER_18	13,334,018
SUPER_Z	51,487,686
SUPER_W	31,610,000
scaffold_32	37,849
scaffold_33_ctg1	10,54
scaffold_34_ctg1	8,553
scaffold_35_ctg1	7,317
scaffold_36_ctg1	6,473
scaffold_37_ctg1	1,042
scaffold_38	358
mtDNA	17,038
TOTAL nuclear genome length	1,513,131,503

Supplementary Table 5: Repeat content of the assemblies of Podarcis raffonei and Podarcis muralis.

	Podarcis raffonei			*Podarcis muralis*		
Type of element	**Number of elements**	**Length (bp)**	**Percentage of assembly**	**Number of elements**	**Length (bp)**	**Percentage of assembly**
SINE	12,150	1,486,878	0.1%	14,867	2,047,452	0.14%
LINE	1,127,892	139,867,006	9.24%	1,131,323	144,604,610	9.57%
PLE	18,974	3,939,366	0.26%	5,846	621,588	0.04%
DIRS	57,643	4,717,364	0.31%	76,239	7,509,802	0.5%
LTR elements	1,917,627	253,438,336	16.75%	1,680,800	203,331,071	13.47%
DNA transposons	2,392,926	227,914,915	15.07%	2,162,929	224,449,166	14.85%
Low complexity	23,792	1,123,846	0.07%	25,704	1,202,995	0.08%
Simple repeat	339,297	14,110,419	0.93%	370,635	15,304,914	1.01%
Unclassified	786,149	83,156,561	5.5%	834,078	81,842,431	5.41%
Total	**6676450**	**729,754,691**	**48.23%**	**6,302,421**	**680,914,029**	**45.07%**

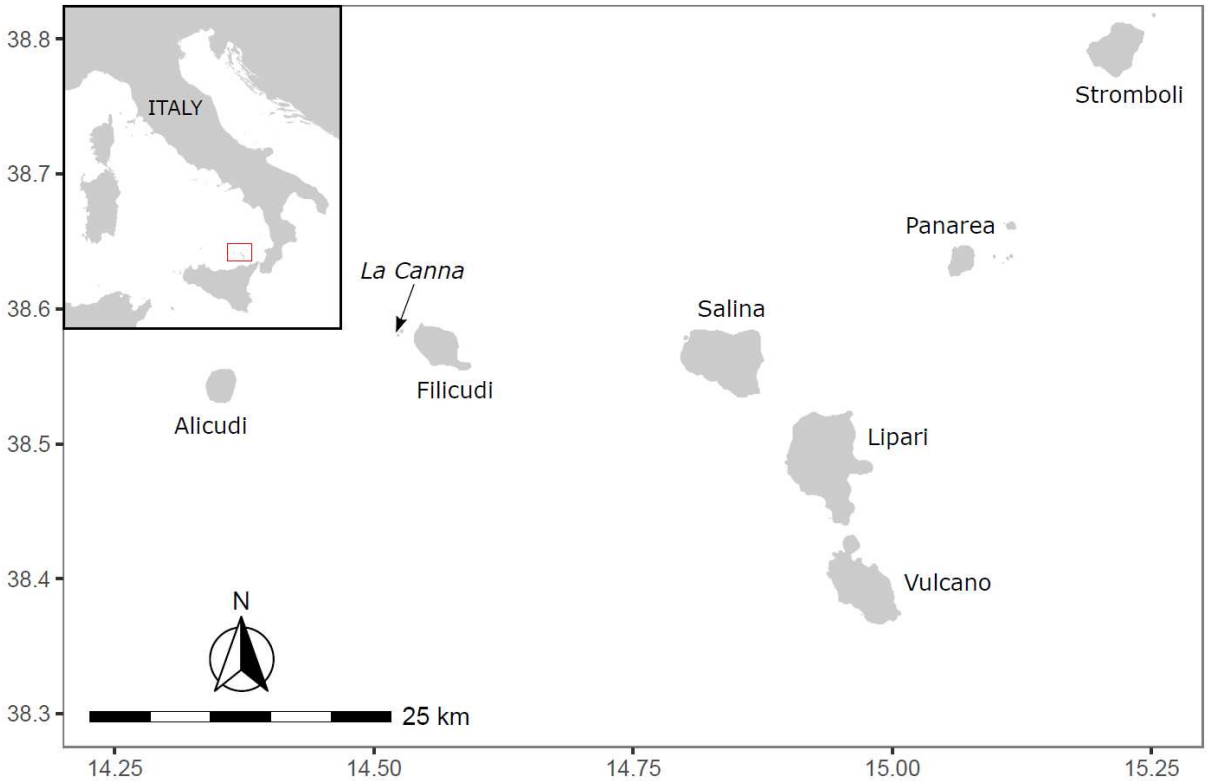
Supplementary Table 6: High-quality genome assemblies currently available for squamate reptiles on NCBI (last access: 01/02/23). High-quality genomes were defined as genomes with a scaffold N50 higher than 10M (standard VGP; Rhie et al. 2021), and a scaffold number less than 5,000. When available, we reported the BUSCO completeness score (Single-copy + Duplicated), repeat content and heterozygosity estimated from Genomescope. Divergence from P. raffonei was estimated with TimeTree.

Species	Divergence (My)	Family	NCBI Accession	Total sequence length (bp)	Number of scaffolds	Scaffold N50	Scaffold L50	BUSCO score (S+D)	BUSCO database
Anolis sagrei	167	Dactyloidae	GCA_025583915.1	1,926,425,113	3,738	253,587,442	4	96.9%	vertebrata
Arizona elegans	167	Colubridae	GCA_022577455.1	1,842,551,953	140	105,945,816	5	95.9%	tetrapoda
Aspidoscelis marmoratus	154	Teiidae	GCA_014337955.1	1,639,530,780	3,826	32,220,929	15	-	-
Aspidoscelis tigris	154	Teiidae	GCA_023333525.1	1,335,668,279	74	9,369,0953	5	-	-
Bungarus multicinctus	167	Elapidae	GCA_023653725.1	1,593,755,901	448	135,406,522	5	94.6%	vertebrata
Charina bottae	167	Boidae	GCA_023362775.1	1,804,939,834	289	97,015,800	5	96.3%	tetrapoda
Crotalus oreganus	167	Viperidae	GCA_024509115.1	1,564,795,203	698	110,762,666	4	-	-
Diadophis punctatus	167	Colubridae	GCA_023053685.1	1,783,023,707	444	83,654,930	5	-	-
Elgaria multicarinata	167	Anguidae	GCA_023053635.1	1,790,509,355	85	107,858,850	7	-	-
Hemicordylus capensis	174	Cordylidae	GCF_027244095.1	2,294,751,221	44	359,646,233	3	95.5%	sauropsidae

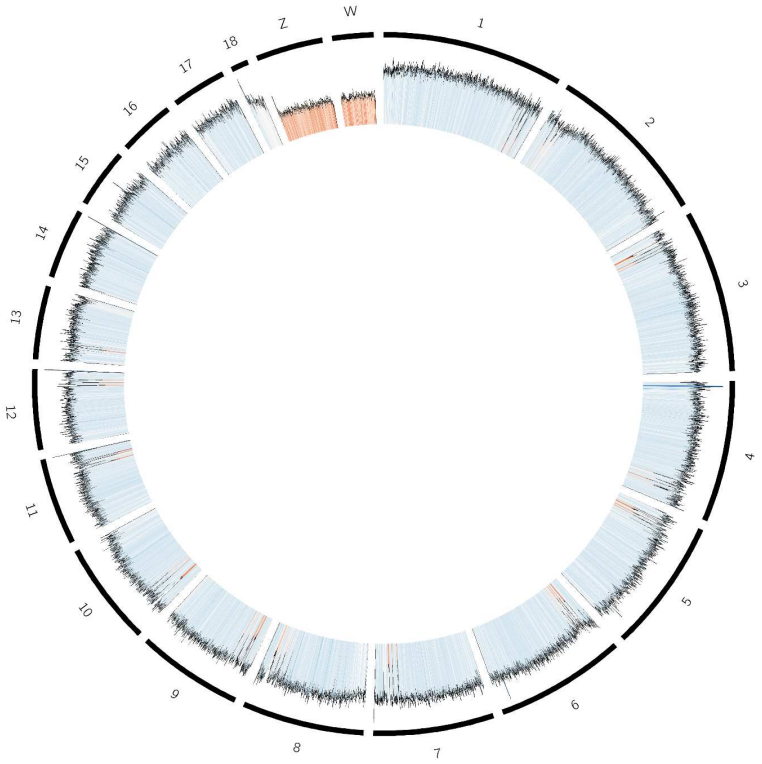
Hydrophis curtus	167	Hydrophiidae	GCA_019472885.1	1,964,827,820	711	266,229,956	3	89.8%	tetrapoda
Hydrophis cyanocinctus	167	Hydrophiidae	GCA_019473425.1	1,980,712,740	1,163	264,245,889	3	90.1%	tetrapoda
Lacerta agilis	NA	Lacertidae	GCF_009819535.1	1,391,404,169	29	86,565,987	7	-	-
Naja naja	167	Elapidae	GCA_009733165.1	1,768,535,092	1,897	224,088,900	3	94.3%	tetrapoda
Paroedura picta	191	Gekkonidae	GCA_003118565.2	1,562,175,643	4,871	109,004,681	6	89.8%	metazoan
Phrynocephalus versicolor	167	Agamidae	GCA_023846285.1	1,603,387,288	4,557	49,226,030	12	90.0%	tetrapoda
Phrynosoma blainvillii	167	Phrynosomatidae	GCA_026167975.1	1,968,358,621	52	352,551,559	3	-	-
Plestiodon gilberti	174	Scincidae	GCA_026170595.1	1,571,222,493	39	231,322,181	3	-	-
Podarcis muralis	12,2	Lacertidae	GCF_004329235.1	1,511,020,169	2,161	92,398,148	7	96.4%, 93.2%	vertebrata, tetrapoda
Podarcis raffonei	0	Lacertidae	GCA_027172205.1	1,513,131,503	28	93,600,000	7	98.3%, 97.3%	vertebrata, tetrapoda
Pseudonaja textilis	167	Elapidae	GCF_900518735.1	1,590,035,073	2,855	14,685,528	31	-	-
Salvator merianae	154	Teiidae	GCA_003586115.2	2,068,170,046	4,512	55,382,274	12	97.4%, 94.4%	vertebrata tetrapoda
Sceloporus occidentalis	167	Phrynosomatidae	GCA_023333645.1	2,856,356,971	608	98,418,489	7	-	-
Shinisaurus	167	Shinisauridae	GCA_021292165.1	2,189,995,079	1,553	296,945,371	4	94.5%	2,586 genes

crocodilurus									
Sphaerodactylus townsendi	191	Sphaerodactylidae	GCF_021028975.2	1,810,846,735	1,742	133,801,376	6	88.3%	tetrapoda
Thamnophis elegans	167	Colubridae	GCF_009769535.1	1,672,190,305	365	100,851,885	6	-	-
Varanus komodoensis	167	Varanidae	GCF_004798865.1	1,507,945,839	1,411	23,831,982	17	96.1%	vertebrata
Varanus salvator	167	Varanidae	GCA_023646645.1	1,702,541,867	858	71,461,993	9	87.5%	vertebrata
Vipera latastei	167	Viperidae	GCA_024294585.1	1,631,568,913	56	222,489,854	3	-	-
Vipera ursinii	167	Viperidae	GCA_947247035.1	1,625,023,540	384	212,821,320	3	-	-

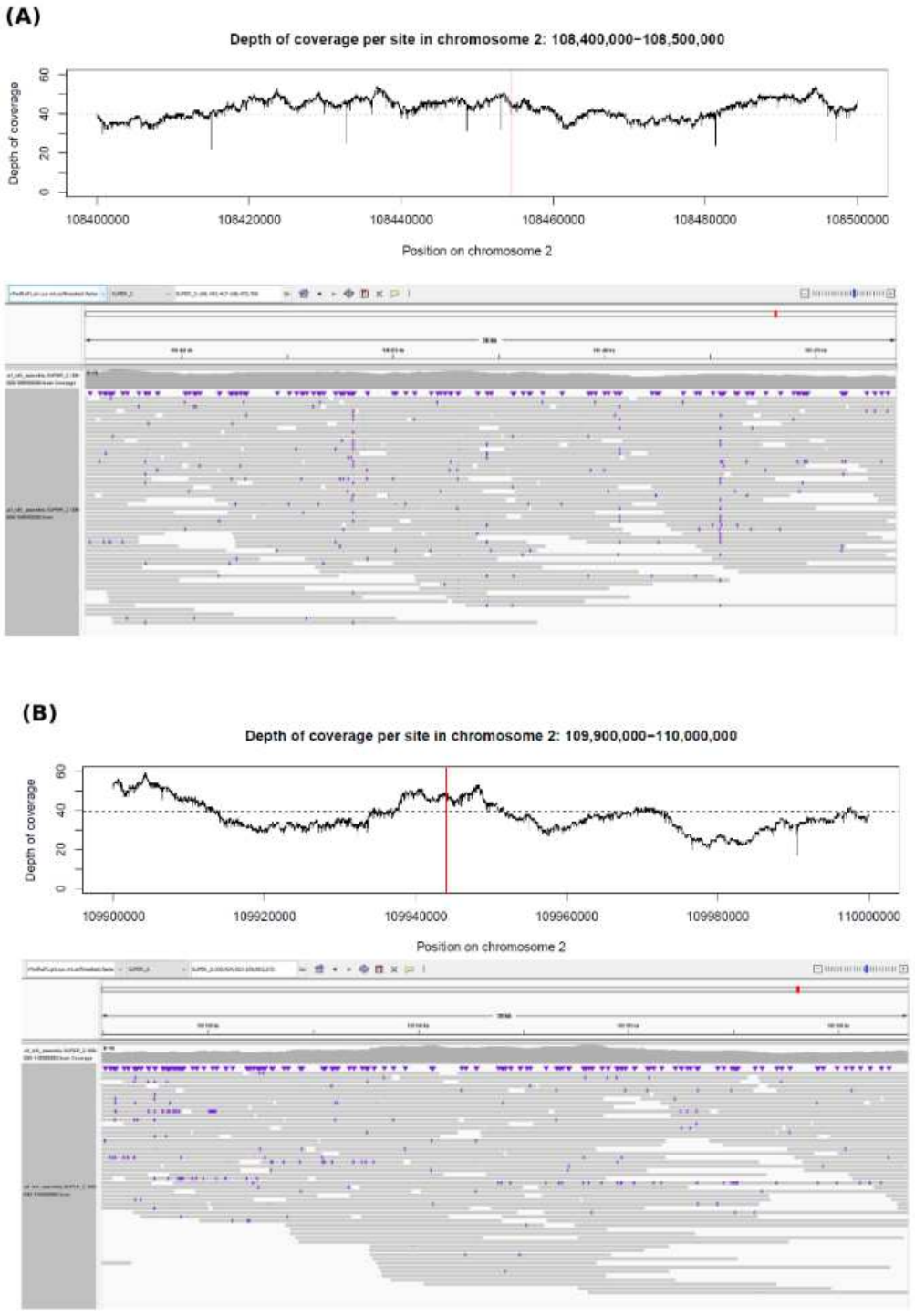
Supplementary Figure 1: Map of the Aeolian islands showing the sampling locality of La Canna.



Supplementary Figure 2: Average depth of coverage in windows of 100 Kb of the HiFi reads mapped against the genome assembly. Only the chromosomal-scale scaffolds are represented. The deep blue colours indicate a high coverage compared to the mean depth of coverage (40X) whereas the red colours indicate a low coverage compared to the mean depth of coverage.



Supplementary Figure 3: Per site depth of coverage (upper plots) and visual representation of the reads with IGV (lower plots) in a 100-Kb window including the beginning (A) and the end (B) of the junction of the chromosome 2 of P. raffonei that mapped to the chromosome 18 of P. muralis.



Chromosome-level reference genome of the Ponza grayling (*Hipparchia sbordonii)*, an Italian endemic and endangered butterfly

Table S1: summary of the manually curated *H. sbordonii* genome assembly, compared with the *H. semele* chromosome-scale assembly.

chromosome number	**number of superscaffolds obtained in *H. sbordonii***	**scaffold size (Mb)**
1	1	17.7
2	1	17.4
3	1	16.6
4	1	16.5
5	1	16.4
6	1	16.5
7	1	15.8
8	1	15.4
9	1	15.3
10	1	14.9
11	1	14.6
12	1	14.6
13	1	14.5
14	1	14.2
15	2	10.0 + 3.9
16	2	11.0 + 2.8
17	1	13.2
18	1	13.1

19	1	14.0
20	1	12.9
21	1	11.7
22	1	11.8
23	2	6.9 + 2.0
24	1	9.3
25	1	8.2
26	2	4.6 + 3.0
27	3	1.8 + 2.4 + 2.5
28	1	6.8
Z	2	4.4 + 11.7
W	missing	/

Table S2: summary of the main features (i.e. location, size and number of 1:1 orthologous genes involved) of the 10 intra-chromosomal inversions larger than 10Kb identified in the comparison between *H. sbordonii* and *H. semele*.

H. semele chromosome	genomic coordinates (Mb)	inversion size (Kb)
6	1,6	104
6	9,8	19
13	0-0,1	79
13	4,8-5	169
13	11,8	19
14	3,4	16
15	5,4	36
18	0-0,6	604
25	5,6-8,3	2741

28	3,8	23

Table S3: Transposable element (TE) content in *Hipparchia sbordonii* genome.

Class	Order	Order or Superfamily	TE_Length (bp)	TE_Percentage Of Assembly (%)
ClassI: Retrotransposons	LTR	LTR (n.s)	949688	0,24
		BEL	1176735	0,3
		Copia	700000	0,18
		ERV	21372	0,01
		Gypsy	15229996	3,92
		ClassI (n.s)	2271495	0,58
		nLTR (n.s)	163025	0,04
	LINE	LINE	8743678	2,25
	SINE	SINE	2557339	0,66
	PLE	PLE	2483403	0,64
ClassII: DNA Transposons	TIR	ClassII (n.s)	6206	0,00
		DNA_CACTA	9346898	2,41
		DNA_Harbinger	860098	0,22
		DNA_hAT	30531378	7,86
		DNA_Mutator	8204691	2,11
		DNA_PiggyBac	51189	0,01
		DNA_P	298899	0,08
		DNA_TcMar	8469121	2,18
	HELITRON	DNA_Helitron	2874114	0,74
		DNA_MITE	2469281	0,64
		DNA_nMITE	26405884	6,79
		Unknown TEs	33418634	8,6

TOTAL_ClassI	34296731	8,83
TOTAL_ClassII	89517759	23,04
TOTAL_TEs	157233124	40,46

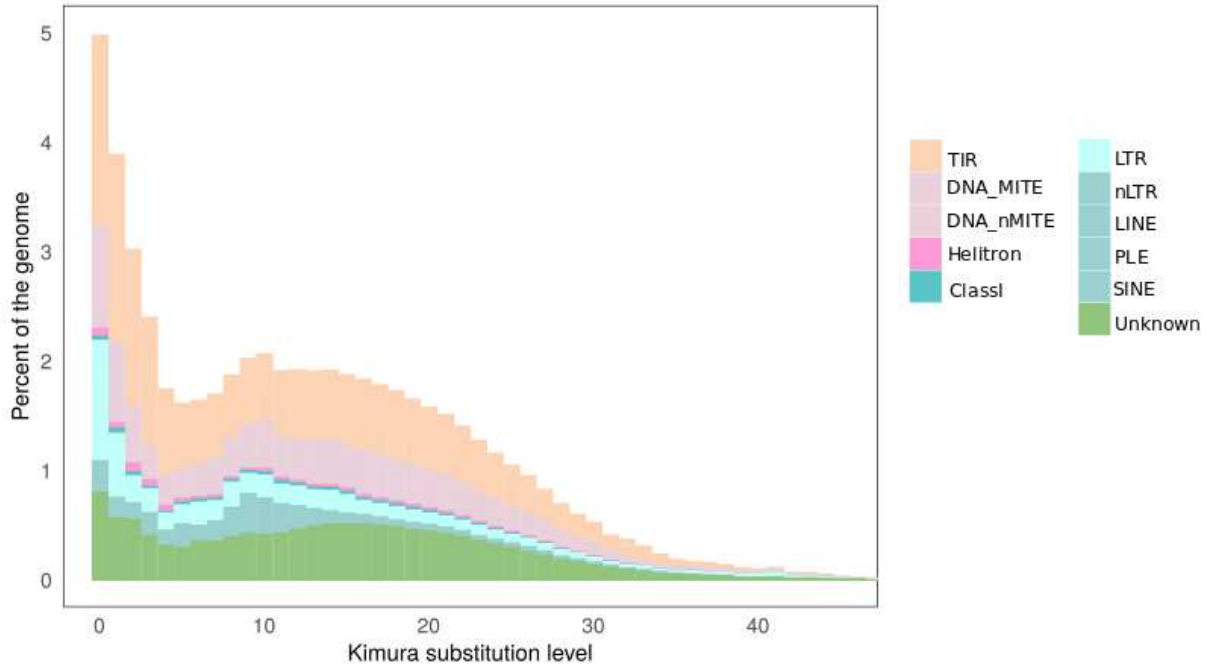


Figure S1: TE landscape of *Hipparchia sbordonii*.

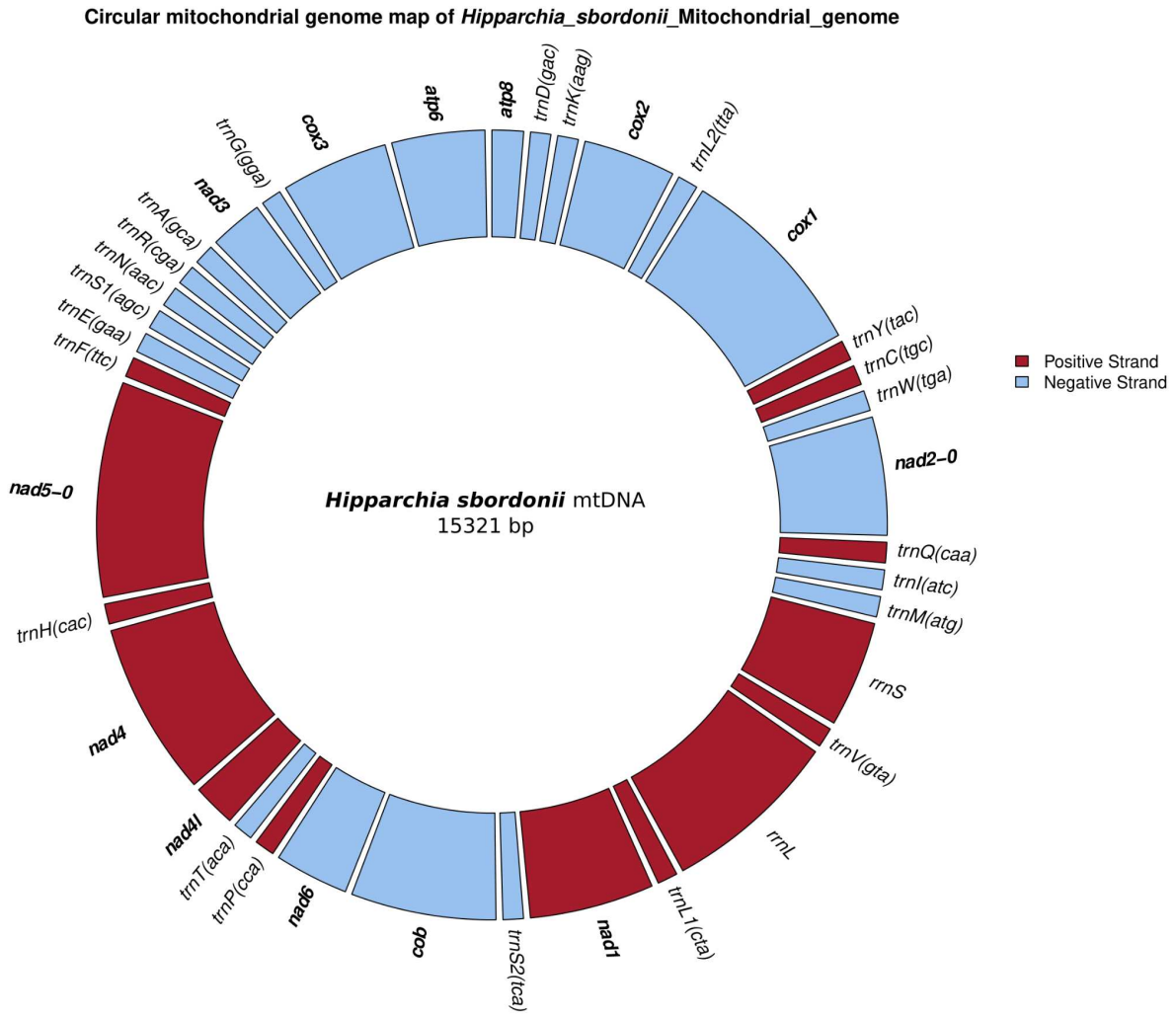
Table S4: Mitochondrial gene in *Hipparchia sbordonii* and their location in the mitochondrial genome.

Hip_sbo_Mit_genome	gene	1	159	-	atp8
Hip_sbo_Mit_genome	tRNA	160	225	-	trnD(gac)
Hip_sbo_Mit_genome	tRNA	227	297	-	trnK(aag)
Hip_sbo_Mit_genome	gene	314	973	-	cox2
Hip_sbo_Mit_genome	tRNA	974	1040	-	trnL2(tta)
Hip_sbo_Mit_genome	gene	1060	2586	-	cox1

Hip_sbo_Mit_genome	tRNA	2579	2642	+	trnY(tac)
Hip_sbo_Mit_genome	tRNA	2643	2706	+	trnC(tgc)
Hip_sbo_Mit_genome	tRNA	2699	2765	-	trnW(tga)
Hip_sbo_Mit_genome	gene	2878	3750	-	nad2-0
Hip_sbo_Mit_genome	tRNA	3827	3895	+	trnQ(caa)
Hip_sbo_Mit_genome	tRNA	3893	3956	-	trnI(atc)
Hip_sbo_Mit_genome	tRNA	3957	4025	-	trnM(atg)
Hip_sbo_Mit_genome	rRNA	4437	5211	+	rrnS
Hip_sbo_Mit_genome	tRNA	5212	5275	+	trnV(gta)
Hip_sbo_Mit_genome	rRNA	5277	6635	+	rrnL
Hip_sbo_Mit_genome	tRNA	6613	6679	+	trnL1(cta)
Hip_sbo_Mit_genome	gene	6687	7610	+	nad1
Hip_sbo_Mit_genome	tRNA	7636	7703	-	trnS2(tca)
Hip_sbo_Mit_genome	gene	7748	8842	-	cob
Hip_sbo_Mit_genome	gene	8872	9381	-	nad6
Hip_sbo_Mit_genome	tRNA	9393	9457	+	trnP(cca)
Hip_sbo_Mit_genome	tRNA	9458	9521	-	trnT(aca)
Hip_sbo_Mit_genome	gene	9557	9811	+	nad4l
Hip_sbo_Mit_genome	gene	9814	11145	+	nad4
Hip_sbo_Mit_genome	tRNA	11153	11218	+	trnH(cac)
Hip_sbo_Mit_genome	gene	11270	12943	+	nad5-0
Hip_sbo_Mit_genome	tRNA	12954	13017	+	trnF(ttc)
Hip_sbo_Mit_genome	tRNA	13016	13080	-	trnE(gaa)
Hip_sbo_Mit_genome	tRNA	13083	13142	-	trnS1(agc)
Hip_sbo_Mit_genome	tRNA	13140	13206	-	trnN(aac)

Hip_sbo_Mit_genome	tRNA	13207	13268	-	trnR(cga)
Hip_sbo_Mit_genome	tRNA	13270	13335	-	trnA(gca)
Hip_sbo_Mit_genome	gene	13345	13683	-	nad3
Hip_sbo_Mit_genome	tRNA	13690	13756	-	trnG(gga)
Hip_sbo_Mit_genome	gene	13765	14547	-	cox3
Hip_sbo_Mit_genome	gene	14556	15221	-	atp6

Figure S2:



Mitochondrial protein gene (in bold), mitochondrial rRNA and tRNA in *Hipparchia sbordonii* and their location in the mitochondrial genome.

Table S5: Table showing the counts, lengths, and types of ncRNAs, present in the genomes of in *Hipparchia semele* and *Hipparchia sbordonii*

ncRNA_type	Counts_*Hipp_semele*	Length_*Hipp_semele* (bp)	Counts_*Hipp_sbordoii*	Length_*Hipp_sbordonii* (bp)
5_8S_rRNA	10	1560	1	156
5S_rRNA	67	7879	52	6109
ACEA_U3	4	814	NA	NA
bantam	1	89	1	89
Histone3	247	11054	67	3034
K_chan_RES	4	455	4	455
let-7	1	76	1	76
LSU_rRNA_archaea	16	39091	NA	NA
LSU_rRNA_bacteria	15	37827	NA	NA
LSU_rRNA_eukarya	16	40136	3	4531
Metazoa_SRP	3	875	3	885
mir-1	1	73	1	73
mir-10	6	417	3	231
mir-1000	1	66	1	66
mir-11	1	69	1	69
mir-1175	1	71	1	71
mir-124	1	79	1	79
mir-133	1	87	1	87
mir-137	1	98	1	98
mir-14	1	59	1	59
mir-184	1	78	1	78
mir-186	NA	NA	2	198

mir-190	1	85	1	85
mir-2	8	504	4	252
mir-210	2	195	2	195
mir-242	2	126	NA	NA
mir-252	1	105	1	105
mir-263	2	177	2	177
mir-274	1	87	1	87
mir-275	1	84	1	84
mir-2755	1	73	1	73
mir-2756	NA	NA	1	75
mir-276	1	89	1	89
mir-2763	1	80	1	80
mir-2765	1	78	1	78
mir-2767	1	79	1	79
mir-277	1	108	1	108
mir-2788	1	91	1	91
mir-2796	1	75	1	75
mir-282	1	93	1	93
mir-305	1	86	1	86
mir-306	1	75	1	75
mir-31	1	79	1	79
mir-317	1	87	1	87
mir-33	1	67	1	67
mir-3327	1	87	1	87
mir-449	1	87	1	87

mir-46	1	68	1	68
mir-67	1	68	1	68
mir-7	1	85	1	85
mir-71	2	116	1	58
mir-745	1	74	1	74
mir-750	1	79	1	79
mir-787	1	89	NA	NA
mir-8	2	146	2	146
mir-9	2	117	2	117
mir-927	1	76	1	76
mir-929	1	71	1	71
mir-932	1	92	1	92
mir-965	1	101	1	101
mir-970	1	74	1	74
mir-971	1	75	1	75
mir-989	1	81	1	81
mir-998	1	79	1	79
mir-iab-4	1	72	1	72
MIR811	1	191	NA	NA
Protozoa_SRP	3	762	NA	NA
R2_retro_el	8	946	12	1414
RNase_MRP	1	233	1	233
RNaseP_nuc	1	294	1	294
snopsi18S-841	2	260	2	258
snopsi28S-1192	1	136	1	136

snoR104	NA	NA	1	118
SNORA16	2	163	2	263
SNORA53	1	134	1	134
SNORD36	1	70	1	70
snosnR60_Z15	2	165	2	165
snosnR61	1	68	1	68
snoU43	1	79	1	79
snoU6-53	1	84	1	84
Sphinx_1	1	99	1	99
Sphinx_2	1	143	1	143
SSU_rRNA_archaea	11	17898	NA	NA
SSU_rRNA_bacteria	9	17208	NA	NA
SSU_rRNA_eukarya	12	17631	1	1902
SSU_rRNA_microsporidia	11	17532	NA	NA
U1	14	2256	14	2258
U11	1	131	1	131
U12	1	148	1	148
U2	19	3284	18	3029
U3	5	895	3	608
U4	5	690	6	771
U4atac	1	133	1	133
U5	8	921	8	920
U6	8	853	8	847
U6atac	3	268	2	179
tRNA	6073	901042	6068	451355

SUM	6663	1129930	6354	486093