Fishing for Junk: investigating the genomic and evolutionary roles of transposable element expansion within neotropical catfish

Christopher L Butler
100247826

A thesis submitted for the degree of Doctor of Philosophy
University of East Anglia, UK
School of Biological Sciences

26/08/22

This thesis is dedicated to the loving memory of Donald Michael Anderson

(1929-2018)

"Every new beginning comes from some other beginning's end"

Table of Contents

List of Tables

List of Figures

List of Supplementary Figures and Tables

List of Figures and Tables found in Appendices

0.1 Acknowledgements

I would like to begin by thanking my supervisory team. Martin, thank you for putting your trust in me from the get-go. Throughout this process you gave me the space to explore this topic and make it 'mine' - yet always provided enough guidance so that I rarely felt lost. I am also extremely grateful for the extra opportunities you co-ordinated for me, including two wonderful trips to Peru and the exciting lobster ageing project. Ellen and Emily, I am so grateful that we were able to collaborate on this project - it really was a joy to explore the confusing world of transposable elements together. Thank you for making the days where I felt useless better. Tracey, your encouragement, and enthusiasm in every progress meeting was infectious and always gave me a timely boost of confidence. This was probably appreciated more than you realised. More than ever my housemates have played an important role during this period of my life. I would like to thank Tom, Elie, Enya, and Leigh, who all spent some time stuck in a lockdown with me. I could not have wished for a better group of people to be working at home with! Thank you for all your laughs, support, and patience. To Midge, thank you for bursting into my life with so much joy and love. The last 18 months would have been a more miserable experience without your (nearly daily) company. Finally, but by no means last, I would like to thank my family. Completing this PhD feels like the culmination of all the enthusiasm, support, and dedication you gave during my upbringing. Some are sadly no longer with me to see this chapter end, but know you helped it to start. To Mum and Dad in particular, I consider this work to be as much your success as mine. Thank you.

0.2 Abstract

Transposable elements (TEs) are short DNA sequences that possess the ability to replicate throughout a genome. They are found ubiquitously across the tree of life and can reach high genomic abundance. Understanding the evolutionary and genomic consequences of TE action is thus of great importance and is aided by a growing ability to annotate TEs within non-model organisms. Teleost genomes have the greatest diversity of TEs of any vertebrate and are therefore ideal study systems to better understand TE biology. This work presents the first detailed foray into the TE biology of the Corydoradinae, a species rich clade of Neotropical catfish with an evolutionary history characterised by rapid colour pattern change, polyploidy and TE proliferation. In-silico modelling was utilised to demonstrate that TE proliferations may be driven by both beneficial insertion effects and whole genome duplications. The latest 'de-novo' TE pipelines were utilised to create a Corydoradinae-specific TE library, with this process increasing estimated TE abundance, altering TE composition, and reducing approximate age of insertions. A significant phylogenetic shift in expressed TE content within the Corydoradinae is also found, though this is somewhat dependent on choice of TE library during annotation. Furthermore, TE insertions appear to accumulate in genic regions at a greater rate within polyploid versus non-polyploid species. For example, a Mariner TE with an amphibian origin has horizontally transferred and inserted within the bone developmental gene '*mmp13*' of multiple polyploid Corydoradinae species, with potential impacts regarding facial shape diversification. Finally, despite many incidences of TE activity inducing changes to colour pattern phenotypes in other organisms, Corydoradinae pigmentation genes were not found to be enriched in TE insertions and evolve under stricter purifying selection pressure than other genes. In summary, this work furthers our understanding of the causes and consequences of TE activity within a non-model system.

0.3 Associated Publications

Below is a list of work found within this thesis that have already been published:

Butler, C.L., Bell, E.A. & Taylor, M.I. Removal of beneficial insertion effects prevent the long-term persistence of transposable elements within simulated asexual populations. *BMC Genomics* 22, 241 (2021). https://doi.org/10.1186/s12864-021-07569-3

Bell EA, **Butler CL***, Oliveira C, Marburger S, Yant L, Taylor MI. Transposable element annotation in non-model species - the benefits of species-specific repeat libraries using semi-automated EDTA and DeepTE de novo pipelines. *Mol Ecol Resour.*22, 823– 833.8 (2022) doi: 10.1111/1755-0998.13489.

***co-first author**

1. General introduction to transposable elements and the Corydoradinae

1.1 What are transposons?

Transposable Elements (TE) are short DNA sequences that possess the potential for replication through their integration into new genomic sites (McClintock 1950). TEs are primarily retained within the population through vertical transmission, distinguishing themselves from other replicative elements (e.g. viruses and phages) which readily insert themselves outside of their genome of origin (Piégu *et al.* 2015). First discovered in Maize (*Zea mays*), TEs are abundant within nearly all known genomes, constituting 85% of the maize genome and 50% of the human genome for example (Lander & IHGSC 2001; Stitzer *et al.* 2021). There are some exceptions, the genomes of filamentous fungi belonging to the *Ashbyra* genus contain no TEs for instance (Dietrich *et al.* 2013). TEs typically fall within the 100-10,000bp range, but can reach extremely large sizes. For example, the teleost DNA element *'teratorn'* can reach a giant 180kb long through a partial fusion with a herpesvirus (Inoue *et al.* 2017) (Arkhipova *et al.* 2019). Historically, TEs have been referred to as 'selfish' or 'parasitic' sequences, with their activity largely considered to negatively correlate with an individual's fitness (Slotkin & Martienssen 2007). However, the ability of transposons to induce large scale genomic modifications has also led to them being increasingly viewed as significant drivers of evolutionary change. The following introduction will begin by exploring both the vast diversity of TEs and the biological impacts that their activity can cause.

Just like the multitude of species on Earth which have evolved in response to different environmental habitats, so different genomes have given rise to a diverse array of TEs. The first attempts to systematically classify the diversity TEs did so based on their mode of transposition (Finnegan 1989). Class I elements (retrotransposons or "copy and paste" elements) replicate via an RNA intermediate, whilst Class II elements (DNA transposons or "cut and paste" elements) mobilise

directly. These two TE classes have been further sub-categorised into a series of hierarchies (subclasses, orders and superfamilies) based on sequence, structural or mechanistic similarities (Wicker *et al.* 2007). Nevertheless, viewing TEs under this simple dichotomy underplays the huge diversity of transposition mechanisms that exist. The enzymatic similarities of encoded transposases have been used to provide a more 'mechanistic' TE classification system (Curcio & Derbyshire 2003). Whilst such proposals have failed to be widely adopted by the wider transposon community, they can provide a better indication on how different modes of transposition were acquired throughout TE history. The need for a TE classification system that better reflects their evolutionary history has been strongly argued (Seberg & Petersen 2009; Piégu *et al.* 2015), with a growing number of studies utilising a phylogenetic framework of different TE types (De Boer *et al.* 2007; Manthey *et al.* 2018; Neumann *et al.* 2019). However, attempts to view TE diversity under a 'universal' phylogenetic framework is challenging; transposons are subject to rapid sequence degradation, possess no single ancestor (Capy 2005; Piégu *et al.* 2015) and phylogenetic analysis may be limited to the use of single "consensus" sequences (Arkhipova 2017). Nevertheless, new classification systems that integrate aspects of both Wicker's and Curcio & Derbyshire's proposals, along with a better appreciation of transposon evolutionary history will provide a more comprehensive overview to the tremendous levels of transposition diversity and should be encouraged.

1.2.1 DNA transposons

DNA transposons do not encode a reverse transcriptase (RT), and instead rely on a single enzyme known as a transposase to integrate themselves at new genomic sites (Pritham *et al.* 2007). DNA transposons have traditionally been divided into two subclasses based on the number of cleaved DNA strands during transposition (Wicker *et al*. 2007). However, Curcio & Derbyshire (2003) grouped DNA transposons into two different clades based on the different enzymatic machinery they encode; namely the DDE transposases and the Y2 transposases. The catalytic domain of these transposases are hypothesised to permit DNA transposons to show considerable target site-specificity during

transposition (Vigdal *et al.* 2002; Feschotte & Pritham 2007). For example, Tc1/Mariner elements demonstrate considerable insertion preference into palindromic AT-repeats, with the malleable structure of these regions being appropriate for DNA disruption (Vigdal *et al*. 2002).

DDE transposases are amongst the most common type of integrase, characterised by the presence of a three-residue amino acid motif; aspartate (D), aspartate (D) and glutamate (E) (Curcio & Derbyshire 2003). Mechanistically, transposition proceeds via a highly conserved three step process; (i) hydrolysis (ii) strand transfer/trans-esterification and (iii) target site duplication (Arkhipova 2017). The repair of the new integration site will result in a signature target-site duplication (2-10bp) (Yuan & Wessler 2011). Under Wicker's (2007) proposal, the DDE transposases are synonymous with terminal inverted repeat (TIR) DNA elements. These 'cut and paste' elements are defined by (i) flanking TIR which may act as the transposase recognition site (ii) the double stranded nature of their cleavage and (iii) a resulting target site duplication at the target DNA site. Despite the initial excision of the transposon from the donor site, TIR elements may still be able to increase in copy number via indirect mechanisms, such as gap repair or preferential transposition to unreplicated chromatids during DNA replication (Feschotte & Pritham 2007). Under Wicker's proposal there are nine different superfamilies of TIR transposons based on both sequence similarity of the inverted repeats and the size of site duplications (Wicker *et al*. 2007).

The other major class II catalytic domain are the Y2 transposases (named as they include a pair of tyrosine (Y) residuals), which permit transposition through an alternative 'rolling circle' (RC) mechanism. DNA transposons which encode a Y2-transposase include eukaryotic Helitrons and various prokaryotic elements such as the insertion sequence 91 (IS91) family (Curcio & Derbyshire 2003). Unlike DDE transposition, Y2-mediated transposition proceeds through single strand cleavage, does not produce target site duplications, and occurs in a replicative rather than a 'cut & paste' mechanism (Feschotte & Wessler 2001; Curcio & Derbyshire 2003). The TE is excised through

the action of encoded Helicase proteins, which assist in the unwinding of the DNA duplex during replication (Curcio & Derbyshire 2003). The mode of replication of Helitron elements means they appear to have a particular propensity to capture genes, giving rise to chimeric transcripts that contain both TE and captured gene fragments (Barbaglia *et al.* 2012) An estimated ~25% of transcripts within Maize (*Zea mays)* are expressed as a result of such process, meaning Helitron elements may play a particularly important role in the evolution of new coding regions (Barbaglia *et al.* 2012). Bacterial Y2-transposases do not encode a Helitron protein, but transposition can still occur in a very similar fashion to eukaryotic Helitron elements (the 'concerted' model) or via an alternative step-wise fashion (the 'sequential' model) whereby a circular intermediate is formed (Curcio & Derbyshire 2003; Thomas & Pritham 2015).

Recently, an increasing number of DNA transposons are being discovered which do not replicate through either a 'cut & paste' or RC mechanism. For example, many prokaryotic DNA transposons encode tyrosine (Y) and serine (S) transposases which, due to their shared mechanistic characters, obtain their name from two families (Y & S) of recombinase proteins (Curcio & Derbyshire, 2003). The mechanisms encoded by Y/S transposases tend not to produce target site duplication (TSDs) and the intermediate dsDNA is circular rather than linear (Curcio & Derbyshire, 2003). Outside of prokaryotes, the tyrosine (Y) transposase can also be found in eukaryotic Crypton elements. Originally discovered in fungi, Crypton elements transpose through the generation of extrachromosomal circular DNA (Goodwin *et al.* 2003; Kojima & Jurka 2011). Finally, the recently discovered Maverick/Polinton elements represent a distinct group of large (15-20 kb) DNA transposons found in a diverse range of non-plant eukaryotes (Wicker *et al*. 2007). Despite possessing TIRs and generating TSDs, Maverick elements are the most complex of all eukaryotic DNA transposons; replicating through a single-strand DNA intermediate and an alternative 'self-synthesising' mechanism (Kapitonov & Jurka 2006; Pritham *et al.* 2007). Within their extensive protein coding domain, Maverick elements possess an integrase (INT), which are proteins otherwise

found within the LTR-retrotransposons. This is suggestive of a possible evolutionary link between class I retrotransposons and class II DNA transposons.

1.2.2 Retrotransposons.

The presence of a reverse transcriptase (RT) within the respective coding region of retrotransposons mean they can be clearly distinguished from DNA transposons (Wicker *et al*. 2007). Transposition occurs through the production of an RNA copy which can subsequently be reverse transcribed, and the resulting double stranded cDNA reinserted into a new genomic site. Highly prevalent in eukaryotic genomes, this 'copy and paste' mechanism allows retrotransposons to reach high copy number, with the doner element remaining intact within its initial location (Sotero-Caio *et al.* 2017). The classification of retrotransposons has traditionally depended on the presence of absence of long terminal repeats (LTR) found within their domain (Wicker *et al*. 2007).

LTR retrotransposons possess terminal repeats either side of a highly conserved ORF containing both polymerase (pol) and group specific antigen (gag) genes (Kazazian 2004). The LTR regions themselves can be used as transcription initiation sites (Neumann *et al.* 2019), whilst the gag domain encodes a protein which acts as a coat for virus-like particles (Kazazian 2004). Proteins required for transposition are found within the pol domain, including a reverse transcriptase (RT), an integrase (INT), an aspartic proteinase (PROT) and an RNAse (Wicker *et al*. 2007). The highly conserved nature of this *pol* domain has permitted comprehensive TE phylogenetic relationships to be generated through sequence alignment, with this approach being most recently applied across numerous *Drosophila* (Bargues & Lerat 2017) and viridiplantae species (Neumann *et al*. 2019). The integrase family which inserts the cDNA copy into the new genomic site is distantly related to the DDE transposase found within class II transposons (Curcio & Derbyshire 2003). Consequently, under Curcio & Derbyshire's proposal, the LTR retrotransposons are grouped together with the DDE transposons, with the former having gained an RT domain at some point in their shared history. This

implies that class I & class II elements have likely had multiple evolutionary origins, with the terms retrotransposon and DNA transposon better viewed as 'TE phenotypes' which may have emerged through evolutionary convergence (Piégu *et al.* 2015).

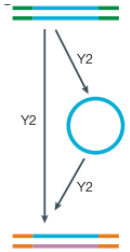
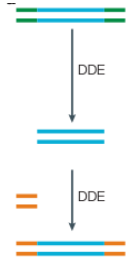
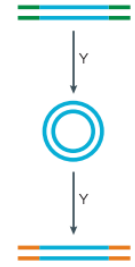
Under Wickers' classification the LTR retrotransposons can be divided into five evolutionarily related superfamilies; namely the (i) retroviruses, (ii) endogenous retroviruses, (iii) Ty3-Gypsy, (iv) Ty1-Copia and (v) Bel/Pao elements. It is worth noting that the use of the word "Gypsy" to describe class of retrotransposons have recently been subject to dispute, primarily because its naming draws on the negative stereotypes of people with Romani heritage (Wei *et al.* 2022). Alternatives such as Ty-3 and *mdg4* elements have been proposed but have not yet been widely adopted by the literature (Wei *et al.* 2022). Retroviruses (e.g. HIV) are RNA transcripts which possess the ability to infect their host DNA through reverse transcription. A novel envelope (env) gene within their ORF allows for viral attachment and penetration; permitting retroviruses to transfer from one genome to another (Gifford & Tristem 2003, Kazazian 2004). Occasionally, a retrovirus may become adopted by the germline of a population generating a so-called exogenous retrovirus (ERVs) (Gifford & Tristem 2003). Degradation of the env gene over time means that the only way these ERVs can be maintained within the population is through vertical transmission (Wicker *et al*. 2007). Approximately 8% of the human genome consists of ERV sequences, with increasing evidence indicating that they may have had important immune roles during human evolution (e.g. co-option for maternal immune tolerance during pregnancy) (Grandi & Tramontano 2018). The evolutionary relationship between ERVs and other LTR retrotransposons remains unclear. They share high sequence similarity with Ty3-Gypsy-like elements, though their age and subsequent order of *env* loss/gain remains unclear (Hayward 2017). Different gene orders within the pol domain separates the Gypsy-like elements with their related Ty1-Copia-like elements (Wicker *et al.* 2007). Large sequence divergences within the Copia RT domain has led to the suggestion that they are the most ancient of all LTR-retroelements (Eickbush & Jamburuthugoda 2008). The final subfamily of LTR

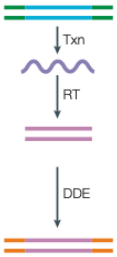
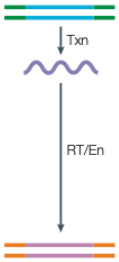
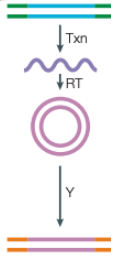
elements are the BEL/Pao elements which to date have only been identified in metazoan genomes, perhaps indicative of their recent evolution (De La Chaux & Wagner 2011). BEL/Pao elements possess an identical gene order as Gypsy elements but are placed within their own superfamily based on sequence dissimilarity within their RT domain (Wicker et al. 2007; De La Chaux & Wagner 2011).

Often grouped as one, Wicker *et al* (2007) divided the non-LTR retrotransposons into four different orders, namely the (i) Dictyostelium intermediate repeat sequences (DIRS) elements, (ii) Penelope-like elements (PLE), (iii) the long-interspersed elements (LINEs) and the (iv) short-interspersed elements (SINEs). Together, the non-LTR retrotransposons are estimated to have resided within eukaryotic genomes for hundreds of millions of years (Han 2010). During that time, they have evolved numerous different transposition mechanisms. The first are the so-called target-primed (TP) retrotransposons which transpose using a combination of reverse transcriptase and endonuclease activities, with the latter responsible for determining the element's target site (Curcio & Derbyshire 2003; Han 2010). Their structure may include a gag-like ORF with RNA binding or nucleic acid chaperone activities, and a second ORF which encodes both the RT and endonuclease domain. Unlike LTR-retrotransposition, reverse transcription occurs at the target DNA site itself and can result in target site deletions or duplications depending on the relative position of the second endonuclease nick (Han 2010). Under Wicker's proposal LINEs and PLEs encode both RT and endonucleases and should thus be classed as TP retrotransposons. PLEs were first described in Drosophila, though they have subsequently been found in the genomes of over 50 diverse species (Wicker *et al*. 2007). The classification of PLEs is a contentious issue because they contain unique RT sequences and contain an unusual endonuclease domain belonging to the Uri family (Pyatkov *et al.* 2004). Nevertheless, phylogenetic studies have concluded that non-LTR retrotransposons and PLEs can be grouped together into a larger 'eukaryotic target-primed' clade of retroelements (Arkhipova 2006). LINEs encode both an endonuclease and RT more similar to that found in the LTR

retrotransposons, with the TP mechanism of their transposition being well documented. Their RT domain also plays a crucial secondary role in allowing 'parasitic' SINEs to transpose across the genome. SINEs are short retroelements which do not exceed a length of 500bp (Wicker *et al*. 2007). Unlike other types of retrotransposons; SINE transcription relies on the action of RNA polymerase III (pol III) rather than polymerase II (pol II), and they do not contain any protein coding domains within their sequence (Carnell & Goodman 2003; Kramerov & Vassetzky 2011). SINEs are consequently referred to as 'non-autonomous', relying on the transposition mechanism of LINE partners of similar sequence. Despite this, approximately 11% of the human genome consisting of the SINE element Alu (Deininger 2011). Together, specific SINE and LINE elements (L1, Alu and SVA) are believed to be the only TE group still active within humans (Belancio *et al.* 2009). The final group of non-LTR retrotransposons are the Y retrotransposons, which includes the anciently diverged DIRS (discovered in the slime mould Dictyostelium) and Ngaro (discovered in *Danio rerio*) elements (Goodwin & Poulter 2004). Unlike other non-LTR retrotransposons, these do not encode an endonuclease, and transpose through the action of a tyrosine (Y) recombinases. This occurs in similar fashion to that earlier described in certain bacterial DNA and Crypton Y transposases, with sequence comparisons suggesting that the two groups may be evolutionarily related (Goodwin *et al.* 2003). Interestingly DIRS elements also possess inverted LTRs but are not classified as LTR retrotransposons because they (i) do not encode a DDE-like integrase, (ii) do not generate TSD upon insertion and (iii) replicate via a dsDNA circular intermediate (Poulter & Goodwin 2005; Malicki *et al.* 2017) (Poulter & Goodwin 2005; Malicki et al. 2017). A summary of the diversity of TE types introduced in this section, and some inferences regarding their possible evolutionary relationships with each other is given in Table 1.1.

Table 1.1. Summary of the different TE types found with within Eukaryotic genomes. Visual representation includes figures previously presented in Curcio & Derbyshire, 2013.

TE Class	Wicker's Proposal	Eukaryotic Elements	Enzymatic Features	Target Site Duplication?	Terminal Inverted Repeats?	Circular DNA intermediate?	Single Stranded Intermediate?	Visual Representation
Class II DNA transposon *(no RNA intermediate)*	Single strand cleavage	*Helitrons*	Y2 transposase (Rolling-Circle)			✓ *(* in sequential model)*	✓	
		Maverick/Polinton	Self-synthesising	✓	✓			
	Double strand cleavage	*Mariner, hAT, Mutator, Transib, P, PiggyBac, PIF-Harbinger, CACTA)*	DDE transposase	✓	✓			
		Crypton	Y transposase	✓		✓		

Class I Retrotransposon *(RNA intermediate)*	LTR retrotransposon	–	Retroviruses	DDE retrotransposition	✓		
		–	ERVs				
		–	Copia-like				
		–	Gypsy-like				
		–	Bel/Pao				
	Non-LTR retrotransposon	–	Penelope-like	TP-retrotransposition			
		–	LINEs				
		–	DIRS	Y-retrotransposition		✓	
		–	Ngaro				
		–	SINEs	Non-autonomous			

1.2 The future of TE classification

Incorporating the 'extra' information provided by 3D protein characteristics, such as a-helices & -sheets may improve the reliability of phylogenetic inference (Arkhipova 2017). There is increasing interest in whether this approach could become a useful tool in future efforts to generate a TE classification system rooted in their evolutionary history. Structural alignments have previously been used to describe the phylogenetic relationship between all retrotransposons (Gladyshev & Arkhipova 2011). The authors used the structural similarities of RT domains to generate four different retrotransposon clades; namely the (i) prokaryotic RTs, (ii) the viral-like LTR retrotransposons and related retroviruses (iii) the non-LTR retrotransposons and related *rvt* genes and finally (iv) the Penelope-like elements and related telomerase reverse transcriptases (TERT). Seemingly, using structural based classifications can infer the evolutionary relationship between distantly related TEs, which could be extended to provide a 'universal TE phylogeny'.

1.3 Host defence against TE activity

Due to their propensity to disrupt genomic stability, the activity of TEs is widely hypothesised to have a net negative impact on host fitness. There are therefore several host defence mechanisms to suppress TE activity at both a pre and post transcriptional level. Amongst higher eukaryotes CG epigenetic DNA methylation at TE loci is the most common strategy for their repression, with the deposition of methylation marks converting the genomic region to a closed 'heterochromatic' state and lowering expression levels. TE sequences are believed to be the most methylated regions of eukaryotic genomes, with heterochromatin itself hypothesised to have originally evolved within the last common ancestor of eukaryotes to silence active TEs (Rigal & Mathieu 2011; Deniz *et al.* 2019; Almeida *et al.* 2022). Amongst eukaryotes methylation is largely controlled by the action of two DNA methyltransferase families; DNMT1 & DNMT3 (Deniz *et al.* 2019). Consequently, the experimental loss of DNMT genes can lead to a subsequent increase in TE activity within eukaryotes (Bourc'his & Bestor 2004; Tsukahara *et al.* 2009; Chernyavskaya *et al.* 2017). Epigenetic TE silencing is also hypothesised to generate conflict between TEs and host. When TEs are in close proximity to host

genes, a closed heterochromatin state will disrupt neighbouring gene expression and possibly limit a host's capacity for TE silencing. Within Maize, donor genes with have been captured by TEs are more heavily methylated than those that are not (Muyle *et al.* 2021). Not only may this conflict potentially allow TE to escape host silencing, but this effect appears to be more pronounced within non-essential vs essential genes. This has led to the suggestion that epigenetic silencing of TEs may provide pseudogenisation routes of nearby genes if they are of low functional importance (Muyle *et al.* 2021).

Across vertebrates, the action of RNA silencing pathways (sometimes referred to as RNA-i/interface pathways) permit the recruitment of methyltransferases (e.g. DNMTs) to epigenetically silence TEs (Matzke & Mosher 2014). Namely this involves both piwi interacting RNAs (not present in plant and fungi) and small interfering RNAs, which are recruited to recognise 'non-self' genetic elements. It is not well understood how TEs are recognised over other 'self' regions of a genome, although structural TE functions (e.g terminal inverted repeats) may allow them to be distinguished (Slotkin & Martienssen 2007). Specifically, these small RNA families can recruit argonaute proteins which post-transcriptionally splice TEs into non-functional copies or DNMTs which pre-transcriptionally add methyl groups to TE sites (Almeida *et al.* 2022). An additional family of proteins known as KRAB (Krüppal-associated box) domain containing zinc finger proteins are found within mammals which recruit DNMTs to silence TEs during rounds of epigenetic depression during embryogenesis (Almeida *et al.* 2022). An arms race between TEs and these proteins mean they are extremely fast evolving, with their abundance highly correlated with the abundance of retrotransposons within mammalian genomes (Almeida *et al.* 2022).

1.4 TE reactivation

In certain circumstances, TEs may evade host defence mechanisms and rapidly proliferate throughout the genome (Figure 1.1; Ungerer *et al.* 2009; Notwell *et al.* 2015; Rogers *et al.* 2018;

Benoit *et al.* 2019). Firstly, TEs may accumulate in regions of the genome where they do not cause a significant change in host fitness, such as areas with low recombination rates or low gene number (Abrusán & Krambeck 2006; González & Petrov 2012). Alternatively, TEs may provide fitness benefits to the host, allowing for their co-option. Secondly, TEs may be able to evade host silencing after periods of environmental stress. This is an effect which has been hypothesised to be induced by changes to genome structure. For example, heat stress with Arabidopsis led to the transcriptional activation of TEs due to heterochromatin decondensation (Pecinka *et al.* 2010). Whilst environmental stresses are likely to be temporary in nature, resulting increases in TE activity may be cross-generational. For example, the progeny of *Arabidopsis* individuals exposed to UV radiation have higher expression of the LTR retrotransposon ONSEN than that recorded within control groups (Migicovsky & Kovalchuk 2014).

A third mechanism by which TEs may evade host silencing involves hybridisation between two parent individuals of different species. Mechanistically, increases in TE activity after hybridisation is likely to occur due to a disruption in epigenetic silencing. For example, with kangaroo hybrids reactivated LTR-retrotransposons are significantly less methylated than within either parental species (O'Neill *et al.* 1998). Furthermore, TE reactivation within lake whitefish (*Coregonus clupeaformis*) hybrids can be attributed to the downregulation of the DNA methyltransferase DNMT1 (Dion-Côté *et al.* 2014). Hybridisation may lead to a subsequent increase in ploidy number, with such 'genomic shocks' being associated with TE activation. For example, increases in ploidy number are associated with a reduction in siRNA production and reactivation of Veju retrotransposons within allopolyploid wheat hybrids (Kenan-Eichler *et al.* 2011). Theoretically, increased TE activity within hybrids could also result from the mismatched inheritance of paternal elements and maternal siRNAs (Crespi & Nosil 2013). The final process that may lead to TE reactivation regards their ability to 'jump' from the genome of one unrelated species to another, in a process known as horizontal transfer. Newly transferred TEs are likely to be able to rapidly

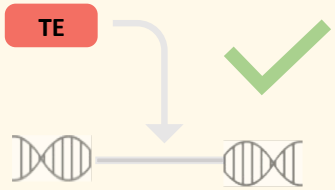
proliferate within a naïve genome, particularly if there has been an absence of co-evolution between host and TE or if silencing is copy number dependent (Schaack *et al.* 2010). LTR retrotransposons and DNA transposons are more likely than non-LTR retrotransposons to undergo horizontal transfer, perhaps due to their more stable double stranded intermediate (Schaack *et al.* 2010). Furthermore, the aquatic environment appears to facilitate the transfer of TEs between species, with 94% of all vertebrate horizontal transfer events involving teleost fish (Zhang *et al.* 2020). The mechanisms by which TEs can transfer between genomes are poorly understood, but are likely to involve either bodily fluids, pathogens, or parasites. For example, the DNA transposon SPIN has undergone a horizontal transfer across multiple tetrapod lineages, which based on sequence similarity may have originated from the ectoparasite *Rhodnius prolixus* and intermediate trematode host, *Lymnaea stagnalis* (Gilbert *et al.* 2010).

1.5 Vertebrate TE abundance & diversity

There is considerable variation (4-60%) in the proportion of TEs within vertebrate genomes; whereby even closely related species can display significant differences in their TE landscape (Sotero-Caio *et al.* 2017) (Figure 1.2). However, the exact features responsible for determining TE abundance are poorly understood. As a general trend, actinopterygian species (the ray finned fishes) have a greater genomic TE abundance than sarcopterygian species (the lobe finned fishes) (Chalopin *et al.* 2015). The emergence of metabolic processes (e.g. flight) within the sarcopterygian lineage may explain the associated decrease in TE abundance, particularly because such processes tend to lead to genome downsizing (Manthey *et al.* 2018). Other features, such as the efficiency of host defence systems and self-regulating feedback loops are also likely important factors in controlling TE abundance.

Co-option:

Preferential TE insertion within non-genic genome regions, or where they may provide a fitness advantage to the host can lead to their proliferation.

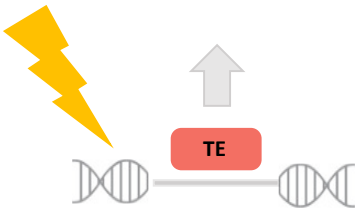


Example:

Within mammals the TE family 'MER130' are enriched within neocortical genes, where they have been co-opted to function as gene enhancers (Notwell et al, 2015).

Stress:

Environmental stress and resulting changes to genome structure may permit transgenerational increases to TE activity.

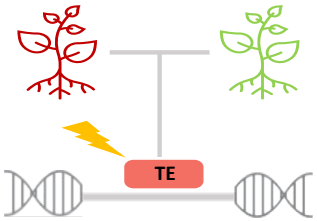


Example:

Tomato individuals experimentally exposed to drought accumulate transcripts of the retrotransposon RIDER, providing evidence of their increased expression after stress (Benoit et al. 2019)

Genome Shocks:

Hybridisation may lead to increased TE activity by disrupting epigenetic marks, causing a mismatch between TE and siRNA silencers or through a whole genome duplication event (allopolyploidy)

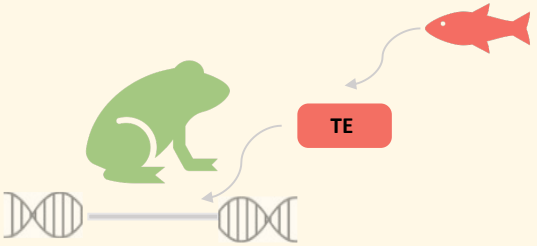


Example:

Within ancient *Helianthus* sunflower hybrids classes of LTR retrotransposon (Ty3/Gypsy) have proliferated leading to substantial increases in genome size (Ungerer et al. 2009).

Horizontal Transfer:

Transfer of TEs from unrelated species may allow for their proliferation within a naïve genome lacking specific TE silencers.



Example:

The large genome of the strawberry frog (*Oophage pumilio*) appears to be due to the ongoing proliferation of Tc1-Mariner elements of teleost origin (Rogers et al. 2018).

Figure 1.1 The four main causes of transposable element (TE) reactivation

A recent meta-analysis reported that high concentration of transposases can lead to a reduction in overall DNA transposon activity, suggesting that proliferation of TEs may be self-regulated via a negative feedback loop (Bire *et al.* 2013). The mechanisms underlying such regulation are not well defined. Recent transposase concentrations have been found to positively correlate with the formation of inhibitory filamentous 'rodlet' structures, whose appearance coincides with a general reduction in transposon activity (Woodard *et al.* 2017).

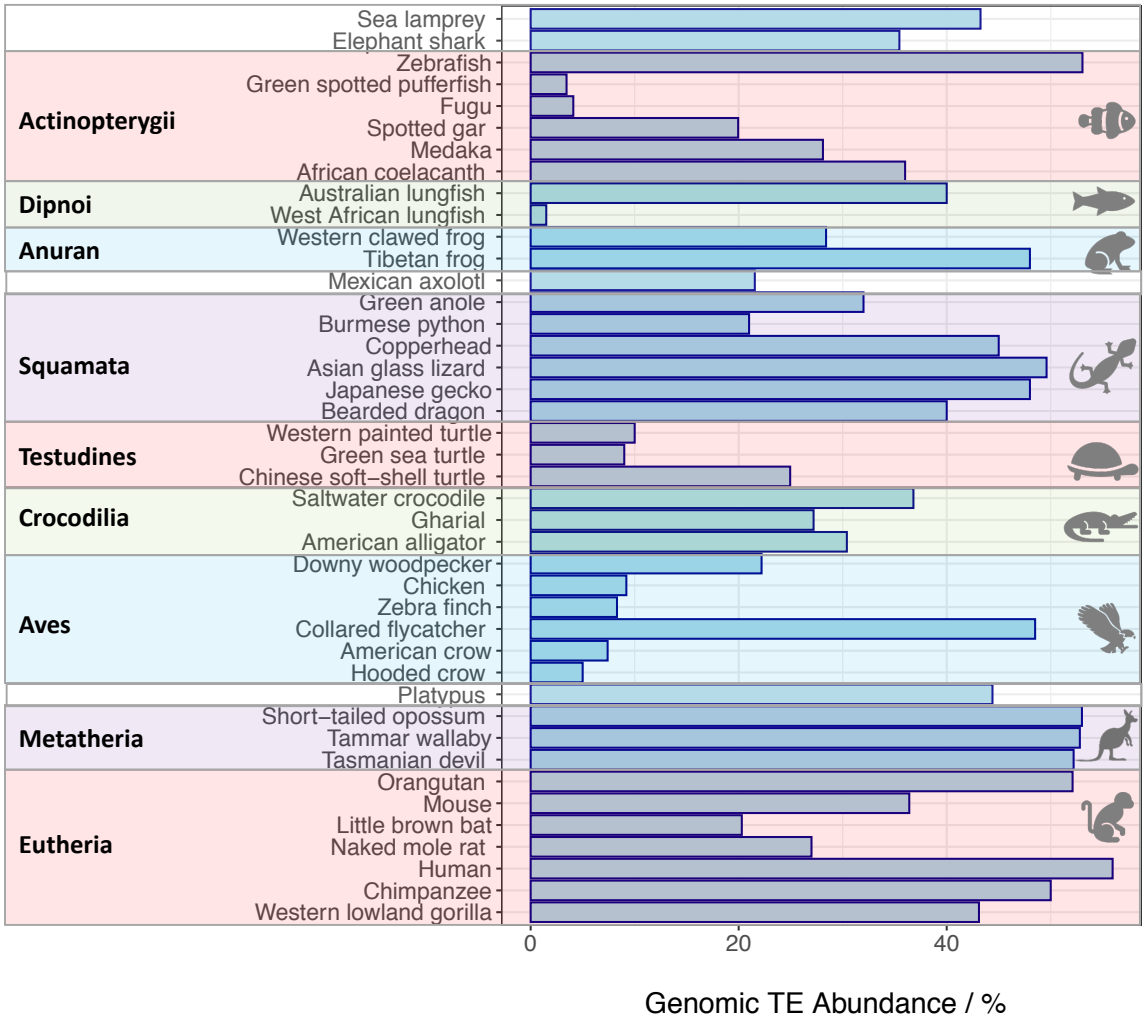


Figure 1.2 Genomic transposable element (TE) abundance across different vertebrate clades. Figure produced using supplementary data from Sotero-Caio *et al.* 2017.

The relationship between TE abundance and diversity is complex. Estimations of TE abundance in humans ranges from 40-90%, yet just two class II retrotransposon families (L1 & Alu) dominate this

make up (González & Petrov 2012). Actinopterygian's possess the largest diversity of transposon types, with all known TE superfamilies being represented in their genomes (Chalopin *et al.* 2015). This holds true, even in the case of the pufferfish genome which is 10x smaller than humans (Chalopin *et al.* 2015). TE diversity has been demonstrated to have an approximately normal distribution with regards to genome size, whereby diversity is lowest in both very small and very large genomes (Elliott & Gregory 2015). Attempts to understand what factors influence TE abundance and diversity are increasing, frequently adopting population dynamic concepts such as competition, mutualism, and parasitism. Non-autonomous elements (e.g. SINEs) rely on the presence of similar elements for their activity; whilst competition between different subfamilies may drive TE diversity. Alternatively, the relationship between TEs and their respective silencers can be treated in a fashion similar to well described predator-prey models. The presence of specific silencing mechanisms has been demonstrated to generate TE diversity, whilst more general 'cross-reactive' silencers decrease TE diversity (Abrusán & Krambeck 2006). Alternatively, the fact that TE proliferation leads to increased ectopic recombination rates can lead to 'intrafamily' competition. As small genomes naturally have higher rates of background ectopic recombination, TE diversification may be strongly selected for as sequence dissimilarity will prevent critical levels of ectopic recombination (González & Petrov 2012).

1.6 Consequences of TE activity

1.6.1 Population level impacts

TE insertions have several large-scale consequences, both at a genomic and population level. A well-documented consequence of TE accumulation is the subsequent increase in genome size. For instance, differential accumulation of LTR retrotransposons across Gossypium and panicoid grass lineages has resulted in a threefold and twofold increase in genome size respectively (Hawkins *et al.* 2006; Estep *et al.* 2013). The positive relationship between TE activity and genome size provides an explanation to the so-called 'C-value paradox'; whereby expansions of non-coding regions can lead to greater genome size without an associated rise in organismal complexity. Secondly, TEs may jump into functional area of the genome and increase rates of mutation. TE insertions generate 0.3% of all

human mutations; and have known associations with over 50 diseases including cystic fibrosis, haemophilia, and various cancer types (Belancio *et al.* 2008). In other organisms this percentage may be much higher; TEs can generate up to 80% of all visible mutations in *Drosophila* for example (González & Petrov 2012). Furthermore, there is correlative and experimental evidence regarding the role that TE activity may have in the biological ageing process. The long-lived naked mole rat (*Heterocephalus glaber*) has a paucity of TEs compared to other rodent species for example, whilst the administering of drugs which prevent the activity of L1 elements have led to significantly longer life spans in both mouse and *Drosophila* species (Kim *et al.* 2011; Wood *et al.* 2016; Simon *et al.* 2019). TEs can also rewire the exome of an individuals, either by inserting within open reading frames, providing alternative poly-A tails or through exonisation (provision of alternatively spliced variants) (Cowley & Oakey 2013). Even when silenced or fragmented, TEs which have lost their own autonomous replicative ability may remain expressed by inserting near genic regions and forming chimeric transcripts (Lanciano & Cristofari 2020).

However, there is growing evidence that occasionally the action of TE insertions may induce selectively neutral or even beneficial effects. Known adaptive phenotypes which are generated through TE insertion include enhanced resistance to viral infection (Chung *et al.* 2007; Magwire *et al.* 2011), increased developmental stability (Sun *et al.* 2014); and the generation of adaptive colour morphs (Sayah *et al.* 2004; Van't Hof *et al.* 2016). More recently, TE induced truncation of the *D. melanogaster* gene 'Veneno' has been implicated in their resistance to the pathogen '*Drosophilia* A virus' (Brosh *et al.* 2022). TEs may also contribute to the creation of novel genes in a process known as exon shuffling, whereby flanking regions of DNA are carried to new genomic locations. LINE mediated retrotransposition of the CypA gene has led to HIV resistance within owl monkeys (*Aotus trivirgatus*) for instance (Sayah *et al.* 2004). TEs may also contribute to the generation of new gene regulatory networks (GRNs), with cis regulatory elements forming (i) 'de novo' after TE degeneration or (ii) through the co-option of transposed TEs containing existing cis-regulatory sequences

(Feschotte 2008). Within the human genome, estimates of known promoter regions containing TE-derived sequences are estimated to be between 25 - 83% for example (Jordan *et al.* 2003; Thornburg *et al.* 2006). Due to the ability of TEs to disperse themselves across the genome, TEs containing cis-regulatory sequences may bring numerous spatially separated genes under a single GRN. For example, the LINE induced duplication of the RE1 transcription binding site has allowed novel target genes to be silenced by the transcriptional repressor REST (Johnson *et al.* 2006). TE proliferation may also result in more 'structural-like' modifications. High sequence similarity between TE copies can lead to ectopic recombination between non-homologous genomic regions. This can result in a wide range of chromosomal rearrangements, such as deletions if recombination occurs between TEs on the same strand, or duplication if recombination occurs between TEs on different strands (González & Petrov 2012). Such rearrangements can dramatically alter individual phenotype. For example, TE induced deletion/duplications of the opsin gene may have had significant consequences regarding the evolution of colour vision within both Old-World primate and amphioxus lineages (Dulai *et al.* 1999; Pantzartzi *et al.* 2018). Finally, recombination between two TEs of opposite orientation may result in the spanning genetic material becoming inverted, with the action of retrotransposons responsible for nearly half of all known inversions between human and chimpanzees (Lee *et al.* 2008).

1.6.2 Population level impacts

The ability of founder populations to rapidly adapt to foreign environments represents somewhat of an evolutionary paradox. Population bottlenecks and resulting reductions in genetic diversity are likely to limit a founder population's ability to generate novel phenotypes (Stapley *et al.* 2015). Nevertheless, these effects may be offset in founder groups by TE proliferation, perhaps stemming from a reduction in overall selection efficiency or exposure to environmental stresses (Stapley *et al.* 2015; Specchia *et al.* 2017). Theoretically, increased mutation rates and novel gene formation will provide TE-enriched founder groups with greater chances to rapidly adapt to new environments.

There are now a growing number of cases in which TE-induced increases in genetic diversity are reported to be higher amongst founder populations. For example, the highly invasive *Cardiocondyla obscurior* ant genome contains a higher proportion of both class I & II TEs clustered within regions of high genetic divergence than other less invasive species (Schrader *et al.* 2014). These 'TE islands' possess a non-random gene distribution; whereby genes associated with lineage divergence (e.g. those affecting body-size, pesticide resistance and production of chemical cues) are more likely to be affected by TE induced duplications and deletions (Schrader *et al.* 2014). Higher TE frequencies have also been reported in migratory *D. melanogaster* populations (González *et al.* 2008). More recently, translocation studies involving chickens (*Gallus gallus*) have suggested that substantial changes in TE expression may facilitate local adaption to different environments in a plastic manner (Liu *et al.* 2022).

The association between TE activity and hybrid sterility, adaptability and chromosomal rearrangements has also led to the suggestion that TEs may play an important role in macroevolutionary processes. The increased adaptability of TE abundant species may mean they are less likely to become extinct during times of environmental change or have enhanced evolutionary potential, as increased mutation rates allow populations to reach new 'adaptive peaks' (Oliver & Greene 2011). This 'TE-thrust' hypothesis remains subject to several theoretical expansions; incorporating ideas on population subdivision/genetic drift (Jurka *et al.* 2011) and the evolutionary 'tug of war' between TEs and epigenetic regulatory mechanisms (Zeh *et al.* 2009). These concepts have been further used to argue for a major update to evolutionary theory, particularly with regards to the significant role that intragenomic conflict may play during the speciation process (Crespi & Nosil 2013). Unlike more classical models of evolution (e.g ecological selection), TE-induced speciation and the associated dynamism between TEs and their respective silencers provides one explanation of how selection can remain perpetual in the light of environmental stasis. Alternatively, the periodic nature of habitat change provides a possible explanation to the 'punctuated

equilibrium' model of evolution; whereby sudden bursts of phenotypic change correlate with stress induced increases in TE activity (Zeh *et al.* 2009). Supporting these hypotheses are findings of positive relationships between TE content and diversification rates amongst *Piciforme* (Manthey *et al.* 2018), *Asteraceae* (Staton & Burke 2015) and *Myotis* bat lineages (Ray *et al.* 2008). Furthermore, a recent analysis across the entirety of the mammalian class found a significant relationship between TE activity levels and rates of speciation (Ricci *et al.* 2018).

The association between increased TE abundance and adaptability may be purely correlative; with increases in TE activity occurring because of a reduction in effective population sizes (i.e. as a result of diversification) (Stapley *et al.* 2015). One way the directionality of these effects could be resolved is to look for patterns of selective sweeps (Stapley *et al.* 2015). Selection acting upon existing genetic variation is known to produce 'soft' selective sweeps; characterised by differences in allele frequencies across many genomic sites, whereas selection acting upon novel genetic variation produce 'hard' selective sweeps resulting in reduced variation in linked loci only. Alternatively, establishing the exact timings of TE proliferation relative to lineage radiation could be an important step to determine if TE expansion occurred before or after diversification. Future efforts must untangle these 'chicken or egg' scenarios to determine whether TEs have a causative or correlative role during speciation and adaptation.

1.7 Bioinformatic approaches for TE detection

1.7.1 Sequence similarity (homology) based approaches

Numerous studies rely on accurate TE detection as a primary step for further analysis. Current methods by which genomic TEs are detected can be classed under two different approaches: homologous (sequence similarity) and de-novo (Bergman & Quesneville 2007). Homologous methods allow TEs to be detected through sequence similarity with an existing TE library, usually through a variation of an extending alignment algorithm (e.g Blast). The most common repository of eukaryotic TE sequence is RepBase (Bao *et al.* 2015), which contains over 44,000 entries from all TE

families (Goerner-Potvin & Bourque 2018). Other databases may be either taxa (e.g. FishTEDB) or TE type specific (e.g SINEBase) (Vassetzky & Kramerov 2013; Shao *et al.* 2018). A major advantage of using sequence similarity methods during TE detection is their ability to detect TEs even at low copy number. RepeatMasker is the most widely used sequence similarity TE detection software, which uses scoring matrices to annotate a given sequence against either a user-generated repeat library or an existing repository; commonly combining RepBase and the eukaryotic specific Dfam library (Wheeler *et al.* 2013; Bao *et al.* 2015). Limitations associated with sequence-similarity detection arise when conducting cross-species analysis. Many TE families are lineage specific, meaning comparisons of study genomes against a reference may bias TE identification against novel insertions (Bergman & Quesneville 2007). Homologous approaches may consequently be less applicable in cases where TEs are largely taxon-specific (Goerner-Potvin & Bourque 2018). Furthermore, a detection bias towards more active TEs may occur if sequence divergence is too large, or elements may be missed altogether (Bergman & Quesneville 2007). In these cases, it may be best to combine a sequence similarity approach with de-novo TE detection.

1.7.2 'De-novo' based approaches

De-novo-based approaches identify TEs using their fundamental characteristics, such as high copy-number or structural features such as LTR or RT domains, and subsequently cluster them based on their similarity (Bergman & Quesneville 2007). A widely used de-novo detection software is RepeatModeler; a sister pipeline to RepeatMasker which combines the multiple alignment clustering approach of RECON (Bao & Eddy 2002) and consensus seed clustering of RepeatScout (Price *et al.* 2005). More recent de-novo TE detection tools include EDTA, which combines multiple 'de-novo' bioinformatic packages to provide a comprehensive pipeline for detecting species-specific TE insertions (Ou *et al.* 2019). Used in conjunction with sequence-similarity methods, the use of de novo detection methods will typically lead to a greater number of TEs being detected. For example, in 2011 de-novo approaches in TE detection led to a 10% increase in estimated transposon

abundance within the human genome (de Koning *et al.* 2011). Whilst de-novo approaches are useful for detecting species-specific TEs, they may fail to detect TEs at low copy number (Bergman & Quesneville 2007). Furthermore, it may be difficult to distinguish TEs from other repetitive genetic components, such as microsatellites or simple repeats. This may be mitigated through looking at TE-specific hit patterns (e.g PILER software), which differ from other repetitive components in the large span length between individual units (Edgar & Myers 2005). Furthermore, de-novo approaches are associated with a diverse range of 'curational' issues, including the failure to capture an element's full length, fragmentation (where one contiguous element is represented by multiple fragmented consensus sequences) and redundancy (where a TE library contains an artificially inflated number of consensus sequences) (Baril *et al.* 2022; Goubert *et al.* 2022). The 'gold standard' for de-novo TE detection would therefore involve a subsequent manual curation step (see Box 1; (Platt *et al.* 2016), though this can be both time consuming and non-replicable due to a low degree of standardisation between different lab groups. There has therefore been a considerable recent effort to provide walkthroughs guides to aid users through the process of manual curation (e.g Goubert *et al.* 2022), and the development of de-novo pipelines that aim to automate such processes (e.g 'EarlGrey') (Baril *et al.* 2022).

Finally, structural characteristics of TEs can provide additional information during de-novo annotation. For example, the software LTRharvest allows for the de-novo detection of LTR retrotransposons through scanning the genome for pairs of similar LTR sequences and flanking TSD (Ellinghaus *et al.* 2008). TEs may also possess specific compositional biases that allow them to be distinguished from other regions of the genome. For example, compared to genes, TEs have a higher A/T nucleotide frequency at the third codon compared to the first (Lerat *et al.* 2002). These compositional characteristics have been used to generate hidden Markov models which can successfully detect and discriminate between class I and class II TEs (Andrieu *et al.* 2004). Nevertheless, such methods may be inaccurate for either elements of short sequence length or

those which are no longer active, particularly if the transposition process provides the selection constraints to generate such compositional biases.

Box 1. Summary of the key principles followed during manual curation of TE consensus libraries which have been generated 'de-novo'. User defined lengths/numbers are designated with an X.

Box 1: Manual curation of 'de-novo' TE consensus sequences

1. De-novo consensus sequences are clustered using 'cd-hit-est' to reduce redundancy. This may incorporate Wicker's (2007) 80-80-80 rule, whereby two elements are considered to belong to the same family if they share 80% sequence similarity, over 80% of the sequence length and if the sequence is longer than 80bp

2. A BLAST, Extract, Extend (BEE) process is then conducted to generate maximum length consensus sequences.

BLAST – top X hits are retrieved between TE consensus (query) and species genome (subject)

EXTRACT – Blast hit sequences are extracted from the genome

EXTEND – The flanking region of each extracted sequence are retrieved by adding X bp to both the 3' and 5' end. Multiple sequence alignment (e.g. MAFFT) is then applied to each extended sequence to obtain a new consensus sequence.

X repeat iterations

3. TE library redundancy is again removed as described in Step 1. Sequences can then be reclassified based on sequence homology to existing TE databases (e.g. RepBase).

1.8 Introduction to the Corydoradinae

1.8.1 Teleost genomes

Teleosts (an infraclass of ray-finned Actinopterygii) represent 30,000 described species: roughly half of all extant vertebrates, and 98% of all fish species (Ravi & Venkatesh 2018). This vast biodiversity is likely to be an outcome of multiple factors, but if often attributed to a lineage-specific whole genome duplication which occurred early in teleost evolution (Glasauer & Neuhauss 2014). Polyploids are hypothesised to have increased mutational robustness (Otto & Whitton 2000) and lineage-specific paralogue loss may also lead to speciation events (Scannell *et al.* 2006). A recent attempt to infer the timing of the teleost specific whole genome duplication, placed it in the late Triassic (235 Ma), thus predating teleost radiation (Davesne *et al.* 2021). This estimate is likely to be the most accurate yet because in addition to extant teleosts, the genome size of extinct species was also incorporated using fossil evidence (Davesne *et al.* 2021). Genomic analysis has shown that fish

species which emerged before the teleost specific duplication event have slower rates of protein and sequence evolution (Amemiya *et al.* 2013; Braasch *et al.* 2016). An additional lineage-specific whole genome duplication has also been inferred within Salmonids, though this occurred significantly earlier than their subsequent diversification, suggesting factors other than polyploidy may have played a more important role in their radiation (Macqueen & Johnston 2014). Whilst many paralogues have been lost from teleost lineages, those that have been selectively retained are likely to have played an important role in teleost evolution. A recent study has shown that within teleosts, pigmentation paralogues are more likely to have been retained than other gene families for example, potentially promoting the high diversity of teleost pigmentation cell types (Lorin *et al.* 2018). Reduced selection pressure on paralogues may also provide the conditions required for neofunctionalisation and evolutionary innovation (Ohno 1970). For example, in electric teleosts of the gymnotiform and mormyroid families the historic duplication of a sodium channel gene (*scn4aa*) allowed one paralogue to become ectopically expressed within the myogenic electric organ and function in the discharge of electrolytes (Arnegard *et al.* 2010).

The genomes of teleosts also contain an abundance of different TE types that may underpin their biological diversity. The reasons as to why teleosts contain such a diversity of transposons is not well understood. Whilst this has been linked to the teleost specific whole genome duplication, the number of TE superfamilies in the spotted gar *(Lepisosteus oculatus)* is equivalent to that of other teleosts, suggesting no link between the whole genome duplication and the diversity of teleost TEs (Chalopin & Volff 2017). Alternatively, teleost TE diversity may be driven by horizontally transferred transposons (HTT), with the majority (93.7% of vertebrates) occurring within ray-finned fish (Zhang *et al*. 2020). The abundance of teleost HTTs may be due to both (i) aquatic environments protecting TEs from UV exposure and (ii) parasite overexposure, which has previously been linked to the transfer of TEs (Metzger *et al.* 2018; Dunemann & Wasmuth 2019). However, untangling these environmental effects from the coupling of species phylogeny is difficult, and further investigation

incorporating aquatic and terrestrial species of multiple evolutionary origins is required (Zhang *et al*. 2020). The consequences of high TE abundance within teleosts are numerous. Bursts of activity amongst the Tc1 TE family have previously been shown to correlate with salmonid speciation for example (De Boer *et al.* 2007). Furthermore, TE abundance is demonstrated to positively correlate with teleost genome size and increase overall genomic GC content (Symonová & Suh 2019).

1.8.2 The Corydoradinae

The Neotropical Corydoradinae (Teleostei; Siluriformes; Callichthyidae) are a diverse subfamily of detritovore catfish found within the benthic freshwater environments of South America (Nijssen 1970; Britto & Lima 2003; Alexandrou *et al.* 2011). With a 150-170 described species and likely many more undescribed, the Corydoradinae genus Corydoras is the most species-rich genus of catfish (Britto & Lima 2003; Alexandrou *et al*. 2011). Corydoras individuals can be easily identified through their laterally compressed head, pair of short mental barbels, and striking colouration (Nijssen 1970). Females are reported to spawn numerous (10-20) egg clutches which are fertilised by multiple males in a promiscuous/random mating system, in which male success is not dependent on any intrinsic traits such as size or courtship frequency (Kohda *et al.* 2002). Unlike other callichthyid species, Corydoras species do not produce foam nests for their offspring, which may lead to higher predation rates, particularly by avian species such as herons and kingfishers (Huysentruyt *et al.* 2009; Alexandrou *et al.* 2011). Individuals therefore reside in large schools of mixed species, with their geographic distribution correlating with the convergence of shared colour patterns (Nijssen 1970; Alexandrou *et al.* 2011). This 'phenotypic sharing' occurs even between evolutionary distinct species, and is likely a form of Mullerian mimicry, whereby common colour patterns (e.g. stripes, spots, and blocks of colour) increase rates of predator learning (Alexandrou *et al.* 2011). In addition to such mimicry, Corydoras defend themselves from predators through the possession of (i) armoured scutes, which have recently been shown to be hyper-mineralised and most resistant to puncturing in regions overlaying vital organs (Sire & Huysseune 1996; Lowe *et al.* 2021), (ii) secretion

of toxic compounds from lockable spines (Wright 2009) and (iii) burst swimming - permitted through high rates of caudal growth (Huysentruyt *et al.* 2009).

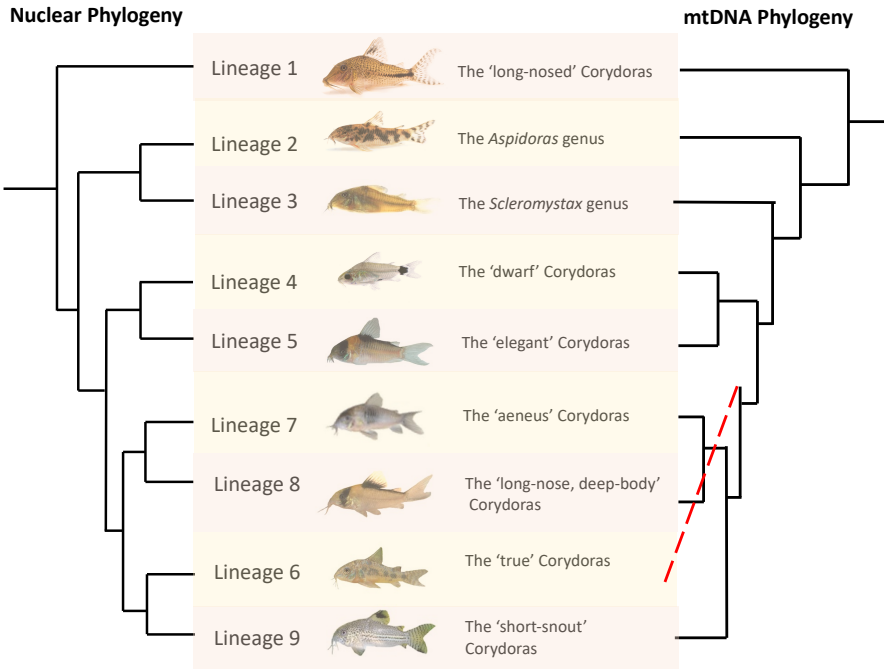


Figure 1.3. The evolutionary relationship between different Corydoradinae lineages. The discordance between the nuclear and mtDNA phylogenies is highlighted in red. Figure produced using data from Marburger *et al.* 2018.

A recent time-calibrated mtDNA phylogeny of the Corydoradinae subfamily, including representatives from all three genera (*Aspidoras, Scleromystax, Corydoras*), reported nine monophyletic lineages, with the most recent common ancestor of the Corydoradinae living 66mya (Marburger *et al.* 2018). This evolutionary relationship was also supported by nuclear RAD sequencing, with the only difference being a shift in the position of lineage 6 to be monophyletic with respect to lineage 9 (Marburger *et al.* 2018) (Figure 1.3). Analysis of haploid genome sizes revealed two significant lineage increases in C-values, with *C. aeneus* of lineage 9 possessing the largest known genome of any teleost species (4.4pg). Evidence of two whole genome duplications were also inferred from SNP ratio and haplotype number per contig, with the earliest occurring at the base of lineage 2-9 (54-66mya) and the most recent (20-30mya) associated with lineage 6 & 9 (Marburger *et al.* 2018). This whole genome duplication has recently been associated with the neo-

functionalisation of immune genes (toll-like receptors) and a reduction of parasite load within polyploid Corydoras species (Bell *et al.* 2020) and a subsequent increase in TE content (Marburger *et al.* 2018). Changes in TE content across the Corydoradinae lineage were found to be largely driven by the proliferation of a single Tc-Mariner DNA transposon element named TC1-IS639-Pogo, which increases from being found in 10% of reads in lineage 1-3 to 70% of reads in lineage 9 (Marburger *et al.* 2018).

1.9 Thesis aims and chapter outlines

The primary aim of this work is to better characterise the evolutionary causes and consequences of TE proliferation within the Corydoradinae. Both in-silico and 'multi-omic' approaches are adopted to explore these themes. Four specific research outcomes are outlined below:

1) How can the use of in-silico modelling assist in exploring the genomic causes of TE proliferation? Can processes widely cited as responsible for TE accumulation, including beneficial insertion effects and whole genome duplication lead to TE proliferation within simulated data sets? This is the research focus of chapter two and three.

2) How does the generation of a Corydoradinae-specific TE library impact downstream analysis? How may the reliance on the well curated zebrafish (*Danio rerio*) TE library during previous Corydoradinae annotation have biased TE-based inferences? This is the research focus of chapter four.

3) How have polyploidy and horizontal transfer impacted the putative TE proliferation during the evolutionary history of the Corydoradinae? Does the expressed TE content of multiple Corydoradinae species align with the previous inference of a TE expansion? How can polyploidy and horizontal transfer impact the genomic distribution of TE insertions and is there a positive relationship between Corydoradinae genome size and expressed TE content? Are there biases between the kind of TE families that are expressed within the Corydoradinae? This is the research focus of chapter five.

4) Are TE insertions particularly abundant within Corydoradinae pigmentation genes? How might this insertion bias have contributed to the rapid evolution of colour patterns? Do Corydoradinae pigmentation genes evolve under more relaxed selection pressure that average? This is the research focus of chapter six.

References

Abrusán, G. & Krambeck, H.-J. (2006). Competition may determine the diversity of transposable elements. *Theor. Popul. Biol.*, 70, 364–375.

Alexandrou, M.A., Oliveira, C., Maillard, M., Mcgill, R.A.R., Newton, J., Creer, S., *et al.* (2011). Competition and phylogeny determine community structure in Müllerian co-mimics. *Nature*, 469, 84–88.

Almeida, M.V., Vernaz, G., Putman, A.L.K. & Miska, E.A. (2022). Taming transposable elements in vertebrates: from epigenetic silencing to domestication. *Trends Genet.*, 38, 529–553.

Amemiya, C.T., Alfoldi, J., Lee, A.P., Fan, S., Philippe, H., MacCallum, I., *et al.* (2013). The African coelacanth genome provides insights into tetrapod evolution. *Nat.* , 496, 311–316.

Andrieu, O., Fiston, A.-S., Anxolabéhère, D. & Quesneville, H. (2004). Detection of transposable elements by their compositional bias. *BMC Bioinformatics*, 5, 94.

Arkhipova, I.R. (2006). Distribution and Phylogeny of Penelope-Like Elements in Eukaryotes. *Syst. Biol.*, 55, 875–885.

Arkhipova, I.R. (2017). Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. *Mob. DNA*, 8.

Arkhipova, I.R., Yushenova, I.A. & Angert, E. (2019). Giant Transposons in Eukaryotes: Is Bigger Better? *Genome Biol. Evol.*, 11, 906–918.

Arnegard, M.E., Zwickl, D.J., Lu, Y. & Zakon, H.H. (2010). Old gene duplication facilitates origin and diversification of an innovative communication system-twice. *PNAS*, 107, 22172–22177.

Bao, Z., and Eddy, S. (2002). Automated de novo identification of repeat sequence families in sequenced genomes. *Gen Res* 12, 1269-1276.

Bao, W., Kojima, K.K. & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA*, 6, 11.

Barbaglia, A.M., Klusman, K.M., Higgins, J., Shaw, J.R., Hannah, L.C. & Lal, S.K. (2012). Gene capture by Helitron transposons reshuffles the transcriptome of maize. *Genetics*, 190, 965–975.

Bargues, N. & Lerat, E. (2017). Evolutionary history of LTR-retrotransposons among 20 Drosophila species. *Mob. DNA*, 8.

Baril, T., Imrie, R.M. & Hayward, A. (2022). Earl Grey: a fully automated user-friendly transposable element annotation and analysis pipeline. *bioRxiv*, 2022.06.30.498289.

Belancio, V.P., Deininger, P.L. & Roy-Engel, A.M. (2009). LINE dancing in the human genome: Transposable elements and disease. *Genome Med.*, 1, 1–8.

Belancio, V.P., Hedges, D.J. & Deininger, P. (2008). Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health. *Genome Res.*, 18, 343–58.

Bell, E.A., Cable, J., Oliveira, C., Richardson, D.S., Yant, L. & Taylor, M.I. (2020). Help or hindrance? The evolutionary impact of whole-genome duplication on immunogenetic diversity and parasite load. *Ecol. Evol.*, 10, 13949–13956.

Benoit, M., Drost, H.-G., Catoni, M., Gouil, Q., Lopez-Gomollon, S., Baulcombe, D.C., *et al.* (2019). Environmental and epigenetic regulation of Rider retrotransposons in tomato. *bioRxiv*, 517508.

Bergman, C.M. & Quesneville, H. (2007). Discovering and detecting transposable elements in genome sequences. *Brief. Bioinform.*, 8, 382–392.

Bire, S., Casteret, S., Arnaoty, A., Piégu, B., Lecomte, T. & Bigot, Y. (2013). Transposase concentration controls transposition activity: Myth or reality? *Gene*, 530, 165–171.
De Boer, J.G., Yazawa, R., Davidson, W.S. & Koop, B.F. (2007). Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. *BMC Genomics*, 8.
Bourc'his, D. & Bestor, T.H. (2004). Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature*, 431, 96–99.
Braasch, I., Gehrke, A.R., Smith, J.J., Kawasaki, K., Manousaki, T., Pasquier, J., *et al.* (2016). The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nat. Genet.* , 48, 427–437.
Britto, M.R. & Lima, F.C.T. (2003). Corydoras tukano, a new species of corydoradine catfish from the rio Tiquié, upper rio Negro basin, Brazil (Ostariophysi: Siluriformes: Callichthyidae). *Neotrop. Ichthyol.*, 1, 83–91.
Brosh, O., Fabian, D.K., Cogni, R., Tolosana, I., Day, J.P., Olivieri, F., *et al.* (2022). A novel transposable element-mediated mechanism causes antiviral resistance in *Drosophila* through truncating the Veneno protein. *Proc. Natl. Acad. Sci.*, 119, e2122026119.
Capy, P. (2005). Diversity of Retrotransposable Elements Classification and nomenclature of retrotransposable elements. *Cytogenet Genome Res*, 110, 457–461.
Carnell, A.N. & Goodman, J.I. (2003). The Long (LINEs) and the Short (SINEs) of It: Altered Methylation as a Precursor to Toxicity. *Toxicol. Sci.*, 75, 229–235.
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.
Chalopin, D. & Volff, J.N. (2017). Analysis of the spotted gar genome suggests absence of causative link between ancestral genome duplication and transposable element diversification in teleost fish. *J. Exp. Zool. Part B Mol. Dev. Evol.*, 328, 629–637.
Chénais, B., Caruso, A., Hiard, S. & Casse, N. (2012). The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. *Gene*, 509.
Chernyavskaya, Y., Mudbhary, R., Zhang, C., Tokarz, D., Jacob, V., Gopinath, S., *et al.* (2017). Loss of DNA methylation in zebrafish embryos activates retrotransposons to trigger antiviral signaling. *Development*, 144.
Chung, H., Bogwitz, M.R., McCart, C., Andrianopoulos, A., Ffrench-Constant, R.H., Batterham, P., *et al.* (2007). Cis-regulatory elements in the Accord retrotransposon result in tissue-specific expression of the Drosophila melanogaster insecticide resistance gene Cyp6g1. *Genetics*, 175, 1071–1077.
Cowley, M. & Oakey, R.J. (2013). Transposable Elements Re-Wire and Fine-Tune the Transcriptome. *PLoS Genet.*, 9, e1003234.
Crespi, B. & Nosil, P. (2013). Conflictual speciation: species formation via genomic conflict. *Trends Ecol. Evol.*, 28, 48–57.
Curcio, M.J. & Derbyshire, K.M. (2003). The Outs and Ins of Transposition: From Mu to Kangaroo. *Nat. Rev. Mol. Cell Biol.*, 4, 865–877.
Davesne, D., Friedman, M., Schmitt, A.D., Fernandez, V., Carnevale, G., Ahlberg, P.E., *et al.* (2021). Fossilized cell structures identify an ancient origin for the teleost whole-genome duplication. *Proc. Natl. Acad. Sci. U. S. A.*, 118, e2101780118.
Deininger, P. (2011). Alu elements: know the SINEs. *Genome Biol.*, 12.
Deniz, Ö., Frost, J.M. & Branco, M.R. (2019). Regulation of transposable elements by DNA modifications. *Nat. Rev. Genet.*
Dietrich, F.S., Voegeli, S., Kuo, S. & Philippsen, P. (2013). Genomes of ashbya fungi isolated from insects reveal four mating-type loci, numerous translocations, lack of transposons, and distinct gene duplications. *G3 Genes, Genomes, Genet.*, 3, 1225–1239.
Dion-Côté, A.-M., Renaut, S., Normandeau, E. & Bernatchez, L. (2014). RNA-seq Reveals

Transcriptomic Shock Involving Transposable Elements Reactivation in Hybrids of Young Lake Whitefish Species. *Mol. Biol. Evol.*, 31, 1188–1199.
Dulai, K.S., von Dornum, M., Mollon, J.D. & Hunt, D.M. (1999). The evolution of trichromatic color vision by opsin gene duplication in New World and Old World primates. *Genome Res.*, 9, 629–638.
Dunemann, S.M. & Wasmuth, J.D. (2019). Horizontal transfer of a retrotransposon between parasitic nematodes and the common shrew. *Mob. DNA*, 10, 1–13.
Edgar, R.C. & Myers, E.W. (2005). PILER: identification and classification of genomic repeats. *Bioinformatics*, 21, 152–158.
Eickbush, T.H. & Jamburuthugoda, V.K. (2008). The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus Res*, 134, 221–234.
Ellinghaus, D., Kurtz, S. & Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, 9, 18.
Elliott, T.A. & Gregory, T.R. (2015). Do larger genomes contain more diverse transposable elements? *BMC Evol. Biol.*, 15.
Estep, M., DeBarry, J. & Bennetzen, J. (2013). The dynamics of LTR retrotransposon accumulation across 25 million years of panicoid grass evolution. *Heredity (Edinb).*, 110, 194–204.
Feschotte, C. (2008). The contribution of transposable elements to the evolution of regulatory networks Life after death: TE exaptation. *Nat. Rev. Genet.*, 9, 397–405.
Feschotte, C. & Pritham, E.J. (2007). DNA Transposons and the Evolution of Eukaryotic Genomes. *Annu. Rev. Genet.*, 41, 331–368.
Feschotte, C. & Wessler, S.R. (2001). Treasures in the attic: rolling circle transposons discovered in eukaryotic genomes. *Proc. Natl. Acad. Sci. U. S. A.*, 98, 8923–4.
Finnegan, D.J. (1989). Eukaryotic transposable elements and genome evolution. *Trends Genet.*, 5, 103–107.
Gifford, R. & Tristem, M. (2003). The Evolution, Distribution and Diversity of Endogenous Retroviruses. *Virus Genes*, 26, 291–315.
Gilbert, C., Schaack, S., Ii, J.K.P., Brindley, P.J. & Feschotte, C. (2010). A role for host-parasite interactions in the horizontal transfer of DNA transposons across animal phyla. *Nature*, 464, 1347–1350.
Gladyshev, E.A. & Arkhipova, I.R. (2011). A widespread class of reverse transcriptase-related cellular genes. *Proc. Natl. Acad. Sci. U. S. A.*, 108, 20311–20316.
Glasauer, S.M.K. & Neuhauss, S.C.F. (2014). Whole-genome duplication in teleost fishes and its evolutionary consequences. *Mol. Genet. genomics*, 289, 1045–1060.
Goerner-Potvin, P. & Bourque, G. (2018). Computational tools to unmask transposable elements. *Nat. Rev. Genet.*, 19, 688–704.
González, J., Lenkov, K., Lipatov, M., Macpherson, J.M. & Petrov, D.A. (2008). High Rate of Recent Transposable Element–Induced Adaptation in Drosophila melanogaster. *PLoS Biol.*, 6, e251.
González, J. & Petrov, D.A. (2012). Evolution of Genome Content: Population Dynamics of Transposable Elements in Flies and Humans. In: *Evolutionary Genomics: Statistical and Computations Methods, Volume 1*. Humana Press, Totowa, NJ, pp. 361–383.
Goodwin, T.J.D., Butler, M.I. & Poulter, R.T.M. (2003). Cryptons: a group of tyrosine-recombinase-encoding DNA transposons from pathogenic fungi. *Microbiology*, 149, 3099–3109.
Goodwin, T.J.D. & Poulter, R.T.M. (2004). A New Group of Tyrosine Recombinase-Encoding Retrotransposons. *Mol. Biol. Evol.*, 21, 746–759.
Goubert, C., Craig, R.J., Bilat, A.F., Peona, V., Vogan, A.A. & Protasio, A. V. (2022). A beginner's guide to manual curation of transposable elements. *Mob. DNA*, 13, 1–19.
Grandi, N. & Tramontano, E. (2018). Human Endogenous Retroviruses Are Ancient Acquired Elements Still Shaping Innate Immune Responses. *Front. Immunol.*, 10.
Han, J.S. (2010). Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions. *Mob. DNA*, 1.

Hawkins, J.S., Kim, H., Nason, J.D., Wing, R.A. & Wendel, J.F. (2006). Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. *Genome Res.*, 16, 1252–61.

Hayward, A. (2017). Origin of the retroviruses: when, where, and how? *Curr. Opin. Virol.*, 25, 23–27.

Huysentruyt, F., Moerkerke, B., Devaere, S. & Adriaens, D. (2009). Early development and allometric growth in the armoured catfish Corydoras aeneus (Gill, 1858). *Hydrobiologia*, 627, 45–54.

Inoue, Y., Saga, T., Aikawa, T., Kumagai, M., Shimada, A., Kawaguchi, Y., *et al.* (2017). Complete fusion of a transposon and herpesvirus created the Teratorn mobile element in medaka fish. *Nat. Commun.*, 8, 1–14.

Johnson, R., Gamblin, R.J., Ooi, L., Bruce, A.W., Donaldson, I.J., Westhead, D.R., *et al.* (2006). Identification of the REST regulon reveals extensive transposable element-mediated binding site duplication. *Nucleic Acids Res.*, 34, 3862–3877.

Jordan, I.K., Rogozin, I.B., Glazko, G. V & Koonin, E. V. (2003). Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet.*, 19, 68–72.

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

Kapitonov, V. V & Jurka, J. (2006). Self-synthesizing DNA transposons in eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.*, 103, 4540–4545.

Kazazian, H.H. (2004). Mobile elements: drivers of genome evolution. *Science (80-.).*, 303, 1626–32.

Kenan-Eichler, M., Leshkowitz, D., Tal, L., Noor, E., Melamed-Bessudo, C., Feldman, M., *et al.* (2011). Wheat hybridization and polyploidization results in deregulation of small RNAs. *Genetics*, 188, 263–72.

Kim, E.B., Fang, X., Fushan, A.A., Huang, Z., Lobanov, A. V., Han, L., *et al.* (2011). Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nat.* , 479, 223–227.

Kohda, M., Yonebayashi, K., Nakamura, M., Ohnishi, N., Seki, S., Takahashi, D., *et al.* (2002). *Male reproductive success in a promiscuous armoured catfish Corydoras aeneus (Callichthyidae)*. *Environ. Biol. Fishes*.

Kojima, K.K. & Jurka, J. (2011). Crypton transposons: identification of new diverse families and ancient domestication events. *Mob. DNA*, 2.

de Koning, A.P.J., Gu, W., Castoe, T.A., Batzer, M.A. & Pollock, D.D. (2011). Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLoS Genet.*, 7, e1002384.

Kramerov, D.A. & Vassetzky, N.S. (2011). SINEs. *Wiley Interdiscip. Rev. RNA*, 2, 772–786.

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

Lander, E. & IHGSC. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409, 860–921.

Lee, J., Han, K., Meyer, T.J., Kim, H.-S. & Batzer, M.A. (2008). Chromosomal Inversions between Human and Chimpanzee Lineages Caused by Retrotransposons. *PLoS One*, 3, e4047.

Lerat, E., Capy, P. & Biémont, C. (2002). Codon Usage by Transposable Elements and Their Host Genes in Five Species. *J. Mol. Evol.*, 54, 625–637.

Liu, Y.-N., Chen, R.-M., Pu, Q.-T., Nneji, L.M. & Sun, Y.-B. (2022). Expression Plasticity of Transposable Elements Is Highly Associated with Organismal Re-adaptation to Ancestral Environments. *Genome Biol. Evol.*, 14.

Lorin, T., Brunet, F.G., Laudet, V. & Volff, J.N. (2018). Teleost Fish-Specific Preferential Retention of Pigmentation Gene-Containing Families After Whole Genome Duplications in Vertebrates. *G3 Genes|Genomes|Genetics*, 8, 1795–1806.

Lowe, A., Summers, A.P., Walter, R.P., Walker, S. & Misty Paig-Tran, E.W. (2021). Scale performance and composition in a small Amazonian armored catfish, Corydoras trilineatus. *Acta Biomater.*, 121, 359–370.

Macqueen, D.J. & Johnston, I.A. (2014). A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proc. Biol.*

Sci., 281, 20132881.
Magwire, M.M., Bayer, F., Webster, C.L., Cao, C. & Jiggins, F.M. (2011). Successive increases in the resistance of Drosophila to viral infection through a transposon insertion followed by a Duplication. *PLoS Genet.*, 7, e1002337.
Malicki, M., Iliopoulou, M. & Hammann, C. (2017). Retrotransposon Domestication and Control in Dictyostelium discoideum. *Front. Microbiol.*, 8.
Manthey, J.D., Moyle, R.G. & Ephane Boissinot, S. (2018). Multiple and Independent Phases of Transposable Element Amplification in the Genomes of Piciformes (Woodpeckers and Allies). *Genome Biol. Evol.*, 10, 1445–1456.
Marburger, S., Alexandrou, M.A., Taggart, J.B., Creer, S., Carvalho, G., Oliveira, C., *et al.* (2018). Whole genome duplication and transposable element proliferation drive genome expansion in Corydoradinae catfishes. *Proc. R. Soc. B*, 285.
Matzke, M.A. & Mosher, R.A. (2014). RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat. Rev. Genet.*, 15, 394–408.
McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci.*, 36, 344–355.
Metzger, M.J., Paynter, A.N., Siddall, M.E. & Goff, S.P. (2018). Horizontal transfer of retrotransposons between bivalves and other aquatic species of multiple phyla. *Proc. Natl. Acad. Sci. U. S. A.*, 115, E4227–E4235.
Migicovsky, Z. & Kovalchuk, I. (2014). Transgenerational changes in plant physiology and in transposon expression in response to UV-C stress in Arabidopsis thaliana Transgenerational changes in plant physiology and in transposon expression in response to UV-C stress in Arabidopsis thaliana. *Plant Signal. Behav.*, 9.
Muyle, A., Seymour, D., Darzentas, N., Primetis, E., Gaut, B.S. & Bousios, A. (2021). Gene capture by transposable elements leads to epigenetic conflict in maize. *Mol. Plant*, 14, 237–252.
Neumann, P., Novák, P., Hoštáková, N. & Macas, J. (2019). Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mob. DNA*, 10.
Nijssen, H. (1970). Revision of the Surinam catfishes of the genus Corydoras (Pisces; Siluriformes; Callichthyidae). *Beaufortia*, 18, 1–75.
Notwell, J.H., Chung, T., Heavner, W. & Bejerano, G. (2015). A family of transposable elements co-opted into developmental enhancers in the mouse neocortex. *Nat. Commun.* , 6, 1–7.
O'Neill, R.J.W., O'Neill, M.J. & Graves, J.A.M. (1998). Erratum: Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. *Nature*, 393, 68–72.
Ohno, S. (1970). *Evolution by Gene Duplication*. *Evol. by Gene Duplic.* Springer Berlin Heidelberg.
Oliver, K.R. & Greene, W.K. (2011). Mobile DNA and the TE-Thrust hypothesis: supporting evidence from the primates. *Mob. DNA*, 2.
Otto, S.P. & Whitton, J. (2000). Polyploid incidence and evolution. *Annu. Rev. Genet.*, 34, 401–437.
Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J., *et al.* (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.*, 20, 1–18.
Pantzartzi, C.N., Pergner, J. & Kozmik, Z. (2018). The role of transposable elements in functional evolution of amphioxus genome: the case of opsin gene family. *Sci. Rep.*, 8, 2506.
Pecinka, A., Dinh, H.Q., Baubec, T., Rosa, M., Lettner, N. & Scheid, O.M. (2010). Epigenetic Regulation of Repetitive Elements Is Attenuated by Prolonged Heat Stress in Arabidopsis W OA. *Plant Cell*, 22, 3118–3129.
Piégu, B., Bire, S., Arensburger, P. & Bigot, Y. (2015). A survey of transposable element classification systems: A call for a fundamental update to meet the challenge of their diversity and complexity. *Mol. Phylogenet. Evol.*, 86, 90–109.
Platt, R.N., Blanco-Berdugo, L., Ray, D.A. & Ray, D.A. (2016). Accurate Transposable Element

Annotation Is Vital When Analyzing New Genome Assemblies. *Genome Biol. Evol.*, 8, 403–10.
Poulter, R.T.M. & Goodwin, T.J.D. (2005). DIRS-1 and the other tyrosine recombinase retrotransposons. *Cytogenet. Genome Res.*, 110, 575–588.
Price, A.L., Jones, N.C. & Pevzner, P.A. (2005). De novo identification of repeat families in large genomes. *Bioinformatics*, 21, 351–358.
Pritham, E.J., Putliwala, T. & Feschotte, C. (2007). Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene*, 390, 3–17.
Pyatkov, K.I., Arkhipova, I.R., Malkova, N. V, Finnegan, D.J. & Evgen'ev, M.B. (2004). Reverse transcriptase and endonuclease activities encoded by Penelope-like retroelements. *Proc. Natl. Acad. Sci. U. S. A.*, 101, 14719–14724.
Ravi, V. & Venkatesh, B. (2018). The Divergent Genomes of Teleosts. *Annu. Rev. Anim. Biosci.*, 6, 47–68.
Ray, D.A., Feschotte, C., Pagan, H.J.T., Smith, J.D., Pritham, E.J., Arensburger, P., *et al.* (2008). Multiple waves of recent DNA transposon activity in the bat, Myotis lucifugus. *Genome Res.*, 18, 717–728.
Ricci, M., Peona, V., Guichard, E., Taccioli, C. & Boattini, A. (2018). Transposable Elements Activity is Positively Related to Rate of Speciation in Mammals. *J. Mol. Evol.*, 86, 303–310.
Rigal, M. & Mathieu, O. (2011). A "mille-feuille" of silencing: Epigenetic control of transposable elements. *Biochim. Biophys. Acta*, 1809, 452–458.
Rogers, R.L., Zhou, L., Chu, C., Márquez, R., Corl, A., Linderoth, T., *et al.* (2018). Genomic takeover by transposable elements in the Strawberry poison frog. *Mol. Biol. Evol.*, 35, 2913–2927.
Sayah, D.M., Sokolskaja, E., Berthoux, L. & Luban, J. (2004). Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. *Nature*, 430, 569–573.
Scannell, D.R., Byrne, K.P., Gordon, J.L., Wong, S. & Wolfe, K.H. (2006). Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature*, 440, 341–345.
Schaack, S., Gilbert, M. & Dric Feschotte, C. (2010). Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends Ecol. Evol.*, 25, 537–546.
Schrader, L., Kim, J.W., Ence, D., Zimin, A., Klein, A., Wyschetzki, K., *et al.* (2014). Transposable element islands facilitate adaptation to novel environments in an invasive species. *Nat. Commun.*, 5, 5495.
Seberg, O. & Petersen, G. (2009). A unified classification system for eukaryotic transposable elements should reflect their phylogeny. *Nat. Rev. Genet.*, 10, 276–276.
Shao, F., Wang, J., Xu, H. & Peng, Z. (2018). FishTEDB: a collective database of transposable elements identified in the complete genomes of fish. *Database*, 2018, 1–9.
Simon, M., Van Meter, M., Ablaeva, J., Ke, Z., Gonzalez, R.S., Taguchi, T., *et al.* (2019). LINE1 Derepression in Aged Wild-Type and SIRT6-Deficient Mice Drives Inflammation. *Cell Metab.*, 29, 871-885.e5.
Sire, J.-Y. & Huysseune, A. (1996). Structure and Development of the Odontodes in an Armoured Catfish, *Corydoras aeneus* (Siluriformes, Callichthyidae). *Acta Zool.*, 77, 51–72.
Slotkin, R.K. & Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.*, 8, 272–285.
Sotero-Caio, C.G., Platt, R.N., Suh, A., Ray, D.A. & Ray, D.A. (2017). Evolution and Diversity of Transposable Elements in Vertebrate Genomes. *Genome Biol. Evol.*, 9, 161–177.

Specchia, V., Janzen, S., Marini, G. & Pinna, M. (2017). The Potential Link between Mobile DNA and the Invasiveness of the Species. *J. RNAi Gene Silenc. |*, 13, 557–561.
Stapley, J., Santure, A.W. & Dennis, S.R. (2015). Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species. *Mol. Ecol.*, 24, 2241–2252.
Staton, S.E. & Burke, J.M. (2015). Evolutionary transitions in the Asteraceae coincide with marked shifts in transposable element abundance. *BMC Genomics*, 16, 623.

Stitzer, M.C., Anderson, S.N., Springer, N.M. & RossIbarra, J. (2021). The genomic ecosystem of transposable elements in maize. *PLOS Genet.*, 17, e1009768.
Sun, W., Shen, Y.-H., Han, M.-J., Cao, Y.-F. & Zhang, Z. (2014). An Adaptive Transposable Element Insertion in the Regulatory Region of the EO Gene in the Domesticated Silkworm, Bombyx mori. *Mol. Biol. Evol.*, 31, 3302–3313.
Symonová, R. & Suh, A. (2019). Nucleotide composition of transposable elements likely contributes to AT/GC compositional homogeneity of teleost fish genomes. *Mob. DNA*, 10.
Thomas, J. & Pritham, E.J. (2015). Helitrons, the Eukaryotic Rolling-circle Transposable Elements. *Microbiol. Spectr.*, 3.
Thornburg, B.G., Gotea, V. & Makayowski, W. (2006). Transposable elements as a significant source of transcription regulating signals. *Gene*, 365, 104–110.
Tsukahara, S., Kobayashi, A., Kawabe, A., Mathieu, O., Miura, A. & Kakutani, T. (2009). Bursts of retrotransposition reproduced in Arabidopsis. *Nature*, 461, 423–426.
Ungerer, M.C., Strakosh, S.C. & Stimpson, K.M. (2009). Proliferation of Ty3/gypsy-like retrotransposons in hybrid sunflower taxa inferred from phylogenetic data. *BMC Biol.*, 7.
Van't Hof, A.E., Campagne, P., Rigden, D.J., Yung, C.J., Lingley, J., Quail, M.A., *et al.* (2016). The industrial melanism mutation in British peppered moths is a transposable element. *Nature*, 534, 102–105.
Vassetzky, N.S. & Kramerov, D.A. (2013). SINEBase: a database and tool for SINE analysis. *Nucleic Acids Res.*, 41, 83–89.
Vigdal, T.J., Kaufman, C.D., Izsvá, Z., Voytas, D.F. & Ivics, Z. (2002). Common Physical Properties of DNA Affecting Target Site Selection of Sleeping Beauty and other Tc1/mariner Transposable Elements. *J. Mol. Evol.*, 323, 441–452.
Wei, K., Aldaimalani, R., Mai, D., Zinshteyn, D., PRV, S., Blumenstiel, J.P., *et al.* (2022). Rethinking the "gypsy" retrotransposon: A roadmap for community-driven reconsideration of problematic gene names. *OSFpreprints*, 10.31219/osf.io/fma….
Wheeler, T.J., Clements, J., Eddy, S.R., Hubley, R., Jones, T.A., Jurka, J., *et al.* (2013). Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res.*, 41, D70.
Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., *et al.* (2007). A Unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.*, 8, 973–982.
Wood, J.G., Jones, B.C., Jiang, N., Chang, C., Hosier, S., Wickremesinghe, P., *et al.* (2016). Chromatin-modifying genetic interventions suppress age-associated transposable element activation and extend life span in Drosophila. *Proc. Natl. Acad. Sci. U. S. A.*, 113, 11277–11282.
Woodard, L.E., Downes, L.M., Lee, Y.-C., Kaja, A., Terefe, E.S. & Wilson, M.H. (2017). Temporal self-regulation of transposition through host-independent transposase rodlet formation. *Nucleic Acids Res.*, 45, 353–366.
Wright, J.J. (2009). Diversity, phylogenetic distribution, and origins of venomous catfishes. *BMC Evol. Biol.*, 9, 282.
Yuan, Y.-W. & Wessler, S.R. (2011). The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proc. Natl. Acad. Sci. U. S. A.*, 108, 7884–9.
Zeh, D.W., Zeh, J.A. & Ishida, Y. (2009). Transposable elements and an epigenetic basis for punctuated equilibria. *BioEssays*, 31, 715–726.
Zhang, H.H., Peccoud, J., Xu, M.R.X., Zhang, X.G. & Gilbert, C. (2020). Horizontal transfer and evolution of transposable elements in vertebrates. *Nat. Commun.*, 11, 1–10.

2 Removal of beneficial insertion effects prevent the long-term persistence of transposable elements within simulated asexual populations

2.1 Chapter Overview and Author Contributions

This chapter modified an in-silico model of TE dynamics developed by Kremer et al (2020) to better understand the evolutionary forces which underpin TE proliferation. We disagreed with the author's original interpretation of their model, which they suggested provided evidence that 'TE engineering' processes may permit their proliferation. In a series of model expansions, we argue that such TE proliferation events may be better explained by the impact of beneficial insertions effects. This work was published in:

Butler CL, Bell EA, Taylor MI. Removal of beneficial insertion effects prevent the long-term persistence of transposable elements within simulated asexual populations. BMC Genomics. 2021 Apr 7;22(1):241.doi: 10.1186/s12864-021-07569-3.

The work from this chapter was conceived by Christopher L Butler, Ellen Bell and Martin Taylor. Analysis was conducted by Christopher L Butler. The final manuscript was written by Christopher L Butler with contribution from Ellen Bell and Martin Taylor. The original model was written by Kremer et al (2020) and is publicly available at https://github.com/stefan-c-kremer/TE_World2.

2.2 Abstract

Background: Transposable elements are significant components of most organism's genomes, yet the reasons why their abundances vary significantly among species is poorly understood. A recent study has suggested that even in the absence of traditional molecular evolutionary explanations, transposon proliferation may occur through a process known as `transposon engineering'. However, their model used a beneficial transposon insertion frequency of 20%, which we believe to be unrealistically high.

Results: Reducing this beneficial insertion frequency, while keeping all other parameters identical, prevented transposon proliferation.

Conclusions: We conclude that the author's original findings are better explained through the action of positive selection rather than 'transposon engineering', with beneficial insertion effects remaining important during transposon proliferation events.

2.3 Background

Transposable elements (TEs) are short regions of non-coding DNA (100-10,000 bp) which can proliferate throughout a genome and are significant genomic components of a taxonomically diverse range of species (Chénais *et al.* 2012; Chalopin *et al.* 2015; Sotero-Caio *et al.* 2017). Understanding the processes which drive TE variation across different species is an important goal in answering the so-called `C-value' paradox (the observed lack of relationship between genome size and organismal complexity) (Elliott & Gregory 2015). TEs are thought to accumulate within a population through a number of evolutionary mechanisms which include (i) positive selection, (ii) genetic drift, (iii) co-evolution with the host, (iv) sexual recombination or (v) horizontal transfer. Kremer *et al* 2020 investigated the fate of TE populations where none of these population genetic scenarios were possible (Kremer *et al.* 2020). Using in-silico modelling, the authors established an asexual population where (i) TE insertions had serious negative effects on host fitness, (ii) TEs could not

evolve insertion site preferences (i.e. no co-evolution with the host) and (iii) TEs were not able to be horizontally transferred. Surprisingly, even in the absence of these evolutionary forces, TEs accumulated in a limited number (3%) of scenarios. The authors concluded that these rare accumulation events may be explained through `TE engineering'; a process in which the activity of TEs significantly alters the landscape of a genome to facilitate further proliferation. Specifically, they suggest that the cycle of TE proliferation and degradation may provide new non-coding regions in which future TEs can insert with little or no consequence on host fitness. Changes in TE abundance which occur through their interactions with either the host genome or other transposons comprise a poorly studied field known as `TE ecology' (Brookfield 2005; Venner *et al.* 2009). Consequently, Kremer et al (2020) appears to have identified a novel mechanism for TE proliferation, with important implications regarding our understanding of TE dynamics.

Here, we highlight some potential issues which question the key findings of Kremer *et al's* study. They claim that their simulations model TE insertions which do not have any beneficial effects on host fitness; crucially ruling out positive selection as an explanation for TE accumulation. However, there was a lack of clarity on the precise meaning of `no beneficial effects', with three explanations on the effects of TE insertions given in their paper. These were: (i) `TEs had a net deleterious effect on host fitness', (ii) TEs had `serious negative effects on host fitness' or, (iii) `violated the assumption that TE insertions are beneficial'. A model in which TE activity had a net deleterious effect could mean that only a very small majority of TE insertions reduce host fitness. Such a model would violate the author's own assumption that TE insertions cannot be beneficial, with fitness increases still occurring during a significant number of insertions. Greater clarity on this issue would have been beneficial in order to help validate the legitimacy of Kremer *et al's* conclusions. The final model used in Kremer et al (2020) included a fixed parameter which simulates a mildly beneficial fitness effect (Insertion effect) during 20% of all TE insertions. Crucially, setting an insertion benefit at this level is not consistent with the author's own conclusion that positive selection is not responsible for driving

the TE accumulation events observed. Theoretically, the level of beneficial TE insertions may not have to be very high for their gradual accumulation. The fact that original TE copies are frequently retained in the genome (i.e. progeny distribution > 1) can provide a buffer to their abundance, even if the majority of new insertions are deleterious. This has been demonstrated in other simple TE dynamic models, whereby increasing the adaptive insertion probability to 0.05% is sufficient to permit TE domestication through positive selection (Le Rouzic *et al.* 2007).

In this study, we repeated Kremer *et al's* (2020) simulations, but explicitly defined 'no beneficial insertion effect' to mean there was no scenario in which TE insertions could generate an increase in host fitness, thus definitely ruling out positive selection as a potential mechanism for TE proliferation.

2.4 Methods

To test whether removing the positive TE insertion effect altered their ability to accumulate, we reanalysed the six parameter scenarios from Kremer et al (2020) where in the majority of cases TEs persisted for the 1,500 generation cut off set by the authors (Table 2.1). We did not change the percentage of TE insertions which led to lethal deleterious (20%) or mildly deleterious (30%) fitness effects. Instead we increased the probability that a TE insertion was neutral (i.e generating no change in host fitness) from 30% to 50%. All other model parameters remained identical. Following Kremer et al, we then repeated each simulation three times independently, which were plotted using the `ggplot2' package in R v3.5.1 (Wickham 2016).

Table 2.1 The parameter scenarios where Kremer et al (2020) reported TE accumulation in the majority of cases

TE Properties				Host Properties				Parameter Scenario
TE Progeny	TE Excision Rate	TE Death Rate	Insertion Bias	Corrected Mutation Rate	Non-Coding DNA	Mutation Effect	Carrying Capacity	
High	Low	Low	Low	Low	High	Low	High	HLLL - LHLH
Low	Low	High	High	Low	High	High	High	LLHH - LHHH
Low	Low	Low	High	Low	High	High	Low	LLLH - LHHL
Low	Low	Low	High	Low	Low	High	High	LLLH - LLHH
Low	Low	Low	Low	Low	High	High	High	LLLL - LHHH
Low	Low	Low	Low	Low	Low	High	Low	LLLL - LLHL

2.5 Results

In our reanalysis, we found that across each of the six parameter scenarios, TEs were better able to accumulate when the model included a beneficial TE insertion effect (Figure 2.1). Indeed, in the 18 iterations we ran, TEs only accumulated in a single instance when the beneficial insertion effects were removed (Run 3 of LLLH-LHHL). We also investigated the frequency of beneficial TE insertions that would be required for TE accumulation. To do this, we chose a parameter scenario in which Kremer et al found TEs to accumulate in every run (namely LLLH-LLHH). When the beneficial TE insertion effect was reduced to 0 (from 20%), the TE population did not proliferate, becoming extinct in under 400 generations in every iteration. We also ran simulations where TE insertions increased host fitness 1%, 5%, 10% and 15% of the time. While this led to incremental increases in TE accumulation, none of the simulations lasted 1,500 generations (Figure 2.2). We therefore conclude that under this TE dynamic model, a significant positive insertion effect is required for TE accumulation in almost all cases.

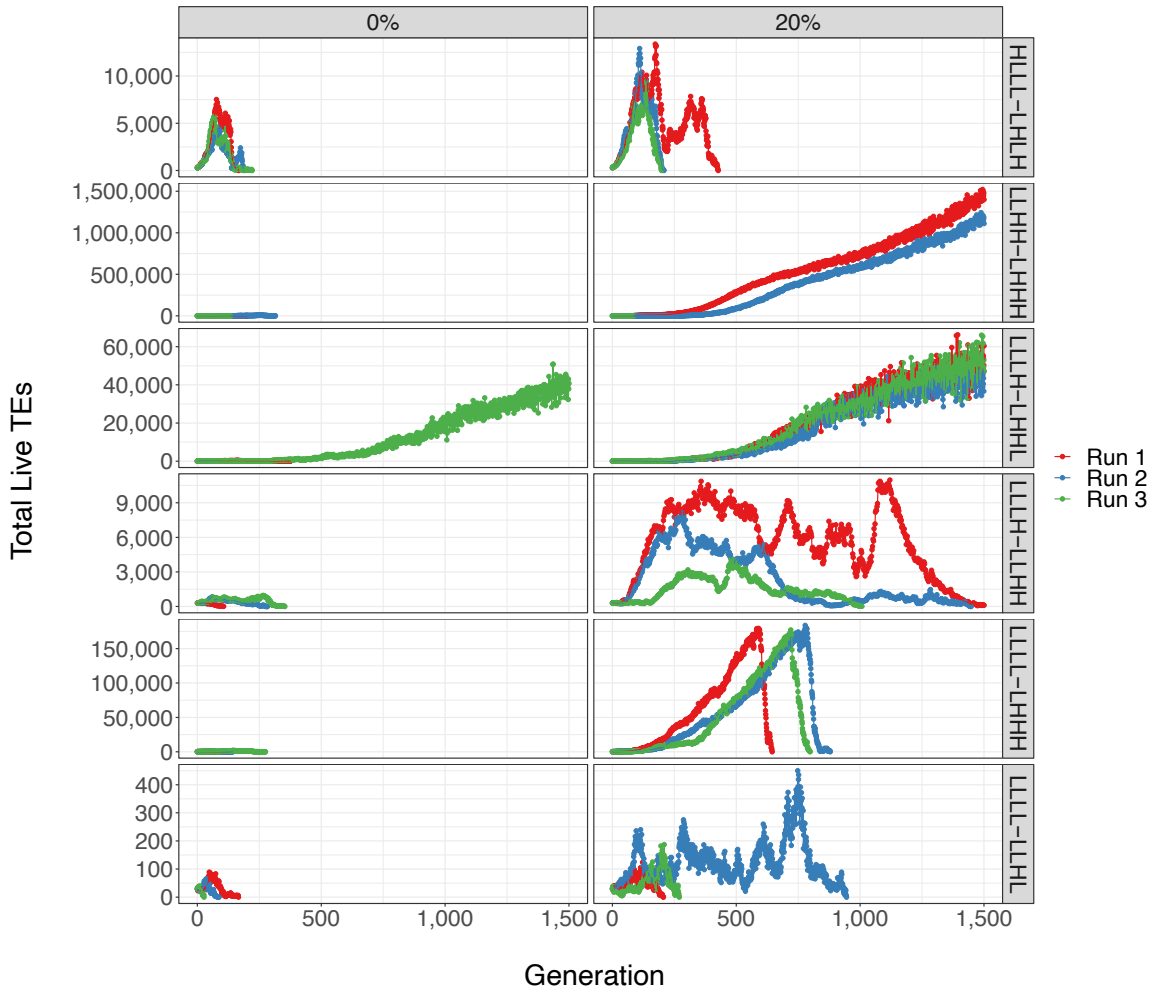


Figure 2.1 TE population dynamics for each of the six parameters where Kremer et al reported TE accumulation in the majority of cases. Beneficial TE insertions were set at either 0% or 20%. When beneficial insertion effects were excluded from the model,

2.6 Discussion

Overall, our findings suggest that Kremer *et al's* (2020) key conclusion of TE accumulation in a significant number of cases can largely be explained by the high beneficial insertion frequency used by the authors. When we removed this effect, TEs did not persist in the overwhelming majority of scenarios. We therefore suggest that the author's original finding of TE accumulation would have been better explained by the action of positive selection instead of a `TE engineering' process. The 20% beneficial insertion effect set by the authors is likely to be a significant overestimate compared to what may be observed in reality. Whilst estimating the frequency of beneficial insertion effects

remains difficult, a recent genome-wide scan of 14,384 human TE polymorphisms concluded that just 1.13% (163) were under positive selection (Rishishwar *et al.* 2018). The true frequency of beneficial TE insertions is likely to be even lower, as many highly deleterious or neutral TEs will lie undetected as they are removed from the genome. Interestingly, we did identify a single combination of parameters in which TEs did accumulate without exhibiting any beneficial fitness effect; namely when TE progeny and excision rate is low, TEs display high insertion bias, and the degree of non-coding regions in the genome are high. This may provide an interesting avenue for understanding genomic characteristics where TEs may accumulate in the absence of positive selection. Many TEs display both an insertion site bias (Sultana *et al.* 2017; Bourque *et al.* 2018) and variable TE activity rates (a product of both TE excision and progeny rates) within the lifetime of a cell (Kim *et al.* 2016). Finally, we wish to emphasise that we are not suggesting that TE ecology explanations should be ruled out when trying to understand the reasons for TE accumulation. On the contrary, when exploring potential hypotheses for TE proliferation it is important to realise that both evolutionary and ecology processes are likely to occur concurrently.

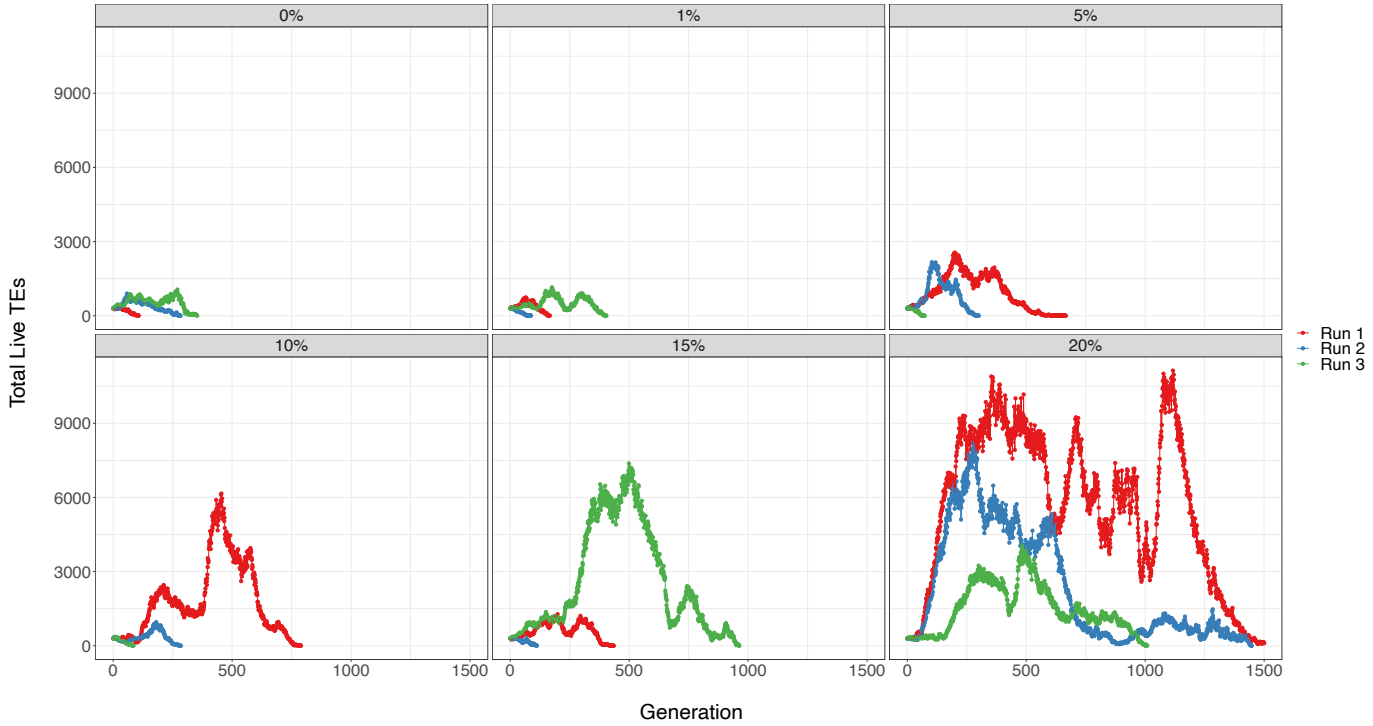


Figure 2.2 TE population dynamics when the beneficial insertion frequency were set at 0%, 1%, 5%, 10%, 15% and 20% respectively. TEs were only able to accumulate for 1,500 generations when the positive insertion frequency was set at 20%. In all other scenarios TE extinction occurred before the simulation ran to completion. The parameter setting for this iteration was LLLH-LLHH

2.7 References

Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., *et al.* (2018). Ten things you should know about transposable elements. *Genome Biol.*, 19, 199.

Brookfield, J.F.Y. (2005). The ecology of the genome - Mobile DNA elements and their hosts. *Nat Rev Genet*, 6.

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.

Chénais, B., Caruso, A., Hiard, S. & Casse, N. (2012). The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. *Gene*, 509.

Elliott, T.A. & Gregory, T.R. (2015). What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philos. Trans. R. Soc. B Biol. Sci.*, 370.

Kim, N.H., Lee, G., Sherer, N.A., Martini, K.M., Goldenfeld, N., Kuhlman, T.E., *et al.* (2016). Real-time transposable element activity in individual live cells. *Proc. Natl. Acad. Sci. U. S. A.*, 113, 7278–7283.

Kremer, S.C., Linquist, S., Saylor, B., Elliott, T.A., Gregory, T.R. & Cottenie, K. (2020). Transposable element persistence via potential genome-level ecosystem engineering. *BMC Genomics*, 21, 367.

Rishishwar, L., Wang, L., Wang, J., Yi, S. V., Lachance, J. & Jordan, I.K. (2018). Evidence for positive selection on recent human transposable element insertions. *Gene*, 675, 69–79.

Le Rouzic, A., Boutin, T.S. & Capy, P. (2007). Long-term evolution of transposable elements. *Proc. Natl. Acad. Sci. U. S. A.*, 104, 19375–19380.

Sotero-Caio, C.G., Platt, R.N., Suh, A., Ray, D.A. & Ray, D.A. (2017). Evolution and Diversity of Transposable Elements in Vertebrate Genomes. *Genome Biol. Evol.*, 9, 161–177.

Sultana, T., Zamborlini, A., Cristofari, G. & Lesage, P. (2017). Integration site selection by retroviruses and transposable elements in eukaryotes. *Nat. Rev. Genet.*, 18, 292–308.

Venner, S., Feschotte, C. & Biémont, C. (2009). Dynamics of transposable elements: towards a community ecology of the genome. *Trends Genet*, 25.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer, New York.

3 Whole genome duplications can promote the proliferation of transposable elements within simulated asexual populations

3.1 Chapter Overview and Author Contributions

This chapter provides an extension of the previously described in-silico TE dynamic model developed by Kremer et al (2020) and described in chapter two. We adjust the model's parameters to simulate whole genome duplication events and observe how this impacts TE abundance. The work from this chapter was conceived by Christopher L Butler and Martin Taylor. Analysis was conducted by Christopher L Butler. The final manuscript was written by Christopher L Butler with contribution from Martin Taylor. The original model was written by Kremer et al (2020) and is publicly available at https://github.com/stefan-c-kremer/TE_World2.

3.2 Background.

Transposable elements (TE), small replicative 'selfish' DNA sequences, are major contributors to the non-coding regions of a genome (Wicker *et al.* 2007; Bourque *et al.* 2018). TE activity is known to induce numerous genomic changes, including increasing (i) rate of mutations, (ii) genome size and (iii) the number of chromosomal rearrangements (Kazazian 2004; Lee *et al.* 2008; González & Petrov 2012; Marburger *et al.* 2018). At a population level, TE activity can also induce changes to rates of speciation and adaptability (Oliver & Greene 2011; Stapley *et al.* 2015). Consequently, an appreciation of the factors which permit TE proliferation is important to better understand the evolution of a genome. Processes that have been attributed to increased TE activity include hybridisation (Ungerer *et al.* 2009), beneficial insertion effects (González *et al.* 2008), environmental stress (Grandbastien *et al.* 2005), and TE ecology (Venner *et al.* 2009). Another factor which has long been associated with the promotion of TE activity regards whole genome duplication events (WGD) (McClintock 1984). WGDs are purported to be relatively common across the tree of life. Vertebrates have been affected by two ancestral rounds of WGD, and a subsequent lineage-specific WGD within teleosts, (Rodriguez &

Arkhipova 2018), whilst WGDs have occurred at least 50 times within the evolutionary history of land plants (Clark & Donoghue 2018). A better understanding of the relationship between WGDs and TE activity is therefore of high importance.

Polyploidy (where an organism possesses more than one complete chromosome set) typically occurs due to a breakdown in meiosis, which can either occur within the same species (autopolyploidy) or as a result of hybridisation (allopolyploidy) (Madlung 2012). There are two theoretical explanations as to why a TE expansion may follow a WGD event. The first is due to a breakdown in TE silencing measures ('Genome Shock Hypothesis'), perhaps due to the mismatched inheritance of TE families and their associated silencers (e.g. siRNAs) (Matzke & Matzke 1998; Parisod & Senerchia 2012). The second is that duplicated (paralogous) genes are subject to a relaxation of selection pressure, meaning there is greater available regions of the genome in which TE insertions can occur without a dramatic loss in host fitness (Matzke & Matzke 1998; Parisod & Senerchia 2012). Numerous attempts have been made to investigate how selection pressure can differ between single-copy 'orthologous' genes vs multi-copy 'paralogous' genes. The ratio of non-synonymous mutations per non-synonymous site (Kn) vs the degree of synonymous mutations per synonymous site (Ks) across 39 genomes (26 bacterial, 6 archaeal and 7 eukaryotic) has suggested that whilst both orthologues and paralogs evolve under purifying selection, Kn/Ks ratios are significantly higher within paralogous genes (Kondrashov *et al.* 2002). Similarly, a comparison between 5,341 human-mouse orthologs demonstrated that genes with a duplicated paralog had a 36% increase in Kn/Ks ratio (Nembaware *et al.* 2002). Crucially, both findings support the theory that a WGD will reduce the degree of purifying selection pressure acting on paralogous genes.

Empirical studies which have investigated whether TE abundance changes after a WGD event have reported a mixture of different outcomes. For example, the spotted gar (*Lepisosteus oculatus*) has similar TE abundance to other post-WGD teleost species, indicative of no causative link between

polyploidy and increased TE abundance (Chalopin & Volff 2017). Similarly, evidence within allopolyploid *Arabidopsis* hybrids suggest that a WGD does not lead to increased TE activity (Beaulieu *et al.* 2009). However, in wheat allopolyploids, a drop in both (i) transposon-related siRNA transcripts and (ii) CpG methylation, are both hallmarks of a disruption to TE silencing measures (Kenan-Eichler *et al.* 2011). Furthermore, relaxed purifying selection in polyploid Arabidopsis lineages has recently been shown to lead to an increase in TE abundance (Baduel *et al.* 2019). However, the relationship between polyploidy and TE activity may frequently be conflated with other ecological processes, with changes to effective population size (Ne), stress tolerance and range expansion all being associated with WGD events (Clo 2022).

One avenue in which TE dynamics can be explored independently of cofounding environmental effects is the use of in-silico modelling. Numerous models have been recently used to better understand TE dynamics, including those that explore the interplay between TE activity and (i) methylation based silencers (e.g RNA interface) (Roessler *et al.* 2018), (ii) genome streamlining (i.e abundance of coding DNA) (Van Dijk *et al.* 2022) and (iii) mode of reproduction (Dolgin & Charlesworth 2006). One simple TE dynamic model recently developed by Kremer et al*,* 2020 simulated an asexual (i.e haploid) single celled organism in which population genetic concepts commonly attributed to TE persistence (e.g. positive selection, genetic drift, co-evolution, recombination, or horizontal transfer) are not included within the simulation. This model has been used to highlight that TE persistence may occur through both (i) TE engineering processes (i.e. a positive feedback loop in which TE proliferation provides genomic substrate where more TE insertions can occur without impacting host fitness) (Kremer *et al.* 2020, 2021) and (ii) beneficial insertion effects (Butler *et al.* 2021). In this study we further develop the modelling framework developed by Kremer et al (2020) to simulate whether a WGD event, and subsequent reduction in the level of purifying selection, can lead to greater TE abundance than would be expected in a haploid genome. This study therefore represents (at least to our understanding) the first in-silico attempt to untangle the possible interplay between WGD and TE dynamics.

3.3 Methods

To investigate how WGD affected TE proliferation we deliberately chose two parameter scenarios found within the original Kremer et al, 2020 model where TEs were shown to persist with expansion (LLLH-LHHL) or without expansion (LLLH-LLHH) within haploid individuals (Table 3.1). In each scenario two conditions were explored. In the first, the simulation was run for 1,000 generations in identical fashion to Kremer *et al* (2020), and therefore represents what would occur in the absence of a WGD event. The second model introduces a WGD at generation 300, and thus the individuals become diploid. Three major genetic changes occur at this point.

1) The number of genes, TEs and amount of junk (non-coding) DNA within the genome double.
2) The chance that a mutation or TE insertion causes a lethal fitness impact (i.e. the individual dies) is halved, instead generating no change in fitness. This not only reflects theoretical predictions that gene duplication provides a buffer from serious fitness impacts, but also closely matches evidence within experimental lines of the yeast *Saccharomyces cerevisiae*, in which the proportion of genes where a deletion is lethal reduces from 29% in orthologues to 12.4% in paralogues (Gu *et al.* 2003).
3) Whilst the fixed probability that a mutation or TE insertion causes a negative fitness impact remains the same in our model, the magnitude (mutation effect) of such activity on the survival likelihood of the individual is halved, matching mathematical predictions theorised by Kondrashov et al, 2002. Importantly, both the frequency and magnitude of beneficial TE insertion/mutations remain identical to the non-WGD scenario and remain below the frequency of negative insertions (i.e. any increase in TE abundance isn't being driven by net positive selection effect).

Table 3.1 The two parameter scenarios used this series of simulations where Kremer et al, 2020 previously reported a TE expansion (LLLH-LHHL) or no TE expansion (LLLH-LLHH)

TE PROPERTIES				HOST PROPERTIES				PARAMETER SCENARIO
TE Progeny	TE Excision Rate	TE Death Rate	Insertion Bias	Corrected Mutation Rate	Non-Coding DNA	Mutation Effect	Carrying Capacity	
Low	Low	Low	High	Low	High	High	Low	LLLH-LHHL
Low	Low	Low	High	Low	Low	High	High	LLLH-LLHH

Three model outcomes were subsequently investigated and plotted under this new model.

1) The number of TEs per individual (LTETOTAL/pop size).
2) The mean genome size (MB) per individual. This was calculating by summing the amount of junk DNA bp, the number of TEs per individual (each with a fixed length of 1 kilobase (kb)) and the number of genes per individual (also with a fixed length of 1 kb).
3) The percentage of the genome which consisted of TEs.

Each simulation was run three times independently and the results were plotted using the 'ggplot2' package in R v3.5.1 (Wickham 2016). The growth rate of the TE population per individual from generation 300-1,000 were also statistically compared between both WGD and non-WGD scenarios using the 'compareGrowthCurves' function within the statmod R package (Giner & Smyth 2016). Specifically, this conducts a permutation test of the difference between two or more groups of growth curves (in this case WGD vs no WGD). Here, we performed the test with 'nsim' (number of permutations) set to 10,000.

3.4 Results

Two important metrics of TE activity after a WDG event were measured in this new series of simulations. The first was growth rate of TE copies per individual between generation 300-1,000 (i.e.

after the potential WDG event occurred). The first set of parameters investigated was one where the TEs were increasing within the haploid population (LLLH-LHHL). The average growth rate within the 'no WGD' scenario was 1.34 and within the WGD scenario it was 3.03, with the growth rate being greater in each diploid vs haploid simulation (Figure 3.1). A permutation test highlighted that this difference in the growth curves of WGD vs no WGD was significant ($t = -31.57$, $P < 0.001$). The second set of parameters investigated was one where the TEs were not increasing within the haploid population (LLLH-LLHH). This is highlighted by the negative growth rate in each of the 'no WGD' iterations, with an average of -0.02 (Figure 3.2). On the contrary, every equivalent 'WGD' simulation yielded a positive growth rate, with an average of 0.23 (Figure 3.2). Once again, a permutation test highlighted that the differences in growth curves between WGD and no WGD was highly significant ($t = -106.25$, $P < 0.001$).

The second metric of TE activity we investigated was TE abundance given as a proportion of genome size, which accounts for the fact that overall genome content (including TE number) doubled at generation 300 in the 'WGD' simulations. Despite raw TE number being higher after a WGD in every run of LLLH-LHHL, TE abundance measured as a proportion of genome size was higher in just one such iteration (run 2) (Figure 3.1). Therefore, in most cases under this parameter scenario, TEs were simply increasing in proportion to the overall genome expansion, with the WGD making little to no impact to the overall proportion of TE content (Figure 3.1). The exception was run 2, whereby TEs made up 5.52% of a genome after a WGD compared to 3.04% of the genome in the absence of a WGD (Figure 3.1). In the second parameter scenario (LLLH-LLHH), TE abundance as a proportion of the genome was substantially higher in every iteration ($\bar{x} = 0.67\%$ after WGD vs $\bar{x} = 0.02\%$ with no WGD) (Figure 3.2). Finally, in both parameter scenarios TE expansion made little difference to overall genome size, which was instead largely influenced by the amount of junk/non-coding DNA (Figure 3.1, Figure 3.2).

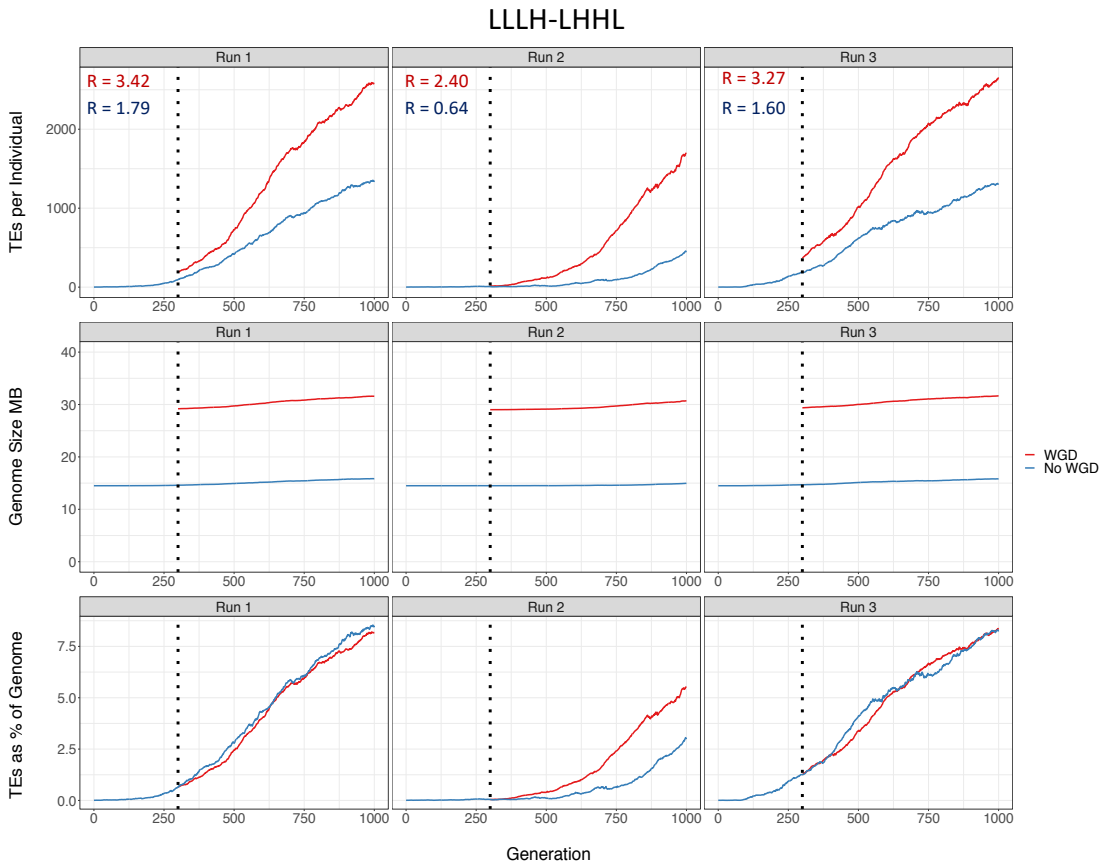


Figure 3.1 Simulated TE and genome-based characteristics within a haploid (no WGD) or diploid (WGD) individual. In the diploid case, the WGD duplication occurred at generation 300. The parameter setting for this iteration was LLLH-LHHL, and each simulation was run three times independently. Growth rates (R) of TEs per individual are highlighted, corresponding to a WDG event either occurring (red) or not occurring (blue).

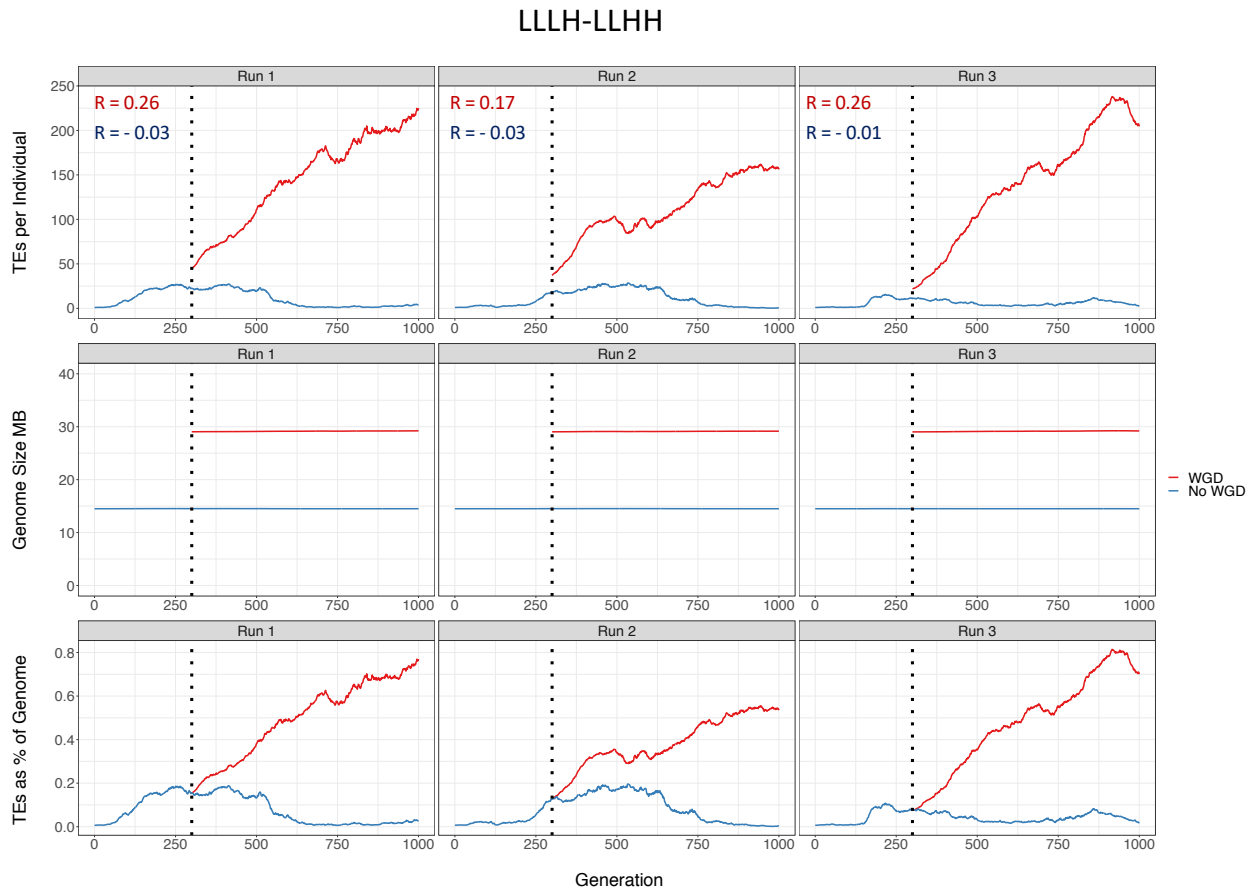


Figure 3.2 Simulated TE and genome-based characteristics within a haploid (no WGD) or diploid (WGD) individual. In the diploid case, the WGD duplication occurred at generation 300. The parameter setting for this iteration was LLLH-LLHH, and each simulation was run three times independently. Growth rates (R) of TEs per individual are highlighted, corresponding to a WDG event either occurring (red) or not occurring (blue).

3.5 Discussion

In these series of in-silico simulations we demonstrate that after a WGD, TEs proliferate at a significantly greater rate across generations than they would have done otherwise, and frequently occupy a greater proportion of a genome than an equivalent non-WGD individual. Two parameter combinations were deliberately chosen during this experiment due to their similarity, differing only in whether the level of non-coding DNA and population size were high or low. We found that the ability of a WGD to spark higher genomic TE content was somewhat dependent on these parameter combinations. In population scenarios where TEs were unable to spread in haploid individuals (low levels of non-coding DNA and high population size - LLLH-LLHH) a WGD led to greater levels of TE activity and abundance in every iteration. However, in population conditions where TEs were able to spread within haploid individuals (high levels of non-coding DNA and low carrying capacity) we found that a WGD led to greater genomic TE abundance in only one of the three runs. These results provide an interesting observation, whereby the degree in which a WGD can promote greater genomic TE abundance may be somewhat dependent on the rate of TE activity pre-duplication. This may explain the variation in previously reported outcomes regarding TE activity after WGD events (Kenan-Eichler *et al.* 2011, Chalopin & Volff 2017, Baduel *et al.* 2019).

The consequence of greater TE activity is a post-WDG organisms are likely to be diverse, causing general genomic and population level changes such as those highlighted earlier. However, there may also be specific consequences to increased TE activity after a WGD (opposed to general TE abundance increases). Gene expression levels have been shown to differ between recent polyploid vs diploid species, particularly within genes located near reactivated TEs (Vicient & Casacuberta 2017). Longer term, sub-genomes with lower TE abundance are likely to display higher expression levels, leading to an effect which may be particularly common in allopolyploids known as (sub)genome dominance (Woodhouse *et al.* 2014). Somewhat paradoxically, TE proliferation may also speed up a return to a diploid state (rediploidisation), as introduced regions of sequence

similarity and subsequent recombination may result in large-scale genomic deletions (Vicient & Casacuberta 2017).

The model used in this study, like any, has a number of assumptions that may not reflect biological reality and are therefore important to discuss. Firstly, the model does not permit any TE based silencing measures within the genome. Any substantial increase in TE abundance is likely to be met by an increased pressure for the emergence of increased host silencing, through the likes of CpG methylation or histone modification (Slotkin & Martienssen 2007). As a result, the degree of increased TE accumulation observed in our model may be somewhat of an overestimate. However, this may be somewhat compensated by a "Genome Shock", in which polyploidy can lead to a reduction in host silencing measures, though these effects are unlikely to last several hundred generations. Furthermore, the influence of a "Genome Shock" in driving increased TE abundance after a WGD has recently been questioned, with studies in polyploid *Capsella* plants indicating that increased TE accumulation is due to a relaxation in selection pressure rather than breakdowns to host silencing (Ågren *et al.* 2016). Secondly, we have assumed the reduction in selection pressure is even across the entire genome, i.e paralogous gene copies do not exhibit asymmetry (no sub-genome dominance or fractionation bias). This assumption is supported by empirical evidence (Kondrashov *et al,* 2002). Finally, the fact that the WGD leads to an exact doubling of junk DNA, TE and gene number means our model may better reflect autopolyploidy rather than allopolyploidy, with hybridisation unlikely to occur between two symmetrical genomes.

In conclusion, this study has demonstrated that a WGD event can induce both greater genomic TE abundance and TE copy growth rates, thus supporting most of the theoretical framework on this matter. Our finding that genomic TE abundance after a WGD event appeared to be dependent on the base level of TE activity may explain why empirical evidence has yielded mixed evidence regarding TE abundance and polyploidy. Future efforts could expand upon these set of simulation to

better refine our understanding of the factors which can generate different TE content across different organisms. Finally, we hope to have demonstrated how in-silico modelling can be used to support theoretical assumptions related to the factors which permit TE abundance variation across the tree of life.

3.6 References

Ågren, J.A., Huang, H.-R. & Wright, S.I. (2016). Transposable element evolution in the allotetraploid Capsella bursa-pastoris. *Am. J. Bot.*, 103, 1197–1202.

Baduel, P., Quadrana, L., Hunter, B., Bomblies, K. & Colot, V. (2019). Relaxed purifying selection in autopolyploids drives transposable element over-accumulation which provides variants for local adaptation. *Nat. Commun.*, 10, 1–10.

Beaulieu, J., Jean, M. & Belzile, F. (2009). The allotetraploid Arabidopsis thaliana–Arabidopsis lyrata subsp. petraea as an alternative model system for the study of polyploidy in plants. *Mol. Genet. Genomics* , 281, 421–435.

Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., *et al.* (2018). Ten things you should know about transposable elements. *Genome Biol.*, 19, 199.

Butler, C.L., Bell, E.A. & Taylor, M.I. (2021). Removal of beneficial insertion effects prevent the long-term persistence of transposable elements within simulated asexual populations. *BMC Genomics*, 22, 241.

Chalopin, D. & Volff, J.N. (2017). Analysis of the spotted gar genome suggests absence of causative link between ancestral genome duplication and transposable element diversification in teleost fish. *J. Exp. Zool. Part B Mol. Dev. Evol.*, 328, 629–637.

Clark, J.W. & Donoghue, P.C.J. (2018). Whole-Genome Duplication and Plant Macroevolution. *Trends Plant Sci.*, 23, 933–945.

Clo, J. (2022). Polyploidization: Consequences of genome doubling on the evolutionary potential of populations. *Am. J. Bot.*, 1–8.

Van Dijk, B., Bertels, F., Stolk, L., Takeuchi, N. & Rainey, P.B. (2022). Transposable elements promote the evolution of genome streamlining. *Philos. Trans. R. Soc. B*, 377.

Dolgin, E.S. & Charlesworth, B. (2006). The Fate of Transposable Elements in Asexual Populations. *Genetics*, 174, 817–827.

Giner, G. & Smyth, G.K. (2016). Statmod: Probability calculations for the inverse Gaussian distribution. *R J.*, 8, 339–351.

González, J., Lenkov, K., Lipatov, M., Macpherson, J.M. & Petrov, D.A. (2008). High Rate of Recent Transposable Element–Induced Adaptation in Drosophila melanogaster. *PLoS Biol.*, 6, e251.

González, J. & Petrov, D.A. (2012). Evolution of Genome Content: Population Dynamics of Transposable Elements in Flies and Humans. In: *Evolutionary Genomics: Statistical and Computations Methods, Volume 1*. Humana Press, Totowa, NJ, pp. 361–383.

Grandbastien, M.-A., Audeon, C., Bonnivard, E., Casacuberta, J.M., Chalhoub, B., Costa, A.-P.P., *et al.* (2005). Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae. *Cytogenet. Genome Res.*, 110, 229–241.

Gu, Z., Steinmetz, L.M., Gu, X., Scharfe, C., Davis, R.W. & Li, W.-H. (2003). Role of duplicate genes in genetic robustness against null mutations. *Nat.* , 421, 63–66.

Kazazian, H.H. (2004). Mobile elements: drivers of genome evolution. *Science.*, 303, 1626–32.

Kenan-Eichler, M., Leshkowitz, D., Tal, L., Noor, E., Melamed-Bessudo, C., Feldman, M., *et al.* (2011). Wheat hybridization and polyploidization results in deregulation of small RNAs. *Genetics*, 188, 263–72.

Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I. & Koonin, E. V. (2002). Selection in the evolution of gene duplications. *Genome Biol.*, 3, 1–9.
Kremer, S.C., Linquist, S., Saylor, B., Elliott, T.A., Gregory, T.R. & Cottenie, K. (2020). Transposable element persistence via potential genome-level ecosystem engineering. *BMC Genomics*, 21, 367.
Kremer, S.C., Linquist, S., Saylor, B., Elliott, T.A., Gregory, T.R. & Cottenie, K. (2021). Long-term TE persistence even without beneficial insertion. *BMC Genomics*, 22, 1–6.
Lee, J., Han, K., Meyer, T.J., Kim, H.-S. & Batzer, M.A. (2008). Chromosomal Inversions between Human and Chimpanzee Lineages Caused by Retrotransposons. *PLoS One*, 3, e4047.
Madlung, A. (2012). Polyploidy and its effect on evolutionary success: old questions revisited with new tools. *Hered.* , 110, 99–104.
Marburger, S., Alexandrou, M.A., Taggart, J.B., Creer, S., Carvalho, G., Oliveira, C., *et al.* (2018). Whole genome duplication and transposable element proliferation drive genome expansion in Corydoradinae catfishes. *Proc. R. Soc. B*, 285.
Matzke, M.A. & Matzke, A.J.M. (1998). Polyploidy and transposons. *Trends Ecol. Evol.*
McClintock, B. (1984). The significance of responses of the genome to challenge. *Science*, 226, 792–801.
Nembaware, V., Crum, K., Kelso, J. & Seoighe, C. (2002). Impact of the Presence of Paralogs on Sequence Divergence in a Set of Mouse-Human Orthologs. *Genome Res.*, 12, 1370–1376.
Oliver, K.R. & Greene, W.K. (2011). Mobile DNA and the TE-Thrust hypothesis: supporting evidence from the primates. *Mob. DNA*, 2.
Parisod, C. & Senerchia, N. (2012). Responses of Transposable Elements to Polyploidy. *Top. Curr. Genet.*, 24, 147–168.
Rodriguez, F. & Arkhipova, I.R. (2018). Transposable elements and polyploid evolution in animals. *Curr. Opin. Genet. Dev.*, 49, 115–123.
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, 803–815.
Slotkin, R.K. & Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.*, 8, 272–285.
Stapley, J., Santure, A.W. & Dennis, S.R. (2015). Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species. *Mol. Ecol.*, 24, 2241–2252.
Ungerer, M.C., Strakosh, S.C. & Stimpson, K.M. (2009). Proliferation of Ty3/gypsy-like retrotransposons in hybrid sunflower taxa inferred from phylogenetic data. *BMC Biol.*, 7.
Venner, S., Feschotte, C. & Biémont, C. (2009). Dynamics of transposable elements: towards a community ecology of the genome. *Trends Genet*, 25.
Vicient, C.M. & Casacuberta, J.M. (2017). Impact of transposable elements on polyploid plant genomes. *Ann. Bot.*, 120, 195–207.
Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., *et al.* (2007). A Unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.*, 8, 973–982.
Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer, New York.
Woodhouse, M.R., Cheng, F., Pires, J.C., Lisch, D., Freeling, M. & Wang, X. (2014). Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. *Proc. Natl. Acad. Sci.*, 111, 5283–5288.

4. Transposable element annotation in non-model species: The benefits of species-specific repeat libraries using semi-automated EDTA and DeepTE de novo pipelines

4.1 Chapter Overview and Author Contributions

This chapter's research focused on both (i) the generation of a Corydoras-specific TE library and (ii) a subsequent assessment of the associated bioinformatic pipeline across several TE based metrics. This work has now been published in a co-first author manuscript with Ellen A Bell and is presented below

Bell EA, **Butler CL***, Oliveira C, Marburger S, Yant L, Taylor MI. Transposable element annotation in non-model species - the benefits of species-specific repeat libraries using semi-automated EDTA and DeepTE de novo pipelines. *Mol Ecol Resour.*22, 823– 833 (2022) doi: 10.1111/1755-0998.13489.***co-first author**

Principally, I developed an Bash/R-based script that parsed TE outputs from RepeatMasker with transcript sequences in mind. This was also modified to run on genomic sequences and subsequently developed into a GitHub page https://github.com/clbutler/RM_TRIPS. I also developed a script to detect the possible origin of TEs (horizontal vs vertical). The data manipulation and plotting required for the production of Figure 4.2 (TE abundance), Figure 4.4b (length distribution) and Figure 4.6 (sequence divergence) were also written by myself. I conducted the quality assessment of the assembled transcriptomes and the running of RepeatMasker on the transcripts. Finally, the infographic for Figure 4.5 (TE fragmentation) was produced by myself. The writing, manuscript development and submission process was shared evenly between myself and Ellen A Bell. This statement has been read and agreed with by Ellen A Bell. An appendix is also included which contains some additional work I conducted which were not included within the published manuscript. Appendix I contains a more detailed insight into the TE-based parse script RMtrips,

whilst Appendix II contains some additional analysis regarding the Corydoras TE library's ability to detect horizontally transferred transposons.

4.2 Abstract

Transposable elements (TEs) are significant genomic components which can be detected either through sequence homology against existing databases or de novo, with the latter potentially reducing the risk of underestimating TE abundance. Here, we describe the semi-automated generation of a de novo TE library using the newly developed EDTA pipeline and DeepTE classifier in a non-model teleost (*Corydoras fulleri*). Using both genomic and transcriptomic data, we assess this de novo pipeline's performance across four TE based metrics: (i) abundance, (ii) composition, (iii) fragmentation, and (iv) age distributions. We then compare the results to those found when using a curated teleost library (*Danio rerio*). We identify quantitative differences in these metrics and highlight how TE library choice can have major impacts on TE-based estimates in non-model species.

Keywords: Corydoras, de novo, genomics, teleost, transcriptomics, transposon annotation

4.3 Introduction

Transposable elements (TEs) are sequences of repetitive, non-coding DNA found in high abundance across the tree of life (Bourque et al., 2018; Wells & Feschotte, 2020; Wicker et al., 2007). Historically overlooked during genomic analysis and annotation, TEs are now recognised as key contributors to genome evolution and regulation, providing alternative promoters, neofunctionalisation, novel exons, and large-scale rearrangements (Bourque et al., 2018; Cowley and Oakey 2013; Hoen & Bureau, 2015). This realisation, coupled with the increased availability of genome sequences, has generated a growing need for both accessible and comprehensive TE annotation in non-model species.

TEs can be detected using either homology or de novo approaches. Homology-based approaches detect TEs through sequence comparisons against existing databases, whilst de novo approaches

identify TEs through signatures such as structure or elevated copy number (Kennedy et al., 2011; Ou et al., 2019). Homology searches may lead to TE underestimates because sequence divergence can render certain TEs unrecognisable, which may be particularly common in non-model organisms where there may be large phylogenetic distances to their closest database entry (Bergman & Quesneville, 2007). Furthermore, due to their potential absence from databases, homology-based searches may bias detection away from species-specific TEs which have inserted since the common ancestor of focal species and library (Platt et al., 2016). This may be particularly true in the case of horizontally transferred TEs which are increasingly recognised to move between vertebrate genomes and may be important for long term TE persistence (Groth & Blumenstiel, 2017; Zhang et al.,2020). Consequently, the generation of TE libraries that do not rely solely on homology-based searching is recommended (Hoen & Bureau, 2015; Platt et al., 2016). However, de novo TE libraries also have disadvantages, as they may fail to detect low-copy number elements or erroneously identify/classify TEs (Bergman & Quesneville, 2007). De novo TE libraries may therefore require a degree of manual curation, which can be both time consuming and labour intensive.

Several semi-automated pipelines for de novo library construction have been created to streamline their development, both in terms of annotation and classification. These include the Extensive de novo TE Annotator (EDTA) (Ou et al., 2019) and DeepTE (Yan et al., 2020). EDTA combines a suite of best-performing packages (LTR_FINDER, LTRharvest, LTR_retriever, Generic Repeat Finder, TIR Learner, HelitronScanner and RepeatMasker) to produce nonredundant TE libraries. EDTA also has an option to use RepeatModeler to do a final sweep for remaining unidentified TEs, thereby utilising two very powerful TE annotating tools (Ou et al., 2019). After initially performing well in rice (*Oryza sativa*, Ou et al., 2019), EDTA has subsequently been run across numerous nonvertebrate genomes, including sweet corn (*Zea mays*, Hu et al., 2021*)*, field mustard (*Brassica rapa*, Cai et al., 2021*)* and sawfly (*Euura lappo*, Michell et al., 2021*)*. DeepTE classifies TEs using machine learning, specifically by using convolutional neural networks to assign TEs to superfamily and order, with good

performance in terms of accuracy and sensitivity against other similar classifiers as well in the assignment of previously unknown TEs (Yan et al., 2020). However, the impacts of EDTA's implementation on TE annotation in non-model, vertebrate genomes, particularly when combined with DeepTE, have yet to be fully explored.

In this study, we describe the use of both EDTA and DeepTE to construct a de novo library for *Corydoras fulleri*, a member of the Corydoradinae which are a species-rich subfamily of Neotropical catfishes with highly variable TE content (Alexandrou et al., 2011, Marburger et al., 2018). Teleost genomes contain the most abundant and diverse TE content of all vertebrates, including numerous horizontally integrated elements, making them interesting organisms to assess de novo pipelines (Sotero-Caio et al., 2017; Zhang et al., 2020). Here, we compare the performance of our *Corydoras*-specific TE library against the RepBase *D*. *rerio* library by estimating TE content within two *Corydoras* species; *Corydoras fulleri* and *Corydoras maculifer*. Specifically, our *Corydoras*-specific TE library was quantitatively assessed using estimates of four key TE-based metrics; (i) abundance, (ii) composition, (iii) fragmentation, (i.e., the likelihood that genomic TE copies have not been captured in a single contiguous manner during library creation), and (iv) sequence divergence distributions. We also use a mixture of both genomic and transcriptomic sequences to test how library type affects TE landscapes across different transposon age groups. Finally, we present this pipeline as a GitHub resource that will be applicable to a diverse range of species in the future (Figure 4.1, https://github.com/ellenbell/FasTE).

4.4 Materials and Methods

Extraction and sequencing of DNA and genome assembly

The genome of *C*. *fulleri* was assembled using both long-read PacBio sequencing and short-read Illumina Sequencing. Genomic DNA was extracted from *C*. *fulleri* using MagAttract HMW DNA Kit (Qiagen) for high molecular weight PacBio sequencing and PureLink Genomic DNA mini kit (by ThermoFisher Scientific) for Illumina Hiseq. Sequencing was performed on two PacBio Sequel cells and one Hiseq lane using 300 bp paired-end reads, which was estimated to generate 60x long-read

PacBio coverage and 100x Illumina Hiseq coverage. All genomic library preparation and sequencing of *C. fulleri* was performed by Novogene Co Ltd.

Genome assembly for *C. fulleri* was performed using wtdbg2 (version 2.5) to create an initial long-read assembly from PacBio data (Ruan & Li, 2019). This first pass assembly was then polished using wtdbg2-racon-pilon.pl v04 script (Schellt, 2019; https://github.com/schellt/wtdbg2-racon-pilon) which performs three iterative corrections, firstly with long-read mapping using minimap2 (version 2.17, Li, 2018) and polishing with Racon (version 1.4.15, Vaser et al., 2017) and then with short-read mapping using bwa mem (version 0.7.17, Li, 2013), merging and sorting using Samtools (version 1.10, Li et al., 2009) and polishing with Pilon (version 1.23, Walker et al., 2014).

The genome of *C. maculifer* was assembled using short-read Illumina based sequencing. Genomic DNA was extracted from *C. maculifer* using Qiagen DNeasy Blood and Tissue extraction kit. Paired-end PCR-free libraries were produced and sequenced on a single lane of an Illumina Hiseq platform using 250 bp paired-end reads, estimated to provide 50X coverage. Twelve Nextera long mate paired (LMP) libraries were also generated and sequenced on a second lane of Illumina Hiseq using 300 bp paired-end reads from which the two libraries with the largest insert size were selected (average insert sizes 8678.2 bp and 8730.0 bp, respectively). These two libraries were then sequenced on an Illumina Hiseq platform with 250 bp paired-end reads to assist with scaffolding. All library preparation and sequencing of *C. maculifer* was performed by the Earlham Institute, Norwich. Paired-end libraries were assembled using w2rap-contigger (Clavijo et al. 2017) under default settings. LMP libraries were cleaned using NextClip (Leggett et al., 2014) and combined with contigs from paired-end assemblies using SOAPdenovo2 (Luo et al., 2012) under default settings but using a kmer size of 19 to produce scaffolds. Genome coverage for both assemblies was assessed using Quast (version 5.0.2, Gurevich et al., 2013) and completeness measured using BUSCO (version 4.1.0, Seppey et al., 2019) (Supplementary Table 4.1).

Extraction and sequencing of RNA and transcriptome assembly

The transcriptome of *C. maculifer* was assembled from short read Illumina based sequencing. RNA extraction (TRIzol Plus RNA Purification Kit) was conducted on somatic muscle tissue. The size selection and integrity of the extracted RNA was confirmed using an Agilent 2100 Bioanalyser (Agilent Technologies) which met internal QC standards of the sequencing provider. Transcriptomic library preparation and sequencing was performed by the Animal Biotechnology Laboratory of Esalq/Piracicaba and the cDNA library was then built using a TruSeq RNA Sample Prep kit (Illumina, Inc). The *C. maculifer* cDNA was sequenced using paired-end sequencing on an Illumina HiSeq (as part of a larger multiplexed run), generating 10.09 million paired reads. The library was then demultiplexed and cleaned using Trimmomatic (version 0.2.36, Bolger et al., 2014) and subsequently assembled using the de novo transcriptome assembler Trinity (version 2.6.9, Grabherr et al., 2013). Transcriptome quality was later assessed using TransRate (version 1.03, Smith-Unna et al., 2016) (Supplementary Table 4.2).

Transposable element annotation

A de novo TE library was generated from the long-read PacBio *C. fulleri* genome using the Extensive de novo TE Annotator (EDTA) (Ou et al., 2019) set to the "others" species parameter. We utilised the inbuilt RepeatModeller (Smit and Hubley, 2008) support which identifies any remaining TEs which might have been overlooked by the EDTA algorithm (--sensitive 1). Classifications within this library were refined using DeepTE using the predefined metazoan model parameter setting (-m) (Yan et al., 2020). TE identification was performed using RepeatMasker (RM; version 1.332) utilising the NCBI/RMBLAST (version 2.6.0+) search engine. This analysis was conducted either against the *Danio rerio* Repbase (26 October 2018) entry, which was also run through DeepTE (to allow for uniformity in TE classification, referred to as the "*D. rerio* library" henceforth), or the *Corydoras*-specific library. RM was run under the most sensitive (-s) parameter setting in all instances. The genomic and transcriptomic RM output files were subsequently cleaned of nondistinct elements by removing overlapping repeats where a match with a higher likelihood score was available. Outputs were then

parsed through a custom R script; RM_TRIPS (which is publically available at https://github.com/clbutler/RM_TRIPS). RM_TRIPS was used to (i) remove repetitive elements not classed as TEs (e.g., microsatellites, simple repeats & sRNAs), (ii) merge elements found on the same contig if they had the same name, orientation, and their combined sequence length was less than or equal to the corresponding reference sequence in the repeat library, (iii) remove merged repeats with a length less than 80 base pairs, and (iv) for transcriptomic data, if multiple identical repeats were found across different transcript isoforms, only one was retained. This was to ensure that each repeat represented a unique genomic locus.

This complete pipeline from de novo library generation through to RM output parsing has been consolidated into the annotated tool, FasTE, which is publically available at https://github.com/ellenbell/FasTE (see Figure 4.1).

Comparative assessment of the performance of Corydoras-specific TE library

TE abundance estimates were calculated from parsed RM output files derived from the *Corydoras*-specific and *D*. *rerio* libraries. These were then standardised across both *Corydoras* species by calculating the percentage of total genome or transcriptome length (bp) represented by TEs. For compositional comparison, TEs were grouped into Helitrons, Maverick elements, DDE DNA elements, long terminal repeat retrotransposons (LTRs) (including dictyosteilium intermediate repeat sequences DIRS), long interspersed nuclear elements (LINEs), Penelope like elements (PLEs) and short interspersed nuclear elements (SINEs) (Wicker et al., 2007). These compositional comparisons were standardised across genomic and transcriptomic sequence data by scaling TE abundance by megabase (MB).

Library fragmentation was assessed firstly by visualising the cumulative abundance estimates of elements against the standardised number of TE entries within both the *Corydoras*-specific library and the *D*. *rerio* library. Second, we compared genomic TE lengths using the *Corydoras*-specific

library against the C. *fulleri* genome and the *D. rerio* library against the *D. rerio* genome (GCF_000002035.6_GRCz11) (Howe et al., 2013).

Age distributions of TEs were compared across library types using their sequence divergence from library entry as a proxy. This made use of the RM outputs which reports the percentage of substitutions in a matching TE compared to its corresponding library hit. Age/sequence divergence distributions were generated for the four major TE classes - DNA transposons, LTR retrotransposons, SINEs and LINEs.

To investigate the potential origin of *C. maculifer* Mariner elements, we extracted every genomic copy with a matching length of >80% against its library hit, and every transcript copy where an element made up >80% of the transcript's length. We subsequently ran a BLASTn search against the RepeatMasker library, with elements potentially horizontally inherited if sequences had both (i) a best match (lowest E value) against a non-teleost species and (ii) following rationale used in (Rogers et al., 2018) a greater than 2% sequence similarity than its best teleost hit. Figures were produced using the ggplot2 package in R (Wickham, 2016).

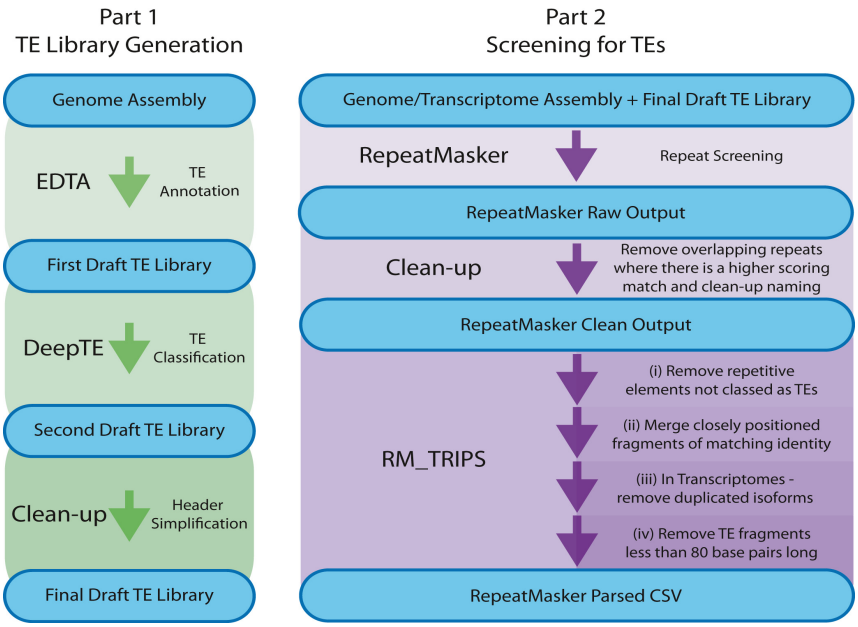


Figure 4.1 FasTE pipeline schematic. Part 1; The three major steps behind de novo TE library generation with EDTA and DeepTE. Part 2; Utilisation of RepeatMasker and de novo libraries to generate estimates of genome wide repeat abundance alongside subsequent parse steps with RM_Trips

4.5 Results

To assess the impact of de novo library creation using EDTA/DeepTE pipelines we generated a de novo TE library (*Corydoras*-specific) from a long-read (PacBio) *Corydoras fulleri* genome assembly and benchmarked it against the *D. rerio* RepBase entry. TE content was then assessed across two *Corydoras* species and sequence types including: (i) a *C. fulleri* genome (ii) a *C. maculifer* genome (another species of the same lineage) and (iii) a *C. maculifer* transcriptome (Figure 4.2a).

Use of the Corydoras-specific TE library led to a 2–3-fold increase in TE abundance estimates

Total TE abundance estimates were higher across both species and sequence types when using the *Corydoras*-specific library. For *C. fulleri*, estimated TE abundance more than doubled from 18.54% of the genome (755.96 hits per MB) using the *D. rerio* RepBase library to 43.45% of the genome (1499.91 hits per MB) using the *Corydoras*-specific library (Figure 4.2b). For the closely related species *C. maculifer*, estimated TE abundance almost tripled from 14.17% of the genome (626.87 hits per MB) using the *D. rerio* RepBase library to 40.23% of the genome (2218.25 hits per MB) using the *Corydoras*-specific library (Figure 4.2b). We then assessed the estimated abundance of TEs across the transcriptome of *C. maculifer*, where TE derived transcripts are expected to represent younger, potentially active, transposons (Lanciano & Cristofari, 2020). Transcriptional TE content was substantially lower than in the *C. maculifer* genome, varying between 1.17% (68.07 hits per MB) and 4.68% (263.22 hits per MB) of the transcriptome when using the *D. rerio* and *Corydoras*-specific library respectively (Figure 4.2b). The substantial increases associated with the use of the *Corydoras*-specific library suggests that the *D. rerio* library missed a large fraction of *Corydoras*-specific elements. We therefore investigated the total number of different TE entries within the *Corydoras*-

specific and *D. rerio* RepBase libraries detected in the *C. fulleri* genome. The *Corydoras*-specific library led to an average fourfold increase in the number of different TEs detected (Supplementary Figure 4.1). Furthermore, across all classes (and particularly for DDE DNA and LTR classes), a number of elements present in the *D. rerio* library were not detected at all within the *C. fulleri* genome (Supplementary Figure 4.1).

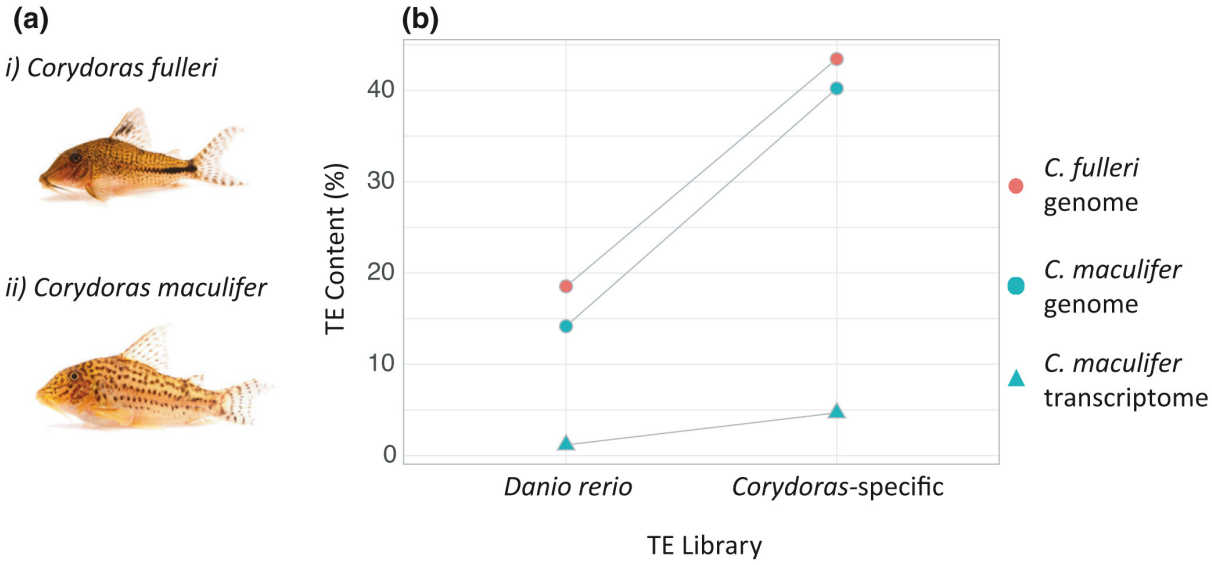


Figure 4.2 TE library type influences TE abundance. (a) the two Corydoras species used in this study (i) *Corydoras fulleri* and (ii) *Corydoras maculifer*. (b) Estimated TE abundance is given as percentage of total genome/transcriptome size for the *C. fulleri* genome and the *C. maculifer* genome and transcriptome

Use of the Corydoras-specific TE library led to substantial changes in estimated TE composition

Using the *Corydoras*-specific library impacted TE composition estimates across both species and sequence types, which we assessed using DeepTE assigned classification. Similar to other teleosts, DDE DNA elements (particularly Tc1 Mariner and hAT transposons) made up substantial proportions of both species genomes and transcriptomes (Figure 4.3). Estimated genomic TE compositions were similar across both genomes investigated, indicating a high level of intralineage TE similarity. TE annotation using the *D. rerio* library detected a similar, relatively high, proportion of SINEs within both genomes, which is in contrast to other teleost species which typically have SINE-depleted genomes (Gao et al., 2016; Shao et al., 2019). On closer inspection however, absolute SINE

abundance (29 MB, 4.57% genome) was similar to that reported in the *D*. *rerio* genome (30.64 MB, 2.24% of genome) (Gao et al., 2016), suggesting that SINE over-representation was a consequence of (i) non-SINE elements being missed when using the *D*. *rerio* library and (ii) SINEs being undetected during de novo library construction and therefore poorly represented in subsequent analyses that depend on the library. Supporting this we found that the number of SINEs detected using the *D*. *rerio* library was largely driven by a single element (HE1 DR1, 84.52% of SINEs in *C*. *maculifer* and 84.12% of SINEs in *C*. *fulleri*) which, following confirmation using BLASTn, was absent in the de novo *Corydoras*-specific library. We also note that the choice of TE library did not generate large compositional changes within transcriptomic sequences (Figure 4.3b iii). To investigate whether any compositional bias had been introduced by DeepTE, we also ran the curated RepBase *D*. *rerio* TE library through DeepTE and compared its classification outputs against the original RepBase library. As expected, the RepBase curated library had a greater range of classifications than the DeepTE classified library (Supplementary Figure 4.2). Although general classification patterns were similar, there was some bias exhibited by DeepTE towards both TIR elements (hAT and Mariner-like) and LTR elements (BEL and Copia) (Supplementary Figure 4.2).

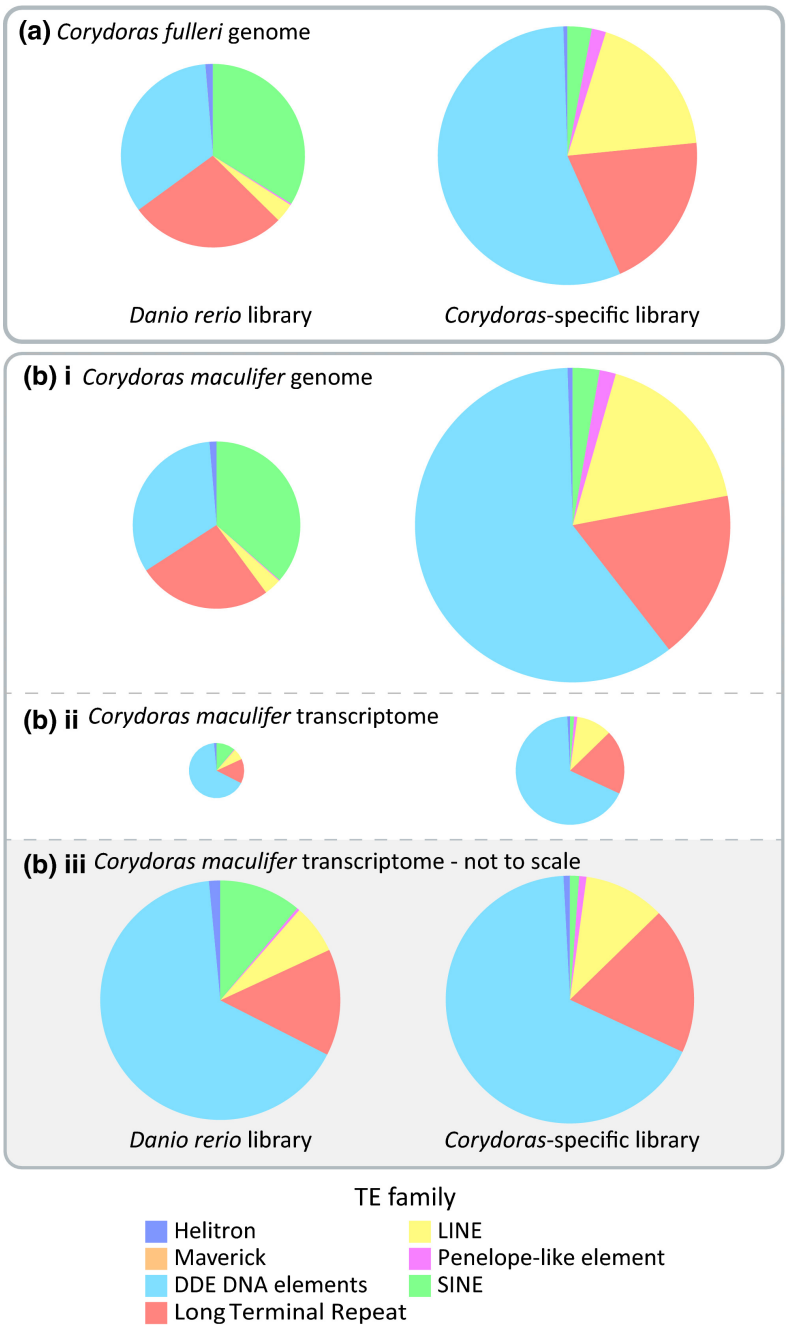


Figure 4.3 TE library type alters TE composition. Estimated TE composition are given in (a) the *C. fulleri* genome, (bi) the *Corydoras maculifer* genome and (bii) the *C. maculifer* transcriptome after using the *Danio rerio* (left) and Corydoras-specific (right) TE libraries. Pie charts are scaled based on TE abundance per MB in all cases apart from (biii) which, for clarity, is the unscaled *C. maculifer* transcriptome composition

The Corydoras-specific library was more fragmented when compared to a curated TE library

We assessed the degree of fragmentation in the *Corydoras*-specific library against the curated *D. rerio* library using (i) cumulative frequency of estimated individual TE abundances and (ii) TE length distributions across the *C. fulleri* genome (Figure 4.4). We define fragmentation as genomic TE copies which have not been captured as a single contiguous unit during library creation (Figure 4.5). An excess of fragmented TE library entries will push a cumulative frequency curve further to the right because many entries will be found at low abundance within the genome (singletons) (Figure 4.5). When standardised by total number of hits, we found little difference between the two libraries (Figure 4.4a) although the *Corydoras*-specific library was inflated with singletons (6.25% of library entries). When looking at the TE length distributions and benchmarking the *Corydoras*-specific library against the RepBase *D. rerio* library (run on their respective genomes) we see markedly similar patterns across all TE classes, with one anomalous peak at c. 350 bp in the LINE distributions (Figure 4.4b). On closer investigation, this peak consisted of a single element (TE_00002410) which, following reanalysis with BLASTx, closely matched a LTR copia element, so is probably a product of misidentification or misclassification by EDTA or DeepTE. We also calculated the average proportion of hits that map back to a single element, with lower values indicating higher degrees of fragmentation. For the *C. fulleri* genome the median number of hits that map to a single element was 0.003% (36 hits per element) and within *D. rerio* it was 0.009% (218 hits per element).

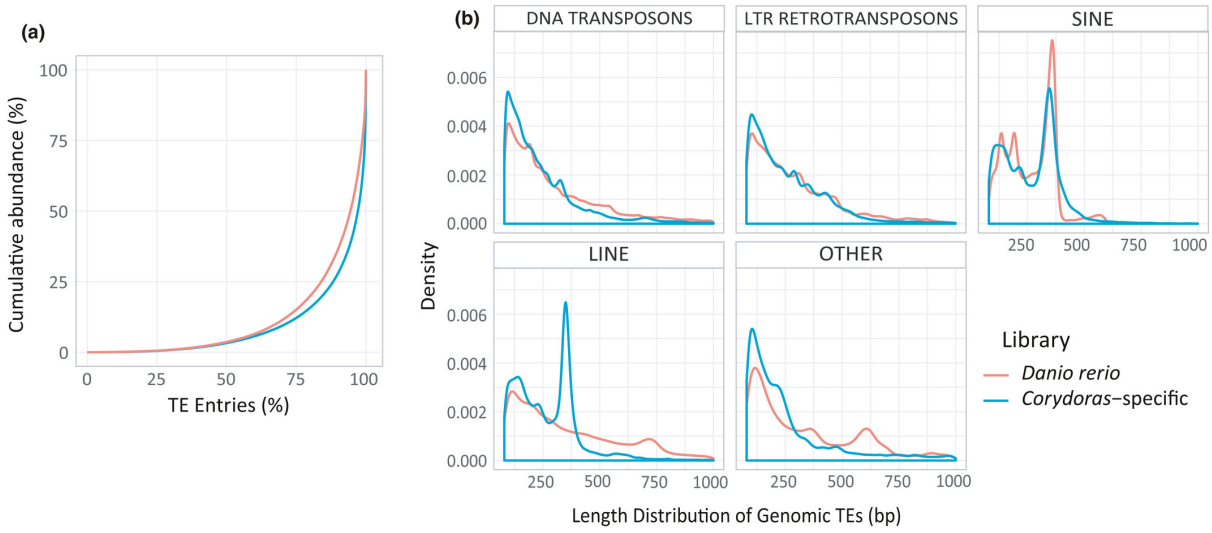


Figure 4.4 Degree of fragmentation is similar for both Corydoras-specific and *Danio rerio* TE libraries. (a) Cumulative frequency of standardised (%) estimated TE abundance using the Corydoras-specific library and the *D. rerio* library. (b) TE length distributions across the Corydoras-specific library and D. rerio library, when run on their respective genomes. For visual purposes the length distribution had a cut off of 1000 bp

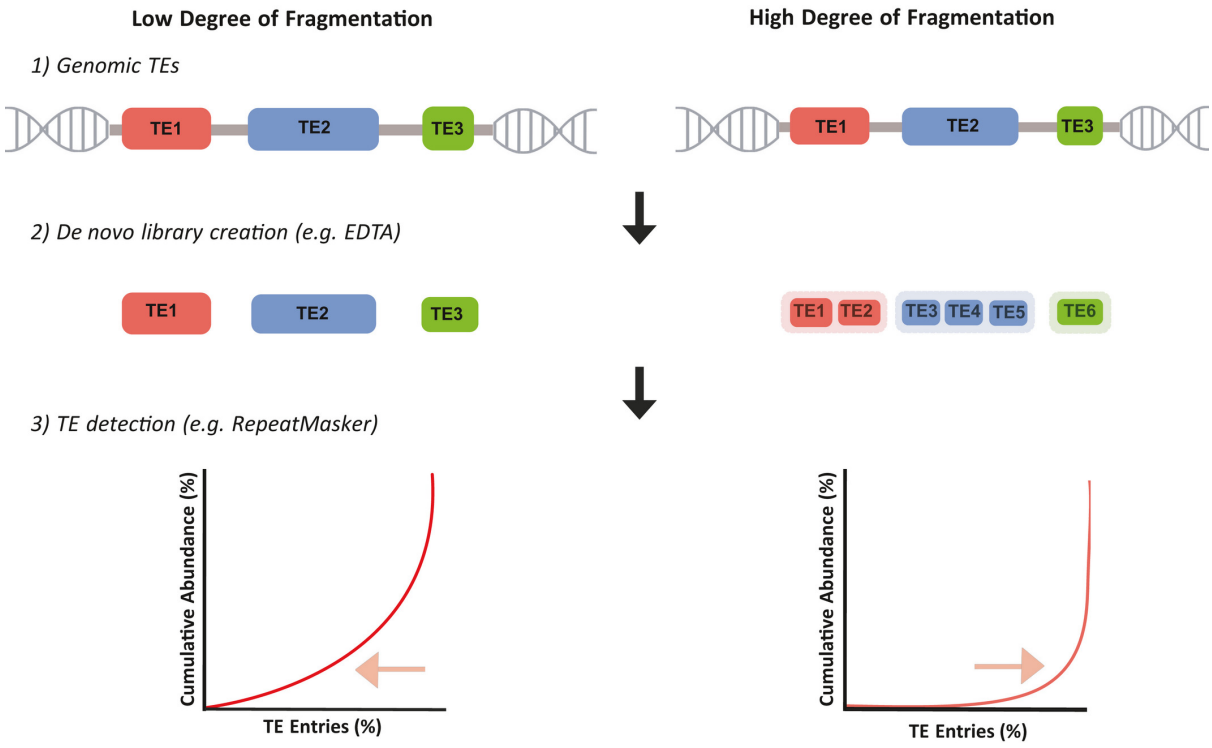


Figure 4.5 Schematic depicting TE fragmentation as a result of de novo library creation. Fragmentation during de novo library creation occurs when single TE copies are detected as multiple fragmented copies. This creates an overinflation of unique library entries, and results in a skewed cumulative frequency curve due to an excess of singletons (TEs detected once only)

The Corydoras-specific TE library reduces average TE age estimates

We investigated the impact the *Corydoras*-specific library had on estimated TE age distributions compared to using the *D. rerio* RepBase library. The *Corydoras*-specific library reduced average sequence divergence against corresponding library entries across all major TE classes and sequence types, suggesting a recent TE accumulation within the *Corydoras* which would have been missed if relying solely on the *D. rerio* library (Figure 4.6). Specifically, the use of the *Corydoras*-specific library significantly reduces the divergence estimates of each element by an average of ~4% (*D. rerio* library 19.90 ± 5.38 sd; *Corydoras*-specific library 15.60 ± 6.71 sd; Welch's t = –519.93, d.f = 1,173,000 p < .001). Finally, the proportion of elements that were very young (estimated to be <5% divergent from its corresponding library entry) was 7.52% within transcriptomic sequences and 5.54% within genomic sequences, suggesting that expressed TEs are on average younger than their genomic counterparts.

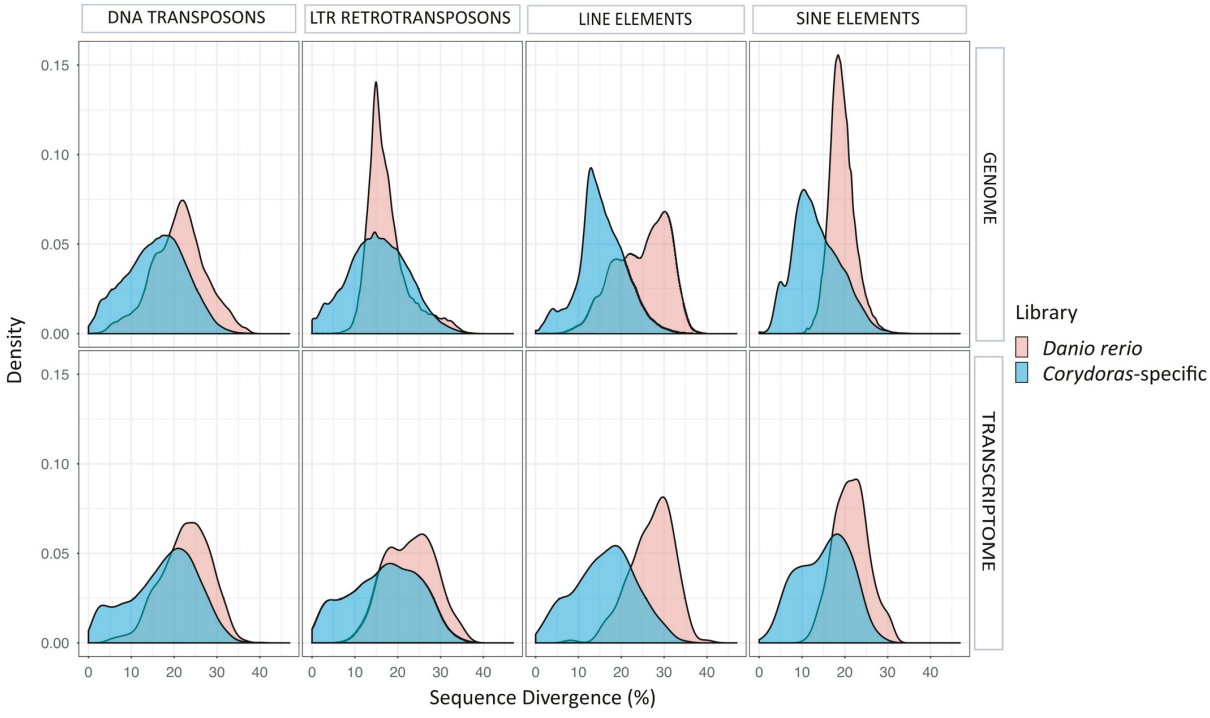


Figure 4.6 TE library type alters TE age distributions. Density plots were used to highlight sequence divergence distributions identified using the Corydoras-specific library and the Danio rerio library. Plots are faceted by the four main TE classes (DNA transposons, LINE elements, LTR Retrotransposons and SINE elements) and sequence type (genome vs transcriptome). All plots are based on *C. maculifer* sequence data

4.6 Discussion

Our results demonstrate how TE library choice can have major implications during TE detection and quantification. The use of the *Corydoras*-specific library led to a 2–3-fold increase in estimated TE abundance in *Corydoras* spp., meaning that ~40% of the two *Corydoras* genomes investigated consist of TEs. TE abundance is highly variable amongst teleosts, ranging between 5% in pufferfish (*Tetraodon nigroviridis*) to 56% in zebrafish (*D. rerio)* (Shao et al., 2019). Use of the *Corydoras*-specific library indicates that TE abundance within these two *Corydoras* genomes (both spp. lineage 1) is comparable to other teleosts, particularly *D. rerio* (~56% of the genome) and *Oryzias latipes* (~33.7% of the genome) (Gao et al., 2016). Theoretically, an inverse relationship between homology-based identification rates and phylogenetic distance exists, in which sequence differences between the species used to develop a library and the target species may provide an obstacle for accurate TE detection. A previous comparison across 40 mammalian genomes demonstrated that TE detection rates exhibit a 'threshold limit', in which TE abundance underestimates are largely avoided until a phylogenetic distance greater than ~90MY is reached, above which homology-based searching may detect as few as 20% of total TEs (Platt et al., 2016). It is therefore no surprise that *Corydoras* TE content was probably underestimated when assessed using the *D. rerio* library given that these species are separated by ~150 million years of evolution (Chen et al., 2013).

In addition, estimated transcriptomic TE abundance was approximately an order of magnitude lower than genomic content, which probably reflects the fact that: (i) TEs may largely be located within non-coding regions of the genome, (ii) many TEs found within *Corydoras* genomes may be degraded and no longer possess the ability to be transposed, or (iii) epigenetic silencing mechanisms (such as CpG methylation and histone modifications) may prevent TE expression (Slotkin & Martienssen, 2007). It is also worth noting that this study used RNA-seq data originating from somatic muscle tissue. TE expression is likely to vary between different tissue types, theoretically evolving to be most active in the germline and comparatively silent in the soma (Haig, 2016).

TE composition estimates within the *Corydoras* were found to be similar to that of other teleosts, with DDE DNA transposons being the most abundant TE class, largely driven by a high abundance of Tc1-Mariner and hAT elements (Shao et al., 2019). Due to their "blurry promoters" Mariner elements appear to have a particular propensity for horizontal transfer across the vertebrate kingdom (Zhang et al., 2020). Despite the suggestion that homology-based methods may miss horizontally transferred elements, a BLASTn search against genomic *C. maculifer* Mariner elements (see methods for full details) demonstrated that the percentage that have a best hit against a non-teleost species differed very little between library type (5.58% for the *Corydoras*-specific and 4.22% for the *D. rerio* library). Interestingly, it appears that the percentage of expressed Mariner elements with a best hit against a non-teleost species within the *C. maculifer* transcriptome was much higher than within the genome (17.14%), suggesting potential horizontally transferred elements may be more likely to be under purifying selection and retain their transposition ability (Zhang et al., 2020). The evolutionary impacts of horizontally transferred TEs are potentially wide-reaching (see Schaack et al., 2010), and thus their accurate annotation is important. More conservative testing across a wider range of elements would be required to fully investigate the role that library type has on the detection of horizontally transferred TEs, and is an important avenue to explore in the future.

The use of the *D. rerio* TE library led to skewed estimates of TE compositions. In particular, homology-based searching inflated the relative proportion of genomic SINE elements to a level equivalent to DDE DNA transposons, which was unexpected given other teleost species contain particularly SINE-depleted genomes (Gao et al., 2016; Shao et al., 2019) and potentially caused by a predisposition for homology-based searching against the detection of certain TE classes (e.g., DNA transposons). Furthermore, the majority of SINEs found by homology searching were represented by a single SINE element, which was not present in the *Corydoras*-specific library, suggesting a failure to comprehensively detect SINE elements during de novo library creation. This finding was, in part, expected: SINE elements have a propensity to be missed during de novo library creation because of

high sequence variation and lack of terminal repeats and supports a prediction made by Ou et al. (2019). Such compositional differences were not observed within transcriptomic sequences, possibly because expressed TEs (which are often younger) tend to have higher levels of sequence similarity (Lanciano & Cristofari, 2020). Over-reliance on homology-based searching may lead to similar inaccuracies during TE abundance and compositional estimates, particularly when working with organisms that are phylogenetically distant from a model organism in which a curated TE library exists.

The substantial increase in individual elements detected in the *Corydoras*-specific library compared to the *D. rerio* library raises the possibility of false discovery and/or fragmentation, whereby libraries contain multiple fragmented entries representing different regions of a single contiguous element (Flynn et al., 2020). Both false discovery and library fragmentation are common pitfalls associated with de novo pipelines (Flynn et al., 2020; Ou et al., 2019,). Without full manual curation, false discovery rate is difficult to assess; however, performance analysis of EDTA within the model rice *Oryza sativa* indicated that EDTA exhibits an overall false-positive rate of ~15% which, even in the unlikely absence of false-negatives, would not explain the degree of estimated TE abundance increase we observed in the *Corydoras* (Ou et al., 2019). We assessed fragmentation by (i) measuring the cumulative percentage abundance of TEs that mapped back to a single element, and (ii) plotting distributions of masked genomic TE lengths across each TE class. The cumulative curve created from the *Corydoras*-specific library suggested that it is likely to be more fragmented than the manually curated RefBase *D. rerio* library. However, when looking at the number of TEs that mapped back to a single element, the differences between the *Corydoras*-specific and *D. rerio* libraries were within the same order of magnitude, and at a similar level to the variation observed when comparing fragmentation levels between de novo generated libraries produced by different pipelines (Flynn et al., 2020). When comparing distributions of TE length between the *Corydoras*-specific and *D. rerio* libraries, the results were remarkably similar with a single anomalous peak within the

LINE class, which on further investigation proved to be evidence of misidentification/misclassification by EDTA and DeepTE. We have provided a framework for comparative analysis against a model species library which may be an efficient way to check de novo libraries for deviations, either in false discovery, fragmentation or missing elements.

When assessing estimated TE age distributions associated with both library types, we found that the *Corydoras*-specific library led to a reduction in average sequence divergence across multiple TE classes, suggestive of a more recent TE accumulation within the *Corydoras*. This mirrors findings in mammals and insects where the use of species-specific TE libraries also reduced relative TE ages (Platt et al., 2016). Analysis of other teleost genomes suggest that TE age distributions can vary considerably among species. Three waves of TE accumulation have been proposed in the evolutionary history of cichlid fish for example, whereas a single, recent insertion peak was identified in the piranha genome (*Pygocentrus nattereri)* (Brawand et al., 2014; Schartl et al., 2019). Furthermore, the variation in estimated sequence divergence associated with the *Corydoras*-specific library was larger than the *D. rerio* library, suggesting that the *Corydoras*-specific library was able to detect TEs of a greater age range. Taken together these findings further support the notion that the majority of TEs detected using the *D. rerio* library are probably those inherited from the common ancestor of the *Corydoras* and *D. rerio*. Additionally, we found that elements found within transcriptomic data had a greater probability of being less divergent than their genomic counterparts, further supporting the hypothesis that expressed TEs tend to be both younger and more intact, with a greater potential for active transposition.

4.7 Conclusion and Outlook

To conclude, we have combined two recent bioinformatic pipelines (EDTA & DeepTE) to generate a novel semi-automated de novo TE library for a non-model group of teleosts (*Corydoras*). We assessed the performance of this *Corydoras*-specific library against a distantly related but highly curated TE library (*D. rerio*). Across both species and sequence types, the use of the *Corydoras*-

specific TE library increased estimated transposon abundance between 2–3x and altered TE composition estimates. We stress that future work on non-model organisms will probably encounter substantial TE underestimates/classification biases if researchers are to rely heavily on homology-based TE detection. Furthermore, we demonstrate that TEs missed by homology methods are likely to be species-specific, and thus elements of most interest if the focal aim of a study is assessing lineage specific TE impacts. Furthermore, use of the de novo library reduced the estimated average sequence divergence/age distributions. This is likely to have important implications for researchers particularly interested in identifying elements that have recently proliferated within a lineage. Many of these TE-based traits varied across sequence type, with expressed TEs being estimated at lower abundance and exhibiting a younger average age. Finally, we provided an assessment of potential fragmentation associated with EDTA generated TE libraries, and whilst the *Corydoras*-specific library exhibited some fragmentation, it is within the same order of magnitude as the manually curated *D. rerio* library. A set of relatively small alterations highlighted by our fragmentation analysis could improve the *Corydoras*-specific library further. By providing both (i) a quantitative assessment of how library choice can influence numerous important TE-based metrics, and (ii) a stepwise pipeline (https://github.com/ellenbell/FasTE) for replication, we hope this study can provide a useful resource for all TE-based researchers, and particularly those who may be new to the field.

4.8 References

Alexandrou, M. A., Oliveira, C., Maillard, M., McGill, R. A. R., Newton, J., Creer, S., & Taylor, M. I. (2011). Competition and phylogeny determine community structure in Müllerian co-mimics. Nature, 469, 84–89.

Bergman C. M., & Quesneville, H. (2007). Discovering and detecting transposable elements in genome sequences. Briefings in Bioinformatics, 8(6), 382–392.

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

Bourque G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., Mager, D. L., & Feschotte, C. (2018). Ten things you should know about transposable elements. Genome Biology, 19(1), 1–12.

Brawand, D., Wagner, C. E., Li, Y. I., Malinsky, M., Keller, I., Fan, S., Simakov, O., Ng, A. Y., Lim, Z. W., Bezault, E., Turner-Maier, J., Johnson, J., Alcazar, R., Noh, H. J., Russell, P., Aken, B., Alföldi, J., Amemiya, C., Azzouzi, N., ... Di Palma, F. (2014). The genomic substrate for adaptive radiation in African cichlid fish. Nature, 513(7518), 375–381.

Cai, X. U., Chang, L., Zhang, T., Chen, H., Zhang, L., Lin, R., Liang, J., Wu, J., Freeling, M., & Wang, X. (2021). Impacts of allopolyploidization and structural variation on intraspecific diversification in Brassica rapa. Genome Biology, 22(166), 1–24.

Chen, W. J., Lavoué, S., & Mayden, R. L. (2013). Evolutionary origin and early biogeography of otophysan fishes (ostariophysi: teleostei). Evolution, 67(8), 2218–2239.

Clavijo, B. J., Garcia Accinelli, G., Wright, J., Heavens, D., Barr, K., Yanes, L., & Di-Palma, F. (2017). W2RAP: A pipeline for high quality, robust assemblies of large complex genomes from short read data, bioRxiv, 1–12.

Cowley, M., & Oakey, R. J. (2013). Transposable Elements Re-Wire and Fine- Tune the Transcriptome. PLoS Genetics, 9(1), 1–7.

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. Proceedings of the National Academy of Sciences, 117(17), 9451–9457.

Gao, B., Shen, D., Xue, S., Chen, C., Cui, H., & Song, C. (2016). The contribution of transposable elements to size variations between four teleost genomes. Mobile DNA, 7(4), 1–16.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., & Regev, A. (2013). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nature Biotechnology, 29(7), 644–652.

Groth, S. B., & Blumenstiel, J. P. (2017). Horizontal transfer can drive a greater transposable element load in large populations. Journal of Heredity, 108(1), 36–44.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. Bioinformatics, 29(8), 1072–1075.

Haig, D. (2016). Transposable element: Self-seekers of the germline, team- players of the soma. BioEssays, 38, 1158–1166.

Hoen, D. R., & Bureau, T. E. (2015). Discovery of novel genes derived from transposable elements using integrative genomic analysis. Molecular Biology and Evolution, 32(6), 1487–1506.

Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., Collins, J. E., Humphray, S., McLaren, K., Matthews, L., McLaren, S., Sealy, I., Caccamo, M., Churcher, C., Scott, C., Barrett, J. C., Koch, R., Rauch, G.-J., White, S., ... Stemple, D. L. (2013). The zebrafish reference genome sequence and its relationship to the human genome. Nature, 496(7446), 498–503.

Hu, Y., Colantonio, V., Müller, B. S. F., Leach, K. A., Nanni, A., Finegan, C., Wang, B. O., Baseggio, M., Newton, C. J., Juhl, E. M., Hislop, L., Gonzalez, J. M., Rios, E. F., Hannah, L. C., Swarts, K., Gore, M. A., Hennen-Bierwagen, T. A., Myers, A. M., Settles, A. M., Tracy, W. F., & Resende, M. F. R. (2021). Genome assembly and population genomic analysis provide insights into the evolution of modern sweet corn. Nature Communications, 12(1), 1–13.

Kennedy, R. C., Unger, M. F., Christley, S., Collins, F. H., & Madey, G. R. (2011). An automated homology-based approach for identifying transposable elements. BMC Bioinformatics, 12, 1–10.

Lanciano, S., & Cristofari, G. (2020). Measuring and interpreting trans-posable element expression. Nature Reviews Genetics, 21(12), 721–736.

Leggett, R. M., Clavijo, B. J., Clissold, L., Clark, M. D., & Caccamo, M. (2014). Next clip: An analysis and read preparation tool for nextera long mate pair libraries. Bioinformatics, 30(4), 566–568.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Available at: http://arxiv.org/abs/1303.3997

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

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

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., & Wang, J. (2012). SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. GigaScience, 1(18), 1–18.
Marburger, S., Alexandrou, M. A., Taggart, J. B., Creer, S., Carvalho, G., Oliveira, C., & Taylor, M. I. (2018). Whole genome duplication and transposable element proliferation drive genome expansion in Corydoradinae catfishes. Proceedings of the Royal Society B, 285, 1–10.
Michell, C., Wutke, S., Aranda, M., & Nyman, T. (2021). Genomes of the willow-galling sawflies Euura lappo and Eupontania aestiva (Hymenoptera: Tenthredinidae): a resource for the research on ecological speciation, adaptation and gall induction. G3, 11(5), 1–7.
Ou, S., Su, W., Liao, Y. I., 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., & Hufford, M. B. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biology, 20(1), 1–18.
Platt, R. N., Blanco-Berdugo, L., & Ray, D. A. (2016). Accurate transposable element annotation is vital when analyzing new genome assemblies. Genome Biology and Evolution, 8(2), 403–410.
Rogers, R. L., Zhou, L., Chu, C., Márquez, R., Corl, A., Linderoth, T., Freeborn, L., MacManes, M. D., Xiong, Z., Zheng, J., Guo, C., Xun, X. U., Kronforst, M. R., Summers, K., Wu, Y., Yang, H., Richards- Zawacki, C. L., Zhang, G., & Nielsen, R. (2018). Genomic takeover by transposable elements in the strawberry poison frog. Molecular Biology and Evolution, 35(12), 2913–2927.
Ruan, J., & Li, H. (2019). Fast and accurate long-read assembly with wtdbg2. bioRxiv. doi: https://doi.org/10.1101/530972
Schaack, S., Gilbert, C., & Feschotte, F. (2010). Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends in Ecology and Evolution, 25(9), 537–546.
Schartl, M., Kneitz, S., Volkoff, H., Adolfi, M., Schmidt, C., Fischer, P., Minx, P., Tomlinson, C., Meyer, A., & Warren, W. C. (2019). The piranha genome provides molecular insight associated to its unique feeding behavior. Genome Biology and Evolution, 11(8), 2099–2106.
Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. In Kollmar, M. (Ed.), Gene Prediction: Methods and Protocols (pp. 227–245). Springer New York.
Shao, F., Han, M., & Peng, Z. (2019). Evolution and diversity of transposable elements in fish genomes. Scientific Reports, 9(1), 1–8.
Slotkin, R. K., & Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. Nature Reviews Genetics, 8(4), 272–285.
Smit, A. F., & Hubley, R. RepeatModeler Open-1.0. Available from: http://www.repeatmasker.org
Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M., & Kelly, S. (2016). TransRate: Reference-free quality assessment of de novo transcriptome assemblies. Genome Research, 26(8), 1134–1144.
Sotero-Caio, C. G., Platt, R. N., Suh, A., & Ray, D. A. (2017). Evolution and diversity of transposable elements in vertebrate genomes. Genome Biology and Evolution, 9(1), 161–177.
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.
Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: An integrated tool for comprehensive microbial vari-ant detection and genome assembly improvement. PLoS One, 9(11), 1–14.
Wells, J. N., & Feschotte, C. (2020). A field guide to transposable elements. Annual Review of Genetics, 54, 1–23.
Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., & Schulman, A. H. (2007). A unified

classification system for eukaryotic transposable elements. Nature Reviews Genetics, 8(12), 973–982.
Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer- Verlag. https://ggplot2.tidyverse.org
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, H. H., Peccoud, J., Xu, M. R. X., Zhang, X. G., & Gilbert, C. (2020). Horizontal transfer and evolution of transposable elements in vertebrates. Nature Communications, 11(1), 1–10.

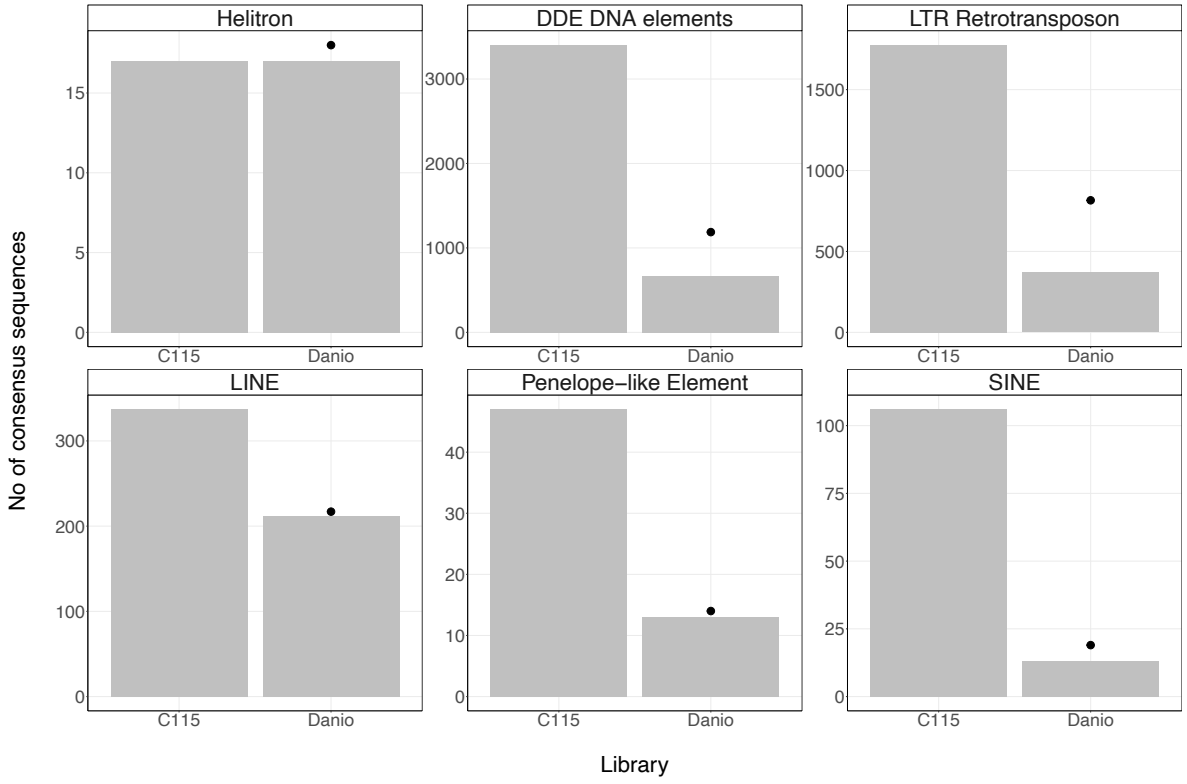
4.9 Supplementary Material and Appendices

Supplementary Table 4.1. *Corydoras fulleri* and *Corydoras maculifer* (Lineage 1) genome assembly metrics.

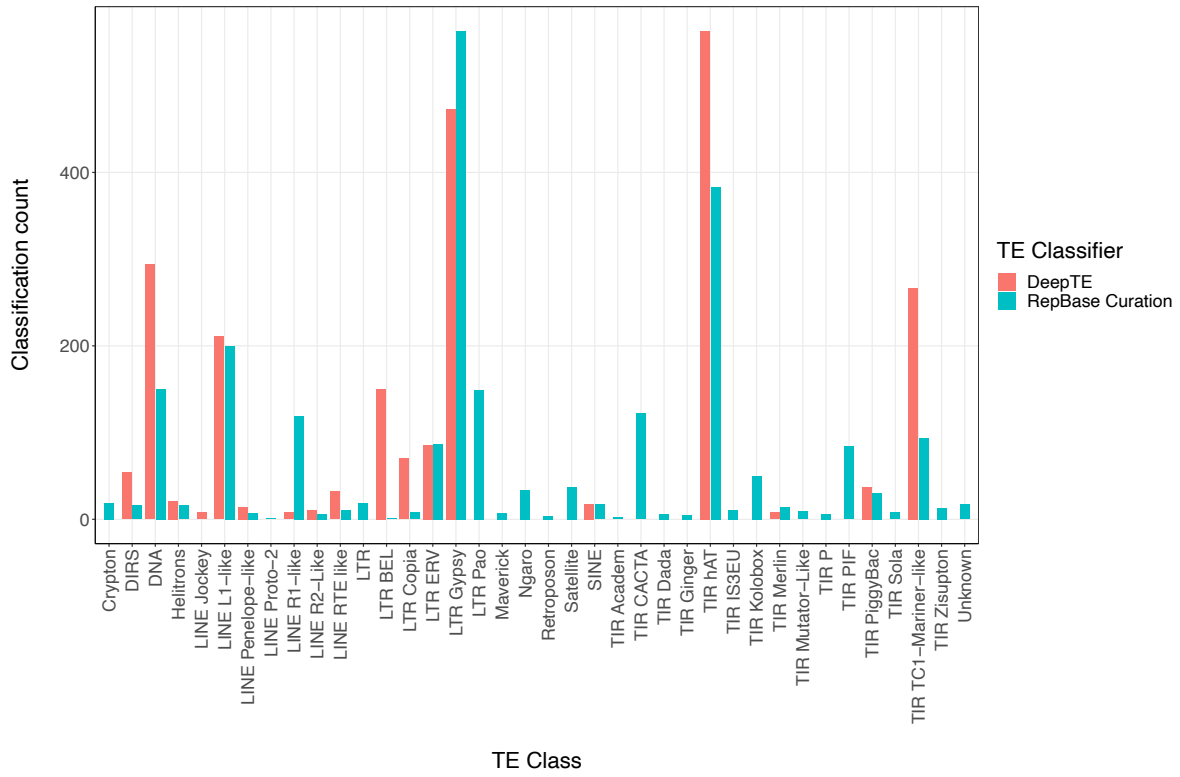
Metric	*Corydoras fulleri*	*Corydoras maculifer*
Contigs	2080	73,026
Total length (MB)	699.66	747.89
GC (%)	42.12	41.35
N50 (MB)	1.56	0.83
Complete Buscos (%)	90.2	88.2

Supplementary Table 4.2. Transcriptome metrics for *C. maculifer*

Metric	
Transcripts	66,579
Total length (MB)	39.46
GC (%)	46.43
N50 (bp)	818
TransRate Assembly Score	0.12



Supplementary Figure 4.1 Number of TEs (as individual library entries) from the RepBase Danio rerio library and the Corydoras-specific EDTA library identified in the *Corydoras fulleri* (formerly C115) genome. Points above D. rerio indicate the total number of TEs entries in the *D. rerio* RepBase database.



Supplementary Figure 4.2 TE counts in the *Danio rerio* library broken down by the classification assigned by DeepTE or RepBase (which has undergone a degree of manual curation).

Appendix I

The annotation of TEs using RepeatMasker ideally requires further parsing whereby masked outputs are further filtered to allow for a biologically accurate account of TE content. Whilst many genomic TE parse scripts exist, the additional issue of isoformic sequences within transcriptome data meant a new parse script was developed and written for this manuscript. This RepeatMasker TrInity based Parse Script (RM-Trips) was a Bash and R-based tool which filters RepeatMasker.out files. It consists of four key steps, which are outlined below. Each step's impact on estimated repeat abundance is given in Table A1. Prior to parsing, a Bash script was run which allowed for distinct repeat retrieval. In cases where the RepeatMasker out file contained multiple overlapping repeats, the element which had a lower scoring match whose domain partly includes the domain of the current match was removed.

Transposable element retrieval: Repeats not classified as TEs, including simple repeats, low complexity repeats, satellites, rRNA, snRNA, and artefacts, are removed. This step caused the greatest decrease in estimated repeat abundance - a drop from 3.51% to 1.87% (Table A1)

Merging of transposable elements: Identical repeats found on the same transcript which had the same orientation and a combined length which was less than or equal to their reference sequence were merged. As expected, but perhaps slightly counter intuitively, this step increases the percentage TE abundance but decreases TE-based hits (Table A1)

Isoformic TE removal: In cases where multiple copies of the same element are found within different transcript isoforms, only one is retained (at random), ensuring each TE corresponds to a unique genomic loci. This step was taken because we were primarily interested in intrinsic TE based activity. TEs co-transcribed with highly spliced genic regions will have highly abundant transcriptional representation. However, this is not be because of any activity level of the TE itself, rather their insertional position within the genome.

Small TE removal: TEs with a length less than 80bp were removed from the analysis. This aligns with the so-called "80-80-80" rule highlighted by Wicker (2007), whereby very short (<80bp) TEs are suggested to be removed from downstream analysis to avoid misclassification of short, random stretches of homology. As expected, removal of small TEs causes a bigger drop in the number of TE hits per MB than their % abundance (Table A1)

Table A1: Breakdown of how parse script RM trips impacts TE abundance estimates, which are given as both % of transcriptome and hits per MB. In this case RepeatMasker was run on the *C. maculifer* transcriptome and masked against the *D. rerio* RepBase library

				RM_trips							
RepeatMasker.out		Distinct repeats		*Transposable Elements (TEs)*		*Merged TEs*		*Isoform TE removal*		*> 80bp TEs*	
TE %	TE hits per MB	TE %	TE hits per MB	TE %	TE hits per MB	TE %	TE hits per MB	TE %	TE hits per MB	TE %	TE hits per MB
3.58	605.95	3.51	580.01	1.87	166.58	1.93	160.63	1.53	138.07	1.17	68.07

In summary, this parse script causes a 3x decrease in estimated TE abundance. Precisely, when the script was run on the *C. maculifer* output (after masking against the *Danio rerio* RepBase library), estimated TE abundance decreased from 3.58% to 1.17% (Table A1). This entire pipeline and its usage is available as an open-source GitHub page which can be found here https://github.com/clbutler/RM_TRIPS

Appendix II

The evolutionary impacts caused by the horizontal transfer of TEs are potentially wide reaching, yet overreliance on homology-based TE searching has been suggested to bias detection against recently horizontally transferred TEs, particularly because such elements will have been inserted since species-divergence from a shared ancestor (Platt et al., 2016). We therefore tested whether TE library type influences the ability to detect potentially horizontally transferred TEs. The Corydoras-specific library detected 17,655 Mariner elements from the genome of *C. maculifer* and 337 from its transcriptome. A protein based BLASTx search against the RepeatMasker peptide database found that 93% and 99% of respective sequences had a hit against a Mariner element. The *D. rerio* library detected 1,268 genomic and 180 transcriptomic Mariner elements, with the respective number of BLASTx Mariner hits being 98% and 99% respectively. Using this pool of extracted Mariner elements, we assessed the proportion of TEs which were potentially horizontally transferred (see methods of Chapter 5 for more details). Use of the Corydoras-specific library increased the proportion of potentially horizontally transferred Mariner elements within genomic sequences, although the difference was slight (5.58% for the Corydoras-specific and 4.22% for the D. rerio library) (Figure A1). Surprisingly, the reverse was true within transcriptomic sequences, with the Corydoras-specific library associated with the detection of fewer potentially horizontally transferred Mariner elements (17.14% for the Corydoras-specific and 27.91% for the *D. rerio* library). Across both library types, the majority of the potential horizontally

transferred Mariner elements (60%) within the *C. maculifer* genome were found to have a closest match against repeats belonging to *Xenopus tropicalis.* We also note that regardless of library type, a greater proportion of horizontally transferred Mariner elements were found within the *C. maculifer* transcriptome than genome, which may reflect the fact that recently introduced elements are more likely to be under purifying selection and retain their ability for active transposition (Zhang et al., 2020). We only tested a subset of TEs found within the Corydoras (Mariner DNA elements), and library choice may have different impacts regarding horizontal transfer detection of other TE types.

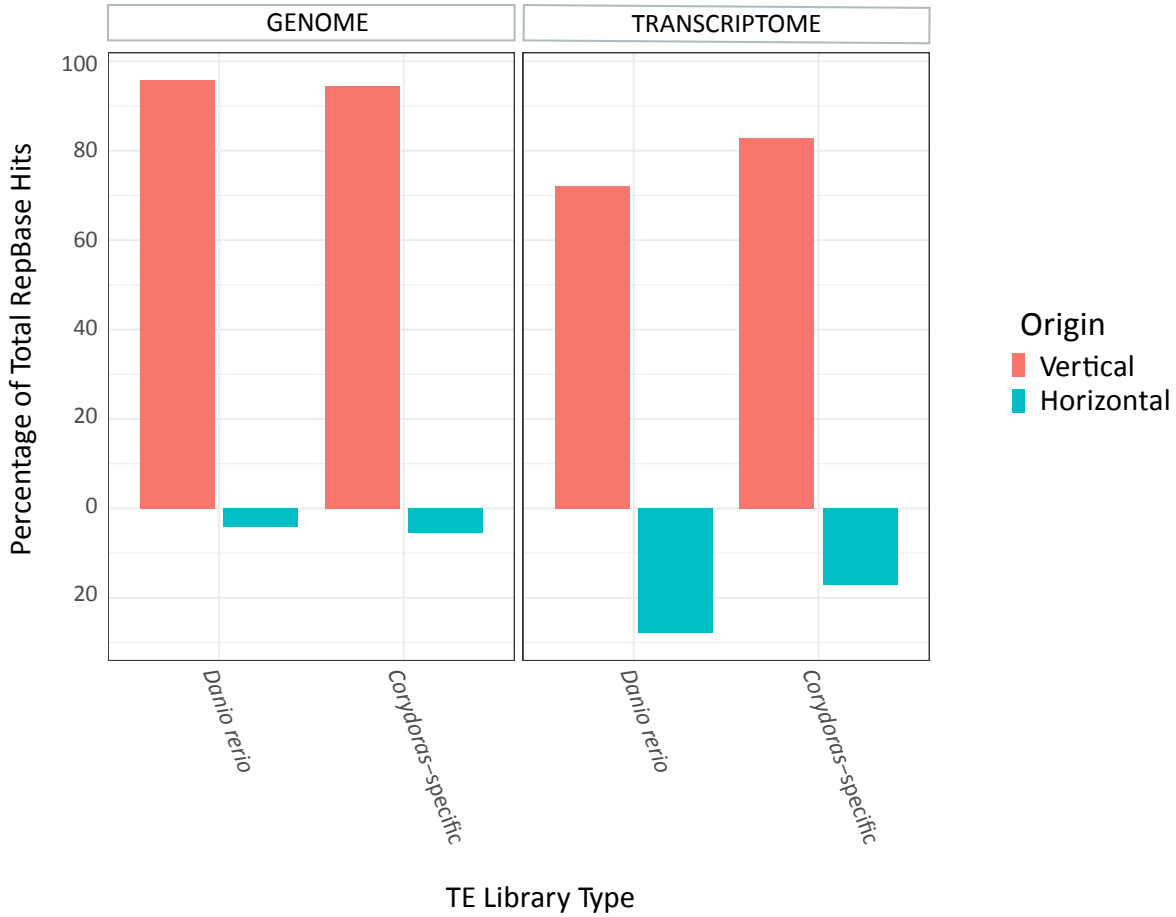


Figure A1 TE library type has marginal impacts on the proportion of horizontally transferred TEs which can be detected. The potential mode of inheritance (vertical and horizontal) is indicated for Mariner element found across both sequence types and using the *Danio rerio* and Corydoras-specific TE libraries. Plots are derived from *C. maculifer* sequence data.

5 Assessing the role of both polyploidy and horizontal transfer when characterising a potential transposable element proliferation event within Neotropical catfish (Corydoradinae)

5.1 Chapter Overview and Author Contributions

This chapter aims to provide a comprehensive description of the transcriptional TE landscape within the Corydoradinae. Primarily, this work investigated the potential role of both polyploidy and horizontal transfer during a putative TE proliferation event. Additional analysis included documenting the gene ontology (GO) terms associated with TE insertions, investigating the association between transcriptional TE abundance and genome size, and seeing whether there are biases regarding the TE classes which are expressed within the Corydoradinae.

The work from this chapter was conceived by Christopher L Butler, Ellen A Bell and Martin Taylor. All analysis was conducted by Christopher L Butler, unless otherwise stated below. RNA extraction and Trinity de-novo assembly of transcriptomes was conducted by Ellen A Bell and Claudio Oliveira. The Corydoradinae-specific TE library used was devised by Ellen A Bell and Christopher L Butler, as earlier described in Chapter 4. Finally, the extraction, assembly and running of RepeatMasker on three additional Corydoradinae genomes was conducted by Ellen A Bell. The chapter was written by Christopher L Butler, with contributions from Ellen A Bell and Martin I Taylor.

5.2 Abstract

Background: Transposable elements (TEs) are short stretches of DNA with the ability to replicate throughout a genome. Somewhat paradoxically, TEs regularly escape host silencing despite their propensity to disrupt genomic stability. A better understanding of the drivers of TE proliferation is therefore of great importance. In this study, we better characterise the impact of both (i) whole genome duplication and (ii) horizontal transfer on a previously reported TE proliferation event within a species-rich clade of Neotropical catfish (Corydoradinae).

Results:

1. Expressed TE content within the Corydoradinae are dominated by class II DNA transposons, particularly those belonging to hAT and Mariner families, though when accounting for phylogenetic signal no relationship between lineage genome size and expressed TE abundance was found. Expressed TE families are found in similar proportion to genomic TE families, suggesting an absence of a bias regarding the type of TE family which can be transcribed.
2. Phylogenetic shifts in expressed TE abundance coincide with a Corydoradinae-specific whole genome duplication, though this is somewhat dependent on the TE library used during annotation. Within polyploid Corydoradinae species, TE insertions within genic regions are enriched, with a Gene Ontology (GO) analysis suggesting TE insertions are frequently located within the transcripts of immune response genes.
3. A horizontally transferred Mariner element of unknown amphibian origin has inserted into the Matrix Metalloproteinase-13 (MMP-13) gene of multiple Corydoradinae species. Given MMP-13's role in bone morphogenesis, we hypothesise this insertion may have impacted Corydoradinae body shape evolution.

Conclusion: This study describes both the causes and impacts of TE proliferation across non-model species, bettering our understanding of their potential evolutionary roles.

Key Words: Phylogenetics; Horizontal Transfer; Gene Ontology; Polyploidy; MMP13

5.3 Background

Non-coding regions represent the majority of genomic content across the tree of life (Gregory 2005). A significant contributor to these regions are transposable elements (TEs); short stretches (100-10,000 bp) of DNA which possess the ability to replicate throughout a genome. Transposons exist in multiple forms, which can largely be characterised based on their mode of transposition, with class I retrotransposons replicating in a "copy and paste" mechanism, and class II DNA transposons replicating in a "cut and paste" mechanism (see Curcio & Derbyshire 2003; Wicker *et al.* 2007 for detailed TE classification). Originally described in Maize (McClintock 1950), TEs are increasingly recognised for their roles in chromosomal rearrangement (Lee *et al.* 2008; Chénais *et al.* 2012) phenotypic change (González *et al.* 2008; Hof *et al.* 2016) and speciation (Oliver & Greene 2011; Ricci *et al.* 2018; Serrato-Capuchina *et al.* 2018). Due to their propensity to disrupt genome stability, TEs are subject to multiple host silencing mechanisms, which can occur at either a pre-transcriptional or post-transcriptional level (Slotkin & Martienssen 2007; Choi & Lee 2020) Somewhat counterintuitively, TE proliferation events appear to be relatively common despite such host defence measures (see Hawkins *et al.* 2006; Charles *et al.* 2008; Plague *et al.* 2008). Consequently, TEs frequently reach high genomic abundance, constituting 50% of the human genome and up to 85% of the maize genome for example (Lander & IHGSC 2001; Stitzer *et al.* 2021). A better understanding of the processes which permit TE proliferation is therefore of great importance in appreciating the full evolutionary scope of TE-induced genetic and phenotypic change.

Many studies have identified the roles that population structure, life history and environmental stress can play during TE (re)-activation. For example, periods of drought led to the proliferation of retrotransposons within tomato *Solanum lycopersicum* (Benoit *et al.* 2019), whilst TE accumulation within founder populations of the highly invasive fire ant (*Cardiocondyla obscurior)* promotes

greater adaptability (Schrader *et al.* 2014). Furthermore, the interplay between large scale genome changes ('genome shocks') and TE proliferation has long been considered, particularly following a whole genome duplication (WDG) event (McClintock 1984). A WGD event occurs due to nondisjunction during meiosis and is typically associated with hybridisation (allopolyploidy), although may also occur within a single species (autopolyploidy). TE proliferation after a WGD may occur either directly (due to breakdown of TE silencing measures) or indirectly (due to either relaxed selection on paralogous genes or a reduction in effective population size) (Matzke & Matzke 1998; Vicient & Casacuberta 2017). Both these direct and indirect mechanisms are unlikely to be mutually exclusive, with each inducing different patterns of TE persistence. A direct breakdown in silencing may cause a discrete but temporary burst of TE activity, whereas reduced selection efficiency is likely to cause continuous TE accumulation until the rediploidisation process has been completed (Parisod *et al.* 2010). To date, empirical evidence of TE proliferation within polyploids has highlighted a spectrum of outcomes, ranging from (i) absence of increased TE activity (Hazzouri *et al.* 2008; Beaulieu *et al.* 2009; Smukowski Heil *et al.* 2021) (ii) restricted accumulation in certain TE families (e.g. Madlung *et al.* 2005)and (iii) widespread TE proliferation (e.g. Baduel *et al.* 2019). The commonality and taxonomic distribution of TE proliferation after a WDG event is therefore poorly understood.

An additional process which has been implicated in driving TE proliferation is the horizontal transfer of genetic content between organisms. Due to their inherent nature for transposition, TEs may be the most common type of genetic material to be horizontally transferred between eukaryotic species, with many known cases of horizontal transfer appearing to involve either parasitic or viral vectors (Peccoud *et al.* 2017). For example, host-parasite interactions involving the triatomine bug *Rhodnius prolixus* have led to horizontally transferred transposons (HTTs) within South American tetrapods, whilst parasitic nematodes are the likely origin of retrotransposon insertion within the common shrew *Sorex araneus* (Gilbert *et al.* 2010; Dunemann & Wasmuth 2019). The DNA Mariner

TE family appear to be particularly susceptible to horizontal transfer, which may be due to "blurry promoters" permitting their replication within naïve genomes (Zhang *et al.* 2020). Naive genomes may be unable to recognise and silence HTTs, which may lead to extensive TE proliferation. For example, numerous HTTs found within the strawberry frog *Oophage pumilo* are found at both high copy number and expression level (Rogers *et al.* 2018). Horizontal transfer of TEs may also transport non-TE genomic material, leading to novel gene formation (Pace *et al.* 2008) or TE assisted gene capture. For example, the transposon mediated horizontal transfer of a toxic *ToxA* gene between fungal wheat pathogens may cause direct negative consequences on crop yield (McDonald *et al.* 2019), whilst the transfer of LINE elements into Hydrophiinae snakes has coincided with their shift from terrestrial to aquatic environments (Galbraith *et al.* 2020). Nevertheless, it remains uncertain as to how common HTT-induced changes may be, with instances of adaptive HTTs possibly being restricted to a narrow taxonomic range or being overestimated because insertions which induce negative fitness effects are likely to be removed from the genome.

In this study, we aim to better characterise both the causes and impact of a previously reported TE proliferation within a teleost clade of Neotropical catfish (Corydoradinae). Though dominated by Class II DNA transposons, TE content across teleosts genomes is more diverse than within other vertebrates, having undergone recent proliferations in both cichlid and piranha species (Brawand *et al.* 2014; Chalopin *et al.* 2015; Schartl *et al.* 2019). Furthermore, teleost TE abundance is positively related to genome size and impacts GC% content (Gao *et al.* 2016; Shao *et al.* 2019; Symonová & Suh 2019). The Corydoradinae (family Callichthyidae) are a species rich group of catfish found in the freshwater environments of South America (Fuller & Evers 2005). Mitochondrial and RAD sequencing data have shown that the Corydoradinae belong to nine distinct phylogenetic lineages, though mtDNA lineages 6 and 9 form one monophyletic group under a nuclear topology (Marburger *et al.* 2018). Haplotype diversity and single nucleotide polymorphism (SNP) read ratios have indicated that there have been at least two WGD events within the evolutionary history of the

Corydoradinae (Marburger et al. 2018). Both methods support a recent WGD affecting nuclear lineage 6 and 9 (20-30MYA), with SNP data suggesting an earlier WGD event also occurred at the base of nuclear lineage 6, 7, 8 & 9 (35-44MYA) and haplotype data suggesting it occurred at the stem of nuclear lineage 2-9 (54-66MYA) (Marburger *et al*. 2018). There is also a significant increase in TE abundance associated with the most recent WGD, largely driven by the proliferation of Mariner DNA transposons (Marburger *et al*, 2018). Here, we investigate the role that whole genome duplication and horizontal transfer may have had in a such TE expansion within the Corydoradinae.

Using nine newly assembles transcriptomes belonging to multiple Corydoradinae species, four long-read (PacBio) genomes, as well as a Corydoradinae-specific TE library (see Bell *et al*. 2022), we aim to provide a comprehensive review of the expressed TE landscape within this species rich group of catfish. Expressed TEs may represent fully autonomous elements, or more likely, partial elements that have become incorporated into the transcripts of other genes (Lanciano & Cristofari 2020; Chang *et al.* 2022). Specifically, we test whether an expansion in transcriptional TE abundance (if any) coincides with the phylogenetic position of the Corydoras-specific WGDs and the role that HTTs have had on such expansion. Furthermore, a 'multi-omic' approach allows us to investigate (i) the location of TE insertions with respect to ploidy level (with the expectation that TEs may insert into coding regions of duplicated genes more frequently if they are under more relaxed purifying selection), (ii) the gene ontology (GO) terms associated with such insertions, (iii) the relationship between genome size and transcriptome TE content and finally (iv) whether certain TE classes are preferentially expressed.

5.4 Materials and Methods

Corydoras transcriptome assembly

An existing transcriptome of Lineage 1 species *Corydoras maculifer* (Bell *et al*. 2022, GenBank accession GJAY00000000.1) was used in this study, in addition to eight new de-novo transcriptome assemblies which were assembled from short read Illumina based sequencing. Namely, these were

Aspidoras fuscoguttatus (Lineage 2), *Scleromystax prionotus* (Lineage 3), *Corydoras hastatus* (Lineage 4)*, Corydoras elegans* (Lineage 5)*, Corydoras paleatus* (Lineage 6)*, Corydoras aeneus* (Lineage 7)*, Corydoras haraldschultzi* (Lineage 8) and *Corydoras araguaiaensis* (Lineage 9). RNA extraction (TRIzol Plus RNA Purification Kit) was conducted on adult somatic muscle tissue, and the cDNA library was then built using a TruSeq RNA Sample Prep kit (Illumina, Inc). cDNA was sequenced using paired-end sequencing on an Illumina HiSeq, with the average sequencing depth of all nine samples being 13.33 million paired reads (ranging from 9.27 to 20.67 million). The libraries were then demultiplexed and cleaned using Trimmomatic (version 0.2.36, Bolger *et al.* 2014) and assembled using Trinity (version 2.6.9, Grabherr *et al.* 2011). Transcriptome assembly quality was assessed using TransRate (version 1.03,Smith-Unna *et al.* 2016) (Supplementary Table 5.1). Transcriptional library preparation and sequencing was performed by the Animal Biotechnology Laboratory of Esalq/Piracicaba.

Corydoradinae genome assembly

In addition to the nine transcriptomes, we also compared transcriptional TE content against genomic TE content using four additional Corydoradinae genomes. Namely, these included an existing *Corydoras fulleri* (Lineage 1) genome (Bell *et al*. 2022) (GenBank accession PRJNA706371), as well as an *Aspidoras CW52* (Lineage 2) (undescribed species), *Corydoras nijsseni* (Lineage 5) and finally a *Corydoras metae* (Lineage 9) genome. The genomes of these additional Corydoradinae species were assembled in similar fashion to that of *C. fulleri* (as described in Bell *et al*. 2022). Briefly DNA was extracted using ThermoFisher Scientific PureLink Genomic DNA mini kit (by ThermoFisher Scientific) for Illumina Hiseq and Qiagen's MagAttract HMW DNA Kit for PacBio sequencing. Sequencing was performed on one Hiseq lane and two PacBio Sequel cells using 300 bp paired-end reads. Genome assembly for these species was conducted using wtdbg2 (v 2.5) and polished using a wtdbg2-racon-pilon.pl v04 script (https://github.com/schellt/wtdbg2-racon-pilon; Ruan & Li 2019) Assembly statistics for these four genomes are provided in Supplementary Table 5.2.

Transposable Element annotation

This study used a Corydoradinae-specific TE library which has previously been created on a Lineage 1 species genome (*Corydoras fulleri)* (Bell *et al.* 2022) (GenBank accession PRJNA706371). The repeat identifier Extensive de novo TE Annotator (EDTA) (Ou *et al.* 2019) and repeat classifier DeepTE (Yan *et al.* 2020) were utilised to generate a Corydoras-specific TE library (as described by Bell *et al*. 2022). This custom library was then used for all subsequent downstream TE annotation, which was performed using RepeatMasker (RM; version 1.332), utilising the NCBI/RMBLAST (version 2.6.0+) search engine. Both transcriptomic and genomic TE annotations were subject to a custom parse script (RM_TRIPS), which is publicly available and fully described at (https://github.com/clbutler/RM_TRIPS) (Bell *et al*. 2022). To standardise for different assembly lengths, TE abundance is given as a percentage of total sequence (bp) length.

Relationship between genome size and transcriptional TE content

Average C value (DNA content in picograms per haploid nucleus) within each nuclear Corydoradinae lineage was obtained from Marburger *et al*. 2018. To account for the influence of phylogenetic signal (i.e. covariance between closely related species), the relationship between lineage genome size and transcriptional TE content was assessed using a phylogenetic generalised least-squares (PGLS) approach (Felsenstein, 1985). We varied the strength of phylogenetic signal through altering Pagel's lambda (λ) between 0 (no phylogenetic signal, equivalent to an ordinary least-square regression) and 1 (strong phylogenetic signal, strict Brownian motion process). We also used the 'nlme' package (Pinheiro & Bates, 2022) within R to select an optimal (λ) value using a maximum likelihood approach.

Detection of phylogenetic shifts in transcriptional TE abundance

We used an Ornstein-Uhlenbeck model implemented within the R package 'L1ou' (Khabbazian *et al.* 2016) to detect phylogenetic shifts in transcriptional TE abundance across different Corydoradinae

species. Confidence support values for each potential phylogenetic shift were generated using a non-parametric bootstrap test in which phylogenetically uncorrelated standardised residuals for each node were assessed and sampled with replacement 100 times. Analysis was conducted on both mtDNA and nuclear Corydoradinae phylogenies to account for the topology discordance when using mtDNA (non-monophyletic lineage 6 + 9) and nuclear (monophyletic lineage 6 + 9) data. The RAD-based phylogeny was converted into an ultrametric tree (i.e. a phylogeny where edge lengths represent time) using the 'chronis' function within the R package 'ape' (Paradis *et al.* 2004), which uses a penalised maximum likelihood method to estimate divergence times (the mtDNA tree was already ultrametric). To account for a possible inflation bias associated with using a Corydoradinae library generated on a Lineage 1 genome (*C. fulleri)* we also ran our TE annotations in identical fashion using a *D. rerio* TE library (RepBase, 2018). Finally, we used the transcriptional (mRNA) sequence of the channel catfish (*Ictalurus punctatus* IpCoco_1.2; Liu *et al.* 2016) as an outgroup. In this instance, TE annotation was conducted in identical fashion to that of the Corydoradinae species, but against a species-specific *I. punctatus* TE library which was created by running EDTA and DeepTE on the publicly available *I. punctatus* genome (GCA_001660625.1_IpCoco_1.2; Liu *et al*. 2016).

Gene ontology (GO) term enrichment and TE insertion location

To assess the broad-scale biological impacts of TE activity across the Corydoradinae, we conducted a Gene Ontology (GO) term enrichment analysis of protein coding transcripts containing a TE (i.e chimeric transcripts) using the Trinotate (v3.1.1) pipeline, an annotation suite designed for the functional annotation of transcriptomes, particularly those belonging to non-model organisms (Bryant *et al.* 2017). Protein coding transcripts were identified using TransDecoder v.5.5.0 (https://github.com/TransDecoder/TransDecoder) which uses log likelihood and a minimum length setting of 100 amino acids to identify open reading frames (ORF). Resulting peptide sequence were then subject to blastx seaches against the UniProt/Swiss-Prot protein database, and the resulting protein hits were automatically assigned biological GO terms through Trinotate. For each

Corydoradinae species, the frequency of GO terms associated with a protein coding transcript containing a TE was standardised by total GO term occurrences across all protein coding transcripts. GO terms that occurred at very low frequency (< 5 per species) throughout the transcriptome were filtered out of our analysis, as were GO terms associated with the coding region of a TE insertion (e.g GO:0006313 "transposition, DNA-mediated", GO:0032197 "transposition, RNA-mediated", GO:0032199 “reverse transcription involved in RNA-mediated transposition") as these transcripts likely represented autonomous, rather than chimeric, TEs. For each species, the top 10 most enriched GO terms associated with transcripts containing a TE were retained. Furthermore, the TransDecoder coordinates of fully intact ORFs (i.e. spanning a start and stop codon) were also used to identify the mean number of TE insertions within gene coding regions for each Corydoradinae species. Where multiple predicted ORFs were given for a single transcript, only the best (maximum log likelihood score) was retained. Corydoradinae species were assigned a ploidy status based on previous inference about the phylogenetic positioning of a whole genome duplication event (SNP vs haplotype data).

Identification of potentially horizontally transferred Mariner elements

The sequences of each Mariner element annotated across the nine Corydoradinae transcriptomes were extracted to test for inferences of horizontal transfer. Mariner elements were chosen over other types of TEs due to their apparent propensity for horizontal transfer within teleosts and their apparent role in the Corydoradinae specific TE proliferation event (Marburger *et al.* 2018; Zhang *et* al. 2020). To ensure chimeric transcripts were not biasing potential instances of horizontal transfer, Mariner sequences were only retained if they were likely autonomous (defined here as a TE representing at least 80% of total transcript length). We employed a framework previously used by Rogers *et al*, (2018), whereby Mariner elements were considered a horizontal transfer candidate if their sequence had (i) a best match (lowest E value) against a non-teleost species and (ii) a greater than 2% sequence similarity than the respective best teleost hit.

Phylogeny of a potentially horizontally transferred Mariner-17 element

A 'Mariner-17' element which we considered to be a good horizontal transfer candidate (see results) was identified across numerous Corydoradinae species. A curated full-length Corydoradinae consensus sequence (Mariner-17_Cory) for this TE was generated by extracting all sequences across all nine transcriptomes with a RepeatMasker hit to the 'Mariner-17' element (Mariner-17_DR using the *D. rerio* RepBase and TE_00001723 using the Corydoradinae-specific EDTA TE library). These extracted sequences were then aligned using Fast Fourier Transform (MAFFT v7.471; Katoh *et al.* 2002) and a parameter which automatically (--auto) selects an appropriate accuracy-oriented alignment based on the input data size. Fragmented Mariner-17 elements were mapped back to its closest homologous full length RepBase entry (TC1-5_Xt) in MAFFT by using the --addfragments function. A consensus sequence was then generated using SeaView v5.0.4 (Gouy *et al.* 2010). We combined and aligned (as described above) the Mariner-17 Corydoradinae consensus with three related RepBase homologues, namely TC1-5_Xt (*Xenopus tropicalis*), Mariner-17_DR (*D. rerio*) and Mariner-12_EL (Northern Pike; *Esox lucius*). A phylogeny with 1,000 UFBoot bootstrap replicates was subsequently generated using IQ-TREE v1.6.12 (Minh *et al.* 2020), using the HKY+F (Hasegawa *et al.* 1985) substitution model selected by ModelFinder (Kalyaanamoorthy *et al.* 2017) The sequence divergence of this Mariner-17_Cory element across each Corydoradinae transcriptomes was assessed against (i) the sequence divergence of every other Mariner element with a consensus within the Corydoradinae-specific TE library and (ii) the Tc1-5_Xt homologue within the *X. tropicalis* transcriptome (GCF_000004195.4_UCB_Xtro_10.0; Mitros *et al.* 2019).

MMP13 transcript and genome extraction

Each transcript identified as MMP13 by the Trinotate pipeline, including those belonging to four additional species transciptomes (*C. nattereri* – Lineage 6, *C. schwartzi* – Lineage 9, *C. julli* – Lineage 9 and *C. cruziensis* – Lineage 9), were extracted. These additional transcripts were assembled in

identical manner to the other Corydoradinae transcriptomes but were not included in the overall TE analysis due to a lower TransRate quality score. In cases where multiple isoforms were retrieved, only the longest isoform was retained for further analysis. Each transcript, along with two *I. punctatus* MMP13 paralogues (MMP13-a; *ENSIPUT00000032163* and MMP13b; *ENSIPUT00000021811*), were subsequently aligned using MAFFT and a gene tree was created using IQ-TREE. The substitution model chosen by ModelFinder was TPM2u+F+I+G4. MMP13 gene regions were also retrieved from the genomes of *C. fulleri* (Lineage 1) and *C. metae* (Lineage 9) by running a blastn search using the *C maculifer* (Lineage 1) MMP13-a transcript. Gene structure was determined using GMAP v.2021-08-25, which aligns and maps cDNA sequences to a genome even in the presence of substantial sequence polymorphisms (Wu & Watanabe 2005). MMP13 gene structure and corresponding TE annotations were subsequently visualised in Interactive Genomics Viewer v2.13.0 (Robinson *et al.* 2011).

All bioinformatic pipelines for this analysis were conducted within R Version 3.6.0 (R Core Team, 2021). All figures were produced using the ggplot2 package in R (Wickham, 2016). Phylogenetic trees were viewed and displayed within FigTreev1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/). Where appropriate, variance statistics are given as ± one standard error.

5.5 Results

TE Content

Transcriptional TE content across the Corydoradinae species is dominated by Class II DNA elements

Across the Corydoradinae, transcriptional TE content ranged from 4.68% in *C. maculifer* (Lineage 1) to 6.87% in *C. hastatus* (Lineage 4). Transcriptional TE content was higher in all Corydoradinae species when compared to the channel catfish *I. punctatus* (3.22%) (Supplementary Table 5.3) Class II DNA transposons were more abundant than Class I retrotransposons within the Corydoradinae (Supplementary Table 5.3). This was largely driven by an abundance of Mariner and hAT elements, which together made up on average ~60% of total transcriptional TE content (Figure 5.1). Whilst

overall less abundant, retrotransposons still represented between ~1.5-1.9% of transcript sequences within the Corydoradinae, with Gypsy LTR and LINE elements being the most abundant class II transposon type (Figure 5.1). The transcriptional TE landscape was largely consistent across different Corydoradinae lineages, with no TE family being present in one species but absent in another (Figure 5.1). Furthermore, the Corydoradinae TE landscape largely aligned with that of their close relative *I. punctatus,* though LINE elements were more abundant within the transcripts of Corydoradinae species (Figure 5.1).

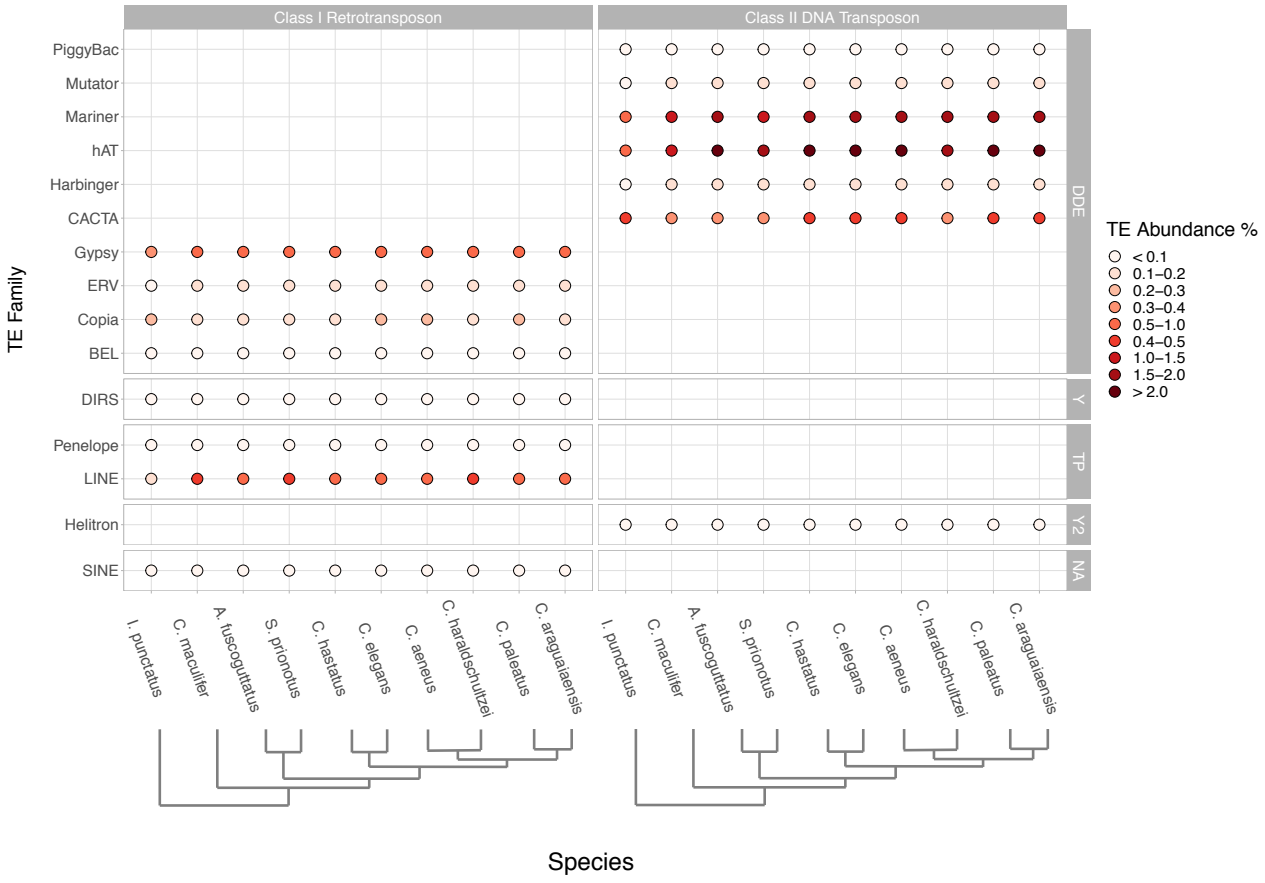


Figure 5.1 Bubbleplot of transcriptional TE content across nine Corydoradinae species and the channel catfish (*I. punctatus*). TE families are divided into Class I Retrotransposons and Class II transposons (columns) as well as transposase family (rows). TE abundance is given as a percentage of total transcriptome length and illustrated using a red colour scale. The nuclear Corydoradinae phylogeny is given on the X axis. TIR = terminal inverted repeat, LTR = long terminal repeat, Y = tyrosine transposase, TP = target-primed retrotransposition, SS = self-synthesising transposition, Y2 = rolling circle transposition & NA = non-autonomous transposition.

When accounting for phylogenetic signal there is no significant relationship between Corydoradine genome size and transcriptional TE abundance. .

There was a positive relationship between lineage genome size and transcriptional TE content within the Corydoradinae (Figure 5.2), however the degree of statistical significance depended on the PGLS model used. When using an ordinary least square regression, which does not account for phylogenetic structure (λ = 0), the relationship between genome size and transcriptional TE content was statistically significant (Table 5.1). However, when accounting for both a strict (λ = 1) and maximum likelihood phylogenetic structure (λ = ML) the relationship between genome size and transcriptional TE content was present, but it was no longer statistically significant at the level of $P < 0.05$ (Table 5.1).

Table 5.1 Statistical summary of the relationship between average Corydoradinae genome size per lineage and transcriptional TE abundance under three different phylogenetic least squared square models; an ordinary least squared regression (λ = 0), a strict Brownian-motion phylogenetic structure (λ = 1), and a maximum likelihood approach to estimating Pagels lambda (λ = ML). ML = maximum likelihood.

PGLS Model	Log-likelihood	Slope	*t*-value	P value
λ = 0	-7.42	0.55	2.88	0.024 (*)
λ = 1	-7.01	0.51	2.33	0.052
λ = ML	1.16	0.66	2.34	0.052

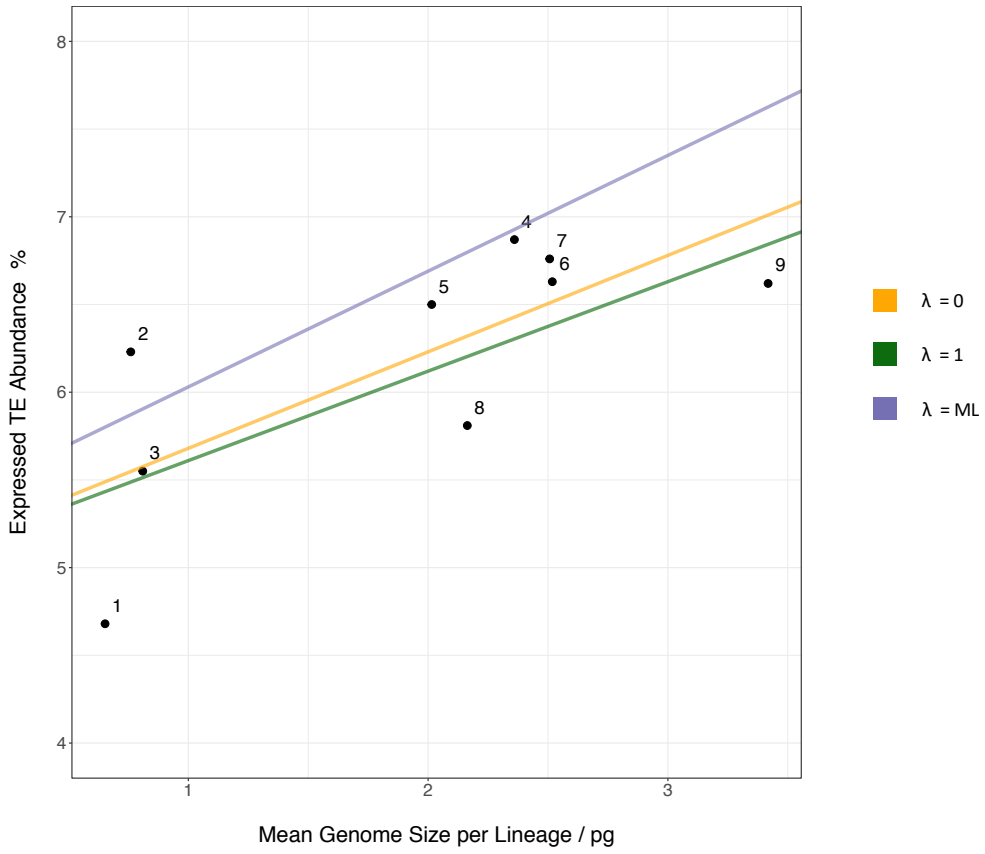


Figure 5.2 The relationship between mean genome size per Corydoradinae lineage and transcriptional TE abundance was assessed using a phylogenetic generalised least-squared approach (PGLS). The degree of phylogenetic signal (Pagel's Lambda, λ) was varied between 0 (an ordinary-least squared approach) and 1 (maximum species co-variance with respect to phylogeny). Finally, a third λ value was selected using a maximum likelihood approach. Numeric labels represent nuclear Corydoradinae lineage. "ML" = maximum likelihood. "pg" = picograms.

There is no bias in transcriptional vs genomic Corydoradinae TE content

Comparison between genomic and transcriptional TE composition was conducted across four different Corydoradinae lineages. As expected, DNA transposons were the most abundant element type across both sequence types, making up an average of 65.7% of total genomic TE sequences and 70.20% of total transcriptional TE sequences (Figure 5.3). TE sequences that did not consist of DNA transposons were almost entirely Class II LTR and LINE elements. Non-autonomous SINE elements were present at very low abundances across both sequence types, making up an average of 1.08% of total genomic TE sequences and 0.70% of total transcriptional sequences respectively (Figure 5.3).

We tested whether there was a bias in expressed TE content, which may occur if certain genomic TE classes are preferentially transcribed over others using a chi-squared goodness of fit (X^2) test across all lineages combined, with a null-hypothesis that genomic and transcriptional TE class proportions had an exact 1:1 ratio. Observed transcriptional TE class composition did not significantly differ from genomic TE class composition (Pearson's $X^2 = 11.50$, df = 15, P = 0.72), suggesting that there was no bias in transcriptional vs genomic TE class composition. We also conducted a chi-squared goodness of fit test on each lineage individually, with all four producing a non-significant result (Supplementary Table 5.4a)

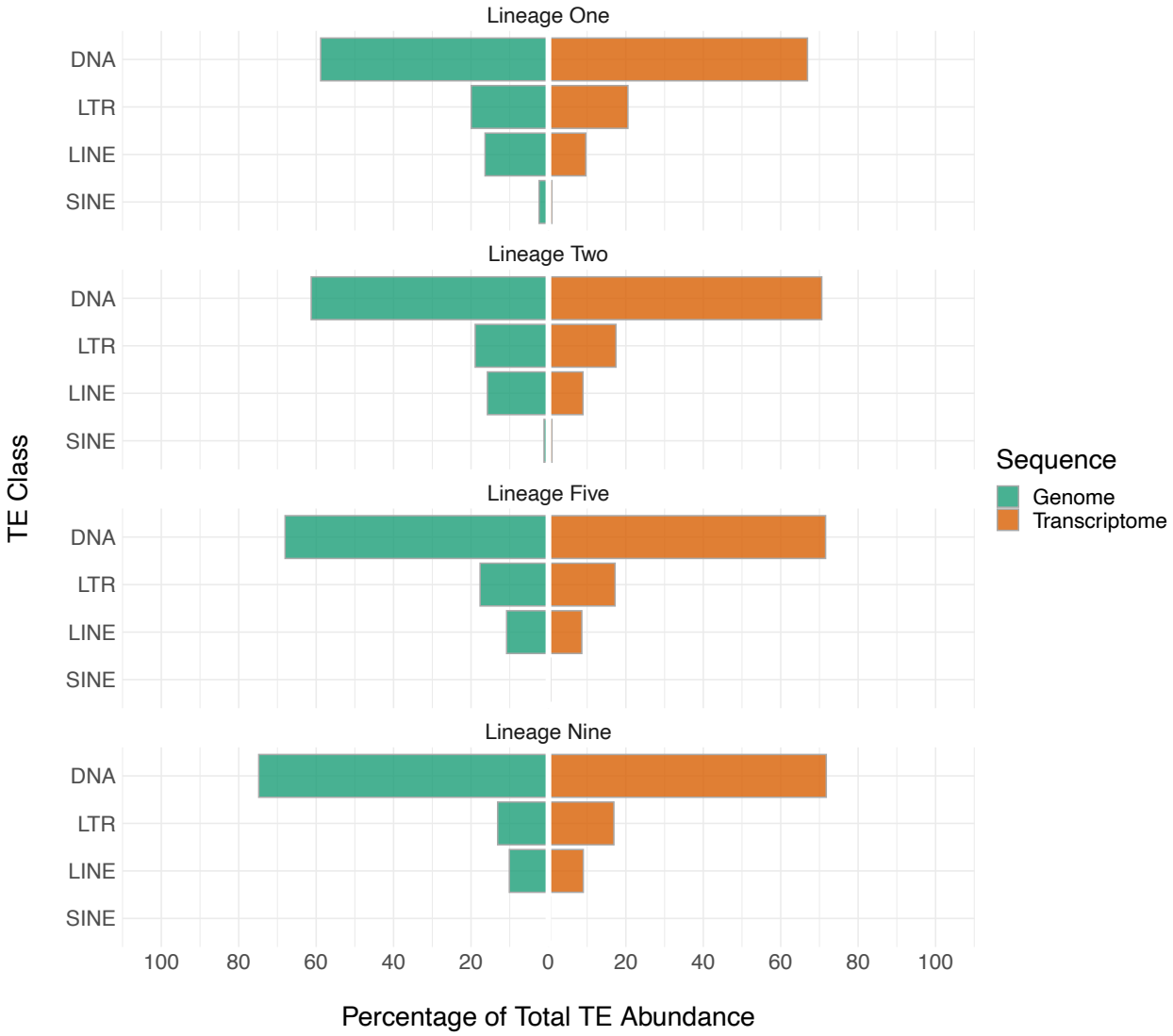


Figure 5.3 Comparison between transcriptional and genomic TE content across four Corydoradinae lineages, with TE abundance given as a percentage of the total TE length of each sequence type. TEs were divided into four main classes: DNA elements, LTR elements, SINEs and LINEs. Comparisons were made between Corydoradinae species belonging to nuclear Lineage 1 (*C. maculifer* transcriptome vs *C. fulleri* genome), Lineage 2 (*A. fuscoguttatus* transcriptome vs A. CW52 genome), Lineage 5 (*C. elegans* transcriptome vs *C. nijsseni* genome) and Lineage 9 (*C. araguaiaensis* transcriptome vs *C. metae* genome).

Given DNA elements are the most abundant TE class within the Corydoradinae, we also investigated whether there were finer biases in transcriptional vs genomic TE compositions across specific Class II DNA transposon families. Across the six DNA element family's present across every species, none were consistently over or under expressed across all four Corydoradinae lineages (Figure 5.4). Other than the highly abundant hAT and Mariner elements, CACTA elements made up ~10% of total DNA element abundance, whilst Mutator, Harbinger and especially PiggyBac elements were present at very low abundance (Figure 5.4). There was no significant combined lineage differences in observed transcriptional DNA element family composition versus genomic DNA element family composition (Pearson's $X^2 = 16.78$, df = 23, P = 0.82). As before, we also conducted a chi-squared goodness of fit test on each lineage individually, with all four producing non-significant results (Supplementary Table 5.4b).

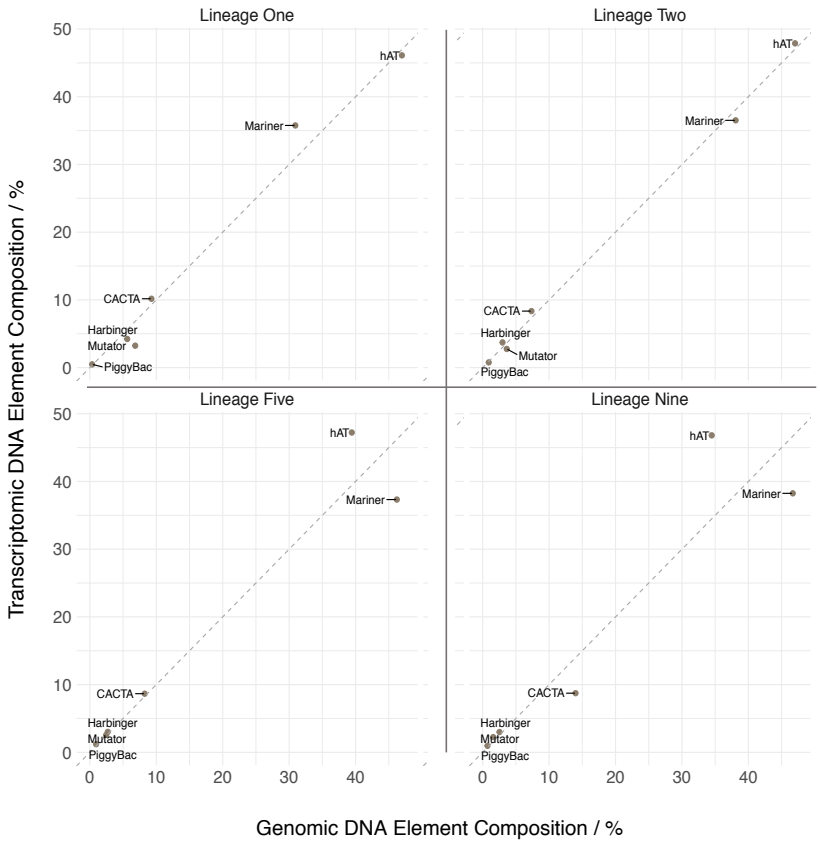


Figure 5.4 Comparison between transcriptional and genomic Class II DNA transposon content across four Corydoradinae lineages, with TE abundance given as a percentage of the total DNA transposon length of each sequence type. DNA families located on the dotted line (y=x) are found in equal proportion across genomic and transcriptomic sequences, whilst those above the line are 'over-expressed' and those under the line are 'under-expressed'. Comparisons were made between Corydoradinae species belonging to nuclear Lineage 1 (*C. maculifer* transcriptome vs *C. fulleri* genome), Lineage 2 (*A. fuscoguttatus* transcriptome vs A. CW52 genome), Lineage 5 (*C. elegans* transcriptome vs *C. nijsseni* genome) and Lineage 9 (*C. araguaiaensis* transcriptome vs *C. metae* genome).

Shifts in TE abundance and distribution

A potential phylogenetic increase in transcriptional TE content occurred at the ancestor of Lineage 4-9, depending on the TE library type used

A significant shift in TE abundance was identified at the base of lineage 4-9 in both the nuclear and mtDNA phylogenies when using the *Danio* TE library, with a bootstrap support value of 100 and 96 respectively (Figure 5.5a). However, when using the Corydoradinae-specific TE library a shift in TE abundance was no longer supported using either the nuclear or mtDNA phylogeny (Figure 5.5b). We investigated the possibility that this finding reflected an upward bias in TE abundance within 'early branching' Corydoradinae lineages, particularly considering the Corydoradinae-specific TE library was

generated using a Corydoras species belonging to Lineage 1 (*C. fulleri*). In support of this hypothesis, the increase in TE abundance associated with the switch from using a *Danio* to Corydoradinae-TE library was greatest in Lineage 1 species *C. maculifer* (4.18x) than other Corydoradinae species (Table 5.2).

Table 5.2 TE abundance (% total transcriptome length) as estimated when using the *Danio* and Corydoradinae-specific TE libraries. The abundance increase associated with using the Corydoradinae library is given within the final column.

Species	TE Abundance / %		Abundance Increase
	Danio Library	Corydoradinae Library	
C. maculifer (L1)	1.12	4.68	4.18 x
A. fuscoguttatus (L2)	2.17	6.23	2.87 x
S. prionotus (L3)	1.84	5.55	3.02 x
C. hastatus (L4)	3.57	6.87	1.92 x
C. elegans (L5)	3.20	6.50	2.03 x
C. aeneus (L7)	3.01	6.76	2.25 x
C. haraldschultzei (L8)	2.77	5.81	2.10 x
C. araguaiaensis (L9)	3.20	6.60	2.06 x
C. paleatus (L6)	3.58	6.63	1.85 x

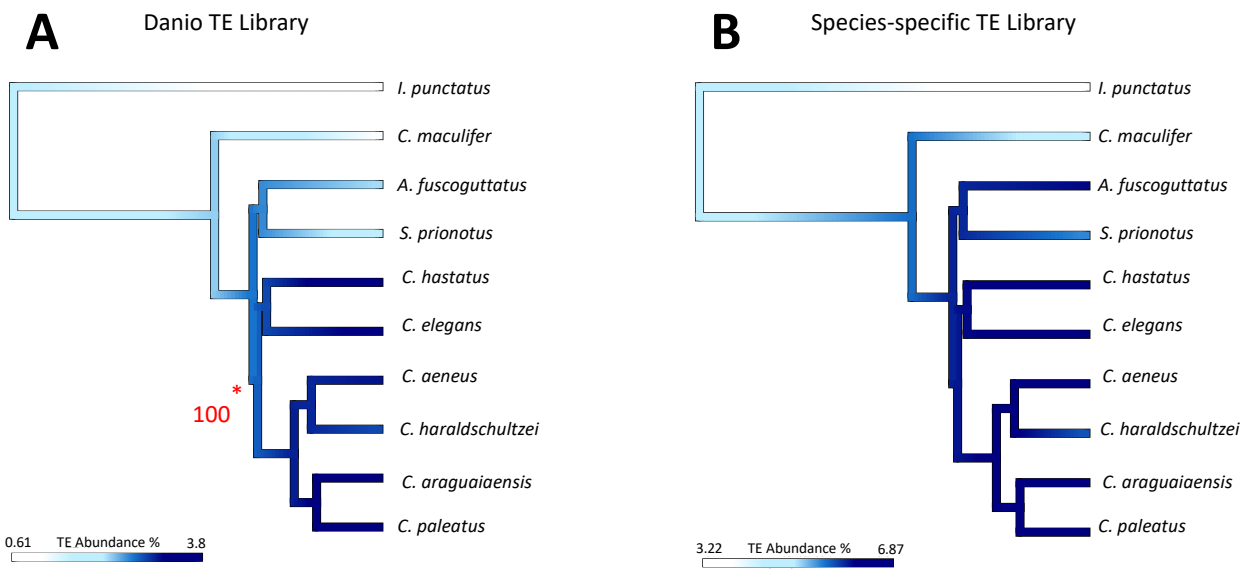


Figure 5.5 Nuclear Corydoradinae phylogeny with transcriptional TE abundance (%) depicted in blue; **A**, when using the D. rerio RepBase entry and **B,** when using species-specific TE libraries (Corydoradinae library for *Corydoras* species and *Ictalurus* library for the outgroup). Significant phylogenetic shifts in transcriptional TE abundance are indicated with a red asterisk, with an associated bootstrap support value also given.

The number of TE insertions located within open reading frames (ORFs) was significantly higher within polyploid than non-polyploid Corydoradinae species.

The number of TE insertions within the open reading frames (ORF) of each Corydoradinae transcriptome was calculated to assess if varying ploidy state impacted the frequency of TE insertions within gene coding regions. Due to the discrepancy in the timing of the earliest Corydoradinae WGD event, two scenarios were imagined. In all cases, *C. araguaiaensis* and *C. paleatus* (Lineage 6 + 9) were assigned 'recent polyploid' and *C. maculifer* (Lineage 1) was assigned diploid. The first alternative is that supported by haplotype data, in which all other Corydoradinae species were assigned 'polyploid'. The second alternative is that supported by SNP data, in which *A. fuscoguttatus* (Lineage 2)*, S. prionotus* (Lineage 3)*, C. hastatus* (Lineage 4) and *C. elegans* (Lineage 5) were assigned 'diploid', and *C. aeneus* (Lineage 7) and *C. haraldschultzei* (Lineage 8) assigned 'polyploid'. We report a significant effect of ploidy level on the mean number of TE insertions per

ORF, both in the haplotype supported scenario (Kruskal-Wallace test: H = 62.53, df = 2, P < 0.001) and the SNP supported scenario (Kruskal-Wallace test: H = 14.12, df = 2, P < 0.001). A post-hoc pairwise comparison using Dunn's test on the haplotype supported scenario indicated that the mean number of TE insertions (per 2kb) within ORFs was significantly lower in the diploid species ($\bar{x}$ = 0.055 ± 0.006) than either the recent polyploid ($\bar{x}$ = 0.082 ± 0.003, Z = -7.34, P < 0.001) or polyploid species ($\bar{x}$ = 0.080 ± 0.002, Z = -7.78, P < 0.001) (Figure 5.6a). There was no significant difference in the mean number of TE insertions within ORFs between 'recent polyploid' and 'polyploid' species. In similar fashion, under the SNP scenario a pairwise comparison indicated that the mean number of TE insertions (per 2kb) within ORFs was lower in the diploid species ($\bar{x}$ = 0.076 ± 0.002) than either the recent polyploid (0.082 ± 0.003; Z = -2.24, P < 0.05) or polyploid species (0.085 ± 0.003; Z = -3.50, P < 0.001), though the degree of this significance was reduced (Figure 5.6b). Again, there was no significant difference in the mean number of TE insertions within ORFs between 'recent polyploid' and 'polyploid' species. It was possible that this effect was driven by the overall greater TE abundance within polyploid Corydoradinae species. We therefore repeated this analysis, but this time standardised the number of TE insertions per ORF by the total number of TEs (per 2kb) found across the whole transcriptome of each Corydoradinae species. In both the SNP ratio and haplotype diversity scenarios, overall patterns of significance remained true leading us to conclude that the finding of greater TE insertions within the ORFs of polyploid species was not simply an artefact of higher TE abundance in 'late branching' Corydoradinae lineages.

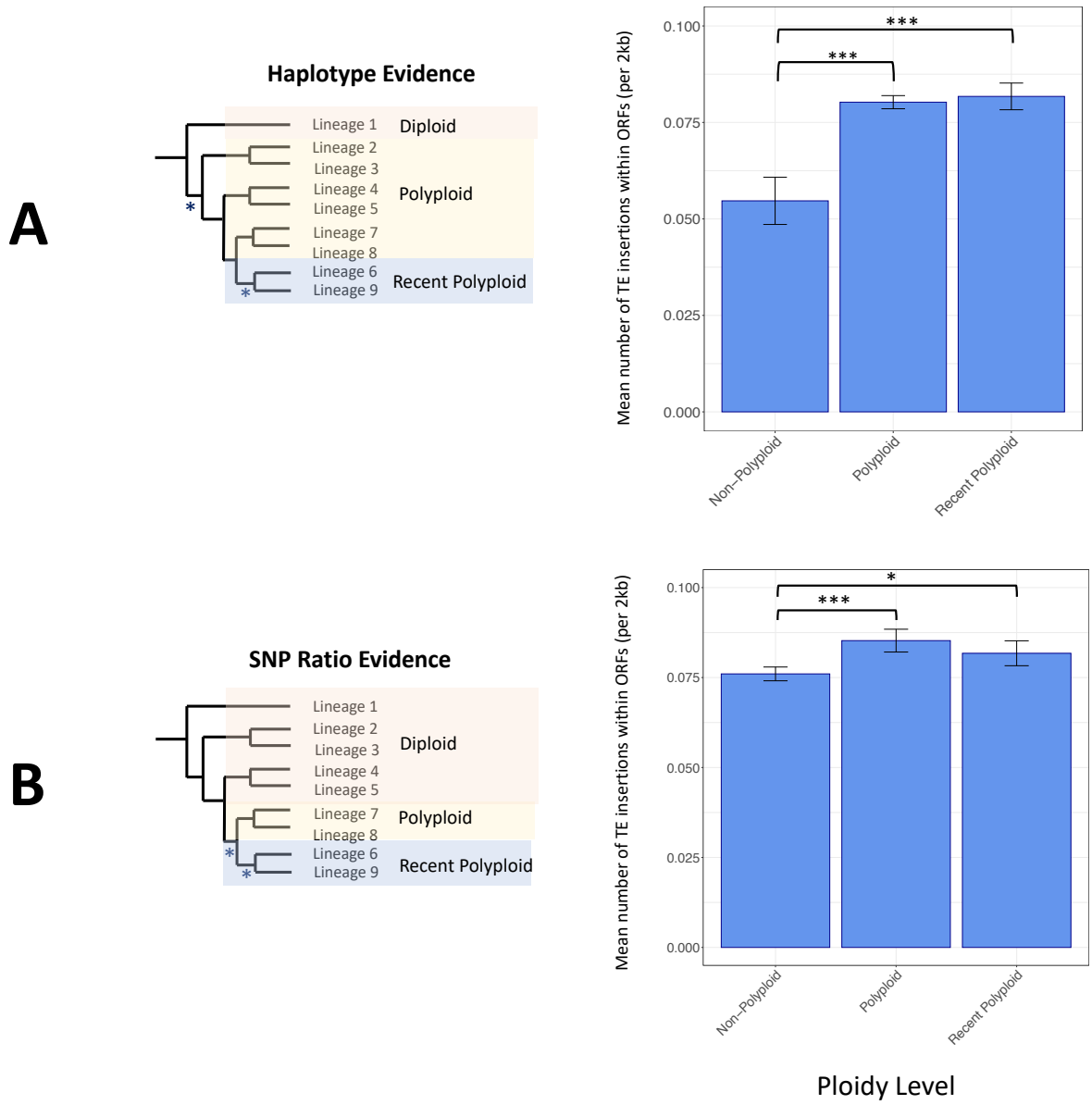


Figure 5.6. The mean number of TEs within the open reading frames (ORFs) of different Corydoradinae species of varying ploidy levels, with scenarios varying based on (**A)** haplotype evidence or (**B)** SNP ratio evidence. Asterisks indicate degree of significant different (*****,** P < 0.001).

Gene Ontology (GO) terms associated with Corydoradinae TE insertions are associated with immune response.

A Gene Ontology (GO) term enrichment analysis of chimeric protein coding transcripts containing a TE for each Corydoradinae species was conducted to ascertain the potential biological impacts of TE insertions. The top 10 most enriched GO terms with respect to TE insertions were selected within each species were selected (see methods) and compared across all nine transcriptomes. The majority of enriched GO terms were present within just one species, with no enriched GO term

being present across more than four of the nine transcriptomes (Figure 5.7). This suggests that the type of biological changes which may be impacted by TE activity within the Corydoradinae may be somewhat lineage specific (Figure 5.7). Across different Corydoradinae species the GO terms most frequently associated with TE insertions appeared to be affiliated with both immune response pathways (e.g. GO:0050701 “interleukin-1 secretion", GO:0045751 “negative regulation of Toll signalling pathway”, GO:1900226 “negative regulation of NLRP3 inflammasome complex assembly”) and blood vessel development (GO:0045765 “regulation of angiogenesis”) (Figure 5.7).

Figure 5.7 Biological Gene Ontology (GO) terms associated with protein coding transcripts containing TE insertions. For each Corydoradinae species the top 10 most enriched GO terms (with respect to TE insertion) are visualised. GO terms are ordered from those enriched in TE insertions within many Corydoradinae species (top) to GO terms associated within single Corydoradinae species (bottom). The number of species each enriched GO term is found in is depicted on the right-hand side (1-4).

Horizontal transfer of Mariner elements within the Corydoradinae.

Despite no evidence of a significant proliferation of horizontally transferred Mariner elements within the Corydoradinae, several show evidence of a potential amphibian origin

Across the nine Corydoradinae transcriptomes we extracted 1,314 Mariner transcripts which had a blast hit against RepBase (see methods). As expected, the majority of these (942) had a best hit against a teleost Mariner element and are therefore most likely to have been vertically inherited (Figure 5.8a). However, we did identify a substantial number of Corydoradinae Mariner elements with best hits against amphibian Mariner elements, particularly those belonging to the *Rana* (154) and *Xenopus* (170) genera (Figure 5.8a). Furthermore, a handful of Mariner elements (<5%) had a best hit against Mariner elements from various taxonomically diverse organisms, including Primate, Anolis and *Drosophila* species (Figure 5.8a). Together, these elements represent a pool of potential horizontally transferred Mariner elements. To determine whether the abundance of these potentially horizontally transferred elements varied between different Corydoradinae lineages, we repeated our earlier phylogenetic shift analysis but using the percentage of total transcriptome length which consisted of potential horizontally transferred Mariner elements. Despite the proportion of HTT vs total Mariner elements with a RepBase match ranging from 17.1% in *C. maculifer* (Lineage 1) and 34.7% in *S. prionotus* (Lineage 3) (Figure 5.8b), as a proportion of total transcriptome length we found no evidence that a significant proliferation of horizontally transferred Mariner elements has occurred across the Corydoradinae.

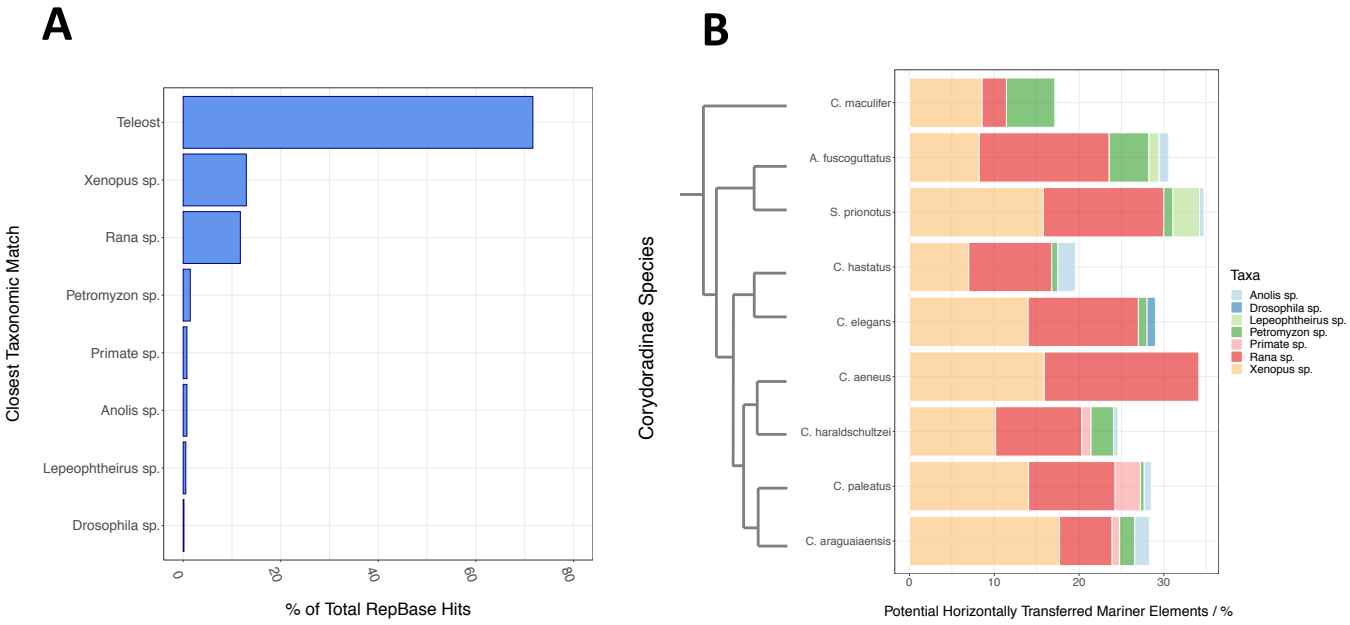


Figure 5.8 A. The distribution of the closest taxonomic hit of all extracted Corydoradinae Mariner elements with a RepBase hit. Those with a teleost hit are likely to be vertically inherited, whilst those with a non-teleost species hit are potential horizontal transfer candidates **B.** The proportion of total extracted Mariner elements which have a non-teleost RepBase hit within each Corydoradinae species. The nuclear Corydoradinae phylogeny is given on the Y axis.

A horizontally transferred Mariner element has inserted within a Matrix Metalloproteinase-13 (MMP-13) paralogue within the common ancestor of Corydoradinae Lineage 6 and 9

To test for evidence of a horizontal transfer event more conclusively we collated a list of candidate Corydoradinae Mariner elements that (I) frequently had a non-teleost closest RepBase hit with high percentage identity (>90%), (ii) were highly abundant across the nine transcriptomes and (iii) found within the majority of Corydoradinae species investigated. One such candidate (TE_00001723) consistently had a closest match with a *X. tropicalis* Mariner element (Tc1-5_Xt), which itself has homologous within Zebrafish (*D. rerio;* Mariner-17_DR) and the Northern Pike (*Esox Lucius;* Mariner-12_EL). An alignment between these three RepBase elements, and a manually curated Corydoradinae TE_00001723 sequence (see methods; hereon in referred to as Mariner-17_Cory) is given in Supplementary Figure 5.1. Pairwise identity between Mariner-17_Cory and Tc1-5_Xt was 92.9%, ~10% greater than with Mariner-17_DR (83.2%) and Mariner-12_EL (80.6%). As expected,

phylogenetic analysis confirmed that the Mariner-17_Cory element's closest relation was the Xenopus homologue (Tc1-1-5_Xt), with a UFboot support value of 100 (Figure 5.9a).

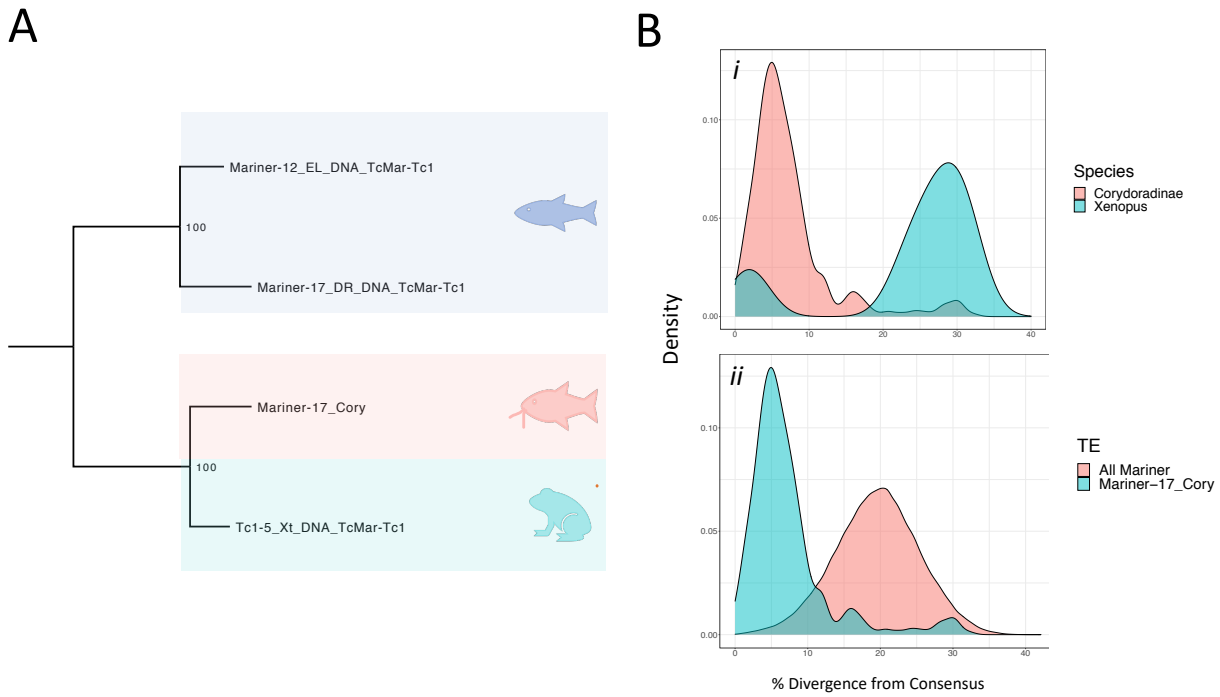


Figure 5.9 A. Phylogenetic tree depicting the relationship between the **'**Mariner-17_Cory' sequence and its three homologues found within RepBase (the *Xenopus tropicalis* Mariner element Tc1-5_Xt, the *Danio rerio* element Mariner-17_DR and the *Esox Lucius* element Mariner-12_EL). The sister relationship between the Xenopus and Corydoradinae element was present in every bootstrap repeat (UFBoot score = 100). IQ-Tree phylogeny was generated using a HKY + F substitution model and 1,000 UFBoot replicates. **B.** Sequence divergence of the Mariner-17_Cory element within the *Corydoras* against (i) the divergence of the homologous element (Tc1-5_XT) within the *Xenopus tropicalis* transcriptome and (ii) every other Mariner element expressed within the Corydoradinae.

High levels of nucleotide similarity alone cannot indicate the directionality of a potential horizontal transfer event. However, given HTT copies within a recipient species will have descended from a common sequence ancestor more recently than vertically inherited TEs, higher levels of sequence divergence between element copies are expected within the donor versus recipient species. We report that the mean sequence divergence of the Tc1-5_Xt element within *X. tropicalis* (23.60 ±

4.49) was significantly higher than the mean sequence divergence of the Mariner-17_Cory element across the nine Corydoradinae species (7.24 ± 0.24) (Mann-Whitney *U*-test: U = 622, n_1=601, n_2=6, P < 0.01) (Figure 5.9bi). To check that this finding was not driven by very low sequence diversity in a limited subset of Corydoradinae species, we checked the divergence within each individual Corydoradinae species. In every instance the mean sequence divergence of 'Mariner-17_Cory' copies were lower than the Tc1-5_Xt homologue within *X. tropicalis*. We also report that the sequence divergence of the Mariner-17_Cory element was significantly lower than the mean sequence divergence of all other Mariner elements expressed across the Corydoradinae transcriptomes (19.50 ± 0.02) (Mann-Whitney *U*-test: U = 79,645,260, n_1 = 601, n_2 = 143,108, P < 0.001) (Figure 5.9bii). Again, this pattern held true when checking such relationship within each Corydoradinae species individually. Taken together, the finding that the Mariner-17_Cory element has a lower sequence diversity (i) than the homologous element (Tc1-5_Xt) found within Xenopus and (ii) other Corydoradinae Mariner elements, suggests that this element was recently transferred from an amphibian species into the Corydoradinae sub-family or a recent common ancestor. We attempted to establish the potential species origin of this horizontal transfer event by running a blastn search of the Mariner-17_Cory sequence against 24 available amphibian genomes on NCBI. 87% (61/70) of the resulting hits were either against *X. tropicalis* or *X. laevis*, with no better matched (< E value) hits outside of these two species. Consequently, we are unable to further narrow down the anuran species involved in this putative horizontal transfer event.

Within *C. paleatus* (Lineage 6), a Mariner-17_Cory element was present within the transcript of the bone developmental gene MMP13, so we aimed to establish whether this insertion was present across multiple Corydoradinae species. We therefore extracted every transcript with a Trinotate-based annotation against MMP13, including those from five additional Corydoradinae species belonging to Lineage 6 & 9 (see methods). The resulting gene tree demonstrated that both teleost

MMP13 paralogues were expressed within all Corydoradinae species (MMP13-a and MMP13-b) (Supplementary Figure 5.2). Furthermore, we found evidence of a ~140bp Mariner-17_Cory insertion within the 3' untranslated (3' UTR) end of multiple MMP13-a transcripts, including both Lineage 6 species (*C. paleatus* & *C. nattereri)* and within 3/5 Lineage 9 species (*C. schwartzi, C. julii* and *C. cruziensis*). The insertion was not present within the MMP13-b paralogue, within any other Corydoradinae lineage, nor the homologs found within *I. punctatus* and *D. rerio*. We therefore concluded that this insertion likely occurred within the common ancestor of nuclear Lineage 6 and 9 (Figure 5.10a).

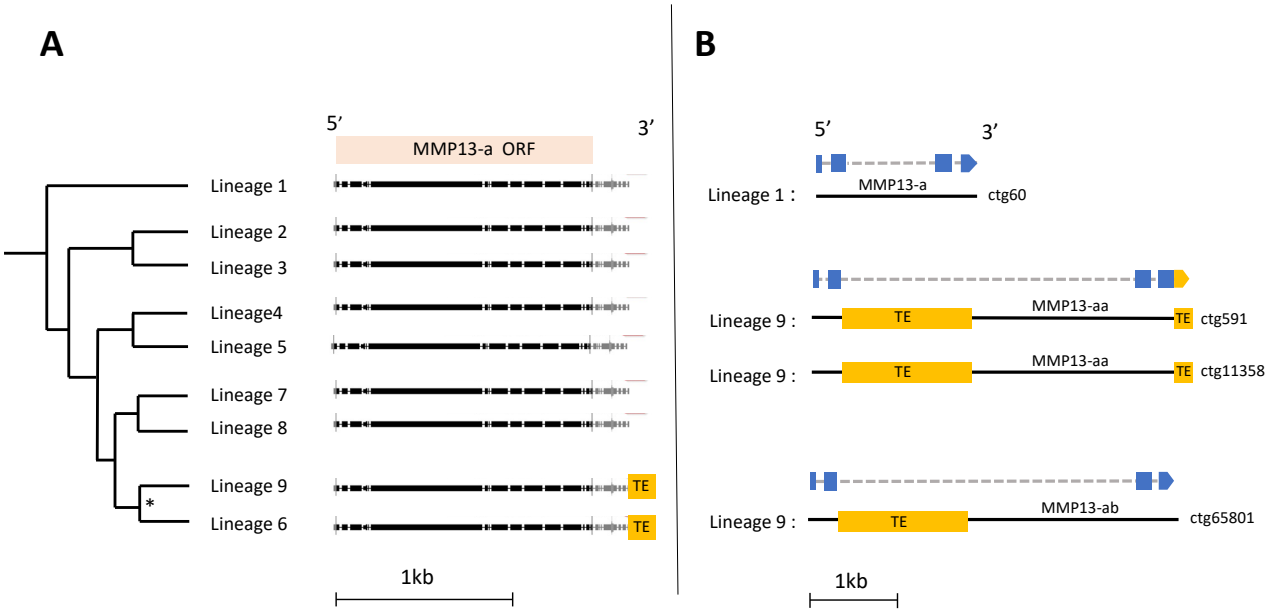


Figure 5.10 A. Graphical representation of the MMP13-a transcript of nine Corydoradinae lineages. A ~140bp fragment of the Mariner-17_Cory element (yellow box) was present within the 3' untranslated region of multiple species belonging to Lineage 6 and 9. The Mariner-17_Cory element insertion is hypothesised to have occurred within the common ancestor of Lineage 6 and 9 (see asterisk on nuclear phylogeny). **B.** Graphical representation of the MMP13-a transcripts aligned to the corresponding genes within the *C. fulleri* (Lineage 1) and *C. metae* (Lineage 9) genome. For visual purposes, only the last 4 exons of the gene are depicted (blue boxes). In both instances, a 1kb bar is shown for scale.

To verify that this insertion was not the result of an assembly error or contamination, we conducted blastn searches (see methods) for the corresponding gene and TE insertion within the respective genomes. We found evidence of one MMP13-a copy within the *C. fulleri* genome (Lineage 1 - ctg60) and three MMP13-a copies within *C. metae* (Lineage 9 – ctg591, ctg11359, ctg65801), with the duplicated paralogues likely resulting from the Corydoradinae-specific whole genome duplication or tandem duplication events. Within the *C. metae* genome (Lineage 9) we identified the ~300bp Mariner-17_Cory insertion which becomes partially incorporated within the 3'UTR of the corresponding transcript within two of the three MMP13-a paralogues (Figure 5.10b). Furthermore, within each *C. metae* MMP13-a paralogue we found evidence of an additional (~1.5kb) Mariner-17_Cory insertion upstream, which has likely contributed to a longer intronic region within *C. metae* (Lineage 9) vs *C. fulleri* (Lineage 1) (3.5kb vs 0.8kb). When extracted, both Mariner-17_Cory insertions align with each other (86% pairwise identity), suggesting these regions represent two independent insertion events. Neither insertion was found within the genome of *C. fulleri* (Lineage 1).

5.6 Discussion

This study characterised the transcriptional TE content of a species rich clade of Neotropical catfish (Corydoradinae), where a previously reported TE proliferation event may have contributed to genome size disparities (Marburger *et al*. 2018). Specifically, this study investigated the contribution that both polyploidy and horizontal transfer may have had during such TE expansion. We report that the majority of transcriptional TEs within the Corydoradinae were Class II DNA transposons, with hAT and Mariner elements being particularly abundant (~60%). This supports the TE landscape reported within Corydoradinae RAD-sequence data, where a Pogo Mariner element (TC1-15630-Pogo) was the most abundant TE found across the Corydoradinae DNA sequences. It also matches the well-established fact that DNA transposons are the dominant element type within teleost genomes, which may be because they often represent older insertions and are shorter in length than

retrotransposons (Brawand *et al.* 2014; Chang *et al*. 2022). Despite being less prevalent, Class I retrotransposons remained expressed across the Corydoradinae, with LINE and LTR Gypsy elements being particularly prevalent within transcript sequences. Non-autonomous SINE elements were present at low abundance across the transcriptomes, with this family of TEs having previously been shown to associate with the life history of different teleost species (catadromous vs anadromous) (Carotti *et al.* 2021). Expressed TE abundance (~4-7%) was found to be an order of magnitude lower than genomic TE abundance, which may reflect either pre-transcriptional silencing or the fact the majority of TE insertions fall within non-transcribed regions of the genome (Bell *et al*. 2022). The transcriptional TE abundance found within the Corydoradinae is slightly higher than that reported in other organisms, including *Helianthus* sunflowers (2.6%) and zebrafish (2.5%) (Gill *et al.* 2014; Chang *et al.* 2022). There was also a remarkable consistency regarding the diversity of TE families detected across the different Corydoradinae transcriptomes, with all TE classes being found in each species, which supports previous reports that TE diversity is maintained across teleost species (Shao *et al*. 2019).

A correlation between TE abundance and genome size have been previously recorded across a range of species, including teleosts (Shao *et al*. 2019), mosquitos (de Melo & Wallau 2020) and the wheat pathogen *Zymoseptoria tritici* (Badet *et al.* 2020). Few studies appear to have investigated this relationship with regards to expressed TE content however, which may reflect a more recent dynamic between transposon activity and increasing genome size. In this study, we found no significant correlation between expressed TE content and the disparate genome size of the Corydoradinae (0.51 – 4.8 pg) when account for phylogenetic signal. It has previously been highlighted that the inclusion phylogenetic signal within ecological studies on average reduces the main effect size, emphasising the importance of including a phylogenetic framework within statistical analyses (Chamberlain *et al.* 2012). The inclusion of more Corydoradinae species data may increase the significance of this relationship. We also compared TE class composition between

Corydoradinae genome and transcriptome sequences and found no evidence of transcriptional bias regarding the type of TE class expressed (i.e. the TE composition of genomic and transcriptomic sequences were equivalent). Previous studies which have compared genomic and transcriptomic TE content in this manner have yielded mixed results. The transcriptional activity of TEs across four different plant species was not found to correspond to their copy number within the genome (Gill *et al*. 2014), whilst the number of expressed sequence tag (EST) with TE homology within the anther smut fungus (*Microbotryum violaceum*) aligned with its genomic composition (Yockteng *et al.* 2007). A better understanding of whether certain TE classes are preferentially incorporated into the exome of an organism is important if we are to fully understand the evolutionary potential of a species genomic TE landscape.

We identified evidence for a shift in expressed TE abundance across the phylogeny of the Corydoradinae, though it depended on the TE library used during their annotation. When using the Danio TE library, we found evidence of a significant shift at the base of Lineage 4-9. However, when using a species-specific TE library we did not find any evidence of a significance phylogenetic shift across the Corydoradinae. Lack of an apparent relationship between polyploidy and TE abundance could be due to (i) no shift, (ii) biased TE annotation or (iii) choice of sequence type. No link between polyploidy and TE abundance has been reported within other species (e.g. Ågren *et al.* 2016; Chalopin & Volff 2017; Carotti *et al.* 2021). However, if there was no link between WGD and TE abundance within the Corydoradinae it would contradict earlier studies (Marburger *et al*. 2018). That studies reliance on RAD sequencing may have introduced biases to estimated TE abundance, with this method purported to being inaccurate at estimating TE abundance within species that have many TE families at low copy numbers (Chak & Rubenstein 2019). Furthermore, a Corydoras specific TE library was not available at the time, which may have led to a TE detection bias, either with species-specific or younger elements being missed (Platt *et al.* 2016; Bell *et al.* 2022). A second possibility is that use of the Corydoradinae-specific TE library introduced a bias in estimated TE

abundance within this study. This hypothesis is supported by (i) the use of the *Danio* TE library detecting a significant phylogenetic increase in transcriptional TE abundance within the base of lineage 4-9 and (ii) comparison between the Corydoradinae-specific and the *Danio* TE library highlighting that estimated TE abundance within 'early-branching' Corydoradinae species was subject to an upward bias, likely because the species specific TE library was generated on a Lineage 1 (*C. fulleri)* genome (Bell *et al*. 2022). Thirdly, it is possible that an increase in TE abundance previously reported within Corydoradinae RAD-data reflects an earlier burst of TE activity which is now subject to transcriptional repression and is unable to be detected within RNA-seq data. Disparate patterns of TE abundance across different sequence types after a whole genome duplication have been reported within allopolyploid and non-polyploid *Spartina* grasses, where genomic TE abundances are equivalent, yet TE groups are transcriptionally repressed within hybrids due to increased DNA methylation (Ainouche *et al.* 2009; Giraud *et al.* 2021).

Next, we investigated the TE distribution with respect to gene coding regions across different Corydoradinae species and found that polyploids had a significantly greater number of TE insertions within ORFs compared to non-polyploid species. This was true even when accounting for variance in total transcriptional TE abundance between species. Increased TE accumulation within genic regions has also been recorded within polyploid *Arabidopsis* and *Capsella* species, which is hypothesised to be driven by relaxed purifying selection after a WGD (Agren *et al*. 2016; Baduel *et al*. 2019). Unlike an immediate increase in TE abundance expected after a direct breakdown of TE silencing measures, the relaxation of selection pressures after a WGD event should lead to a continuous accumulation of TEs (Parisod *et al*. 2010). This may explain why the more recent Corydoradinae WGD event (20-30 mya) did not lead to greater TE accumulation within the genic regions of Lineage 6 + 9 species ("recent polyploid"), as simply not enough time has passed for a potential change in TE insertion distribution to be detected. We also wished to ascertain the potential broad-scale impact of such genic TE insertions within the Corydoradinae, so conducted a biological GO analysis and found that

genes involved in immune pathways appeared to be the most consistently enriched in TEs. Similar findings have been found within mammals, where higher TE occurrence within immune gene regions have been hypothesised to accelerate the evolution of immune regulatory networks and contribute to the rapid adaptation of disease resistance (Ye *et al.* 2020). Future work may wish to uncover how widespread this pattern of TE enrichment may be.

The horizontal transfer of TEs may have contributed to the potential transposon expansion within the Corydoradinae, which we may expect to be particularly likely due to the (i) low UV and low dry air exposure within the aquatic environment they inhabit (Metzger *et al*. 2018; Zhang *et al.*, 2020) and (ii) their frequent exposure to parasitism (Bell *et al.* 2020). Whilst we found no evidence of a significant phylogenetic expansion in horizontally transferred Mariner elements, a substantial proportion of Corydoradinae insertions appeared to have an amphibian origin. Horizontal transfer between teleost and amphibian species is purported to be particularly common and has been reported within both salmonids and the strawberry poison frog *Oophaga pumilio* (De Boer *et al.* 2007; Rogers *et al.* 2018). We therefore wished to verify potential horizontal transferred TE insertions with amphibians using four key lines of evidence. Namely, (i) phylogenetic incongruence between host and TE phylogenies, (ii) patchy TE distributions within phylogenies (iii) high sequence similarity between TEs from distantly related species and (iv) lower sequence divergence between the TE copies of recipient vs donor species (Schaack *et al.* 2010; Panaud 2016). Under this framework, we found strong evidence of a horizontally transferred Mariner element within the Corydoradinae (Mariner-17_Cory) with a very high percentage similarity (>90%) to the Tc1-5_Xt repeat within the *Xenopus tropicalis* genome. Owing to their streamlined genetic make-up, class II 'cut and paste' DNA transposons are more regularly transferred than retrotransposons, with such elements containing an intron-less transposase gene and requiring no specific host factors during transposition (Schaack *et al.* 2010; Zhang *et al.* 2020). Given propensity for horizontal transfer is positively correlated with both phylogenetic similarity and geographical proximity (Peccoud *et al.*

2017), we hypothesise this insertion has likely transferred from a yet unassembled South American frog genome, perhaps via a parasitic or viral host.

Within Corydoradinae Lineage 6 and 9, this horizontally transferred element has inserted within the 3' UTR of duplicated MMP13-a copies, with transcriptomic evidence suggesting it is also incorporated within corresponding transcripts. MMP13 has previously been implicated in zebrafish morphological development, with knock out mutants displaying a range of abnormal phenotypes, including those related to body axis formation and craniofacial shape (Hillegass *et al.* 2007). Given craniofacial shape variation within the Corydoradinae plays an important role in the maintenance of sympatry through disparate food acquisition, it is possible that MMP13 divergence has played an important role during their evolutionary history (Alexandrou *et al.* 2011). Whilst the exact phenotypic impact of this insertion remains uncertain, TEs located within 3' UTR gene regions are known to play an important in role in post-transcriptional regulation. For example, translational repression via a TE insertion within the 3' UTR of the developmental gene *GHD2* impacts several agronomically important phenotypes within rice (Shen *et al.* 2017). Future experimental work may be able to reveal the exact phenotypic and potential evolutionary impact that this horizontal TE has had within the Corydoradinae.

5.7 Concluding Remarks

Understanding the factors that lead to increased TE abundance are vital in gaining a better appreciation of the ability of TEs to induce large scale genomic and evolutionary change. Using new genome and transcriptomic assemblies, this study provides a detail description of the transcriptional TE landscape within a clade of species-rich Neotropical catfish (Corydoradinae), as well as providing a detailed assessment of both the evolutionary causes and consequences of a putative TE proliferation event. Using a new Corydoradinae-specific TE library we found no evidence that either polyploidy or horizontal transfer has led to a TE expansion. However, we do find evidence that whole

genome duplications events may impact genomic TE distributions, with a greater number of TE insertions within the genic regions of polyploid Corydoradinae species. Furthermore, the horizontal transfer of a Mariner element with an amphibian origin within multiple Corydoradinae MMP13 genes may have had an important role during their evolutionary history, with future experimental analyses required to elucidate the possible phenotypic impact of these insertions.

5.8 References

Ågren, J.A., Huang, H.-R. & Wright, S.I. (2016). Transposable element evolution in the allotetraploid Capsella bursa-pastoris. *Am. J. Bot.*, 103, 1197–1202.

Ainouche, M.L., Fortune, P.M., Salmon, A., Parisod, C., Grandbastien, M.A., Fukunaga, K., *et al.* (2009). Hybridization, polyploidy and invasion: Lessons from Spartina (Poaceae). *Biol. Invasions*, 11.

Alexandrou, M.A., Oliveira, C., Maillard, M., Mcgill, R.A.R., Newton, J., Creer, S., *et al.* (2011). Competition and phylogeny determine community structure in Müllerian co-mimics. *Nature*, 496, 84–88.

Badet, T., Oggenfuss, U., Abraham, L., McDonald, B.A. & Croll, D. (2020). A 19-isolate reference-quality global pangenome for the fungal wheat pathogen Zymoseptoria tritici. *BMC Biol.* , 18.

Baduel, P., Quadrana, L., Hunter, B., Bomblies, K. & Colot, V. (2019). Relaxed purifying selection in autopolyploids drives transposable element over-accumulation which provides variants for local adaptation. *Nat. Commun.*, 10, 1–10.

Beaulieu, J., Jean, M. & Belzile, F. (2009). The allotetraploid Arabidopsis thaliana–Arabidopsis lyrata subsp. petraea as an alternative model system for the study of polyploidy in plants. *Mol. Genet. Genomics* , 281, 421–435.

Bell, E.A., Butler, C.L., Oliveira, C., Marburger, S., Yant, L. & Taylor, M.I. (2022). Transposable element annotation in non-model species: The benefits of species-specific repeat libraries using semi-automated EDTA and DeepTE de novo pipelines. *Mol. Ecol. Resour.*, 22, 823–833.

Benoit, M., Drost, H.G., Catoni, M., Gouil, Q., Lopez-Gomollon, S., Baulcombe, D., *et al.* (2019). Environmental and epigenetic regulation of Rider retrotransposons in tomato. *PLOS Genet.*, 15, e1008370.

De Boer, J.G., Yazawa, R., Davidson, W.S. & Koop, B.F. (2007). Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. *BMC Genomics*, 8.

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

Brawand, D., Wagner, C.E., Li, Y.I., Malinsky, M., Keller, I., Fan, S., *et al.* (2014). The genomic substrate for adaptive radiation in African cichlid fish. *Nat.* , 513, 375–381.

Bryant, D.M., Johnson, K., DiTommaso, T., Tickle, T., Couger, M.B., Payzin-Dogru, D., *et al.* (2017). A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. *Cell Rep.*, 18, 762–776.

Carotti, E., Carducci, F., Canapa, A., Barucca, M., Greco, S., Gerdol, M., *et al.* (2021). Transposable Elements and Teleost Migratory Behaviour. *Int. J. Mol. Sci.*, 22, 1–12.

Chak, S.T.C. & Rubenstein, D.R. (2019). TERAD: Extraction of transposable element composition from RADseq data. *Mol. Ecol. Resour.*, 19, 1681–1688.

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.

Chalopin, D. & Volff, J.N. (2017). Analysis of the spotted gar genome suggests absence of causative link between ancestral genome duplication and transposable element diversification in teleost fish. *J. Exp. Zool. Part B Mol. Dev. Evol.*, 328, 629–637.

Chamberlain, S.A., Hovick, S.M., Dibble, C.J., Rasmussen, N.L., Van Allen, B.G., Maitner, B.S., *et al.* (2012). Does phylogeny matter? Assessing the impact of phylogenetic information in ecological meta-analysis. *Ecol. Lett.*, 15, 627–636.

Chang, N.C., Rovira, Q., Wells, J.N., Feschotte, C. & Vaquerizas, J.M. (2022). Zebrafish transposable elements show extensive diversification in age, genomic distribution, and developmental expression. *Genome*

Res., gr.275655.121.
Charles, M., Belcram, H., Just, J., Huneau, C., Viollet, A., Couloux, A., *et al.* (2008). Dynamics and Differential Proliferation of Transposable Elements During the Evolution of the B and A Genomes of Wheat. *Genetics*, 180, 1086.
Chénais, B., Caruso, A., Hiard, S. & Casse, N. (2012). The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. *Gene*, 509, 7–15.
Choi JY & Lee YCG. (2020). Double-edged sword: The evolutionary consequences of the epigenetic silencing of transposable elements. *PLOS Genet.*, 16, e1008872.
Curcio, M.J. & Derbyshire, K.M. (2003). The outs and ins of transposition: from mu to kangaroo. *Nat. Rev. Mol. Cell Biol.*, 4, 865–877.
Dunemann, S.M. & Wasmuth, J.D. (2019). Horizontal transfer of a retrotransposon between parasitic nematodes and the common shrew. *Mob. DNA*, 10, 1–13.
Felsenstein J. (1985). Phylogenies and the Comparative Method . *Am. Nat.*, 125, 1–15.
Fuller, I.A.M. & Evers, H.G. (2005). Identifying corydoradinae catfish : Aspidoras-Brochis-Corydoras-Scleromystax & C-numbers, 384.
Galbraith, J.D., Ludington, A.J., Suh, A., Sanders, K.L. & Adelson, D.L. (2020). New Environment, New Invaders—Repeated Horizontal Transfer of LINEs to Sea Snakes. *Genome Biol. Evol.*, 12, 2370–2383.
Gao, B., Shen, D., Xue, S., Chen, C., Cui, H. & Song, C. (2016). The contribution of transposable elements to size variations between four teleost genomes. *Mob. DNA*, 7.
Gilbert, C., Schaack, S., Pace, J.K., Brindley, P.J. & Feschotte, C. (2010). A role for host–parasite interactions in the horizontal transfer of transposons across phyla. *Nat.* , 464, 1347–1350.
Gill, N., Buti, M., Kane, N., Bellec, A., Helmstetter, N., Berges, H., *et al.* (2014). Sequence-Based Analysis of Structural Organization and Composition of the Cultivated Sunflower (Helianthus annuus L.) Genome. *Biology* 295–319.
Giraud, D., Lima, O., Rousseau-Gueutin, M., Salmon, A. & Aïnouche, M. (2021). Gene and Transposable Element Expression Evolution Following Recent and Past Polyploidy Events in Spartina (Poaceae). *Front. Genet.*, 12, 352.
González, J., Lenkov, K., Lipatov, M., Macpherson, J.M. & Petrov, D.A. (2008). High Rate of Recent Transposable Element–Induced Adaptation in Drosophila melanogaster. *PLoS Biol.*, 6, e251.
Gouy, M., Guindon, S. & Gascuel, O. (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.*, 27, 221–224.
Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., *et al.* (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* , 29, 644–652.
Gregory, T.R. (2005). Synergy between sequence and size in large-scale genomics. *Nat. Rev. Genet.*, 6, 699–708.
Hasegawa, M., Kishino, H. & Yano, T. aki. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol. 1985 222*, 22, 160–174.
Hawkins, J.S., Kim, H.R., Nason, J.D., Wing, R.A. & Wendel, J.F. (2006). Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. *Genome Res.*, 16, 1252.
Hazzouri, K.M., Mohajer, A., Dejak, S.I., Otto, S.P. & Wright, S.I. (2008). Contrasting patterns of transposable-element insertion polymorphism and nucleotide diversity in autotetraploid and allotetraploid Arabidopsis species. *Genetics*, 179, 581–592.
Hillegass, J.M., Villano, C.M., Cooper, K.R. & White, L.A. (2007). Matrix Metalloproteinase-13 Is Required for Zebra fish (Danio rerio) Development and Is a Target for Glucocorticoids. *Toxicol. Sci.*, 100, 168–179.
Hof, A.E.V. t., Campagne, P., Rigden, D.J., Yung, C.J., Lingley, J., Quail, M.A., *et al.* (2016). The industrial melanism mutation in British peppered moths is a transposable element. *Nat.* , 534, 102–105.
Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., Von Haeseler, A. & Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* , 14, 587–589.
Katoh, K., Misawa, K., Kuma, K.I. & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, 30, 3059–3066.
Khabbazian, M., Kriebel, R., Rohe, K. & Ané, C. (2016). Fast and accurate detection of evolutionary shifts in Ornstein–Uhlenbeck models. *Methods Ecol. Evol.*, 7, 811–824.
Lanciano, S. & Cristofari, G. (2020). Measuring and interpreting transposable element expression. *Nat. Rev. Genet.*, 21, 721–736.
Lander, E. & IHGSC. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409, 860–921.
Lee, J., Han, K., Meyer, T.J., Kim, H.-S. & Batzer, M.A. (2008). Chromosomal Inversions between Human and Chimpanzee Lineages Caused by Retrotransposons. *PLoS One*, 3, e4047.

Liu, Z., Liu, S., Yao, J., Bao, L., Zhang, J., Li, Y., *et al.* (2016). The channel catfish genome sequence provides insights into the evolution of scale formation in teleosts. *Nat. Commun.* , 7, 1–13.
Madlung, A., Tyagi, A.P., Watson, B., Jiang, H., Kagochi, T., Doerge, R.W., *et al.* (2005). Genomic changes in synthetic Arabidopsis polyploids. *Plant J.*, 41, 221–230.
Marburger, S., Alexandrou, M.A., Taggart, J.B., Creer, S., Carvalho, G., Oliveira, C., *et al.* (2018). Whole genome duplication and transposable element proliferation drive genome expansion in Corydoradinae catfishes. *Proc. R. Soc. B Biol. Sci.*, 285.
Matzke, M.. & Matzke, A.J.. (1998). Polyploidy and transposons. *Trends Ecol. Evol.*, 13, 241.
McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci.*, 36, 344–355.
McClintock, B. (1984). The significance of responses of the genome to challenge. *Science*, 226, 792–801.
McDonald, M.C., Taranto, A.P., Hill, E., Schwessinger, B., Liu, Z., Simpfendorfer, S., *et al.* (2019). Transposon-mediated horizontal transfer of the host-specific virulence protein ToxA between three fungal wheat pathogens. *MBio*, 10.
de Melo, E.S. & Wallau, G.L. (2020). Mosquito genomes are frequently invaded by transposable elements through horizontal transfer. *PLOS Genet.*, 16, e1008946.
Metzger, M.J., Paynter, A.N., Siddall, M.E. & Goff, S.P. (2018). Horizontal transfer of retrotransposons between bivalves and other aquatic species of multiple phyla. *Proc. Natl. Acad. Sci. U. S. A.*, 115, E4227–E4235.
Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Von Haeseler, A., *et al.* (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.*, 37, 1530–1534.
Mitros, T., Lyons, J.B., Session, A.M., Jenkins, J., Shu, S., Kwon, T., *et al.* (2019). A chromosome-scale genome assembly and dense genetic map for Xenopus tropicalis. *Dev. Biol.*, 452, 8–20.
Oliver, K.R. & Greene, W.K. (2011). Mobile DNA and the TE-Thrust hypothesis: supporting evidence from the primates. *Mob. DNA*, 2.
Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J., *et al.* (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.*, 20, 1–18.
Pace, J.K., Gilbert, C., Clark, M.S. & Feschotte, C. (2008). Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. *Proc. Natl. Acad. Sci. U. S. A.*, 105, 17023–17028.
Panaud, O. (2016). Horizontal transfers of transposable elements in eukaryotes: The flying genes. *C. R. Biol.*, 339, 296–299.
Paradis, E., Claude, J. & Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, 20, 289–290.
Parisod, C., Alix, K., Just, J., Petit, M., Sarilar, V., Mhiri, C., *et al.* (2010). Impact of transposable elements on the organization and function of allopolyploid genomes. *New Phytol.*, 186, 37–45.
Peccoud, J., Loiseau, V., Cordaux, R. & Gilbert, C. (2017). Massive horizontal transfer of transposable elements in insects. *Proc. Natl. Acad. Sci. U. S. A.*, 114, 4721–4726.
Pinheiro, J., Bates, D. (2022). nlme: Linear and NonLinear Mixed Effects Models. R package version 3.1-159
Plague, G.R., Dunbar, H.E., Tran, P.L. & Moran, N.A. (2008). Extensive Proliferation of Transposable Elements in Heritable Bacterial Symbionts †. *J. Bacteriol.*, 190, 777–779.
Platt, R.N., Blanco-Berdugo, L., Ray, D.A. & Ray, D.A. (2016). Accurate Transposable Element Annotation Is Vital When Analyzing New Genome Assemblies. *Genome Biol. Evol.*, 8, 403–10.
Ricci, M., Peona, V., Guichard, E., Taccioli, C. & Boattini, A. (2018). Transposable Elements Activity is Positively Related to Rate of Speciation in Mammals. *J. Mol. Evol.*, 86, 303–310.
Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., *et al.* (2011). Integrative genomics viewer. *Nat. Biotechnol.* , 29, 24–26.
Rogers, R.L., Zhou, L., Chu, C., Márquez, R., Corl, A., Linderoth, T., *et al.* (2018). Genomic takeover by transposable elements in the Strawberry poison frog. *Mol. Biol. Evol.*, 35, 2913–2927.
Ruan, J. & Li, H. (2019). Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* , 17, 155–158.
Schaack, S., Gilbert, M. & Dric Feschotte, C. (2010). Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends Ecol. Evol.*, 25, 537–546.
Schartl, M., Kneitz, S., Volkoff, H., Adolfi, M., Schmidt, C., Fischer, P., *et al.* (2019). The Piranha Genome Provides Molecular Insight Associated to Its Unique Feeding Behavior. *Genome Biol. Evol.*, 11, 2099–2106.
Schrader, L., Kim, J.W., Ence, D., Zimin, A., Klein, A., Wyschetzki, K., *et al.* (2014). Transposable element islands facilitate adaptation to novel environments in an invasive species. *Nat. Commun.*, 5, 5495.
Serrato-Capuchina, A., Matute, D., Serrato-Capuchina, A. & Matute, D.R. (2018). The Role of Transposable

Elements in Speciation. *Genes.*, 9, 254.
Shao, F., Han, M. & Peng, Z. (2019). Evolution and diversity of transposable elements in fish genomes. *Sci. Reports* , 9, 1–8.
Shen, J., Liu, J., Xie, K., Xing, F., Xiong, F., Xiao, J., *et al.* (2017). Translational repression by a miniature inverted-repeat transposable element in the 3′ untranslated region. *Nat. Commun.* , 8, 1–10.
Slotkin, R.K. & Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.* , 8, 272–285.
Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J.M. & Kelly, S. (2016). TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Res.*, 26, 1134–1144.
Smukowski Heil, C., Patterson, K., Hickey, A.S.M., Alcantara, E. & Dunham, M.J. (2021). Transposable Element Mobilization in Interspecific Yeast Hybrids. *Genome Biol. Evol.*, 13.
Stitzer, M.C., Anderson, S.N., Springer, N.M. & RossIbarra, J. (2021). The genomic ecosystem of transposable elements in maize. *PLOS Genet.*, 17, e1009768.
Symonová, R. & Suh, A. (2019). Nucleotide composition of transposable elements likely contributes to AT/GC compositional homogeneity of teleost fish genomes. *Mob. DNA*, 10.
Vicient, C.M. & Casacuberta, J.M. (2017). Impact of transposable elements on polyploid plant genomes. *Ann. Bot.*, 120, 195–207.
Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., *et al.* (2007). A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet*, 8, 973–982.
Wu, T.D. & Watanabe, C.K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21, 1859–1875.
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.
Ye, M., Goudot, C., Hoyler, T., Lemoine, B., Amigorena, S. & Zueva, E. (2020). Specific subfamilies of transposable elements contribute to different domains of T lymphocyte enhancers. *Proc. Natl. Acad. Sci. U. S. A.*, 117, 7905–7916.
Yockteng, R., Marthey, S., Chiapello, H., Gendrault, A., Hood, M.E., Rodolphe, F., *et al.* (2007). Expressed sequences tags of the anther smut fungus, Microbotryum violaceum, identify mating and pathogenicity genes. *BMC Genomics*, 8.
Zhang, H.H., Peccoud, J., Xu, M.R.X., Zhang, X.G. & Gilbert, C. (2020). Horizontal transfer and evolution of transposable elements in vertebrates. *Nat. Commun.*, 11, 1–10.

5.9 Supplementary Figures and Tables.

Supplementary Table 5.1 Transcriptome Assembly statistics for each of the nine Corydoradinae species assembled de-novo using Trinity.

Species	Transcripts	Total Length (MB)	GC (%)	N50 (bp)	TransRate Assembly Score
C. maculifer	66,579	39.46	46.43	818	0.12
A. fuscoguttatus	78,164	73.95	45.52	1,688	0.13
S. prionotus	105,310	110.33	46.06	2,025	0.08
C. hastatus	69,798	34.04	43.59	584	0.06
C. elegans	113,590	94.86	44.80	1,492	0.08
C. paleatus	118,445	91.16	44.09	1,326	0.07
C. aeneus	78,189	55.98	43.90	1,081	0.10
C. haraldschultzei	115,020	89.20	45.58	1,320	0.11
C. araguaiaensis	65,992	35.62	43.18	706	0.10

Supplementary Table 5.2 Genome assembly statistics for each of the four Corydoradinae species assembled using wtdbg.

	Corydoras fulleri	*Aspidoras CW52*	*Corydoras nijsseni*	*Corydoras metae*
Contigs	2,080	1,570	40,243	124,831
Total Length (MB)	699.66	808.15	1,677.66	5,707.67
GC %	42.12	40.53	39.55	39.58
N50 (MB)	1.56	4.36	0.07	0.12
Complete BUSCO %	90.2	92.4	59.9	76.7

Supplementary Table 5.3 Transcriptional TE abundance within each of the nine Corydoradinae species and the channel catfish (*I. punctatus).* Abundances are given as a percentage of total transcriptome length (%) and are divided into both Class I Retrotransposon families and Class II DNA transposon families.

Species	Class I Retrotransposons								Total Class I	Class II DNA Transposons							Total Class II	Total TE Abundance %
	BEL	Copia	ERV	Gypsy	DIRS	LINE	Penelope	SINE		CACTA	Harbinger	hAT	Mariner	Mutator	PiggyBac	Helitron		
I. punctatus	0.02	0.22	0.05	0.39	0.00	0.18	0.06	0.00	**0.92**	0.41	0.07	0.98	0.65	0.06	0.04	0.08	**2.30**	**3.22**
C. maculifer	0.04	0.17	0.13	0.62	0.01	0.45	0.05	0.04	**1.50**	0.32	0.13	1.44	1.12	0.10	0.02	0.04	**3.17**	**4.68**
A. fuscoguttatus	0.03	0.19	0.17	0.69	0.01	0.56	0.07	0.06	**1.77**	0.37	0.16	2.10	1.61	0.12	0.03	0.06	**4.46**	**6.23**
S. prionotus	0.03	0.19	0.14	0.58	0.01	0.46	0.06	0.04	**1.51**	0.37	0.12	1.92	1.47	0.11	0.01	0.05	**4.05**	**5.55**
C. hastatus	0.02	0.18	0.14	0.80	0.01	0.56	0.06	0.03	**1.81**	0.49	0.14	2.22	1.98	0.12	0.06	0.05	**5.06**	**6.87**
C. elegans	0.03	0.21	0.14	0.72	0.01	0.56	0.08	0.04	**1.79**	0.40	0.14	2.20	1.74	0.12	0.06	0.05	**4.70**	**6.50**
C. paleatus	0.03	0.22	0.16	0.73	0.01	0.59	0.08	0.03	**1.86**	0.43	0.13	2.16	1.78	0.14	0.06	0.06	**4.77**	**6.63**
C. aeneus	0.04	0.22	0.16	0.75	0.01	0.53	0.07	0.03	**1.81**	0.44	0.17	2.30	1.77	0.13	0.07	0.07	**4.96**	**6.76**
C. haraldschultzei	0.03	0.20	0.15	0.65	0.01	0.48	0.06	0.04	**1.61**	0.37	0.12	1.95	1.55	0.11	0.05	0.05	**4.21**	**5.81**
C. araguaiaensis	0.04	0.20	0.14	0.73	0.01	0.59	0.06	0.03	**1.81**	0.41	0.14	2.22	1.81	0.11	0.05	0.06	**4.80**	**6.60**

Supplementary Table 5.4. A. Pearson's Chi-Squared (X^2) results when testing whether there was a significant bias in TE class expression across four Corydoradinae lineages. In each case, the null hypothesis of a 1:1 ratio between transcriptomic and genomic TE abundance was accepted, suggesting there was no significant bias in the type of TE class expressed within any Corydoradinae lineage. TE classes included in analysis were DNA, LTR, LINE and SINE. Degrees of freedom was 3 in each case. **B.** Pearson's Chi-Squared (X^2) results when testing whether there was a significant bias in DNA transposon family expression across four Corydoradinae lineages. In each case, the null hypothesis of a 1:1 ratio between transcriptomic and genomic TE abundance was accepted, suggesting there was no significant bias in the DNA transposon family expressed within any Corydoradinae lineage. DNA transposon families included Mariner, hAT, CACTA, Harbinger, Mutator and PiggyBac elements. Degrees of freedom was 5 in each case.

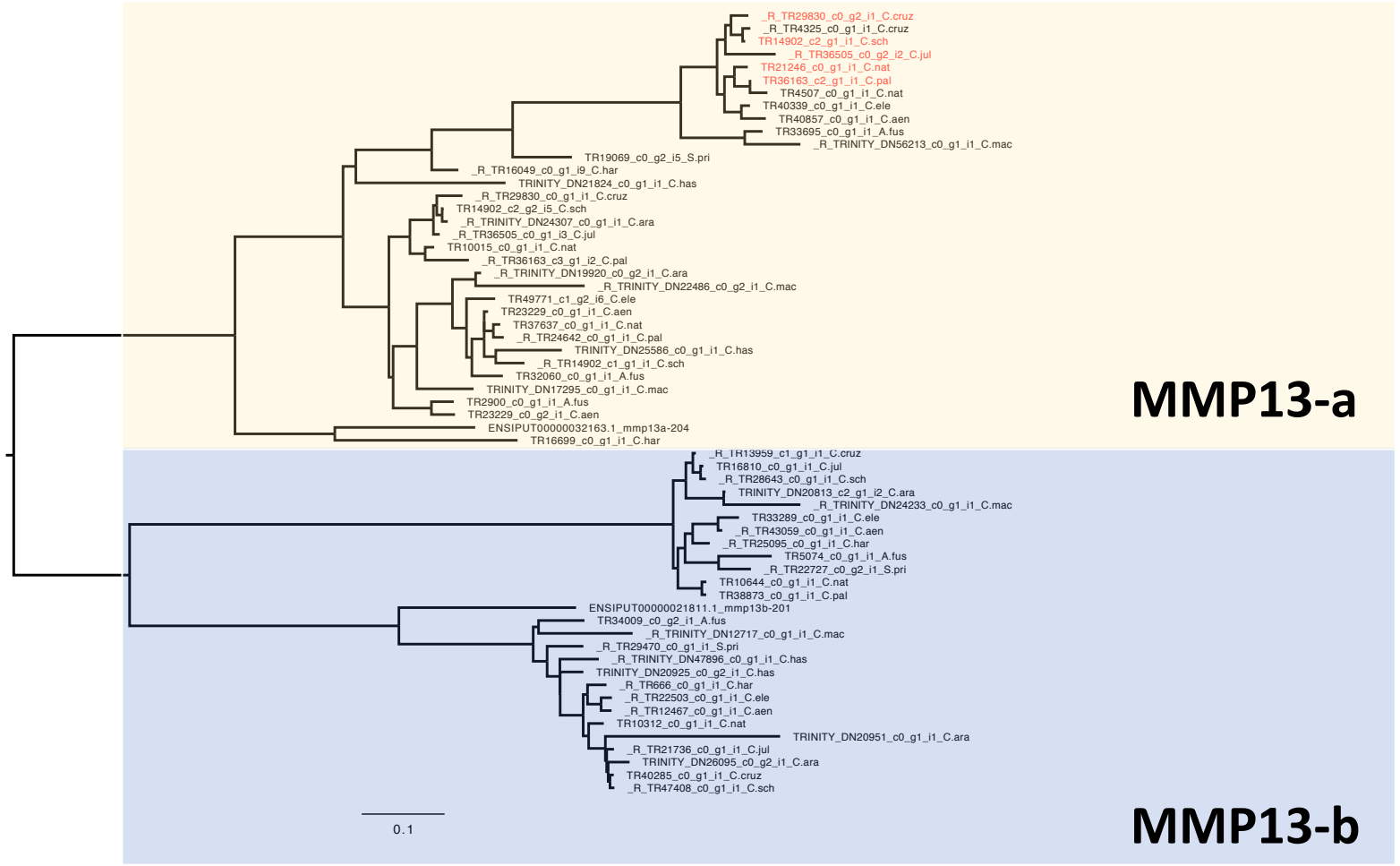
A

Lineage	Person's X^2 (chi-squared)	P value
One	4.85	0.18
Two	4.50	0.21
Five	0.68	0.88
Nine	1.41	0.70

B

Lineage	Person's X^2 (chi-squared)	P value
One	3.15	0.68
Two	0.62	0.99
Five	3.38	0.64
Nine	9.62	0.09

Supplementary Figure 5.1 Alignment of the consensus Mariner-17_Cory element against its three homologues within RepBase (Mariner-17-DR, *Danio,* Mariner-12_ES, *Esox* and TC1-5_Xt, *Xenopus).* The alignment was conducted using MAFFT v7.471 and visualised using JalView v2.11.2.2



Supplementary Figure 5.2 Gene tree of each Corydoradinae MMP13 transcript built using IQ-TREEv1.6.12. Based on the inclusion of two *Ictalurus punctatus* paralogues (MMP13-a; ENSIPUT00000032163 and MMP13b; ENSIPUT00000021811), we shade the two monophyletic clades as either MMP13a or MMP13b. The MMP13 transcripts highlighted in red contained the horizontally transferred Mariner-17_Cory element.

6 Colour and the Corydoras: the pigmentation genes of a mimetic clade of Neotropical catfish evolve under stricter than average selection pressure and with no associated enrichment in transposable element activity.

6.1 Chapter Overview and Author Contributions

This chapter attempted to answer whether rapid change within pigmentation phenotypes may be disproportionately impacted by the activity of TEs. Primarily, this work attempted to answer if TE insertions are upwardly inflated within pigmentation genic regions, and whether this may have contributed to the rapid colour pattern changes which have occurred within the Corydoradinae. This involved looking at whether TE insertions were more prevalent within pigmentation transcripts and respective cis-regulatory regions than within non-pigmentation genes. Additional analyses included in this chapter involved estimating the degree of purifying selection pressure acting on Corydoradinae pigmentation genes through calculating their Kn/Ks ratios and looking at expression differences of pigmentation genes belonging to different Corydoradinae species.

The study was conceived and developed by Christopher L. Butler, Emily Phelps and Martin I. Taylor. Genome annotation of *C. fulleri*, extraction of up- and downstream gene regions and reciprocal blasting of pigmentation transcripts was conducted by Emily Phelps with assistance by Christopher L Butler. Transcriptomic sequence data was extracted and sequenced by Claudio Oliveira and assembled by Ellen A. Bell and Christopher L. Butler. The Corydoras de novo TE library was constructed by Ellen A. Bell, and subsequently parsed using a custom script written by Christopher L. Butler (as described in Chapter 4). The estimation of rate of colour pattern change was based on existing data curated by Alexandrou *et al*, 2011. All further analysis found within this chapter was conducted by Christopher L. Butler, including downstream TE analysis, pigmentation candidate curation, transcript alignment, Kn/Ks ratio calculation, statistical frameworks and expression analysis. The chapter was written by Christopher L. Butler with contributions from Martin I. Taylor, Emily Phelps and Ellen Bell.

6.2 Abstract

Pigmentation has important evolutionary roles across the animal kingdom, though the genetic processes which drive colour pattern diversity are yet to be fully understood. One process with a historical association in generating pigmentation diversity is the activity of transposable elements (TEs), small replicative DNA sequences which are well-established drivers of genetic change. Under a 'gene disruption' model, TEs are predicted to accumulate near non-essential genes at a greater rate than essential genes. Consequently, the diversification of colour patterns may be disproportionately impacted by TE insertions within pigmentation loci, though this hypothesis has yet to be formally tested. In this study, we explore the association between pigmentation genes and TE insertions across the Corydoradinae; a clade of Neotropical catfish which display both remarkable colour pattern diversity and mimetic relationships. Surprisingly, we found that Corydoradinae pigmentation appear to evolve under stricter than average purifying selection pressure. In support of the 'gene disruption' model of TE insertion distribution, we find that (i) the likelihood of TE presence was greater within non-pigmentation vs pigmentation transcripts and (ii) the cis-regulatory regions upstream of Corydoradinae pigmentation genes are not enriched in TE insertions. We also found no correlation between TE insertion likelihood and rate of colour pattern evolution across the Corydoradinae. In conclusion, this study provides no evidence that rapid pigmentation change within the Corydoradinae is underpinned by increased TE activity within pigmentation genes. We suggest that many of the previously identified incidences of colour pattern change being caused by TE insertions may be due to an easily identifiable phenotype change rather than biased insertion patterns.

Key words: Pigmentation, Vertebrates, Transcriptome, Corydoradinae, Transposons

6.3 Background

Across the animal kingdom, colour diversification occurs at both inter- and intra-specific levels. This diversity provides a range of evolutionary benefits, including (i) reducing unwarranted aggression from conspecifics, (ii) interspecific mate recognition and intraspecific mate choice, and (iii) the avoidance of predation through mimicry or crypsis (Alexandrou *et al.* 2011; Merrill *et al.* 2014; Nishikawa *et al.* 2015; Hemingson *et al.* 2019). Rapid bursts of colour pattern evolution may induce speciation events, and have been recorded in the adaptive radiation of both African Great lake cichlids and parasitic feather lice (Seehausen *et al.* 1999; Bush *et al.* 2019). Numerous molecular processes may lead to pigmentation changes, ranging from single nucleotide polymorphisms (SNPs), genomic inversions, to the activity of transposable elements (TEs); short parasitic DNA sequences which possess the ability to replicate and disrupt a genome (Joron *et al.* 2011; Van't Hof *et al.* 2016; Lin *et al.* 2018). However to date, variation in vertebrate pigmentation diversity has been linked to relatively few genomic loci in a limited number of predominantly model species (Lewis & Van Belleghem 2020). An improved understanding of the molecular processes which underpin rapid colour pattern evolution therefore remains an important goal within the field of evolutionary genetics.

Historically, the potential for pigmentation pattern change to be driven by TE activity may be relatively under-explored, not only because the inherent repetitive nature of TEs makes them underrepresented within many genome assemblies (Peona *et al.* 2018), but also because their insertions can cause a multitude of downstream impacts that are only now beginning to be fully appreciated. TE activity may increase rates of mutagenesis, induce gene duplication or cause genomic deletions and inversions through the ectopic recombination of newly introduced units of high sequence similarity(Bourque *et al.* 2018). Consequently, TE activity is a powerful driver of rapid phenotypic change, with multiple studies highlighting their significant contribution towards both environmental adaptation and speciation (Stapley *et al.* 2015; Ricci *et al.* 2018; Schrader & Schmitz

2019). Host silencing means a high proportion of genomic TEs are either degraded or inactive, opposed to autonomous entities, yet even these non-active TEs can still indirectly impact host gene evolution. For example, host TE silencing via the deposition of epigenetic methylation marks may lead to the inadvertent suppression of neighbouring genes and can contribute to the expression divergence of multi-species orthologues (Zeng *et al.* 2018; Choi JY & Lee YCG 2020). Furthermore, many non-autonomous TEs remain transcribed if they insert themselves within or close to gene coding regions, forming chimeric transcripts in which fragmented TEs are incorporated into mature mRNA (Lanciano & Cristofari 2020). For instance, the majority (two-thirds) of TE transcripts found within zebrafish (*Danio rerio)* are not self-autonomous but remain expressed through their association with host gene promoters (Chang *et al.* 2022). These TEs can significantly modify the transcriptome by increasing transcript diversity via open reading frame (ORF) insertion, exonisation (formation of new exons from previously intronic sequences) or the provision of alternative polyadenylation signals (Cowley & Oakey 2013; Lanciano & Cristofari 2020).

The association between TE activity and pigmentation diversity was first established in the 1950s with the discovery that TE insertions underpinned diversity in maize kernel colouration (McClintock 1950). Further examples of TE activity modifying an organism's pigmentation patterns have been subsequently identified. Within both the *Ipomoea* genus of flowering plants and the 'Hanfu' apple cultivar, alterations in anthocyanin pigment expression are caused by nearby TE insertions and results in the emergence of a wide variety of colour patterns (Iida *et al.* 2004; Telias *et al.* 2011). In mice, methylation associated with the silencing of a TE insertion upstream of the *agouti* locus has resulted in variation in coat pattern (Waterland & Jirtle 2003). TE induced pigmentation changes have also been shown to be adaptive to the host organism. For example, the insertion of an unknown TE into the first intron of the cortex gene within the peppered moth *(Biston betularia)* led to an adaptive melanistic '*carbonaria*' phenotype (Hof *et al.* 2016), and the insertion of a TE into the promoter region of the teleost specific gene *fhl2b* is thought to have contributed to the sexually

selected ‘egg spot’ phenotype of haplochromine cichlids (Santos *et al.* 2014). However, it is not well established whether TEs insertions within pigmentation loci occur more frequently than in other gene types or whether pigmentation-related TE changes are merely easier to identify due to their intrinsic nature of producing a visible phenotypic change. Under a ‘gene disruption model’, TEs are expected to accumulate near non-essential genes at a greater rate than essential genes (Correa *et al.* 2021). This is likely the outcome of both (i) host selection against TE insertions within conserved genomic regions and (ii) insertion site preferences exhibited by TEs towards regions where the chances of subsequent silencing is lower (Pereira 2004; Zhang *et al.* 2020). Given pigmentation genes may not be under the same strict purifying selection constraints as other gene types (Sturm & Duffy 2012; Wilde *et al.* 2014, Lorin *et al.* 2018), TE driven pigmentation change may be particularly common. Conversely, non-essential genes are likely to be younger than essential genes and may have had less time to accumulate TE insertions (‘gene-age model’) (Correa *et al*. 2021). Transposon-gene insertion dynamics is likely driven by the outcome of these two opposing forces (‘gene disruption’ versus ‘gene age’ model), of which surprisingly little is known.

In this study we investigate the role TE activity may have had in generating rapid pigmentation change within the Corydoradinae (Teleostei; Siluriformes; Callichthyidae), a species rich group of Neotropical catfish. Teleosts (an infraclass of ray-finned Actinopterygii) represent nearly half of all vertebrate species (~30,000 described species) and possess a striking repertoire of both TE and pigmentation cell (chromatophore) diversity (Ravi & Venkatesh 2018). Teleosts therefore represent an ideal group to investigate both the frequency and tempo of TE induced pigmentation changes. The genomes of teleosts are abundant in class II DNA elements, particularly those belonging to the Mariner and hAT families (Shao *et al.* 2019). With a total of 24 TE superfamilies present, TE repertoire with teleosts is also more diverse than any other vertebrate group (Sotero-Caio *et al.* 2017; Shao *et al.* 2019). Total TE abundance varies between 6% (*Tetraodon nigroviridis*) and 55% (*Danio rerio)* and is positively correlated with teleost genome size (Shao *et al.* 2019). Teleosts have

eight known chromatophores, which are organised in different layers between the epidermis and muscle to create complex colour patterns (Lorin *et al.* 2018; Irion & Nüsslein-Volhard 2019). These include the black melanocytes, reflective silver iridophores, yellow-orange xanthophores, red erythrophores, white leucophores, blue cyanophores, red-violet ertho-irodophores, blue-red cyano-erythorphores and the newly described fluorescent chromatophore and two dichromatic chromatophores, red-violet ertho-irodophores and the blue-red cyano-erythorphores (Salis *et al.* 2019). The phenotypes resulting from the arrangement of these chromatophores are controlled by a complex network of pigmentation genes, which have been preferentially retained compared to non-pigmentation genes after the teleost specific whole genome duplication (Braasch *et al.* 2008; Lorin *et al.* 2018). The Corydoradinae consist of nine monophyletic lineages based on a mitochondrial DNA phylogeny, with a nuclear restricted site associated DNA (RAD) based phylogeny largely supporting this topology, though mtDNA lineages 6 and 9 are instead shown to form one monophyletic group (Marburger *et al.* 2018). A significant expansion of TEs has occurred during the Corydoradinae's 66MY evolution, largely driven by the proliferation of Class II Tc1/Mariner DNA elements which have contributed to their disparate genome sizes (0.6 to 4.4 picograms per haploid cell) (Marburger *et al.* 2018). This expansion has occurred concurrently with a putative whole genome duplication event, which has promoted immunogenetic diversity within polyploid species (Bell *et al.* 2020). Under a gene disruption model, we may expect a greater proportion of TE insertions to fall close to and disrupt pigmentation gene regions within putative polyploid species, particularly as paralogous gene copies evolve under more relaxed purifying selection pressure (Kondrashov *et al.* 2002). Nevertheless, the evolutionary consequences of both high and variable TE abundance within the Corydoradinae have remained largely unexplored.

Here, we utilise a newly annotated *Corydoras fulleri* genome (mtDNA Lineage 1) and nine Corydoradinae transcriptomes to test the hypothesis that their pigmentation genes may have a greater likelihood of TE insertions. Specifically, we assess whether (i) pigmentation genes evolve

under more relaxed selection pressure than non-pigmentation genes (ii) more TE insertions are associated with Corydoradinae pigmentation transcripts and cis-regulatory regions and finally (iii) whether inter-species differences in the expression of pigmentation genes are greater than that of non-pigmentation genes. A multi-species approach also allows us to test whether differences in the number of TE insertions associated with pigmentation genes may be driving disparate rates of colour pattern changes within different Corydoradinae lineages.

6.4 Materials and Methods

Estimating Rates of Colour Pattern Change.

Calculating the rate of colour pattern evolution within each Corydoradinae lineage was achieved using an existing manual description of the pigmentation pattern of 203 different Corydoradinae species (Alexandrou *et al*, 2011). Body outlines were divided into 20 sections and manually scored on their presence/absence of different pigmentation pattern. Possible patterns included 'spotty', 'light grey', 'banded', 'brown', 'marbled', 'blotched', 'reticulated', 'triple banded' or 'ancestral' (pale with darkened dorsal colouration). Additional colour traits included 'lateral stripe', 'eye/tail band' and 'blocks of colour'. The number of unique patterns per lineage were then converted to colour patterns per millions year, using the known clade age from a fossil calibrated mtDNA phylogeny provided in Marburger *et al.* 2018. Colour pattern diversity was set within a phylogenetic context, specifically by using an Ornstein-Uhlenbeck model within the R package 'l1ou' to detect shifts in the rate of colour pattern evolution without the need for pre-existing assumptions of their location (Khabbazian *et al.* 2016). The RAD-based phylogeny was converted into an ultrametric tree (i.e. a phylogeny where edge lengths represent time) using the 'chronos' function in the R package 'ape, (Paradis *et al*. 2004). This uses a penalised maximum likelihood method to estimate divergence times A non-parametric bootstrap test was used to compute confidence support values for each shift position. Phylogenetically uncorrelated standardised residuals were calculated for each node and then sampled with replacement 100 times.

Corydoras Transcriptome Assembly

An existing transcriptome of Lineage 1 species *Corydoras maculifer* (Bell *et al*. 2022, GenBank accession GJAY00000000.1) was used in this study, in addition to eight new de-novo transcriptome assemblies which were assembled from short read Illumina based sequencing. Namely, these were *Aspidoras fuscoguttatus* (Lineage 2), *Scleromystax prionotus* (Lineage 3), *Corydoras hastatus* (Lineage 4)*, Corydoras elegans* (Lineage 5)*, Corydoras paleatus* (Lineage 6)*, Corydoras aeneus* (Lineage 7)*, Corydoras haraldschultzi* (Lineage 8) and *Corydoras araguaiaensis* (Lineage 9). RNA extraction (TRIzol Plus RNA Purification Kit) was conducted on adult somatic muscle tissue, and the cDNA library was then built using a TruSeq RNA Sample Prep kit (Illumina, Inc). cDNA was sequenced using paired-end sequencing on an Illumina HiSeq, with the average sequencing depth of all nine samples being 13.33 million paired reads (ranging from 9.27 to 20.67 million). The libraries were then demultiplexed and cleaned using Trimmomatic (version 0.2.36, Bolger *et al.* 2014) and assembled using Trinity (version 2.6.9, Grabherr *et al.* 2011). Transcriptome assembly quality was assessed using TransRate (version 1.03, Smith-Unna *et al.* 2016) (Supplementary Table 6.1). Transcriptomic library preparation and sequencing was performed by the Animal Biotechnology Laboratory of Esalq/Piracicaba.

Transposable Element Annotation

A Corydoradinae-specific TE library was previously created by running two bioinformatic pipelines on a Lineage 1 *C. fulleri* genome (Bell *et al.* 2022) GenBank accession PRJNA706371). Namely, the repeat identifier Extensive de novo TE Annotator (EDTA) (Ou *et al.* 2019) and repeat classifier DeepTE (Yan *et al.* 2020) were utilised to generate a Corydoras-specific TE library (as described by Bell *et al*. 2022). This custom library was then used for all subsequent downstream TE annotation, which was performed using RepeatMasker (RM; version 1.332), utilising the NCBI/RMBLAST (version 2.6.0+) search engine. TE annotations were subject to a custom parse script (RM_TRIPS), which is publicly available and fully described at (https://github.com/clbutler/RM_TRIPS) (Bell *et al*. 2022). All steps of

this pipeline were followed, though identical TEs found across transcript isoforms were retained for downstream analysis.

Candidate Pigmentation Genes

Identification of Corydoradinae pigmentation genes was achieved using a candidate list of 198 vertebrate pigmentation genes (VPGs) detailed in Lorin *et al.* 2018. Broadly, this was a list of genes which had a (i) Gene Ontology (GO) term of "pigmentation" (GO: 0043473) and (ii) met the definition of "genes involved in the differentiation of a neural-crest derived pigment cell in at least one vertebrate species''. In addition to the 198 VPGs, a subset of pigmentation genes was manually curated for this study, with the primary purpose of discarding genes where evidence of an experimental link to colour pattern function was missing within a teleost. Specifically, pigmentation genes were only included in this subset when (i) they were found to be differentially expressed across teleost individuals of varying colour morphs or (ii) knock out gene experiments within teleosts demonstrated a colour pattern change without inviability. This filtering aimed to remove highly pleiotropic pigmentation genes, whereby functions beyond colour pattern determination will likely skew the final selection pressure of such genes. This final subset consisted of 28 different gene families and, whilst not representing an exhaustive list of teleost-associated pigmentation genes, will be referred to as the 'teleost pigmentation gene' set (TPGs) from hereon (Supplementary Table 6.2). The *C. fulleri* genome was annotated for protein coding regions using the Genome Sequence Annotation Server (GenSAS, https://www.gensas.org) (Humann *et al.* 2019), which incorporated *ab initio* structural and functional annotation, alongside evidenced based prediction of protein coding regions utilising multi-species transcriptomic data. Specifically, the transcriptomes of the nine Corydoradinae species (see above) as well as those of *Ictalurus punctatus* (GCA_001660625.1 IpCoco_1.2; Liu *et al.* 2016), *Danio rerio* (GCA_000002035.4 GRCz11, Howe *et al.* 2013) and *Hemibagrus wyckioides* (GCA_019097595.1, Shao *et al.* 2021) were used. A blastp search against the SwissProt protein sequence of each of the VPGs was conducted to retrieve a list of candidate

pigment genes within the *C. fulleri* genome. Genic regions were filtered to remove sequences below 100 bp. Pigmentation regions were later verified using a reciprocal blast with transcriptomic data.

Extraction of 5' Upstream and 3' Downstream Gene Regions

The flanking regions of *C. fulleri* pigmentation genes were extracted using the gff co-ordinate output from GenSas draft annotation. Sequences were filtered for early stop codons, with missing start/stop codons added using AGAT (https://github.com/NBISweden/AGAT). Sequences which did not contain a complete coding sequence were discarded, with open reading frames detected using TransDecoder v5.5.0 (https://github.com/TransDecoder/TransDecoder). This generated a final subset of 119 pigmentation gene loci. Specifically, the 2kb upstream (5') and 1kb downstream (3') flanking gene coding regions were extracted using Seqkit (Shen *et al.* 2016). These lengths were chosen as they represent the flanking regions of a gene where a TE insertion is most likely to impact its function, with 90% of human cis-regulatory polymorphisms located within the 2kb upstream region of the first gene exon (Sinnett *et al.* 2006) and the average 3' UTR length within teleosts being ~750bp (Xiong *et al.* 2018). The flanking regions of a random subset of 100 genes not belonging to our pigmentation subset (and thus assumed to be non-pigmentation genes) were also extracted. A Gene Ontology (GO) enrichment analysis was then conducted on this subset, to ensure that no GO term was overrepresented in the non-pigmentation gene set. Briefly, this was conducted by collecting the Panther Ontology Terms associated with the gene names. This utilized the Panther GO-Slim molecular function database in the species *Danio rerio, Homo sapiens, Rattus norvegicus, Mus musculus, Xenopus tropicalis* and *Bos taurus*. No GO terms were overrepresented in the subset, and thus this constituted the final non-pigmentation gene set.

Pigmentation and Non-Pigmentation Transcript Retrieval

Pigmentation transcripts across our nine assembled Corydoradinae transcriptomes were obtained by running a blastn search between the annotated pigmentation gene regions in the *C. fulleri* genome and a softmasked (for TEs and low complexity regions) version of each transcriptome. Each

candidate pigment transcript was subject to a reciprocal blastn search against NCBI to verify each transcript's identity. Non-pigmentation transcripts were retrieved using the Trinotate_Trinotate-v3.1.1 pipeline (https://github.com/Trinotate/Trinotate.github.io/blob/master/index.asciidoc), with each CDS sequence subjected to a blastp search against the Uniprot/SwissProt protein database. Transcripts solely aligning with a transposase protein were discarded in downstream analysis because these likely represent autonomous, rather than co-transcribed TEs. To check that both pigmentation and non-pigmentation transcripts had been well assembled, we ensured that an open reading frame (of minimum length 100 amino acids) was present and spanned either a start or stop codon using TransDecoder v5.5.0. During downstream analysis of TE insertion rate, the longest isoform of each species gene was retained. When investigating the location of TE insertions within a transcript possessing multiple candidate coding regions, the open reading frame with the highest log-likelihood score was retained. When investigating TE insertion rate differences across the Corydoradinae, comparisons were only made between genes with representatives in both the low and high rate of colour pattern evolution groups, with the longest transcript being retained in each instance.

Estimation of Kn/Ks Ratios

The non-synonymous (Kn) vs synonymous (Ks) substitution rates across aligned pigmentation and non-pigmentation protein coding regions (CDS) were estimated using the 'KnKs()' function of the 'ape' package in R, which implements Li's method of substitution rate estimation (Li 1993; Paradis *et al.* 2004). Pigmentation and non-pigmentation transcripts were aligned against the CDS of their orthologue within the channel catfish *Ictalurus punctatus* (IpCoco_1.2; Liu *et al.* 2016), whereby in cases of multiple isoforms the longest transcript was used. A protein coding alignment was obtained using a combination of MAFFT (v7.471, Katoh *et al.* 2002) (utilising both the *--adjustdirection* to check for reverse complement alignments and *--auto* which automatically optimises the refinement algorithm during alignment) and MASCE (v2.05, Ranwez *et al.* 2011) which accounts for underlying

codon structure to ensure CDS alignments are given 'in frame'. Each resulting alignment was then manually verified, with regions outside of the CDS being trimmed. Comparison of Kn/Ks ratio of each pigmentation gene between Corydoradinae species of differing rates of colour pattern evolution used the longest transcript per group (i.e low and high).

Log-fold Expression Divergence between Species

Pairwise expression divergence of both pigmentation and non-pigmentation genes were estimated between each of the nine Corydoradinae species. Extracted RNA-seq reads used in each species transcriptome assembly were aligned to the annotated pigmentation/non-pigmentation *C. fulleri* genes using the 'Rsubread' R package, with count matrices being generated from alignment files using the 'featureCounts' function (Liao *et al.* 2019). Genes expressed at low levels across all species were filtered before differential expression analysis using the 'filterByExpr' function in the 'edgeR' R package (Robinson *et al.* 2010). Counts also underwent a TMM (weighted trimmed mean of M values) normalisation, which accounts for differences in RNA-seq library size and composition (i.e proportion of genes that are highly expressed) between samples (Robinson *et al.* 2010). Log fold expression differences between species was subsequently conducted within the R package 'NOIseq', which better adjusts to differences in sequence read depths between samples, as well as working in the absence of replication (Tarazona *et al.* 2015). Genes were considered to be differentially expressed between species using a commonly used >2 log-fold cut off value (Bigler *et al.* 2013).

Statistical Analysis

TE insertions data within transcript data was found to be positively skewed, with a high number of absences (i.e zero counts). Non-parametric tests (e.g Mann-Whitney U statistic) may not be suitable for comparing means under highly skewed count distributions, particularly when there are multiple ties in the data (see McElduff *et al.* 2010). We therefore analysed TE insertions within transcripts under three different generalised linear models within R, namely (i) Poisson (ii) Negative Binomial ('glm.nb', MASS Ripley & Venables 2002) and a (iii) Binomial Logistic Regression for

presence/absence data. A zero-inflation model was not considered, because whilst the data had numerous zero entries, these were likely to be 'true' zeros.. To account for different transcript lengths all models including an offset term accounting for different sequence lengths. Genomic TE insertion count within the up and downstream regions of genes were less skewed and therefore non-parametric permutation tests were considered appropriate.

All bioinformatic pipelines for the analysis of this paper were conducted within R Version 3.6.0 (R Core Team, 2021). All figures were produced using the ggplot2 package in R (Wickham, 2016). Where appropriate, variance statistics are given as ± one standard error.

6.5 Results

The Corydoradinae exhibit varying rates of colour pattern evolution.

Table 6.1 The number of colour patterns (MY) per nuclear Corydoradinae lineage

Nuclear Lineage	Number of Species	Clade Age/ MY	Number of Unique Colour Patterns	Colour Patterns per MY
1	33	48.54	15	0.31
2	11	30.59	2	0.07
3	8	33.86	3	0.09
4	4	33.36	3	0.09
5	13	23.8	7	0.29
7	14	15.94	8	0.50
8	39	25.62	18	0.70
9 + 6	81	19.77	22	1.11

The rate of colour pattern evolution exhibited by each Corydoradinae lineage was highly variable, with nuclear Lineage 9 + 6 possessing 22 different colour patterns and lineage 2 displaying just 2 (Table 6.1). Overall rate of pigmentation change varied by over two orders of magnitude, between

0.07 and 1.11 colour patterns per million years. Mapping these phylogenetic shifts onto an existing Corydoradinae phylogeny, we identified a significant increase in colour pattern diversity at the base of lineage 7-9 with an 91% bootstrap support value. (Figure 6.1). During downstream analysis Corydoradinae species were subsequently classified as having either a “low” (lineages 1-5) or “high” rate of colour pattern evolution (lineage 6-9), as this permitted adequate representation of pigmentation transcripts between species of different rates of colour pattern evolution.

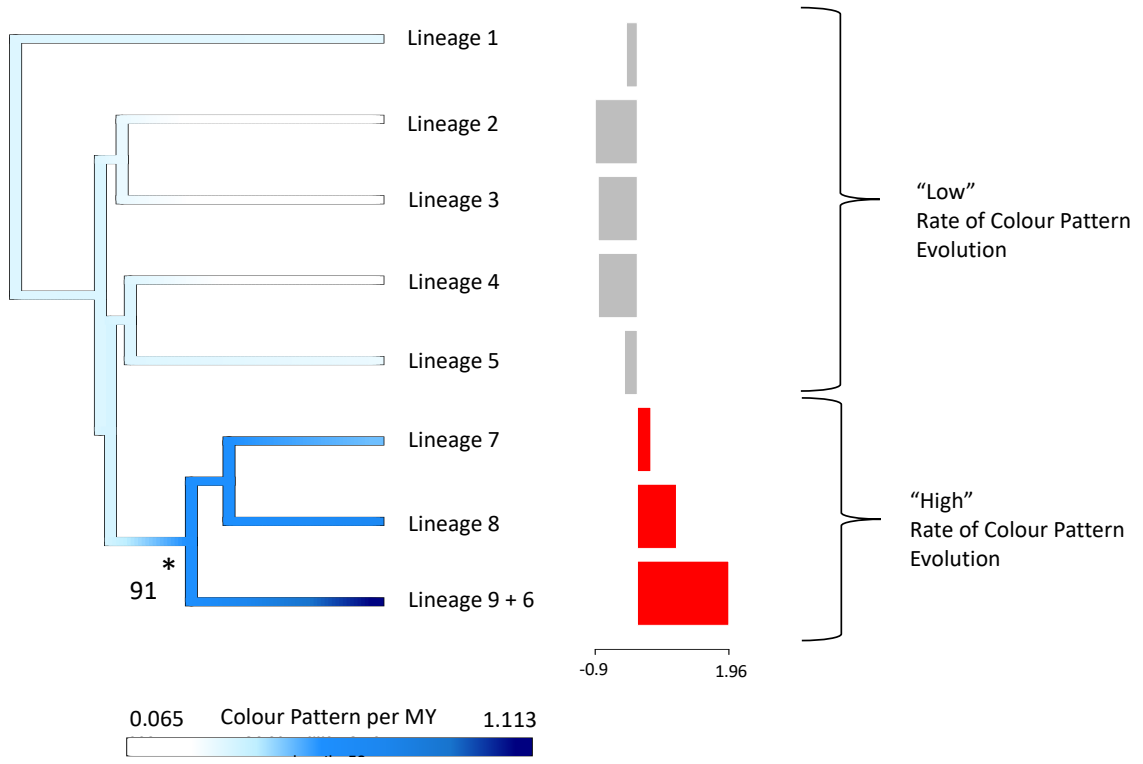


Figure 6.1 Nuclear Corydoradinae phylogeny with colour patterns per million years mapped in colour. The statistical shift in number of colour patterns is indicated with an asterisk and associated bootstrap support.

Corydoradinae pigmentation genes evolve under stricter than average purifying selection

A total of 489 full-length pigmentation transcripts (377 genes and 112 isoforms) were retrieved across the nine Corydoradinae transcriptomes, representing 103 of the 198 possible vertebrate pigmentation genes (VPGs) (Table 6.2). In the low colour pattern group 282 pigmentation transcripts were extracted (214 gene matches and 68 isoforms) and in the high colour pattern group 207

pigmentation transcripts were extracted (163 genes and 44 different isoforms) (Table 6.2). Between the high and low rate of colour pattern evolution groups there were 72 shared pigmentation genes with a full-length transcript representative.

Table 6.2 The number of vertebrate pigmentation transcripts retrieved for each of the nine Corydoradinae species, which are divided into those with a low and high rate of colour pattern evolution.

Rate of Colour Pattern Evolution	Species	Reciprocal Pigmentation Blast Hits		
		Gene	Isoform	Total
Low	*C. maculifer*	19	3	22
	A. fuscoguttatus	61	14	75
	S. prionotus	54	18	72
	C. hastatus	9	3	12
	C. elegans	71	30	101
	Subtotal:	**214**	**68**	**282**
High	*C. aeneus*	25	6	31
	C. haraldschultzei	66	20	86
	C. araguaiaensis	17	0	17
	C. paleatus	55	18	73
	Subtotal:	**163**	**44**	**207**
	TOTAL	**377**	**112**	**489**

Under a gene disruption model of TE insertion, TE enrichment within pigmentation genes is theorised to be a result of relaxed purifying selection. To evaluate whether this hypothesis was supported we first compared the Kn/Ks ratio of 57 Corydoradinae pigmentation transcripts and 50 non-pigmentation equivalents (see methods). The mean Kn/Ks ratio of the VPGs ($\bar{x}$ = 0.11 ± 0.01) was significantly lower than that of non-pigmentation genes ($\bar{x}$ = 0.16 ± 0.01) (Welch's t = 2.59, d.f. = 90.76, $P < 0.05$), suggesting that contrary to our hypothesis, Corydoradinae pigmentation genes are subject to stricter than average purifying selection (Figure 6.2a). When using the reduced TPG

dataset there was no significant difference in the Kn/Ks ratio between pigmentation and non-pigmentation genes (Supplementary Figure 6.1a). Across Corydoradinae species with a high and low rate of colour pattern evolution, we compared the Kn/Ks ratio of 57 paired VPGs (Figure 6.2b). The estimated Kn/Ks ratio of VPGs varied by two orders of magnitude, between those likely to be under particularly strong purifying selection (e.g. *ap1m1*, $\bar{x}$ = 0.005) and those under more relaxed selection (e.g. *osmt1*, $\bar{x}$ = 0.40). No VPGs within our subset showed evidence of being under positive selection (Kn/Ks > 1). The paired Kn/Ks ratios for the low ($\bar{x}$ = 0.11 ± 0.01) and high ($\bar{x}$ = 0.11 ± 0.01) rate of colour pattern evolution groups were not significantly different from one another (Paired T = 0.80, df = 56, P = 0.43). However, we note two genes involved in melanocyte development had different Kn/Ks ratios between the two colour pattern diversity groups, namely *gli3* and *dock7*. Within the high rate of colour pattern evolution group, *dock7* was found to be under reduced purifying selection pressure (Kn/Ks ratio of 0.16 vs 0.03) whilst *gli3* was under stricter purifying selection pressure (Kn/Ks ratio of 0.12 vs 0.25) (Figure 6.2b). These findings were also consistent when using the reduced TPG subset (Supplementary Figure 6.1b). Overall, we report that VPGs across Corydoradinae species with different rates of colour pattern evolution appear to be under similar selection pressure, with a lower rate of molecular evolution than non-pigmentation genes.

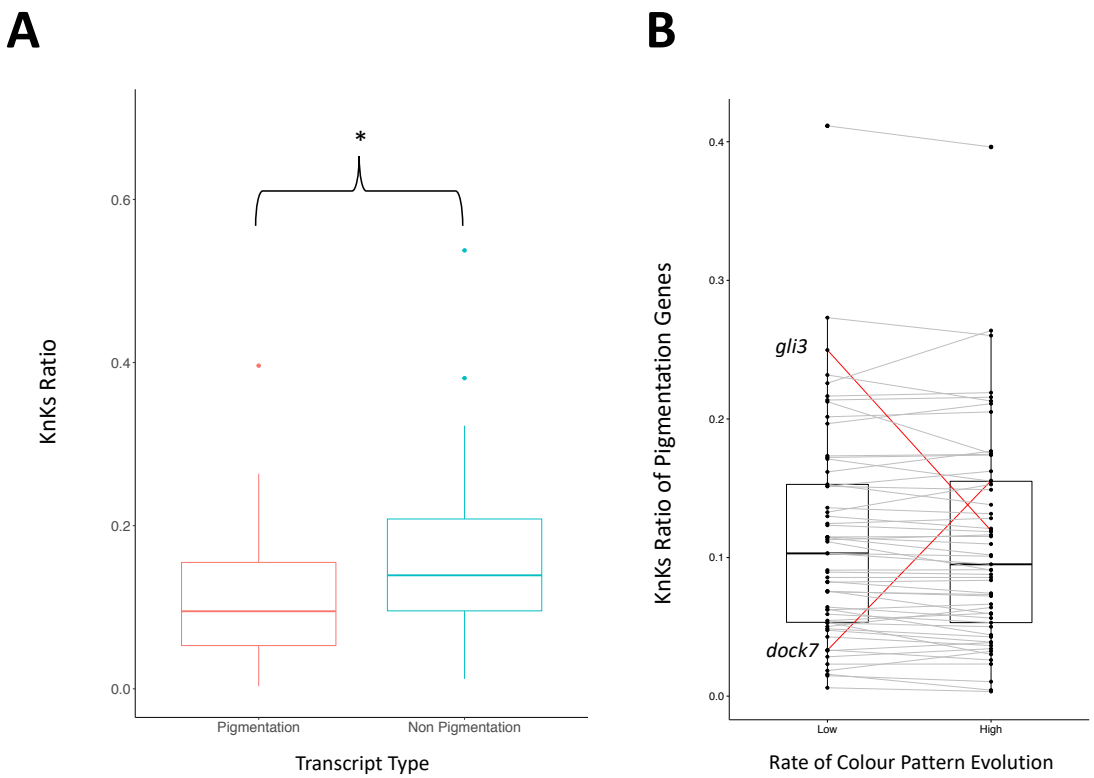


Figure 6.2 A. Corydoradinae pigmentation genes (VPGs) had significantly lower Kn/Ks ratios than non-pigmentation genes. **B.** There was no significant difference in the Kn/Ks ratio of VPGs shared between Corydoradinae species with a low and high rate of colour pattern evolution. Two genes are highlighted (gli3 and dock7) which demonstrated evidence of being under varying selection pressure within Corydoradinae species with differing rates of colour pattern evolution. All gene alignments were conducted against orthologues found within the channel catfish *(Ictalurus punctatus*). Asterisks indicate degree of significant different (* $P < 0.05$).

No evidence of enriched or variable number of transposable element insertions within Corydoradinae pigmentation transcripts

The distributions of TE insertions within both pigmentation and non-pigmentation transcripts of all Corydoradinae species were highly positively skewed (Figure 6.3a). All generalised linear models indicated that transcript length had a highly significant ($P < 0.001$) positive relationship with TE insertion number (Table 6.3a). To account for any differences in size between pigmentation and non-pigmentation transcripts, we therefore offset TE insertions by transcript length within our models, with a log-likelihood test indicating that a negative binomial regression was the best fitting model. We found no significant difference in the number of TE insertions per log transcript length

between VPGs ($\bar{x}$ = 0.031 ± 0.0034) and non-pigmentation genes ($\bar{x}$ = 0.034 ± 0.0002) [Relative Rate (RR): 1.04, P = 0.39] (Table 6.3a). When comparing the likelihood of TE presence or absence across the two transcript types, we did find a significant effect, with a greater likelihood of a TE insertion being present within non-pigmentation versus pigmentation transcripts [RR: 1.45, P < 0.01] (Table 6.3a).

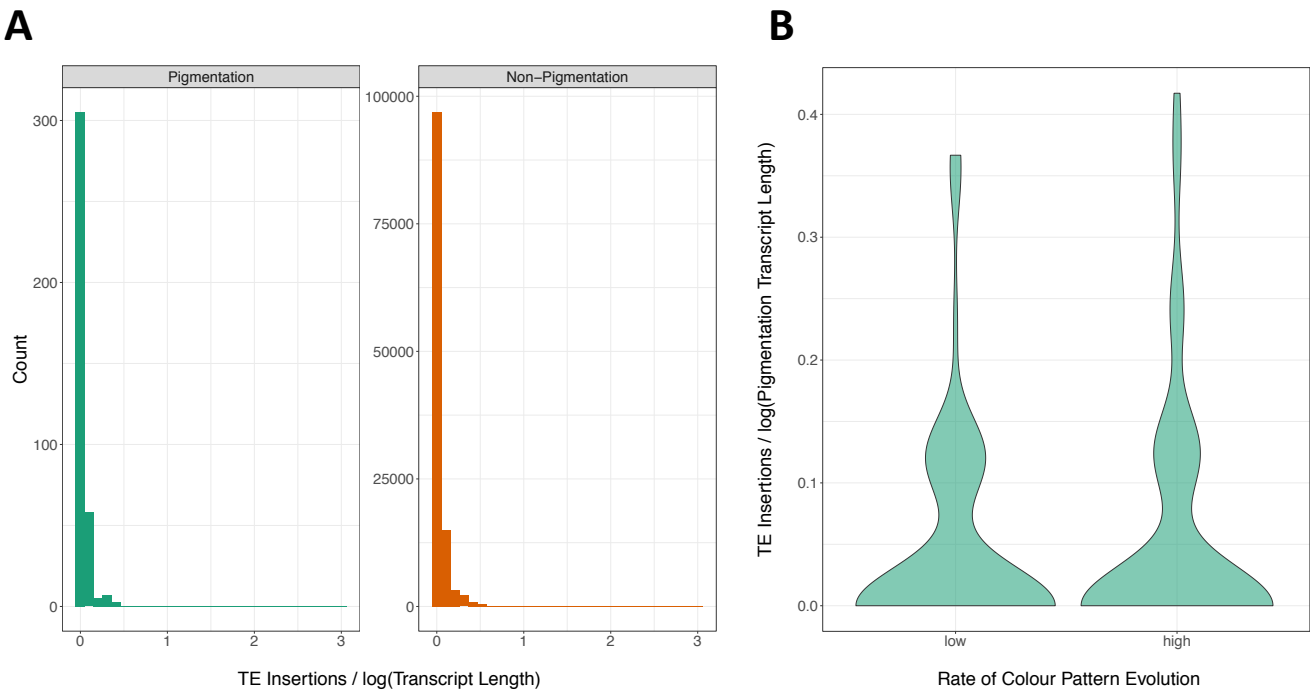


Figure 6.3. Number of TE insertions within Corydoradinae transcripts. **A.** Histogram comparing TE insertion number between pigmentation and non-pigmentation transcripts. **B.** Comparison of the number of TE insertions within pigmentation transcripts between Corydoradinae species with a low and high rate of colour pattern evolution. All insertions are given per log transcript length.

We also investigated whether there were cross-species variability in the number of TE insertions within VPGs, and whether this was associated between species with high or low rates of colour pattern evolution (Figure 6.3b). When using the longest transcript representatives of each shared VPG, there was no significant difference in the mean number of TE insertions per pigmentation transcript length between the low ($\bar{x}$ = 0.045 ± 0.010) and high ($\bar{x}$ = 0.052 ± 0.012) rate of colour

pattern Corydoradinae groups [RR:1.11, P = 0.35] (Table 6.3b). This non-significance remained when accounting for the paired gene structure between the two groups (Paired Wilcoxon signed rank: V = 246, n = 72, P = 0.38). Finally, the likelihood of TE presence/absence within pigmentation transcripts was not significantly different between both the low and high rate of colour pattern evolution groups [RR:1.28, P = 0.53] (Table 6.3b).

We also repeated all analyses on the reduced TPG gene set. The relationship between likelihood of TE presence/absence between non-pigmentation and pigmentation (VPG) transcript was no longer significant (RR 1.08, P = 0.8]. However, the number of TE insertions per TPG transcript length was found to be significantly greater within the higher (0.073 ± 0.028) than the lower rate (0.034 ± 0.019) of colour pattern evolution groups (RR 1.89, P < 0.01) (Supplementary Table 6.3). All other findings remained consistent. In summary, we found no clear evidence that the number of TE insertions is greater within Corydoradinae pigmentation versus non-pigmentation transcripts. In fact, the likelihood of TE presence may be higher within non-pigmentation transcripts, supporting the 'gene disruption' model of TE insertion dynamics (given the earlier finding that non-pigmentation Corydoradinae genes evolve under more relaxed selection pressure). We also report no consistent evidence that the number of TE insertions within pigmentation transcripts significantly differed across Corydoradinae species with variables rates of colour pattern evolution. However, when using the reduced TPG subset, we did find evidence that pigmentation transcripts contain a greater number of TE insertions within Corydoradinae with a high rate of colour pattern evolution.

Table 6.3 Output from multiple generalised lineage models (GLMs)investigating the relationship between (**A)** TE insertion number and transcript type (pigmentation versus non-pigmentation) & length **(B)** TE insertion number and rate of colour pattern (high or low) within the Corydoradinae.

A

Model	No Offset				Offset (*log(Transcript Length)*)			
		Estimates	P	AIC		Estimates	P	AIC
Poisson	Intercept	-2.296e+00	<0.001	153,510	Intercept	-1.334	<0.001	1,214,663
	Transcript type - Non-pigmentation	3.857e-01	<0.001					
	Transcript Length	2.713e-04	<0.001		Transcript type - Non-pigmentation	0.039	0.28	
Negative Binomial	Intercept	-2.591e+00	<0.001	144,695	Intercept	-1.334	<0.001	1,127,829
	Transcript type - Non-pigmentation	4.691e-01	<0.001					
	Transcript Length	3.710e-04	<0.001		Transcript type - Non-pigmentation	0.039	0.39	
Binomial Logistic Regression (Presence/Absence)	Intercept	-2.482e+00	<0.001	109,350	Intercept	-9.235	<0.001	108,156
	Transcript type - Non-pigmentation	3.337e-01	<0.05					
	Transcript Length	3.459e-04	<0.001		Transcript type – Non-pigmentation	0.374	<0.01	

B

Model	Offset (*log(Pigment Transcript Length)*			
		Estimates	P	AIC
Negative Binomial	Intercept	-0.95	<0.001	1949.3
	Colour pattern - High	0.10	0.35	
Binomial Logistic Regression (Presence/Absence)	Intercept	-9.12	<0.001	
	Colour pattern - High	0.24	0.53	163.33

There is a significant effect of pigmentation transcript region and likelihood of a TE insertion

Across Corydoradinae pigmentation transcripts, we investigated whether transcript location had a significant effect on the likelihood of a TE insertion (Figure 6.4). A negative binomial regression indicated that TE insertion rate across all three regions were significantly different from one other (Table 6.4). Both the mean number of TE insertions per CDS length (0.013 ± 0.003) and 3' length (0.029 ± 0.005) was significantly greater than the number of insertions per 5' length (0.003 ± 0.001) [RR 3.93, 8.60; P < 0.001]. The mean number of TE insertions per 3' region was also significantly greater than per CDS region [RR 2.19; P < 0.001] (Table 6.4).

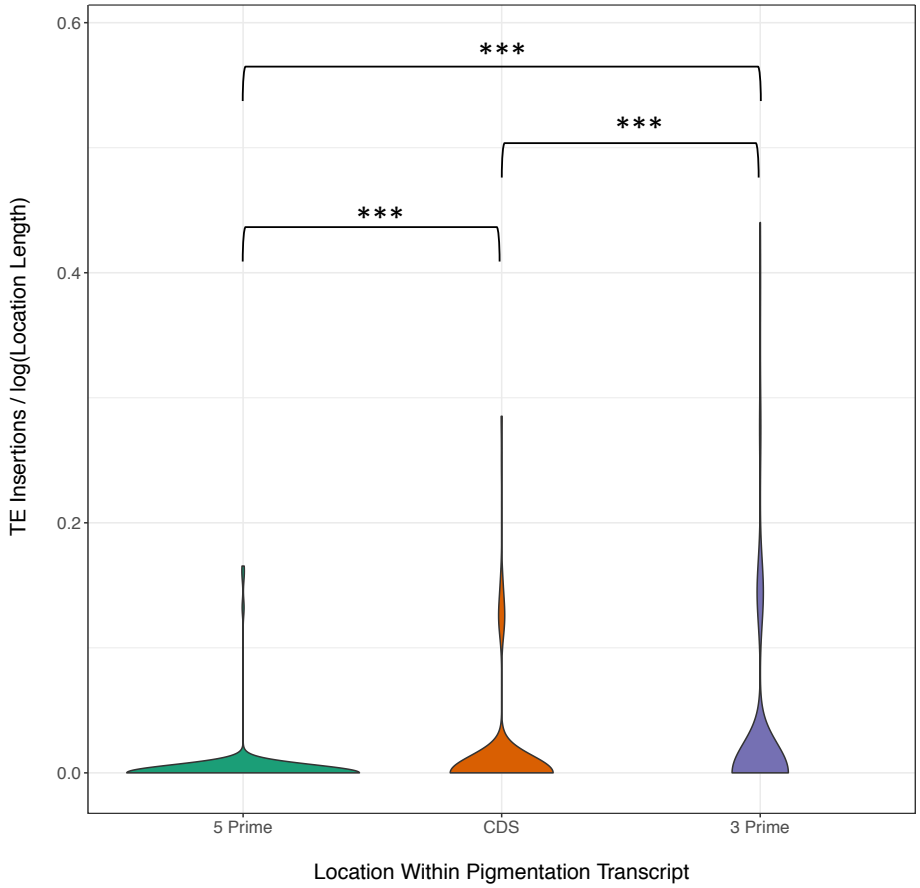


Figure 6.4 Number of TE insertions within different regions of Corydoradinae pigmentation transcripts (standardised per region length). Asterisks indicate degree of significant different (*** P < 0.001). CDS = coding sequence.

Table 6.4 Output from negative binomial regression indicating pigmentation transcript location (3', 5' CDS) has a significant effect on TE insertion number (offset by region length). CDS = coding sequence.

Model		**Offset** *(log(transcript location length))*		
		Estimate	P	AIC
Negative Binomial	Intercept	-3.62	< 0.001	3318.7
	Location - CDS	1.37	< 0.001	
	Location – 3'	2.15	< 0.001	

The upstream promoter region of Corydoradinae pigmentation genes are not enriched in transposable elements

TE abundance within gene promoter regions (2kb upstream) was compared between all 119 *C. fulleri* VPGs and 100 random non-pigmentation genes. We found no difference between the average number of TE insertions per VPG promoter ($\bar{x}$ = 2.80 ± 0.18) and non-pigmentation promoters ($\bar{x}$ = 2.63 ± 0.18) (Mann-Whitney U test: U = 5,727, n_1 = 119, n_2 = 100, P = 0.62) (Figure 6.5a). A non-significant difference in TE abundance was also observed within the equivalent 1kb downstream (3' UTR) region, whereby the mean number of insertions within VPGs and non-pigmentation genes was 2.25 ± 0.26 and 2.48 ± 0.26 respectively (standardised per 2kb) (Figure 6.5a). (Mann-Whitney U test: U = 6374.5, n_1 = 119, n_2 = 100, P = 0.34). Combined, there was a significant effect of gene position (upstream and downstream) and type (pigmentation and non-pigmentation) on number of TE insertions (Kruskal-Wallace test: H = 12.21, df = 3, P < 0.01), with a post-hoc Dunn's test indicating this was driven by higher TE insertions in upstream vs downstream regions within pigmentation genes (Z = 3.13, P < 0.05). All other pairwise comparisons were non-significant (Figure 6.5a). It was also noted that the number of TE insertions within both flanking pigmentation gene regions were over an order of magnitude higher than the equivalent observed within pigmentation transcript regions ($\bar{x}$ = 0.18 ± 0.03 per 2kb length).

In addition to abundance, we investigated the distribution of TE insertions across both gene types and found no significant difference in average distance from the 5' start codon within VPGs ($\bar{x}$ = 1,037 bp ± 30.5 bp) and non-pigmentation genes ($\bar{x}$ = 1,052 bp ± 31.9 bp) (Mann-Whitney U test of TE midpoints: U = 44,618, n_1= 333, n_2 = 263, P = 0.69) (Figure 6.5b). No significant difference in TE insertion distribution was also found within the downstream (3'UTR) sequences of both gene types (Mann-Whitney U test of TE midpoints: U = 7,747, n_1= 134, n_2 = 124, P = 0.35). These findings remained true when using the reduced TPG subset, although there was no significant difference between the upstream vs downstream location on TE insertion number within pigmentation genes was gone (Supplementary Figure 6.2a, 6.2b).

The majority of TEs within the promoter region of VPGs were Class II DNA elements (185/333, 56%), which is in almost exact proportion to their abundance across the entire *C. fulleri* genome (589,374/1,049,430, 56%). This suggests there is no bias in the class of TE that insert within the promoter region of Corydoradinae VPGs. However, the sequence divergence from the consensus of the DNA elements present within the promoter regions of VPGs (13.97 ± 0.45) was significantly lower than the genome-wide average (15.70 ± 0.01) (Welch's t = 3.84, df = 184.14, P < 0.001), suggesting that the TEs associated with VPGs may represent younger, more recent insertions (Figure 6.5c). In summary, although TEs found within the promoter region of Corydoradinae pigmentation genes are younger than the genome-wide average, there is no evidence that they are present in greater number or positioned closer to the gene start codon than that of non-pigmentation genes.

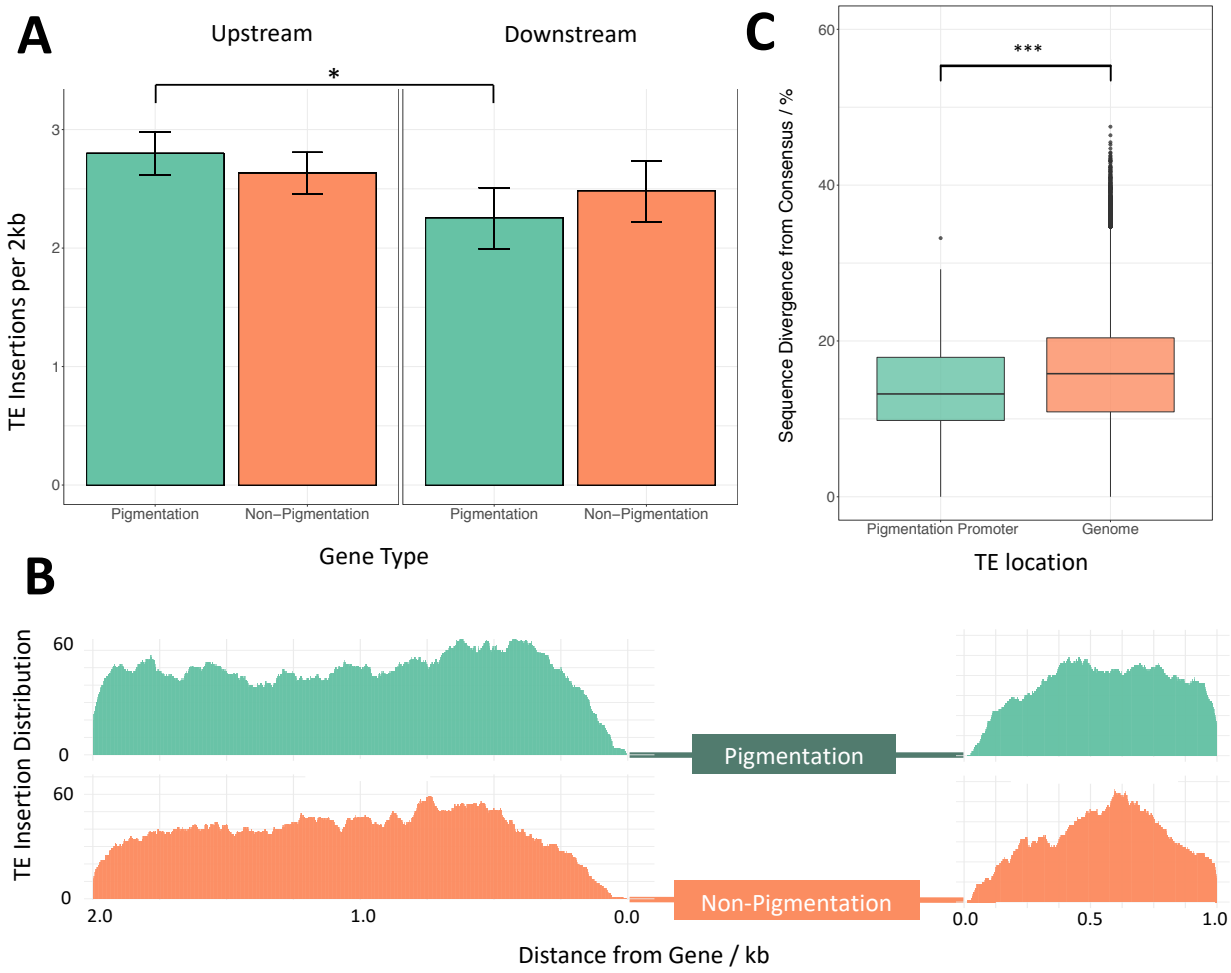


Figure 6.5. A. Mean number of TE insertions within the 5' promoter (upstream) and 3' UTR (downstream) regions of VPG and non-pigmentation genes. **B.** TE insertion count at each base pair position within the 5' upstream and 3' downstream region of pigmentation and non-pigmentation genes. **C.** The age of DNA elements found within VPG promoters vs the genome wide average (using sequence divergence from TE consensus as a proxy). Asterisks indicate significance level (* $P < 0.05$,*** $P < 0.001$). Upstream and downstream regions for both gene types were extracted from the genome of *C. fulleri.*

Pairwise log fold difference in Corydoradinae pigmentation gene expression is not greater than that of non-pigmentation genes.

We also investigated if TE insertions within the promoter regions of the Corydoradinae may be driving differential pigmentation gene expression between species. However, we found no significant relationship between the number of TE insertions within the upstream region of individual pigmentation genes in *C. fulleri* and their mean log-fold expression difference across Corydoradinae species (Supplementary Figure 6.3). Furthermore, we investigated whether pigmentation genes may have greater inter-species expression variation, driving pigmentation

diversity across the Corydoradinae independently of TE insertion patterns. However, cross-species comparisons found no significant difference in the mean pairwise log2 fold difference of VPG expression (0.96 ± 0.02) versus non-pigmentation gene expression (1.05 ± 0.02) (Mann-Whitney *U*-test: U=1,976,988, n_1 = 1,800, n_2 = 2,268, P = 0.08) (Figure 6.6a). Finally, the proportion of VPGs that met a log2 fold change cut off > 2 in comparisons within species with a low (low vs low) and high (high vs high) rate of colour pattern evolution was 0.14 ± 0.03 and 0.07 ± 0.01 respectively (Figure 6.6b). In comparisons between low and high rate of colour pattern, this proportion was 0.10 ± 0.01 (Figure 6.6b). Overall, pairwise species comparisons indicated that there was no significant effect of rate of colour pattern differences/similarity on the proportion of differentially expressed VPGs (ANOVA: F = 2.44; d.f = 2, 33; P = 0.10). These findings were also matched within the reduced TPG subset (Supplementary Figure 6.4a, 6.4b). In summary, there was no evidence that Corydoradinae VPGs are any more likely to be differentially expressed than non-pigmentation genes (through TE insertions or otherwise), or that differences in pairwise expression level may be driving disparate rates of colour pattern evolution among lineages.

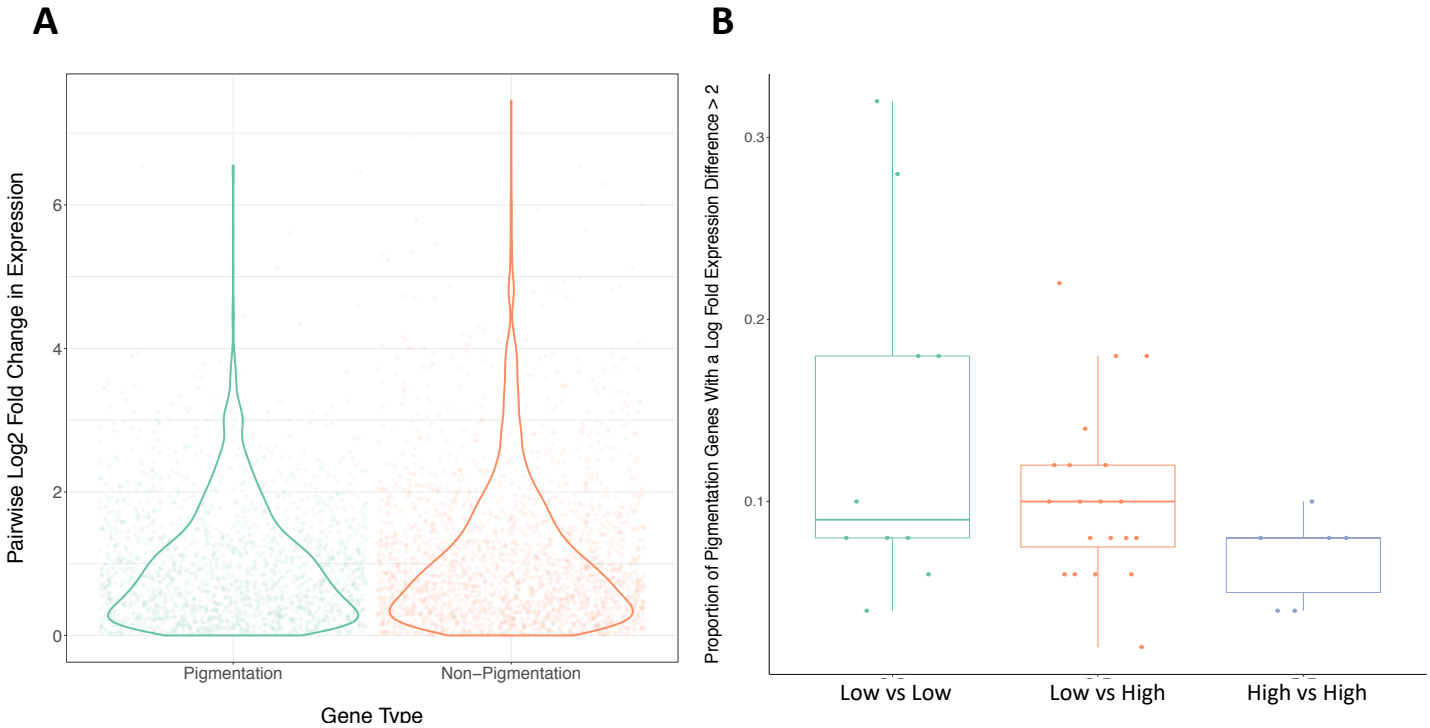


Figure 6.6 Differential expression of Corydoradinae vertebrate pigmentation genes (VPGs) in pairwise species comparisons. **A.** Overall log2-fold differences in Corydoradinae VPG vs non-pigmentation genes. **B.** Proportion of VPGs with a log2-fold difference > 2 between Corydoradinae species of high or low rates of colour pattern evolution.

6.6 Discussion

To investigate the potential for TE insertions to induce colour pattern change under a 'gene disruption' model, this study primarily investigated (i) the Kn/Ks ratio and (ii) the number of TE insertions within the pigmentation loci of a colour pattern diverse clade of catfish (the Corydoradinae). Rate of colour pattern evolution across different Corydoradinae lineages was found to be highly variable, with nuclear DNA lineages 6-9 exhibiting a greater number of unique pigmentation patterns per MY than lineages 1-5. This difference in rate of colour pattern evolution may partially be explained by the varying species richness of each lineage, though because both colour pattern diversity and species number are tightly correlated these factors are difficult to disentangle. Alternatively, disparity in colour pattern richness may be driven by a greater number of TE induced changes within pigmentation loci. Given that the Corydoradinae lineages with a higher rate of colour pattern change are those with greater TE abundance it is possible that these two variables causally interact (Marburger *et al.* 2018).

Within the Corydoradinae, Kn/Ks ratio comparisons between pigmentation and non-pigmentation genes indicated that their pigmentation genes are under stricter purifying selection than the transcriptome-wide average, with none being under positive selection (Kn/Ks > 1). This finding contrasts with that reported in cichlids, whereby differentially expressed pigmentation genes involved in egg spot development are found to have a higher Kn/Ks ratio than the transcriptome-wide average (Santos *et al.* 2016). Furthermore, the Kn/Ks ratio of pigmentation genes did not appear to vary between Corydoradinae lineages, indicating that differing rates of colour pattern evolution between species is not driven by varying rates of molecular evolution within pigmentation genes. We hypothesise that this finding may be the result of Müllerian mimicry, whereby the fact that Corydoradinae species have converged on similar colour patterns may mean their pigmentation genes under strong purifying selection to maintain protection from predators (colour

monomorphism) (Briolat *et al.* 2019). Comparisons between human-mouse orthologous have demonstrated that genes under stricter selection pressure are significantly less likely to be associated with TE insertions (Mortada *et al.* 2010). We would therefore predict under a gene disruption model of TE distribution that Corydoradinae pigmentation genes may be depleted in TE insertions.

Supporting this, TE presence was significantly more likely within non-pigmentation versus pigmentation transcripts. This suggests the 'gene-disruption' model of TE insertion may be more accurate than the 'gene-age' model (Correa *et al*. 2021). The proportion of pigmentation transcripts containing a TE (~20%) was not only being concordant with the complete set of Corydoradinae gene-coding transcripts, but also to that reported within humans (~21%) (Babarinde *et al.* 2021). We did identify a significant greater number of TEs insertion within the pigmentation transcripts of Corydoradinae species with a high (lineage 6-9) rate of colour pattern evolution versus low (lineage 1-5), although this difference was only significant when using the reduced TPG set. However, combined with the lack of relationship with the full pigmentation gene set and the absence of differences in TE absence/presence leads us to conclude that processes other than TE accumulation are likely to drive disparate rates of colour pattern evolution across the Corydoradinae. We also report that TE insertion likelihood depended on transcript region. TEs were significantly more likely to be present within the 3' UTR of a pigmentation gene transcript than either the 5' UTR or CDS. This mirrors the insertion distribution pattern of Alu elements within humans, whereby the percentage of transcripts containing a TE insertion within the 3' UTR is higher than those with an insertion within the 5' UTR or CDS (Moolhuijzen *et al.* 2010). This may reflect the fact that biological processes associated with 3' UTR insertions (e.g. mRNA decay or localisation effects) are under more relaxed selection than those associated with 5' UTR insertions (e.g. the modulation of mRNA translation efficiency) (Chuong *et al.* 2016). It has also been suggested that TEs may be predisposed for 3' UTR

exonisation, though the mechanistic reasons why this may be the case remain unknown (Kapusta *et al.* 2013).

The average number of TE insertions within the promoter region of *C. fulleri* pigmentation genes were over an order of magnitude higher than the estimated rate within pigmentation CDS regions (2.8 vs 0.18 insertions per 2kb). This is similar to TE insertion patterns within multiple plant species, including, Maize (*Zea mays*) bamboo (*Phyllostachys edulis)* and *Capsella* sp. (Zhou *et al.* 2016; Niu *et al.* 2019; Zhang & Qi 2019), and is to be expected if TEs which disrupt a gene coding region are purged by host defence mechanisms. However, similar to pigmentation transcripts, we report no evidence that TE insertions are more prevalent within the promoter region (or downstream 3' UTR) of Corydoradinae pigmentation genes when compared to non-pigmentation genes. Most TE insertions within *C. fulleri* pigmentation promoters were class II DNA elements, which were on average younger than the genome-wide average and may therefore represent more recent insertions. DNA elements may be found within or nearby gene regions at a greater rate than other TE types, potentially because they do not carry their own internal cis-promoter sequences and have less potential to affect nearby gene expression (Chang *et al*. 2022).

Gene expression may be impacted by the proximity of nearby TEs, with increasing distance to the nearest insertion being positively related to gene expression level within *Populus* plants (Zhao *et al.* 2022). Pairwise species comparisons in log-fold gene expression variation within the Corydoradinae found that pigmentation genes are on average no more differentially expressed than non-pigmentation genes. This is somewhat expected given the equal number of TE insertions upstream and within the same distance to coding start sites across both gene types. At a finer scale, we found no relationship between the number of TE insertions within the upstream region of Corydoradinae pigmentation genes and the magnitude of log-fold expression change between species. Furthermore, the proportion of differentially expressed pigmentation genes did not vary between

Corydoradinae lineages of varying colour pattern rates. This finding may reflect the fact that the TE insertions found within the promoter region of pigmentation loci within *C. fulleri* are inherited from a common ancestor and shared across multiple Corydoradinae species, a hypothesis that could be tested with a greater number of annotated Corydoradinae genomes available. TEs also appear to play an overall marginal effect in driving expression divergence between human-chimp orthologues (Warnefors *et al.* 2010). It is also possible that the age or tissue type where RNA was extracted from our Corydoradinae samples may not best capture the developmental stages where pigmentation differentiation is occurring.

The list of pigmentation genes used in this study is likely to be neither exhaustive nor balanced with regards to their contribution to pigmentation development versus other biological processes. Pleiotropic effects mean that several pigmentation genes may have important roles beyond colour pattern formation, with 75% of human pigmentation genes have demonstrable roles outside of pigmentation development for example (Baxter *et al.* 2019). This issue is further complicated within teleosts because a whole genome duplication early in their diversification has meant many duplicated genes may have undergone non, neo, or sub-functionalisation (Braasch *et al*. 2008). However, when repeating our analysis on a subset of vertebrate pigmentation genes where experimental evidence has highlighted an important role within teleost colour pattern development and that gene knock-outs do not cause host inviability, our results remain almost entirely consistent. One exception was that the Kn/Ks ratio of this reduced set of pigmentation genes was not significantly lower than non-pigmentation genes, suggesting that this subset has filtered out those genes with crucial functions outside of pigmentation development. Despite our finding that Corydoradinae pigmentation sequences are not enriched in TE insertions, it remains possible that colour pattern change within the Corydoradinae may be driven by TE polymorphisms within just one of the many pigmentation genes investigated, particularly those which are part of larger gene networks. The interaction between single TE insertions and pigmentation change within teleosts has

been documented in a handful of cases. An exonic TE insertion is the known determinant of albinism in the medaka fish *(Oryzias latipes*) for example, whilst a gold/dark polymorphism is maintained within *Amphilophus* cichlid species through a TE derived insertion within the *'goldentouch'* gene (Koga & Hori 1997; Kratochwil *et al.* 2022).

6.7 Conclusion and Outlook

TEs are abundant components of many species' genomes and powerful agents of genetic change. Whilst several studies have focussed on the phenotypic impacts of TE insertions within a single gene type, little is known about broader patterns of TE activity and whether certain gene types may be disproportionately affected by their insertions. This study aimed to address this by investigating whether pigmentation genes within a colour-diverse clade of Neotropical catfish contain a higher-than-average number of TE insertions. Surprisingly, Corydoradinae pigmentation genes were found to evolve under stricter than average purifying selection pressure, and, as expected under a gene disruption model of TE distribution, the upstream, downstream and transcript sequences of these genes were not found to be inflated in TE insertions. This study therefore suggests that any inferences of a link between biased TE activity and pigmentation change may simply reflect the fact that such insertions cause an obvious phenotypic change. Future work may wish to focus on (i) how phylogenetically comparable these patterns are, (ii) the exact functional impacts of TE insertions within Corydoradinae colour pattern genes and (iii) whether other gene families are TE enriched.

6.8 References

Alexandrou, M.A., Oliveira, C., Maillard, M., Mcgill, R.A.R., Newton, J., Creer, S., *et al.* (2011). Competition and phylogeny determine community structure in Müllerian co-mimics. *Nature*, 496, 84–88.

Babarinde, I.A., Ma, G., Li, Y., Deng, B., Luo, Z., Liu, H., *et al.* (2021). Transposable element sequence fragments incorporated into coding and noncoding transcripts modulate the transcriptome of human pluripotent stem cells. *Nucleic Acids Res.*, 49, 9132–9153.

Baxter, L.L., Watkins-Chow, D.E., Pavan, W.J. & Loftus, S.K. (2019). A curated gene list for expanding the horizons of pigmentation biology. *Pigment Cell Melanoma Res.*, 32, 348–358.

Bell, E.A., Butler, C.L., Oliveira, C., Marburger, S., Yant, L. & Taylor, M.I. (2022). Transposable element annotation in non-model species: The benefits of species-specific repeat libraries using semi-automated EDTA and DeepTE de novo pipelines. *Mol. Ecol. Resour.*, 22, 823–833.

Bell, E.A., Cable, J., Oliveira, C., Richardson, D.S., Yant, L. & Taylor, M.I. (2020). Help or hindrance? The evolutionary impact of whole-genome duplication on immunogenetic diversity and parasite load. *Ecol. Evol.*, 10, 13949–13956.

Bigler, J., Rand, H.A., Kerkof, K., Timour, M. & Russell, C.B. (2013). Cross-Study Homogeneity of Psoriasis Gene Expression in Skin across a Large Expression Range. *PLoS One*, 8, e52242.
Bolger, A.M., Lohse, M. & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30, 2114–2120.
Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., *et al.* (2018). Ten things you should know about transposable elements. *Genome Biol.* , 19, 1–12.
Braasch, I., Volff, J.N. & Schartl, M. (2008). The evolution of teleost pigmentation and the fish-specific genome duplication. *J. Fish Biol.*, 73, 1891–1918.
Briolat, E.S., Burdfield-Steel, E.R., Paul, S.C., Rönkä, K.H., Seymoure, B.M., Stankowich, T., *et al.* (2019). Diversity in warning coloration: selective paradox or the norm? *Biol. Rev.*, 94, 388–414.
Bush, S.E., Villa, S.M., Altuna, J.C., Johnson, K.P., Shapiro, M.D. & Clayton, D.H. (2019). Host defense triggers rapid adaptive radiation in experimentally evolving parasites. *Evol. Lett.*, 3, 120–128.
Chang, N.-C., Rovira, Q., Wells, J.N., Feschotte, C. & Vaquerizas, J.M. (2022). Zebrafish transposable elements show extensive diversification in age, genomic distribution, and developmental expression. *Genome Res.*, Epub ahead of print.
Choi JY & Lee YCG. (2020). Double-edged sword: The evolutionary consequences of the epigenetic silencing of transposable elements. *PLOS Genet.*, 16, e1008872.
Chuong, E.B., Elde, N.C. & Feschotte, C. (2016). Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* , 18, 71–86.
Correa, M., Lerat, E., Birmelé, E., Samson, F., Bouillon, B., Normand, K., *et al.* (2021). The Transposable Element Environment of Human Genes Differs According to Their Duplication Status and Essentiality. *Genome Biol. Evol.*, 13.
Cowley, M. & Oakey, R.J. (2013). Transposable Elements Re-Wire and Fine-Tune the Transcriptome. *PLOS Genet.*, 9, e1003234.
Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., *et al.* (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* , 29, 644–652.
Hemingson, C.R., Cowman, P.F., Hodge, J.R. & Bellwood, D.R. (2019). Colour pattern divergence in reef fish species is rapid and driven by both range overlap and symmetry. *Ecol. Lett.*, 22, 190–199.
Hof, A.E.V. t., Campagne, P., Rigden, D.J., Yung, C.J., Lingley, J., Quail, M.A., *et al.* (2016). The industrial melanism mutation in British peppered moths is a transposable element. *Nat.* , 534, 102–105.
Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M., *et al.* (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nat.* , 496, 498–503.
Humann, J.L., Lee, T., Ficklin, S. & Main, D. (2019). Structural and Functional Annotation of Eukaryotic Genomes with GenSAS. *Methods Mol. Biol.* , 1962, 29–51.
Iida, S., Morita, Y., Choi, J.D., Park, K. Il & Hoshino, A. (2004). Genetics and epigenetics in flower pigmentation associated with transposable elements in morning glories. *Adv. Biophys.*, 38, 141–159.
Irion, U. & Nüsslein-Volhard, C. (2019). The identification of genes involved in the evolution of color patterns in fish. *Curr. Opin. Genet. Dev.*, 57, 31–38.
Joron, M., Frezal, L., Jones, R.T., Chamberlain, N.L., Lee, S.F., Haag, C.R., *et al.* (2011). Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nat.* , 477, 203–206.
Kapusta, A., Kronenberg, Z., Lynch, V.J., Zhuo, X., Ramsay, L.A., Bourque, G., *et al.* (2013). Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. *PLOS Genet.*, 9, e1003470.
Katoh, K., Misawa, K., Kuma, K.I. & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, 30, 3059–3066.

Khabbazian, M., Kriebel, R., Rohe, K. & Ané, C. (2016). Fast and accurate detection of evolutionary shifts in Ornstein–Uhlenbeck models. *Methods Ecol. Evol.*, 7, 811–824.
Koga, A. & Hori, H. (1997). Albinism Due to Transposable Element Insertion in Fish. *Pigment Cell Res.*, 10, 377–381.
Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I. & Koonin, E. V. (2002). Selection in the evolution of gene duplications. *Genome Biol.*, 3, 1–9.
Kratochwil, C.F., Kautt, A.F., Nater, A., Härer, A., Liang, Y., Henning, F., *et al.* (2022). An intronic transposon insertion associates with a trans-species color polymorphism in Midas cichlid fishes. *Nat. Commun.* , 13, 1–8.
Lanciano, S. & Cristofari, G. (2020). Measuring and interpreting transposable element expression. *Nat. Rev. Genet.*, 21, 721–736.
Lewis, J.J. & Van Belleghem, S.M. (2020). Mechanisms of Change: A Population-Based Perspective on the Roles of Modularity and Pleiotropy in Diversification. *Front. Ecol. Evol.*, 8.
Li, W.H. (1993). Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* , 36, 96–99.
Liao, Y., Smyth, G.K. & Shi, W. (2019). The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.*, 47.
Lin, M., Siford, R.L., Martin, A.R., Nakagome, S., Möller, M., Hoal, E.G., *et al.* (2018). Rapid evolution of a skin-lightening allele in southern African KhoeSan. *Proc. Natl. Acad. Sci. U. S. A.*, 115, 13324–13329.
Liu, Z., Liu, S., Yao, J., Bao, L., Zhang, J., Li, Y., *et al.* (2016). The channel catfish genome sequence provides insights into the evolution of scale formation in teleosts. *Nat. Commun.* , 7, 1–13.
Lorin, T., Brunet, F.G., Laudet, V. & Volff, J.N. (2018). Teleost Fish-Specific Preferential Retention of Pigmentation Gene-Containing Families After Whole Genome Duplications in Vertebrates. *G3 Genes|Genomes|Genetics*, 8, 1795–1806.
Marburger, S., Alexandrou, M.A., Taggart, J.B., Creer, S., Carvalho, G., Oliveira, C., *et al.* (2018). Whole genome duplication and transposable element proliferation drive genome expansion in Corydoradinae catfishes. *Proc. R. Soc. B Biol. Sci.*, 285.
McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci.*, 36, 344–355.
McElduff, F., Cortina-Borja, M., Chan, S.K. & Wade, A. (2010). When t-tests or Wilcoxon-Mann-Whitney tests won't do. *Adv. Physiol. Educ.*, 34, 128–133.
Merrill, R.M., Chia, A. & Nadeau, N.J. (2014). Divergent warning patterns contribute to assortative mating between incipient Heliconius species. *Ecol. Evol.*, 4, 911–917.
Moolhuijzen, P., Kulski, J.K., Dunn, D.S., Schibeci, D., Barrero, R., Gojobori, T., *et al.* (2010). The transcript repeat element: The human Alu sequence as a component of gene networks influencing cancer. *Funct. Integr. Genomics*, 10, 307–319.
Mortada, H., Vieira, C. & Lerat, E. (2010). Genes Devoid of Full-Length Transposable Element Insertions are Involved in Development and in the Regulation of Transcription in Human and Closely Related Species. *J. Mol. Evol.*, 71, 180–191.
Nishikawa, H., Iijima, T., Kajitani, R., Yamaguchi, J., Ando, T., Suzuki, Y., *et al.* (2015). A genetic mechanism for female-limited Batesian mimicry in Papilio butterfly. *Nat. Genet.* , 47, 405–409.
Niu, X.M., Xu, Y.C., Li, Z.W., Bian, Y.T., Hou, X.H., Chen, J.F., *et al.* (2019). Transposable elements drive rapid phenotypic variation in Capsella rubella. *Proc. Natl. Acad. Sci. U. S. A.*, 116, 6908–6913.
Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J., *et al.* (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.*, 20, 1–18.
Paradis, E., Claude, J. & Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, 20, 289–290.
Peona, V., Weissensteiner, M.H. & Suh, A. (2018). How complete are "complete" genome

assemblies?—An avian perspective. *Mol. Ecol. Resour.*, 18, 1188–1195.
Pereira, V. (2004). Insertion bias and purifying selection of retrotransposons in the Arabidopsis thaliana genome. *Genome Biol.*, 5, 1–10.
Ranwez, V., Harispe, S., Delsuc, F. & Douzery, E.J.P. (2011). MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons. *PLoS One*, 6, e22594.
Ravi, V. & Venkatesh, B. (2018). The Divergent Genomes of Teleosts. *Annu. Rev. Anim. Biosci.*, 6, 47–68.
Ricci, M., Peona, V., Guichard, E., Taccioli, C. & Boattini, A. (2018). Transposable Elements Activity is Positively Related to Rate of Speciation in Mammals. *J. Mol. Evol.*, 86, 303–310.
Ripley, B. & Venables, B. (2002). *Modern Applied Statistics with S*. 4th edn. Springer, New York.
Robinson, M.D., McCarthy, D.J. & Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26, 139–140.
Salis, P., Lorin, T., Lewis, V., Rey, C., Marcionetti, A., Escande, M.L., *et al.* (2019). Developmental and comparative transcriptomic identification of iridophore contribution to white barring in clownfish. *Pigment Cell Melanoma Res.*, 32, 391–402.
Santos, M.E., Baldo, L., Gu, L., Boileau, N., Musilova, Z. & Salzburger, W. (2016). Comparative transcriptomics of anal fin pigmentation patterns in cichlid fishes. *BMC Genomics*, 17, 1–16.
Santos, M.E., Braasch, I., Boileau, N., Meyer, B.S., Sauteur, L., Böhne, A., *et al.* (2014). The evolution of cichlid fish egg-spots is linked with a cis-regulatory change. *Nat. Commun.*, 5.
Schrader, L. & Schmitz, J. (2019). The impact of transposable elements in adaptive evolution. *Mol. Ecol.*, 28, 1537–1549.
Seehausen, O., Mayhew, P.J. & Van Alphen, J.J.M. (1999). Evolution of colour patterns in East African cichlid fish. *J. Evol. Biol.*, 12, 514–534.
Shao, F., Han, M. & Peng, Z. (2019). Evolution and diversity of transposable elements in fish genomes. *Sci. Reports* , 9, 1–8.
Shao, F., Pan, H., Li, P., Ni, L., Xu, Y. & Peng, Z. (2021). Chromosome-Level Genome Assembly of the Asian Red-Tail Catfish (Hemibagrus wyckioides). *Front. Genet.*, 12.
Shen, W., Le, S., Li, Y. & Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS One*, 11, e0163962.
Sinnett, D., Beaulieu, P., Bélanger, H., Lefebvre, J.F., Langlois, S., Théberge, M.C., *et al.* (2006). Detection and characterization of DNA variants in the promoter regions of hundreds of human disease candidate genes. *Genomics*, 87, 704–710.
Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J.M. & Kelly, S. (2016). TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Res.*, 26, 1134–1144.
Sotero-Caio, C.G., Platt, R.N., Suh, A. & Ray, D.A. (2017). Evolution and Diversity of Transposable Elements in Vertebrate Genomes. *Genome Biol. Evol.*, 9, 161–177.
Stapley, J., Santure, A.W. & Dennis, S.R. (2015). Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species. *Mol. Ecol.*, 24, 2241–2252.
Sturm, R.A. & Duffy, D.L. (2012). Human pigmentation genes under environmental selection. *Genome Biol.*, 13, 1–15.
Tarazona, S., Furió-Tarí, P., Turrà, D., Di Pietro, A., Nueda, M.J., Ferrer, A., *et al.* (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res.*, 43.
Telias, A., Lin-Wang, K., Stevenson, D.E., Cooney, J.M., Hellens, R.P., Allan, A.C., *et al.* (2011). Apple skin patterning is associated with differential expression of MYb10. *BMC Plant Biol.*, 11, 1–15.
Van't Hof, A.E., Campagne, P., Rigden, D.J., Yung, C.J., Lingley, J., Quail, M.A., *et al.* (2016). The industrial melanism mutation in British peppered moths is a transposable element. *Nature*, 534, 102–105.
Warnefors, M., Pereira, V. & Eyre-Walker, A. (2010). Transposable Elements: Insertion Pattern and Impact on Gene Expression Evolution in Hominids. *Mol. Biol. Evol.*, 27, 1955–1962.
Waterland, R.A. & Jirtle, R.L. (2003). Transposable Elements: Targets for Early Nutritional Effects on

Epigenetic Gene Regulation. *Mol. Cell. Biol.*, 23, 5293–5300.
Wilde, S., Timpson, A., Kirsanow, K., Kaiser, E., Kayser, M., Unterländer, M., *et al.* (2014). Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. *Proc. Natl. Acad. Sci. U. S. A.*, 111, 4832–4837.
Xiong, P., Hulsey, C.D., Meyer, A. & Franchini, P. (2018). Evolutionary divergence of 3' UTRs in cichlid fishes. *BMC Genomics*, 19, 1–11.
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.
Zeng, L., Pederson, S.M., Kortschak, R.D. & Adelson, D.L. (2018). Transposable elements and gene expression during the evolution of amniotes. *Mob. DNA*, 9, 1–9.
Zhang, X. & Qi, Y. (2019). The Landscape of Copia and Gypsy Retrotransposon During Maize Domestication and Improvement. *Front. Plant Sci.*, 10.
Zhang, X., Zhao, M., McCarty, D.R. & Lisch, D. (2020). Transposable elements employ distinct integration strategies with respect to transcriptional landscapes in eukaryotic genomes. *Nucleic Acids Res.*, 48, 6685–6698.
Zhao, Y., Li, X., Xie, J., Xu, W., Chen, S., Zhang, X., *et al.* (2022). Transposable Elements: Distribution, Polymorphism, and Climate Adaptation in Populus. *Front. Plant Sci.*, 13.
Zhou, M., Tao, G., Pi, P., Zhu, Y., Bai, Y. & Meng, X. (2016). Genome-wide characterization and evolution analysis of miniature inverted-repeat transposable elements (MITEs) in moso bamboo (Phyllostachys heterocycla). *Planta*, 244, 775–787.

6.9 Supplementary Tables and Figures

Supplementary Table 6.1. Transcriptome Assembly statistics for each of the nine Corydoradinae species assembled de-novo using Trinity.

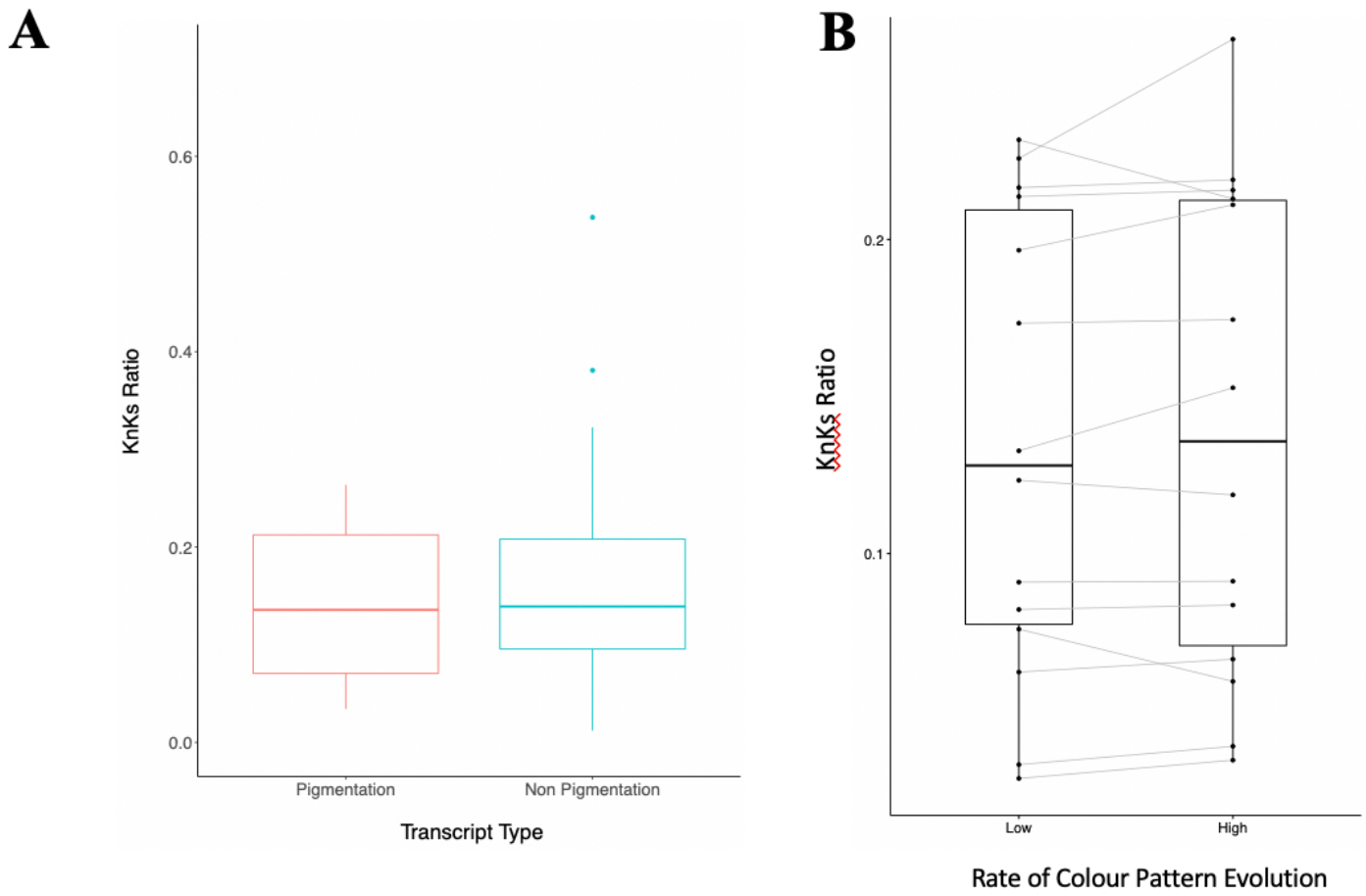
Species	Transcripts	Total Length (MB)	GC (%)	N50 (bp)	TransRate Assembly Score
C. maculifer	66,579	39.46	46.43	818	0.12
A. fuscoguttatus	78,164	73.95	45.52	1,688	0.13
S. prionotus	105,310	110.33	46.06	2,025	0.08
C. hastatus	69,798	34.04	43.59	584	0.06
C. elegans	113,590	94.86	44.80	1,492	0.08
C. paleatus	118,445	91.16	44.09	1,326	0.07
C. aeneus	78,189	55.98	43.90	1,081	0.10
C. haraldschultzei	115,020	89.20	45.58	1,320	0.11
C. araguaiaensis	65,992	35.62	43.18	706	0.10

Supplementary Table 6.2 Subset of Lorin's et al (2018) vertebrate pigmentation gene candidates where experimental and/or differential expression evidence indicates that the gene has a role in altering pigmentation patterns within teleosts. Here on in this subset is referred to as teleost pigmentation gene (TPG) candidates.

Pigmentation Gene	Species	Reference
bnc2	*Danio rerio*	Patterson, L. B., & Parichy, D. M. (2013). Interactions with iridophores and the tissue environment required for patterning melanophores and xanthophores during zebrafish adult pigment stripe formation. *PLoS Genetics*, *9*(5), e1003561.
dct	*Salmo marmoratus*	Sivka, U., Snoj, A., Palandačić, A., & Bajec, S. S. (2013). Identification of candidate genes involved in marble color pattern formation in genus *Salmo*. *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics*, *8*(3), 244-249.
ece2	*Danio rerio*	Singh, A. P., & Nüsslein-Volhard, C. (2015). Zebrafish stripes as a model for vertebrate colour pattern formation. *Current Biology*, *25*(2), R81-R92.
edn3	*Danio rerio*	Spiewak, J. E., Bain, E. J., Liu, J., Kou, K., Sturiale, S. L., Patterson, L. B., ... & Parichy, D. M. (2018). Evolution of Endothelin signaling and diversification of adult pigment pattern in Danio fishes. *PLoS genetics*, *14*(9), e1007538
ednrb	*Danio rerio*	Spiewak, J. E., Bain, E. J., Liu, J., Kou, K., Sturiale, S. L., Patterson, L. B., ... & Parichy, D. M. (2018). Evolution of Endothelin signaling and diversification of adult pigment pattern in Danio fishes. *PLoS genetics*, *14*(9), e1007538
erbb3	*Danio rerio*	Budi, E. H., Patterson, L. B., & Parichy, D. M. (2008). Embryonic requirements for ErbB signaling in neural crest development and adult pigment pattern formation. *Development* (2008) 135 (15): 2603–2614.
fhl2	*Astatotilapia burtoni, Cynotilapia pulpican*	Santos, M. E., Braasch, I., Boileau, N., Meyer, B. S., Sauteur, L., Böhne, A., ... & Salzburger, W. (2014). The evolution of cichlid fish egg-spots is linked with a cis-regulatory change. *Nature Communications*, *5*(1), 1-11.

foxd3	*Carassius carassius*	Zhang, Y., Liu, J., Fu, W., Xu, W., Zhang, H., Chen, S., ... & Xiao, Y. (2017). Comparative Transcriptome and DNA methylation analyses of the molecular mechanisms underlying skin color variations in Crucian carp (*Carassius carassius L.*). *BMC Genetics*, *18*(1), 1-12.
gfpt1	*Danio rerio*	Yang, C. T., Hindes, A. E., Hultman, K. A., & Johnson, S. L. (2007). Mutations in gfpt1 and skiv2l2 cause distinct stage-specific defects in larval melanocyte regeneration in zebrafish. *PLoS Genetics*, *3*(6), e88.
gnaq	*Sinocyclocheilus* spp.	Li, C., Chen, H., Zhao, Y., Chen, S., & Xiao, H. (2020). Comparative transcriptomics reveals the molecular genetic basis of pigmentation loss in *Sinocyclocheilus* cavefishes. *Ecology and evolution*, *10*(24), 14256-14271.
hps5	*Gasterosteus aculeatus*	Hart, J. C., & Miller, C. T. (2017). Sequence-based mapping and genome editing reveal mutations in stickleback Hps5 cause oculocutaneous albinism and the casper phenotype. *G3: Genes, Genomes, Genetics*, *7*(9), 3123-3131
kita	*Poecilia reticulata*	Kottler, V. A., Fadeev, A., Weigel, D., & Dreyer, C. (2013). Pigment pattern formation in the guppy, Poecilia reticulata, involves the Kita and Csf1ra receptor tyrosine kinases. *Genetics*, *194*(3), 631-646.
med12	*Danio rerio*	Rau, M. J., Fischer, S., & Neumann, C. J. (2006). Zebrafish Trap230/Med12 is required as a coactivator for Sox9-dependent neural crest, cartilage and ear development. *Developmental biology*, *296*(1), 83-93.
mpv17	*Danio rerio*	Krauss, J., Astrinides, P., Frohnhöfer, H. G., Walderich, B., & Nüsslein-Volhard, C. (2013). *transparent*, a gene affecting stripe formation in Zebrafish, encodes the mitochondrial protein Mpv17 that is required for iridophore survival. *Biology Open*, *2*(7), 703-710.
myc	*Danio rerio*	Le Guyader, S., Maier, J., & Jesuthasan, S. (2005). Esrom, an ortholog of PAM (protein associated with c-myc), regulates pteridine synthesis in the zebrafish. *Developmental Biology*, *277*(2), 378-386
nf1	*Danio rerio*	Shin, J., Padmanabhan, A., De Groh, E. D., Lee, J. S., Haidar, S., Dahlberg, S., ... & Look, A. T. (2012). Zebrafish neurofibromatosis type 1 genes have redundant functions in tumorigenesis and embryonic development. *Disease models & mechanisms*, *5*(6), 881-894.
nrg1	*Danio rerio*	Brown, D., Samsa, L. A., Ito, C., Ma, H., Batres, K., Arnaout, R., ... & Liu, J. (2018). Neuregulin-1 is essential for nerve plexus formation during cardiac maturation. *Journal of cellular and molecular medicine*, *22*(3), 2007-2017.
pax7	*Danio rerio*	Minchin, J. E., & Hughes, S. M. (2008). Sequential actions of Pax3 and Pax7 drive xanthophore development in zebrafish neural crest. *Developmental biology*, *317*(2), 508-522.

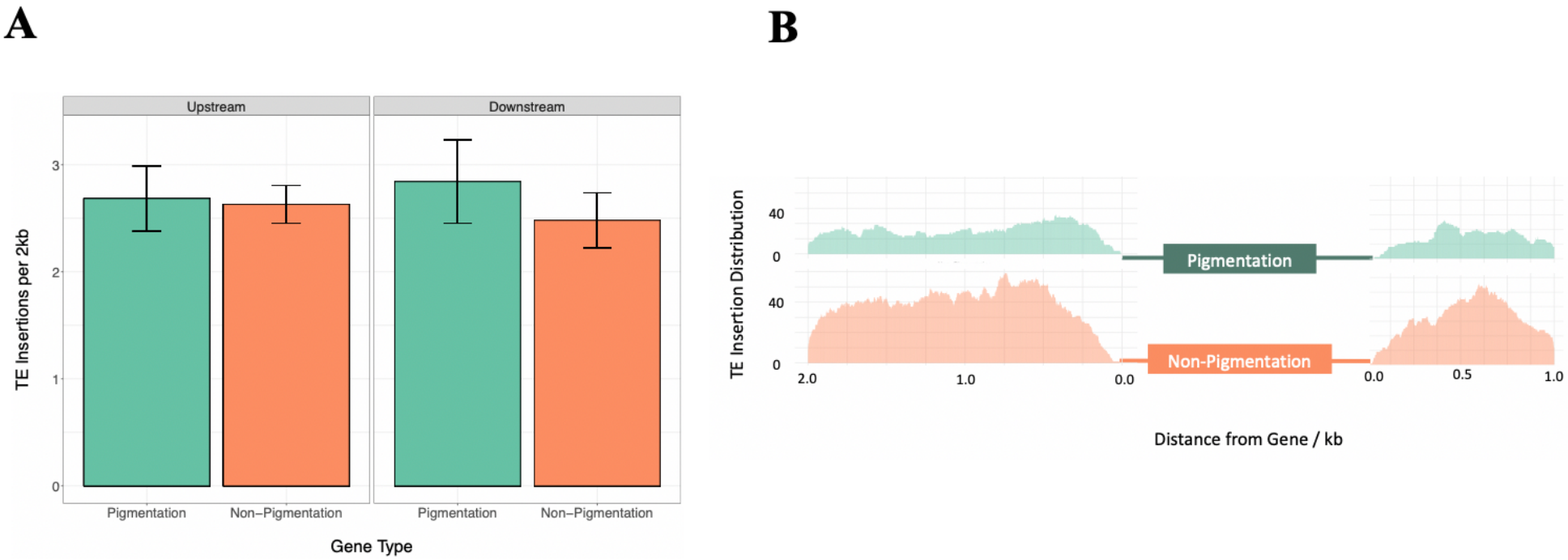
rab32	*Danio rerio*	Coppola, U., Annona, G., D'Aniello, S., & Ristoratore, F. (2016). Rab32 and Rab38 genes in chordate pigmentation: an evolutionary perspective. *BMC evolutionary biology*, *16*(1), 1-14.
rab38	*Danio rerio*	Coppola, U., Annona, G., D'Aniello, S., & Ristoratore, F. (2016). Rab32 and Rab38 genes in chordate pigmentation: an evolutionary perspective. *BMC evolutionary biology*, *16*(1), 1-14.
sox5	*Danio rerio & Oryzias latipes*	Nagao, Y., Takada, H., Miyadai, M., Adachi, T., Seki, R., Kamei, Y., ... & Hashimoto, H. (2018). Distinct interactions of Sox5 and Sox10 in fate specification of pigment cells in medaka and zebrafish. *PLoS Genetics*, *14*(4), e1007260.
sox9	*Danio rerio*	Rau, M. J., Fischer, S., & Neumann, C. J. (2006). Zebrafish Trap230/Med12 is required as a coactivator for Sox9-dependent neural crest, cartilage and ear development. *Developmental biology*, *296*(1), 83-93.
trpm1	*Danio rerio*	Kastenhuber, E., Gesemann, M., Mickoleit, M., & Neuhauss, S. C. (2013). Phylogenetic analysis and expression of zebrafish transient receptor potential melastatin family genes. *Developmental Dynamics*, *242*(11), 1236-1249.
trpm7	*Danio rerio*	McNeill, M. S., Paulsen, J., Bonde, G., Burnight, E., Hsu, M. Y., & Cornell, R. A. (2007). Cell death of melanophores in zebrafish trpm7 mutant embryos depends on melanin synthesis. *Journal of Investigative Dermatology*, *127*(8), 2020-2030.
tyrp1	*Oreochromis* spp.	Zhu, W., Wang, L., Dong, Z., Chen, X., Song, F., Liu, N., ... & Fu, J. (2016). Comparative transcriptome analysis identifies candidate genes related to skin color differentiation in red tilapia. *Scientific reports*, *6*(1), 1-12.
tyrp2	*Oncorhynchus mykiss*	Wu, S., Huang, J., Li, Y., Liu, Z., Zhang, Q., Pan, Y., & Wang, X. (2021). Cloning, sequence analysis, and expression of tyrp1a and tyrp2 genes related to body colour in different developmental stages and tissues of rainbow trout Oncorhynchus mykiss. *Aquaculture International*, *29*(3), 941-961.
vps11	*Oreochromis* spp.	Fang, W., Huang, J., Li, S., & Lu, J. (2022). Identification of pigment genes (melanin, carotenoid and pteridine) associated with skin color variant in red tilapia using transcriptome analysis. *Aquaculture*, *547*, 737429.
vps18	*Danio rerio*	Maldonado, E., Hernandez, F., Lozano, C., Castro, M. E., & Navarro, R. E. (2006). The zebrafish mutant vps18 as a model for vesicle-traffic related hypopigmentation diseases. *Pigment cell research*, *19*(4), 315-326.



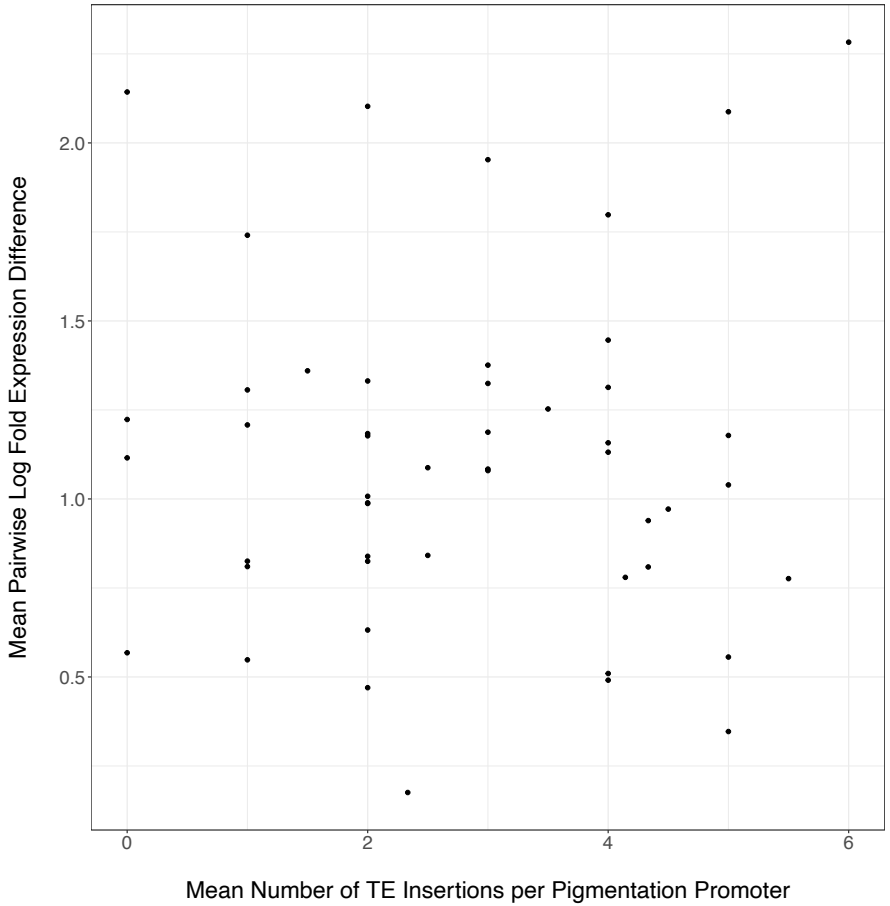
Supplementary Figure 6.1 (A) The mean Kn/Ks ratio of the TPGs ($\bar{x}$ = 0.14 ± 0.02) and non-pigmentation genes (0.16 ± 0.01) (Welch's t = 0.66, d.f. = 26.49, P = 0.51) **(B)** The paired Kn/Ks ratios of TPGs across the high ($\bar{x}$ = 0.14 ± 0.02) and low ($\bar{x}$ = 0.14 ± 0.02) rate of colour pattern evolution groups were not significantly different from one another (Paired T = 1.05, df = 13, P = 0.31). All gene alignments were conducted against orthologues found within the channel catfish *(Ictalurus punctatus).*

Supplementary Table 6.3 Negative binomial regression output investigating whether the number of TE insertions (offet per transcript length) within the reduced teleost pigmentation (TPG) subset differs between Corydoradine species with a low or high rate of colour pattern evolution

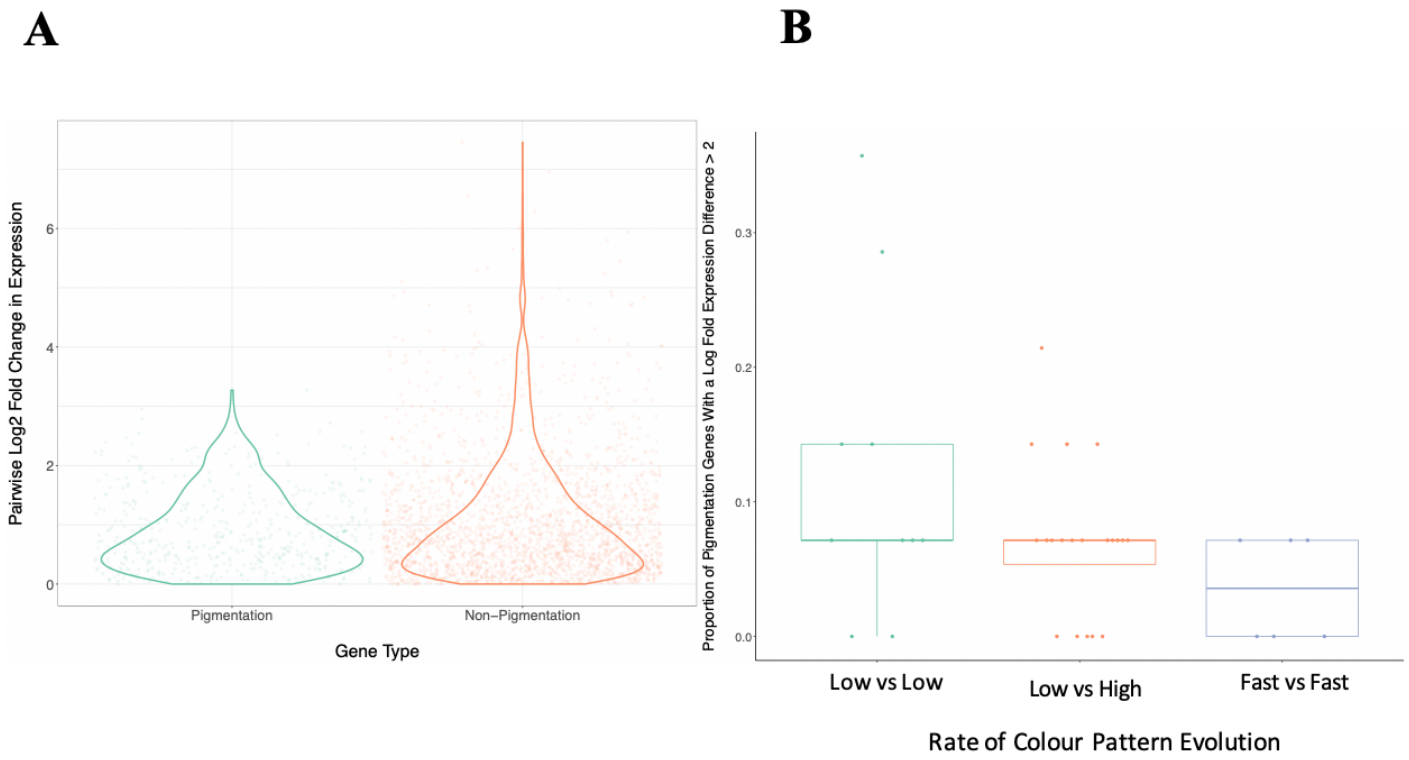
Model	**Offset** (*log(Pigment Transcript Length)*)			
		Estimates	**P**	**AIC**
Negative Binomial	Intercept	-1.18	<0.001	575.81
	Colour pattern - High	0.64	<0.01	



Supplementary Figure 6.2 A Within the Corydoradinae, the mean number of TE insertions (per 2k) within the upstream region of each teleost pigmentation gene (TPG) (2.68 ± 0.30) and non-pigmentation gene (2.63 ± 0.18) were not significantly different from each other (Mann-Whitney U test: U = 1877.5, n_1 = 38, n_2 = 100, P = 0.91). There was also no significant difference between the mean number of TEs within the downstream region of TPGs (2.84 ± 0.39) and non-pigmentation genes (2.48 ± 0.39) (Mann-Whitney U test: U = 1699.5, n_1 = 38, n_2 = 100, P = 0.32). **B.** The distance from the 5' start codon of TEs within the promoter regions of TPGs ($\bar{x}$ = 959 bp ± 57.1 bp) and non-pigmentation gene ($\bar{x}$ = 1,052 bp ± 31.9 bp) were not significantly different from one another (Mann-Whitney U test of TE midpoints: U = 14,876, n_1 = 102, n_2 = 263, P 0.11). Similarly, no significant difference in TE insertion distribution was found within the downstream (3'UTR) sequences of both gene types (Mann-Whitney U test of TE midpoints: U = 3,127.5, n_1 = 56, n_2 = 124, P = 0.29).



Supplementary Figure 6.3 There was no significant relationship between the number of TE insertions within the upstream promoter region of pigmentation genes within the genome of *C. fulleri* (Lineage 1) and their mean cross-species pairwise log fold expression differences (F = 0.01, df = 1, 48, P = 0.91). Pairwise comparisons were conducted using the RNA-seq data of a Lineage 1 species (*C. maculifer*) against a species representative from each other lineage representative



Supplementary Figure 6.4 A. There was no significant difference in the mean pairwise log2 fold difference of Corydoradinae TPG expression (0.89 ± 0.03) versus non-pigmentation gene expression (1.05 ± 0.02) (Mann-Whitney U-test: U=545,954, n_1 = 504, n_2 = 2,268, P = 0.10). **B.** Pairwise species comparisons indicated that there was no significant effect of rate of colour pattern differences/similarity on the proportion of differentially expressed TPGs (ANOVA: F = 2.61; d.f = 2, 33; P = 0.09).

7 Synthesis

For over 20 years, the fact that most genomes across the tree of life largely consist of non-coding DNA has become well established, leading to a shift away from a purely genic view of evolution (Castillo-Davis 2005; Ahnert *et al.* 2008; Perenthaler *et al.* 2019). A major component of non-coding DNA are transposable elements (TEs), short DNA sequences that possess the ability to replicate throughout a genome (Wicker *et al.* 2007). Consequently, understanding both the causes and impacts of TE proliferation has become a major focus within the field of evolutionary genomics, which is only recently uncovering the diverse consequences of TE activity, both at a genomic and a population level (Chénais *et al.* 2012; Cowley & Oakey 2013; Stapley *et al.* 2015; Choi JY & Lee YCG 2020). Furthermore, the increased availability of whole genome sequences (The Darwin Tree of Life Consortium, 2022) and a growing toolkit of bioinformatic pipelines which aid in the process of TE annotation (O'Neill *et al.* 2020) has meant that their study is more accessible than ever. To date, TE research has largely fallen within two approaches (i) experimental manipulation, usually within a single model species (Min *et al.* 2020; Sun *et al.* 2020) and (ii) comparative approaches, normally across multiple non-model species (e.g. Hawkins *et al.* 2006; Baril & Hayward 2022). This thesis has focussed on the latter, investigating both the causes and consequences of TE variability across numerous Corydoradinae catfish species.

7.1 Summary of key thesis findings.

Chapter 2 and 3 manipulated an existing model of TE dynamics to better understand the role of (i) beneficial insertion effects and (ii) whole genome duplication on the maintenance and proliferation of TEs within asexual populations. The ability of TEs to persist across many

generations was proportional to the percentage likelihood of a TE insertion causing a fitness increase within the host. Whilst this may not be particularly surprising, it was interesting to note that in the complete absence of beneficial insertion effects, the ability of TEs to persist across even a few was generations was limited. This suggests TE induced adaptations may be widespread, benefitting both the host and TE population. Whilst many incidences of TE activity inducing beneficial phenotypes exist (Santos *et al.* 2016; Van't Hof *et al.* 2016), one suspects that many more will be uncovered as the number of TE annotated genomes increases. Furthermore, this adds to the growing evidence that TEs may be important drivers of phenotypic adaption (González *et al.* 2008). Secondly, we found that whole genome duplication events, and the subsequent relaxation of purifying selection pressure, can spark TE proliferation, though this appeared to be limited to cases where TE populations were static within the pre-duplicated individuals. The fact that TE proliferation after a whole genome duplication was somewhat dependent on existing TE dynamics may explain why existing evidence regarding the interplay between TE number and polyploidy has been mixed (Chalopin & Volff 2017; Baduel *et al.* 2019). Like any model, these simulations do not perfectly represent the reality of nature. The model of TE dynamics used for these chapters could be improved in the future by incorporated the (i) role of sexual recombination, (ii) the role of TE silencers (e.g. siRNAs) and finally (iii) the ability of TEs of different lengths or transposition methods to accumulate at different rates.

Chapter four described the process of creating an initial 'de-novo' Corydoradinae-specific TE library. This was conducted to obtain more accurate TE based annotations during downstream analyses. Two recent bioinformatic packages (EDTA/DeepTE) were utilised to help us achieve these aims. The results from this chapter can be divided into two broad

categories, namely (i) the impact EDTA/Deep TE had on TE annotation and (ii) an initial genomic TE description within the Corydoradinae. We found that the use of a de-novo TE library increased estimated TE abundance, altered TE composition, and reduced estimated TE ages. Substantial changes to these TE metrics highlighted the importance of TE library choice during TE studies. Using the new Corydoradinae-specific TE library we were also able to obtain a description of the TE landscape within a Lineage 1 Corydoras (*C. fulleri*) for the first time. Calculated genomic TE content (~43%) lies within the higher estimates of TE abundance within teleosts, with the landscape dominated by type II DNA transposons (particularly Tc1-Mariner and hAT elements) (Shao *et al.* 2019). Finally, we observed important differences between sequence types, with TEs found in the transcriptome representing younger elements present in lower abundance. Though the focus of this chapter was to test the bioinformatic pipelines 'as is', it would have been interesting to spend some time manually curating this initial Corydoradinae-specific TE library, particularly considering the well-established issues surrounding de-novo creation (e.g. library fragmentation, multiple consensus sequences etc) and its potential impact on downstream TE analysis (Platt *et al.* 2016; Baril *et al.* 2022).

Chapter five aimed to utilise the previously described Corydoradinae-specific TE library to provide the first detailed description of the expressed TE landscape across multiple Corydoradinae species. This chapter presented multiple results, but the key findings were that (i) a phylogenetic increase in expressed TE abundance has occurred within the Corydoradinae, though it is somewhat dependent on the TE library used, (ii) polyploid Corydoradinae species have a greater number of TE insertions within gene coding regions and (iii) when accounting for phylogenetic signal, there is no significant positive relationship

between lineage genome size and expressed TE content. The confidence we have in all these findings would increase with a greater number of species representatives, particularly as only one was available per lineage. Furthermore, we found evidence that a Mariner DNA element with an unknown amphibian origin has inserted within the development gene *MMP13* of multiple Corydoradinae species via horizontal transfer. A greater number of genome assemblies of South American amphibian species would have allowed us to better ascertain the likely origin of this horizontally transferred TE. Furthermore, whilst MMP13s role in bone development means we can hypothesise that this insertion may have impacted Corydoradinae body shape evolution, future experimental analysis would be needed to confirm this.

Chapter six assessed the hypothesis that biased TE insertions may induce pigmentation change at a higher rate compared to other gene types. This hypothesis is based on the numerous studies which have identified a link between colour pattern change and TE insertions (e.g. Iida *et al.* 2004; Van't Hof *et al.* 2016; Kratochwil *et al.* 2022). Furthermore, a 'gene disruption' model of TE insertion patterns provides a theoretical framework as to why TE insertions may frequently occur within pigmentation genes (due to their reduced essentially compared to other gene types). However, contrary to our hypothesis, we found no evidence that TE insertions are inflated within Corydoradinae pigmentation genes (both in transcripts and in upstream promoter regions), with the likelihood of TE presence being greater within non-pigmentation transcripts. Furthermore, pigmentation genes were found to evolve under stricter purifying selection pressure than non-pigmentation genes, suggesting pigmentation gene sequences are highly conserved between Corydoradinae species. This leads to numerous outstanding questions, including (i) what has driven rapid

pigmentation change within the evolutionary history of the Corydoradinae? (ii) what are the phenotypic consequences of the TE insertions we do find within the promoter and transcript regions of Corydoradinae pigmentation genes and (iii) what role, if any, does Müllerian mimicry play in driving high purifying selection rates on pigmentation genes within the Corydoradinae?

7.2 Highlights

- Simulations of TE dynamics highlight the importance of beneficial insertion effects and whole genome duplication on TE proliferation events
- Use of a species-specific TE library may generate significance differences to estimates of TE traits in downstream analyses (e.g. abundance, composition and age)
- Late branching (and likely polyploid) Corydoradinae species have greater expressed TE abundance and increased TE accumulation within genic regions
- A horizontally transferred Mariner elements from an unknown amphibian origin has inserted within the bone development gene '*MMP13*' of multiple Corydoradinae species, with potential impacts regarding body shape evolution
- Within the Corydoradinae, TEs do not insert within pigmentation genes or their promoter regions at a higher rate than in non-pigmentation genes.

7.3 Future Directions

There are numerous future research projects that could be sparked from the work contained within this thesis. In this final section, three further research ideas are briefly highlighted, chosen because they would build upon this existing work and provide a significant advancement to the field of TE biology (Figure 7.1).

1) **Measuring changes in direct expression of TEs within polyploid and non-polyploid Corydoradinae species.** Determining expression level of TEs involves establishing the number of RNA-seq reads that map back to unique TE loci found within a genome

(Lanciano & Cristofari 2020). At the beginning of the research project neither a Corydoradinae genome nor TE library existed, which meant the initial focus was to investigate transcriptional TE content (i.e. the proportion of the exome that consists of TE sequences). Whilst this is a useful metric in assessing the TE content that has been established (e.g. through co-transcription) within the transcriptome of different Corydoradinae species, it likely reflects past TE activity. In comparison, estimating the expression level of each TE represents the degree of activity that is still occurring (Lanciano & Cristofari 2020). Within the *Spartina* group of grasses almost all TE families are more highly expressed in hexaploids than tetraploids (Giraud *et al.* 2021). It would be interesting to see if these results are consistent within the Corydoradinae, whereby polyploid species may have greater levels of TE expression than non-polyploid species, indicating TE accumulation is still occurring and at disproportionate rates across different Corydoradinae species (Figure 7.1a)

2) **Is there a significant relationship between the diversity of genomic TEs and their respective silencers (e.g siRNAs)?** Small interfering RNAs (siRNAs) play an important role in TE silencing via RNA-interference, contributing to both pre- and post-transcriptional silencing via the recruitment of methyltransferases or argonaute proteins respectively (Slotkin & Martienssen 2007). The research found within this thesis has largely focussed on the variation TE abundance between Corydoradinae species, which provides only one side of the coin in their dynamic arms race with respective silencers. It would be interesting to investigate whether genomic TE diversity correlates with siRNA diversity among different Corydoradinae species (Figure 7.1b). Similar studies have been conducted within wheat, where TEs which are more transcriptionally active are associated with greater siRNA abundance,

though this comparison was made between different tissue types rather than different species (Sun *et al.* 2013).

3) **Does hybridisation induce TE reactivation within the Corydoradinae, and provide a potential barrier to inter-specific breeding?** Hybridisation is a potentially important process in TE activation, whether that be via 'Genome Shocks', the mismatched inheritance of TEs and their associated silencers or a reduction in purifying selection pressure in autopolyploids (Parisod & Senerchia 2012; Crespi & Nosil 2013). Given the net impact of increased TE activity is likely to reduce host fitness, this process may provide a significant barrier to inter-specific breeding within sympatric species (Crespi & Nosil, 2013). TE induced hybrid incompatibilities may have played a disproportionately large role in the rapid speciation of the Corydoradinae, particularly considering TE content is highly variable between even closely related species. Corydoradinae hybridisation is purported to be relatively common within the aquatic pet trade, meaning hybrid individuals may be produced relatively easily within an aquarium setting (van der Walt *et al.* 2017). RNA-seq or PCR quantification could test whether Corydoradinae hybrids actually have increased TE activity/copy number compared to parent species (Figure 7.1c). This would contribute to existing reports of teleost hybrids with increased TE activity. For example, hybrid individuals of the teleost genus *Cottus* have higher TE copy numbers than parent species (Dennenmoser *et al.* 2017), whilst TE expression is upregulated within hybrid individuals of Lake Whitefish (Dion-Côté *et al.* 2014).

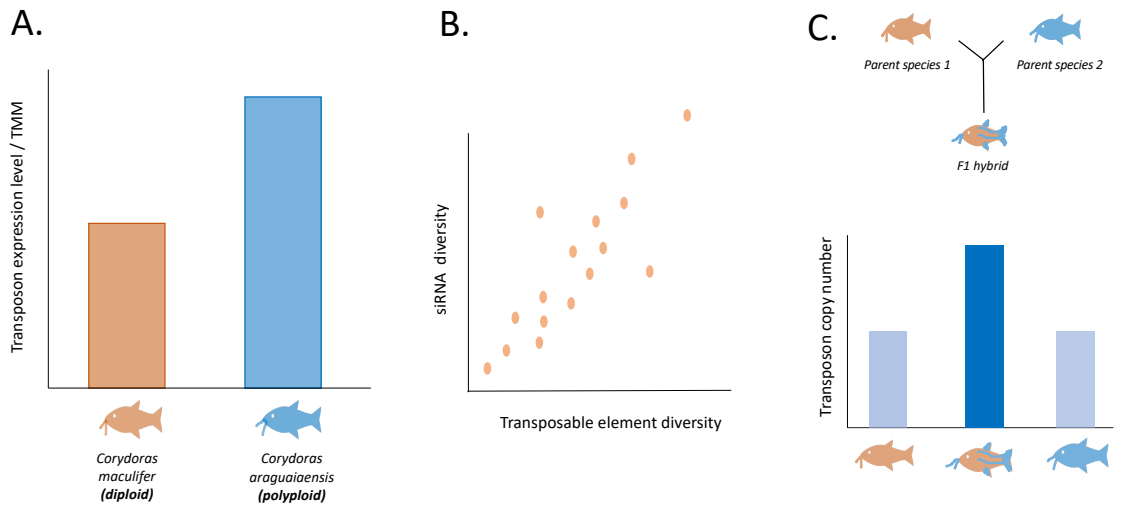


Figure 7.1 Three future research projects regarding TE activity within the Corydoradinae. **A.** Is TE expression greater within polyploid versus diploid species? **B.** Is there a relationship between the diversity of genomic TE and associated silencers (e.g. siRNA)? **C.** Does hybridisation provide a barrier to inter-specific breeding by increasing TE activity levels?

7.4 Personal Reflections

The following section outlines the personal reflections I have on the completion of a thesis. The purpose of this activity is to highlight to any prospective students what key takeaways I have learnt about the nature of scientific research. Naturally, this content may appear slightly grandiose and self-indulgent, but most of the learning I have had during this time has been on the nature of conducting research, rather than the research content itself, and it therefore seems a shame not to share. Here is a summary of such reflections.

1) **Visualisation is an important intermediate step**. I was often guilty of thinking the purpose of visualisation is the production of shiny, exciting figures, with the end goal of their inclusion in scientific publication. However, visualisations can be a fantastic intermediate step to reassure oneself that the coding and analysis being conducted is

correct. This is particularly important when dealing with large data sets where checking that the outcomes of several lines of code makes biological sense is extremely tricky.

2) **Bottom-up science is not as straightforward as one may think.** In many ways I was extremely fortunate to embark on this PhD where the transcriptomic data was already available. However, unlike top-down science, where one begins with the hypothesis and then goes out and collects data to best answer it, bottom-up science has numerous pitfalls I was in danger of falling into. Firstly, one can spend a long time in a rabbit hole exploring data and seeing what stories are 'spat out'. How easy is it to ignore all the data analysis that doesn't support the story you want to tell when you are exploring data without a guiding hypothesis? I wish I had spent more time formulating testable hypotheses early on in the PhD process. I think this is crucial to ensure you are conducting scientific research in an ethical manner. Secondly, I wish I had spent more time reflecting about what scientific questions could (and crucially *could not!*) be answered by understanding the potential limitation the data I had in front of me.

3) **Automate your code!** By the end of my PhD I had probably reran some analysis upwards of 50 times. Automation made this entire process much less painful and should be prioritised above everything else (yes even the presentation of a cool result in a PI meeting!!). Automation should be encouraged as much as possible, ranging from data input, file path structure to the saving of figure outputs. I wish I had learnt about the power of SnakeMake early on, which can bring together code from multiple sources, e.g. Bash, R and Python. This is particularly useful in cases where you find you are jumping between multiple coding languages to create bioinformatic pipelines.

4) **Listen to your mind.** After a difficult or stressful day, it is a good idea to go home, relax and do something different the next workday. Work on a different project until the initial emotional reaction has disappeared, and you can then conduct the work in an objective rather than subjective frame of mind. This may be particularly applicable in times of paper rejections and unearthing coding errors for example.

Furthermore, I wish to spend a couple of sentences outlying my own personal reflections on where I feel the field of transposable elements is currently placed. As the number of assembled genomes increases, so does the number of (i) descriptive papers highlighting TE landscapes and (ii) bioinformatic tools available which offer (often marginal) improvements over existing ones. Whilst there is undeniable worth in these works (indeed much of my thesis falls under that category), I cannot help but feel the field is becoming bloated with TE descriptions and an ever-increasing pool of bioinformatic pipelines - which arguably makes the field even more daunting for new researchers! Instead, I believe the field needs to start answering *hypothesis driven* questions surrounding TEs role in evolution, where experimental evidence can help prove causation, not just correlation. The Mullerian peppered moth is a beautiful example of this but is sadly all too rare. Until that changes, I can't help but feel the outdated view that TEs are simply regions of 'junk DNA' will continue to persist.

7.5 References

Ahnert, S.E., Fink, T.M.A. & Zinovyev, A. (2008). How much non-coding DNA do eukaryotes require? *J. Theor. Biol.*, 252, 587–592.

Baduel, P., Quadrana, L., Hunter, B., Bomblies, K. & Colot, V. (2019). Relaxed purifying selection in autopolyploids drives transposable element over-accumulation which provides variants for local adaptation. *Nat. Commun.*, 10, 1–10.

Baril, T. & Hayward, A. (2022). Migrators within migrators: exploring transposable element

dynamics in the monarch butterfly, Danaus plexippus. *Mob. DNA 2022 131*, 13, 1–19.
Baril, T., Imrie, R.M. & Hayward, A. (2022). Earl Grey: a fully automated user-friendly transposable element annotation and analysis pipeline. *bioRxiv*, 2022.06.30.498289.
Castillo-Davis, C.I. (2005). The evolution of noncoding DNA: how much junk, how much func? *Trends Genet.*, 21, 533–536.
Chalopin, D. & Volff, J.N. (2017). Analysis of the spotted gar genome suggests absence of causative link between ancestral genome duplication and transposable element diversification in teleost fish. *J. Exp. Zool. Part B Mol. Dev. Evol.*, 328, 629–637.
Chénais, B., Caruso, A., Hiard, S. & Casse, N. (2012). The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. *Gene*, 509, 7–15.
Choi JY & Lee YCG. (2020). Double-edged sword: The evolutionary consequences of the epigenetic silencing of transposable elements. *PLOS Genet.*, 16, e1008872.
Cowley, M. & Oakey, R.J. (2013). Transposable Elements Re-Wire and Fine-Tune the Transcriptome. *PLoS Genet.*, 9, e1003234.
Crespi, B. & Nosil, P. (2013). Conflictual speciation: species formation via genomic conflict. *Trends Ecol. Evol.*, 28, 48–57.
Dennenmoser, S., Sedlazeck, F.J., Iwaszkiewicz, E., Xiang, |, Li, Y., Altm€ Uller, J., *et al.* (2017). Copy number increases of transposable elements and protein-coding genes in an invasive fish of hybrid origin. *Mol. Ecol.*, 26, 4712–4724.
Dion-Côté, A.-M., Renaut, S., Normandeau, E. & Bernatchez, L. (2014). RNA-seq Reveals Transcriptomic Shock Involving Transposable Elements Reactivation in Hybrids of Young Lake Whitefish Species. *Mol. Biol. Evol.*, 31, 1188–1199.
Giraud, D., Lima, O., Rousseau-Gueutin, M., Salmon, A. & Aïnouche, M. (2021). Gene and Transposable Element Expression Evolution Following Recent and Past Polyploidy Events in Spartina (Poaceae). *Front. Genet.*, 12, 352.
González, J., Lenkov, K., Lipatov, M., Macpherson, J.M. & Petrov, D.A. (2008). High Rate of Recent Transposable Element–Induced Adaptation in Drosophila melanogaster. *PLoS Biol.*, 6, e251.
Hawkins, J.S., Kim, H.R., Nason, J.D., Wing, R.A. & Wendel, J.F. (2006). Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. *Genome Res.*, 16, 1252–1261.
Iida, S., Morita, Y., Choi, J.D., Park, K. Il & Hoshino, A. (2004). Genetics and epigenetics in flower pigmentation associated with transposable elements in morning glories. *Adv. Biophys.*, 38, 141–159.
Kratochwil, C.F., Kautt, A.F., Nater, A., Härer, A., Liang, Y., Henning, F., *et al.* (2022). An intronic transposon insertion associates with a trans-species color polymorphism in Midas cichlid fishes. *Nat. Commun.* , 13, 1–8.
Lanciano, S. & Cristofari, G. (2020). Measuring and interpreting transposable element expression. *Nat. Rev. Genet.*, 21, 721–736.
Min, B., Park, J.S., Jeong, Y.S., Jeon, K. & Kang, Y.K. (2020). Dnmt1 binds and represses genomic retroelements via DNA methylation in mouse early embryos. *Nucleic Acids Res.*, 48, 8431–8444.
O'Neill, K., Brocks, D. & Hammell, M.G. (2020). Mobile genomics: tools and techniques for tackling transposons. *Philos. Trans. R. Soc. B*, 375.
Parisod, C. & Senerchia, N. (2012). Responses of Transposable Elements to Polyploidy. In: *Topics in Current Genetics 24*. pp. 147–168.

Perenthaler, E., Yousefi, S., Niggl, E. & Barakat, T.S. (2019). Beyond the Exome: The Non-coding Genome and Enhancers in Neurodevelopmental Disorders and Malformations of Cortical Development. *Front. Cell. Neurosci.*, 13, 352.

Platt, R.N., Blanco-Berdugo, L., Ray, D.A. & Ray, D.A. (2016). Accurate Transposable Element Annotation Is Vital When Analyzing New Genome Assemblies. *Genome Biol. Evol.*, 8, 403–10.

Santos, M.E., Baldo, L., Gu, L., Boileau, N., Musilova, Z. & Salzburger, W. (2016). Comparative transcriptomics of anal fin pigmentation patterns in cichlid fishes. *BMC Genomics*, 17, 1–16.

Shao, F., Han, M. & Peng, Z. (2019). Evolution and diversity of transposable elements in fish genomes. *Sci. Reports* , 9, 1–8.

Slotkin, R.K. & Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.*, 8, 272–285.

Stapley, J., Santure, A.W. & Dennis, S.R. (2015). Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species. *Mol. Ecol.*, 24, 2241–2252.

Sun, F., Guo, W., Du, J., Ni, Z., Sun, Q. & Yao, Y. (2013). Widespread, abundant, and diverse TE-associated siRNAs in developing wheat grain. *Gene*, 522, 1–7.

Sun, L., Jing, Y., Liu, X., Li, Q., Xue, Z., Cheng, Z., *et al.* (2020). Heat stress-induced transposon activation correlates with 3D chromatin organization rearrangement in Arabidopsis. *Nat. Commun.* , 11, 1–13.

The Darwin Tree of Life Consortium. (2022). Sequence locally, think globally: The Darwin Tree of Life Project. *PNAS*, 119.

Van't Hof, A.E., Campagne, P., Rigden, D.J., Yung, C.J., Lingley, J., Quail, M.A., *et al.* (2016). The industrial melanism mutation in British peppered moths is a transposable element. *Nature*, 534, 102–105.

van der Walt, K.A., Mäkinen, T., Swartz, E.R. & Weyl, O.L.F. (2017). DNA barcoding of South Africa's ornamental freshwater fish–are the names reliable? *African J. Aquat. Sci.*, 42, 155–160.

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