The Evolution and Adaptive Effects of Transposable Elements in Birds and Elapids

Mr. James Douglas Galbraith

This thesis is presented for the degree of

Doctorate of Philosophy

Supervisors:
Prof. David L. Adelson, Univerity of Adelaide
Asst. Prof Alexander Suh, University of East Anglia

The University of Adelaide
School of Biological Sciences

25th of June 2021

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Mr. James Douglas Galbraith

July 19, 2021

Abstract

Transposable elements (TEs) are genetic sequences able to copy or move themselves across their host genome. As TEs move within their host they can act as a source of genetic novelty, and hence are often described as "drivers of evolution". This novelty includes contributing or altering regulatory and coding regions, and promoting non-allelic homologous recombination and, in turn, major structural rearrangements. In some cases, TEs can further contribute to genomic change by jumping between organisms in a process known as horizontal transposon transfer (HTT). HTT is the passing of TEs between organisms by means other than parent to offspring, and has been well described across vertebrates, with multiple events noted in both birds and squamates.

Birds are the most diverse class of reptiles, encompassing over 10,000 species, however studies in TE evolution in birds have focused on single lineages. Early findings from the chicken genome led to the assumption that avian TEs are largely stable and inactive. More recent studies have similarly focused on single lineages of birds, revealing some variation in TE activity across birds. In contrast to birds, few studies have explored the evolution of TEs in squamates (lizards and snakes) at a class or family level, instead examining their evolution either across the order or comparing two long diverged species. As such, it is unknown whether patterns seen across all squamates occur at shorter time scales. At lower levels many squamate families are highly diverse, rapidly adapting to new environments and ecological niches. One such family is Hydrophiinae, a family of elapid snakes containing ∼100 terrestrial snakes, ∼60 marine sea snakes and 6 amphibious sea kraits.

In this thesis I investigate the evolution of TEs in two diverse groups of reptiles: birds and Australo-Melanesian elapid snakes (Hydrophiinae). I provide the first comprehensive study of TE activity across all orders of birds, focusing on the dominant superfamily, Chicken Repeat 1 (CR1) retrotransposons. By performing comparative genomic analyses I have identified significant variation in the rate of TE expansion both between and within avian orders. Clades including parrots, kiwis and waterfowl show high diversity and large, recent expansions of CR1 retrotransposons, while in various ratites and songbirds CR1s have been near inactive for tens of millions of years.

The rest of the chapters focus on the evolution of TEs in hydrophiines, finding repeated HTT events into marine hydrophiines from other marine organisms. TEs in hydrophiines that were acquired via HTT appear to have played a role in their

adaptation to the marine environment, with insertions found throughout regulatory regions. In the sea kraits, one horizontally transferred TE has rapidly expanded to make up 8-12% of the sea krait genome in a timespan of just 15-25 million years, the fastest known expansion of TEs in amniotes following a HTT event.

Together this thesis presents bioinformatic analyses of two diverse clades of reptiles, Aves and Hydrophiine, finding that to truly understand TEs, their evolution and the potential adaptive effects they can cause, we must examine life on both a broad and fine scale.

Acknowledgements

This work would not have been possible without a very long list of people, far too long to list here.

Firstly, Dave and Alex, you have been the most wonderful and supportive supervisors. The time, effort and concern you've had for my work and wellbeing has been amazing. Dave, thank you for your continuous optimism and encouragement to investigate what others would have brushed off as mere coincidence. Thank you Alex for providing me with confidence, constructive criticism and literally saving my life by telling me to stop being stoic and go to a doctor. Together you've helped me transform my curiosity into proper scientific inquiry.

To the many members of the Adelson lab over the past five years, thank you. Dan, thank you for insightful discussions of science and philosophy, and very valid criticisms when I needed it. Reuben, Atma, Lu and Brittany, you may not have been here in person for a quite while but you're like family to me. Reuben, you showed me that to be a scientist wasn't to be a mad workaholic, and that late starts and late nights are actually okay. Atma and Lu, your work laid the bedrock for everything I've worked on, and have provided encouragement when I thought my findings were meaningless. Last, but not least, Brittany. Thank you for being a wonderful best friend and emotional support beam, even from the other side of the planet.

To the postgrads and postdocs at EBC, without you I wouldn't have survived my time in Sweden. Ghazal, Aaron and Mercè, from the day I stepped in the door I felt included thanks to you. Vale, you've been like a PhD sister to me, going through this rollercoaster together. You and Diem sneaking into the ICU and then visiting basically everyday made me realise I already had family in Uppsala.

To the Sanders lab, thank you for opening my eyes to the fascinating world of reptiles. Thank you Kate for being like an unofficial supervisor. You filled me with optimism and intrigue for all things reptilian. Al, I'm incredibly grateful for the bioinformatic help and the laughs along the way.

Thank you also to my assessors Professor Huaijun Zhou and Dr Clément Gilbert. Your comments have provided me with encouragement and confidence going forward into a career in research.

And finally to my family, you've dropped everything multiple times to help me cope, whether I was here in Adelaide going through a rough spot mentally or over in Sweden recovering from surgery. Mum and Dad, you might not understand all the ins and outs of my research, but you've nurtured my curiosity and "but why" attitude from the very beginning. Kate, you're the perfect sister. If I needed someone to talk to you were there, and understood both the science and the emotions.

Contents

Introduction

Transposable elements (TEs), sometimes referred to as "jumping genes", are genetic elements able to move and copy themselves throughout genomes. TEs were first observed and described by Barbara McClintock in maize (1950). In analysing chromosome 9 she observed "mutable loci", which caused changes in phenotype and enabled major structural rearrangements. McClintock predicted that the "behavior of these new mutable loci in maize cannot be considered peculiar to this organism", and in the following 70 years TEs have been found in almost all eukaryotes (Almojil et al. 2021).

TEs largely fall into two classes, DNA transposons and retrotransposons (Wicker et al. 2007). DNA transposons cut and paste using a transposase protein, while retrotransposons copy and paste using a reverse transcriptase (Feschotte and Pritham 2007; Eickbush and Malik 2002). Retrotransposons are further divided into LTR elements, which are reverse transcribed outside of the nucleus before insertion, and non-LTR retrotransposons, which are reverse transcribed at the site of insertion. In addition to autonomous elements which move or copy themselves, each class contains non-autonomous elements which rely on the machinery of autonomous elements to be moved across the genome.

As TEs move throughout the genome, their insertion into functional elements can lead to frameshifts, missense and nonsense mutations, and their retrotransposition can promote non-allelic homologous recombination, leading to exon shuffling and sequence duplication and deletion (Underwood and Choi 2019; Krasileva 2019). In the lifetime of an organism this can be harmful, with TE caused modifications being linked to altered gene expression and diseases including numerous cancers, haemophilia and cystic fibrosis (Teugels et al. 2005; Jiang and Upton 2019; Ostertag and Kazazian 2001). While likely to be nearly neutral or negative to an individual, changes caused by TEs can prove beneficial to the species at macroevolutionary scales.

TEs can be exapted as coding or regulatory elements or enable duplication of coding sequences. The impact of such changes range from the insertion of a TE fragment beneficially altering the expression of a gene through to increased numbers of TEs causing ectopic recombination and, in turn, reproductive isolation (Belyayev, 2014). For example, TEs played a key role in the evolution of pregnancy in both mammals and viviparous skinks (Lynch et al. 2011; Cornelis et al. 2017). Like other genetic sequences, TEs are normally vertically inherited, however many instances of TEs being horizontally transferred between species have been reported (Ivancevic et al. 2018; Gilbert and Feschotte 2018; Peccoud et al. 2017).

Horizontal transfer is the transfer of genetic material between species or populations by methods other than vertical inheritance (parental inheritance). Horizontal gene transfer (HGT) is well characterised in bacteria through plasmids, and has major implications in

antibiotic resistance (see reviews (Sørensen et al. 2005; Koonin, Makarova, and Aravind 2001)). In contrast, few instances of HGT have been identified in eukaryotes, with the majority noted in plants and fungi (Graham and Davies 2021; Wickell and Li 2020; Bredeweg and Baker 2020). While HGT is rare in animals, many instances of horizontal transposon transfer (HTT) between distant lineages have been reported. The AviRTE non-LTR retrotransposon was transferred between long-diverged birds and nematodes $\sim$ 50 Mya, and the BovB non-LTR retrotransposon transferred between numerous lineages of mammals and squamates, likely by a parasitic tick (Ivancevic et al. 2018; Walsh et al. 2013; Peccoud et al. 2017). The hAT superfamily of DNA transposons has similarly been found to have been horizontally transferred between mammals, amphibians and squamates (Pace et al. 2008).

While the evolution of TEs has been well described in mammals, and to a lesser extent in birds, very few comparative studies have compared their evolution in non-avian reptiles. The pattern of TE evolution seen in birds and mammals is highly similar. The TE landscape of both mammals and birds is dominated by a single lineage of non-LTR retrotransposons, L1s in therian mammals and CR1s in birds, with some lineages containing recent expansions of endogenous retroviruses (Ivancevic et al. 2016; Zhang et al. 2014; Mager and Stoye 2015). Additionally, in both orders there is little interspecific variation in genome size and total repeat content, with most avian genomes being 7-10% TEs and mammalian genomes 30-50%. This relatively constant TE content appears to be the result of an "accordion" model, with expansions in genome size due to TE activity being counterbalanced by extensive DNA loss (Kapusta, Suh, and Feschotte 2017). In contrast, the few studies which have investigated TE evolution in squamates have found significant variation in both TE diversity and total TE content (Castoe et al. 2011; Pasquesi et al. 2018).

Little research has yet investigated TE evolution within families of squamates, instead comparing long diverged species. As such it is unclear if the pattern of variation at a higher level holds true within families. Conversely, past research into TEs in birds has focused on a single order, resulting in the overall pattern of avian TE evolution being unclear. Here I aim to address these discrepancies, first by examining the evolution of the dominant avian TE family, CR1 retrotransposons, across all birds, and second by closely focusing on a diverse family of elapid snakes, Hydrophiinae.

References

Almojil, Dareen, Yann Bourgeois, Marcin Falis, Imtiyaz Hariyani, Justin Wilcox, and Stéphane Boissinot. 2021. "The Structural, Functional and Evolutionary Impact of Transposable Elements in Eukaryotes." *Genes* 12 (6): 918.

Bredeweg, Erin L., and Scott E. Baker. 2020. "Horizontal Gene Transfer in Fungi." In *Grand Challenges in Fungal Biotechnology*, edited by Helena Nevalainen, 317–32. Cham: Springer International Publishing.

Castoe, Todd A., Kathryn T. Hall, Marcel L. Guibotsy Mboulas, Wanjun Gu, A. P. Jason de Koning, Samuel E. Fox, Alexander W. Poole, et al. 2011. "Discovery of Highly Divergent Repeat Landscapes in Snake Genomes Using High-Throughput Sequencing." *Genome Biology and Evolution* 3 (May): 641–53.

Cornelis, Guillaume, Mathis Funk, Cécile Vernochet, Francisca Leal, Oscar Alejandro Tarazona, Guillaume Meurice, Odile Heidmann, et al. 2017. "An Endogenous Retroviral Envelope Syncytin and Its Cognate Receptor Identified in the Viviparous Placental Mabuya Lizard." *Proceedings of the National Academy of Sciences of the United States of America* 114 (51): E10991–0.

Eickbush, Thomas H., and Harmit S. Malik. 2002. "Origins and Evolution of Retrotransposons." In *Mobile DNA II*, 1111–44. American Society of Microbiology.

Feschotte, Cédric, and Ellen J. Pritham. 2007. "DNA Transposons and the Evolution of Eukaryotic Genomes." *Annual Review of Genetics* 41: 331–68.

Gilbert, Clément, and Cédric Feschotte. 2018. "Horizontal Acquisition of Transposable Elements and Viral Sequences: Patterns and Consequences." *Current Opinion in Genetics & Development* 49 (April): 15–24.

Graham, Laurie A., and Peter L. Davies. 2021. "Horizontal Gene Transfer in Vertebrates: A Fishy Tale." *Trends in Genetics: TIG* 37 (6): 501–3.

Ivancevic, Atma M., R. Daniel Kortschak, Terry Bertozzi, and David L. Adelson. 2016. "LINEs between Species: Evolutionary Dynamics of LINE-1 Retrotransposons across the Eukaryotic Tree of Life." Genome Biology and Evolution 8 (11): 3301–22.

Ivancevic, Atma M., R. Daniel Kortschak, Terry Bertozzi, and David L. Adelson. 2018. "Horizontal Transfer of BovB and L1 Retrotransposons in Eukaryotes." *Genome Biology* 19 (1): 85.

Jiang, Jiayue-Clara, and Kyle R. Upton. 2019. "Human Transposons Are an Abundant Supply of Transcription Factor Binding Sites and Promoter Activities in Breast Cancer Cell Lines." *Mobile DNA* 10 (April): 16.

Kapusta, Aurélie, Alexander Suh, and Cédric Feschotte. 2017. "Dynamics of Genome Size Evolution in Birds and Mammals." *Proceedings of the National Academy of Sciences of the United States of America* 114 (8): E1460–69.

Koonin, E. V., K. S. Makarova, and L. Aravind. 2001. "Horizontal Gene Transfer in Prokaryotes: Quantification and Classification." *Annual Review of Microbiology* 55: 709–42.

Krasileva, Ksenia V. 2019. "The Role of Transposable Elements and DNA Damage Repair Mechanisms in Gene Duplications and Gene Fusions in Plant Genomes." *Current Opinion in Plant Biology* 48 (April): 18–25.

Lynch, Vincent J., Robert D. Leclerc, Gemma May, and Günter P. Wagner. 2011. "Transposon-Mediated Rewiring of Gene Regulatory Networks Contributed to the Evolution of Pregnancy in Mammals." *Nature Genetics* 43 (11): 1154–59.

Mager, Dixie L., and Jonathan P. Stoye. 2015. "Mammalian Endogenous Retroviruses." *Microbiology Spectrum 3 (1):* MDNA3–0009 – 2014.

McClintock, B. 1950. "The Origin and Behavior of Mutable Loci in Maize." *Proceedings of the National Academy of Sciences of the United States of America* 36 (6): 344–55.

Ostertag, E. M., and H. H. Kazazian Jr. 2001. "Biology of Mammalian L1 Retrotransposons." Annual Review of Genetics 35: 501–38. Pace, John K., 2nd, Clément Gilbert, Marlena S. Clark, and Cédric Feschotte. 2008. "Repeated Horizontal Transfer of a DNA Transposon in Mammals and Other Tetrapods." *Proceedings of the National Academy of Sciences of the United States of America* 105 (44): 17023–28.

Pasquesi, Giulia I. M., Richard H. Adams, Daren C. Card, Drew R. Schield, Andrew B. Corbin, Blair W. Perry, Jacobo Reyes-Velasco, et al. 2018. "Squamate Reptiles Challenge Paradigms of Genomic Repeat Element Evolution Set by Birds and Mammals." *Nature Communications* 9 (1): 2774.

Peccoud, Jean, Vincent Loiseau, Richard Cordaux, and Clément Gilbert. 2017. "Massive Horizontal Transfer of Transposable Elements in Insects." *Proceedings of the National Academy of Sciences of the United States of America* 114 (18): 4721–26.

Sørensen, Søren J., Mark Bailey, Lars H. Hansen, Niels Kroer, and Stefan Wuertz. 2005. "Studying Plasmid Horizontal Transfer in Situ: A Critical Review." *Nature Reviews. Microbiology* 3 (9): 700–710.

Teugels, Erik, Sylvia De Brakeleer, Guido Goelen, Willy Lissens, Erica Sermijn, and Jacques De Grève. 2005. "De Novo Alu Element Insertions Targeted to a Sequence Common to the BRCA1 and BRCA2 Genes." *Human Mutation* 26 (3): 284.

Underwood, Charles J., and Kyuha Choi. 2019. "Heterogeneous Transposable Elements as Silencers, Enhancers and Targets of Meiotic Recombination." *Chromosoma* 128 (3): 279–96.

Walsh, Ali Morton, R. Daniel Kortschak, Michael G. Gardner, Terry Bertozzi, and David L. Adelson. 2013. "Widespread Horizontal Transfer of Retrotransposons." *Proceedings of the National Academy of Sciences of the United States of America* 110 (3): 1012–16.

Wickell, David A., and Fay-Wei Li. 2020. "On the Evolutionary Significance of Horizontal Gene Transfers in Plants." *The New Phytologist* 225 (1): 113–17.

Wicker, Thomas, François Sabot, Aurélie Hua-Van, Jeffrey L. Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, et al. 2007. "A Unified Classification System for Eukaryotic Transposable Elements." *Nature Reviews. Genetics* 8 (12): 973–82.

Zhang, Guojie, Cai Li, Qiye Li, Bo Li, Denis M. Larkin, Chul Lee, Jay F. Storz, et al. 2014. "Comparative Genomics Reveals Insights into Avian Genome Evolution and Adaptation." *Science* 346 (6215): 1311–20.

Genome stability is in the eye of the beholder: retrotransposon activity varies significantly across avian diversity

"Birds were flying from continent to continent long before we were. They reached the coldest place on Earth, Antarctica, long before we did. They can survive in the hottest of deserts. Some can remain on the wing for years at a time. They can girdle the globe." - David Attenborough, *The Life of Birds*

Birds the only dinosaur lineage to survive the K-T mass extinction, and have since evolved into the most diverse lineage of reptiles. Despite their high diversity, many conclusions about avian evolution have been based on findings in one species, the chicken. For example, until the sequencing of the zebra finch genome, the repetitive portion of all bird genomes was assumed to be roughly the same and change little. This was also due to karyotyping indicating the overall structure of bird genomes being highly conserved. As a greater diversity of avian genomes have been sequenced it has become clear that the repetitive content of the avian genome is likely not stable, with large expansions of CR1 retrotransposons noted in woodpeckers and their allies, and expansions of endogenous retroviruses seen in songbirds. While these findings clearly suggest the repetitive portion of the avian genome is not stable, past comparative studies were limited to single families of birds. Since the sequencing of the chicken genome in 2007 over 500 additional species of bird have been sequenced and made publicly available, allowing for class wide comparative genomics. To gain a greater understanding of how TEs have evolved in birds I set out to characterise how CR1s, the dominant family of TEs in birds, has evolved since extant birds' divergence approximately 100 Mya.

All supplementary data for this chapter can be found at
github.com/jamesdgalbraith/thesis_supplementary_material/tree/main/Chapter_1

Statement of Authorship

Title of Paper	**Genome stability is in the eye of the beholder: recent retrotransposon activity varies significantly across avian diversity**
Publication Status	☐ Published ☐Accepted for Publication ☒ Unpublished and Unsubmitted work written in manuscript style ☐ Submitted for Publication
Publication Details	James D. Galbraith, R. Daniel Kortschak, Alexander Suh, David L. Adelson (2021) Genome stability is in the eye of the beholder: recent retrotransposon activity varies significantly across avian diversity. Submitted as a research article to Genome Biology and Evolution

Principal Author

Name of Principal Author (Candidate)	James D. Galbraith		
Contribution to the Paper	Designed and performed analysis, interpreted results and wrote manuscript		
Overall percentage (%)	75%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	24/06/2021

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	R. Daniel Kortschak		
Contribution to the Paper	Supervised the development of the work and assisted in analysing the results and writing the manuscript		
Signature		Date	2021-06-23

Name of Co-Author	Alexander Suh		
Contribution to the Paper	Supervised the development of the work and assisted in analysing the results and writing the manuscript		
Signature		Date	2021-06-21

Please cut and paste additional co-author panels here as required.

Statement of Authorship

Name of Co-Author	David L. Adelson		
Contribution to the Paper	Supervised the development of the work and assisted in analysing the results and writing the manuscript		
Signature		Date	2021-06-22

Genome stability is in the eye of the beholder: recent retrotransposon activity varies significantly across avian diversity

James D. Galbraith[1], R. Daniel Kortschak[1], Alexander Suh[2,3,*], David L. Adelson[1,*].

1)School of Biological Sciences, The University of Adelaide, Adelaide, South Australia, Australia

2)School of Biological Sciences, University of East Anglia, Norwich, UK

3)Department of Organismal Biology, Evolutionary Biology Centre (EBC), Science for Life Laboratory, Uppsala University, Uppsala, Sweden

*)Corresponding author

Short title: Variable retrotransposon activity in birds

Abstract:

Since the sequencing of the zebra finch genome it has become clear the avian genome, while largely stable in terms of chromosome number and gene synteny, is more dynamic at an intrachromosomal level. A multitude of intrachromosomal rearrangements and significant variation in transposable element content have been noted across the avian tree. Transposable elements (TEs) are a source of genome plasticity, because their high similarity enables chromosomal rearrangements through non-allelic homologous recombination, and they have potential for exaptation as regulatory and coding sequences. Previous studies have investigated the activity of the dominant TE in birds, CR1 retrotransposons, either focusing on their expansion within single orders, or comparing passerines to non-passerines. Here we comprehensively investigate and compare the activity of CR1 expansion across orders of birds, finding levels of CR1 activity vary significantly both between and with orders. We describe high levels of TE expansion in genera which have speciated in the last 10 million years including kiwis, geese and Amazon parrots; low levels of TE expansion in songbirds across their diversification, and near inactivity of TEs in the cassowary and emu for millions of years. CR1s have remained active over long periods of time across most orders of neognaths, with activity at any one time dominated by one or two families of CR1s. Our findings of higher TE activity in species-rich clades and dominant families of TEs within lineages mirror past findings in mammals.

Author Summary:

Transposable elements (TEs) are mobile, self replicating DNA sequences within a species' genome, and are ubiquitous sources of mutation. The dominant group of TEs within birds are chicken repeat 1 (CR1) retrotransposons, making up 7-10% of the typical avian genome. Because past research has examined the recent inactivity of CR1s within model birds such as the chicken and the zebra finch, this has fostered an erroneous view that all birds have low or no TE activity on recent timescales. Our analysis of numerous high quality avian genomes across multiple orders identified both similarities and significant differences in how CR1s expanded. Our results challenge the established view that TEs in birds are largely inactive and instead suggest that their variation in recent activity may contribute to lineage-specific changes in genome structure. Many of the

patterns we identify in birds have previously been seen in mammals, highlighting parallels between the evolution of birds and mammals.

Introduction:

Following rapid radiation during the Cretaceous-Paleogene transition, birds have diversified to be the most species-rich lineage of extant amniotes (Jarvis et al. 2014; Ericson et al. 2006; Wiens 2015). Birds are of particular interest in comparative evolutionary biology because of the convergent evolution of traits seen in mammalian lineages, such as vocal learning in songbirds and parrots (Bradbury and Balsby 2016; Petkov and Jarvis 2012; Pfenning et al. 2014), and potential consciousness in corvids (Nieder et al. 2020). However in comparison to both mammals and non-avian reptiles, birds have much more compact genomes (Gregory et al. 2007). Within birds, smaller genome sizes correlate with higher metabolic rate and the size of flight muscles (Hughes and Hughes 1995; Wright et al. 2014). However, the decrease in avian genome size occurred in an ancestral dinosaur lineage over 200 Mya, well before the evolution of flight (Organ et al. 2007). A large factor in the smaller genome size of birds in comparison to other amniotes is a big reduction in repetitive content (Zhang et al. 2014).

The majority of transposable elements (TEs) in the chicken (*Gallus gallus*) genome are degraded copies of one superfamily of retrotransposons, chicken repeat 1 (CR1) (International Chicken Genome Sequencing Consortium 2004). The chicken has long been used as the model avian species, and typical avian genomes were believed to have been evolutionarily stable due to little variation in chromosome number and chromosomal painting showing little chromosomal rearrangement (Burt et al. 1999; Shetty et al. 1999). These initial, low resolution comparisons of genome features, combined with the degraded nature of CR1s in the chicken genome, led to the assumption of a stable avian genome both in terms of karyotype and synteny but also in terms of little recent repeat expansion (International Chicken Genome Sequencing Consortium 2004; Wicker et al. 2005). The subsequent sequencing of the zebra finch (*Taeniopygia guttata*) genome supported the concept of a stable avian genome with little repeat expansion, but revealed many

intrachromosomal rearrangements and a significant expansion of endogenous retroviruses (ERVs), a group of long terminal repeat (LTR) retrotransposons, since divergence from the chicken (Warren et al. 2010; Ellegren 2010). The subsequent sequencing of 48 bird genomes by the Avian Phylogenomics Project confirmed CR1s as the dominant TE in all non-passerine birds, with an expansion of ERVs in oscine passerines following their divergence from suboscine passerines (Zhang et al. 2014). The TE content of most avian genomes has remained between 7-10% not because of a lack of expansion, but due to the loss and decay of repeats and intervening non-coding sequence through non-allelic homologous recombination, cancelling out genome size expansion that would have otherwise increased with TE expansion (Kapusta et al. 2017). Since then, hundreds of bird species have been sequenced, revealing variation in karyotypes, and both intrachromosomal and interchromosomal rearrangements (Hooper and Price 2017; Damas et al. 2018; Feng et al. 2020; Kretschmer et al. 2020a, 2020b). This massive increase in genome sequencing has similarly revealed TEs to be highly active in various lineages of birds. Within the last 10 million years ERVs have expanded in multiple lineages of songbirds, with the newly inserted retrotransposons acting as a source of structural variation (Suh et al. 2018; Boman et al. 2019; Weissensteiner et al. 2020). Recent CR1 expansion events have been noted in woodpeckers and hornbills, leading to strikingly more repetitive genomes than the “typical” 7-10%. Between 23% to 30% of woodpecker and hoopoe genomes are CR1s, however their genome size remains similar to that of other birds (Feng et al. 2020; Manthey et al. 2018; Zhang et al. 2014). While aforementioned research focusing on the chicken suggested CR1s have not recently been active in birds, research focusing on individual avian lineages has used both recent and ancient expansions of CR1 elements to resolve deep nodes in a wide range of orders including early bird phylogeny (Suh et al. 2011; Matzke et al. 2012; Suh et al. 2015), flamingos and grebes (Suh et al. 2012), landfowl (Kriegs et al. 2007; Kaiser et al. 2007), waterfowl (St John et al. 2005), penguins (Watanabe et al. 2006), ratites (Haddrath and Baker 2012; Baker et al. 2014; Cloutier et al. 2019) and perching birds (Treplin and Tiedemann 2007; Suh et al. 2017). These studies largely exclude terminal branches and, with the exception of a handful of CR1s in grebes (Suh et al. 2012) and geese (St John et al. 2005), the timing of very recent insertions across multiple species remains unaddressed.

An understanding of TE expansion and evolution is important as they generate genetic novelty by promoting recombination that leads to gene duplication and deletion, reshuffling of genes and major structural changes such as inversions and chromosomal translocations (Zhou and Mishra 2005; Bailey et al. 2003; Lim and Simmons 1994; Underwood and Choi 2019; Lee et al. 2008; Chuong et al. 2017). TEs also have the potential for exaptation as regulatory elements and both coding and noncoding sequences (Warren et al. 2015; Wang et al. 2017; Barth et al. 2020). *Ab initio* annotation of repeats is necessary to gain a true understanding of genomic repetitive content, especially in non-model species (Platt et al. 2016). Unfortunately, many papers describing avian genomes (Cornetti et al. 2015; Jaiswal et al. 2018; Laine et al. 2016) only carry out homology-based repeat annotation using the Repbase (Bao et al. 2015) library compiled from often distantly related model avian genomes (mainly chicken and zebra finch. This lack of *ab initio* annotation can lead to the erroneous conclusion that TEs are inactive in newly sequenced species (Platt et al. 2016). Expectations of low repeat expansion in birds inferred from two model species, along with a lack of comparative TE analysis between lineages is the large knowledge gap we addressed here. As CR1s are the dominant TE lineage in birds, we carried out comparative genomic analyses to investigate their diversity and temporal patterns of activity.

Results

Identifying potential CR1 expansion across birds

From all publicly available avian genomes, we selected 117 representative assemblies not under embargo and with a scaffold N50 above 20,000 bp (available at July 2019) for analysis (SI Table 1). To find all CR1s that may have recently expanded in the 117 genomes, we first used the CARP *ab initio* TE annotation tool. From the output of CARP, we manually identified and curated CR1s with the potential for recent expansion based on the presence of protein domains necessary for retrotransposition, homology to previously described CR1s, and the presence of a distinctive 3' structure. To retrotranspose and hence expand, CR1s require endonuclease and reverse transcriptase domains within a single ORF, and a 3' structure containing a hairpin and microsatellite which potentially acts as a recognition site for the reverse transcriptase (Suh et al.

2014; Suh 2015). If a CR1 identified from homology contained both protein domains and the distinctive 3' structure, we classified it as a "full length" CR1. We next classified a full length CR1 as "intact" CR1 if the endonuclease and reverse transcriptase were within a single intact ORF. Using the full length CR1s and previously described avian and crocodilian CR1s in Repbase as queries (Green et al. 2014; International Chicken Genome Sequencing Consortium 2004; Warren et al. 2010), we performed iterative searches of the 117 genomes to identify divergent, low copy number CR1s which may not have been identified by *ab initio* annotation. We ensured the protein domains and 3' structures were present throughout the iterative searches. Assemblies with lower scaffold N50s generally contained fewer full length CR1s and none in the lowest quartile contained intact CR1s (Figure 1). Outside of the lowest quartile, assembly quality appeared to have little impact on the proportion of intact, full length repeats. The correlation of the low assembly quality with little to no full length CR1s was seen both across all species and within orders.

Our iterative search identified high numbers of intact CR1s in kiwis, parrots, owls, shorebirds and waterfowl (Figures 1 and 2). Only 2 of the 22 perching bird (Passeriformes) genomes contained intact CR1s, and all contained 10 or fewer full length CR1s. Similarly, of the 7 landfowl (Galliformes) genomes, only the chicken contained intact CR1s and contained fewer than 20 full length CR1s. High numbers of full length and intact repeats were also identified in two woodpeckers, Anna's hummingbird, the chimney swift and the hoatzin, however, due to a lack of other genome sequences from their respective orders, we were unable to perform further comparative within order analyses of these species to look for recent TE expansion, i.e., within the last 10 million years. Of all the lineages we examined, only four have high quality assemblies of genera which have diverged within the last 10 million years and, based on the number of full length CR1s identified, the potential for very recent CR1 expansion: ducks (*Anas*), geese (*Anser*), Amazon parrots (*Amazona*) and kiwis (*Apteryx*) (Silva et al. 2017; Mitchell et al. 2014; Sun et al. 2017). While the large number of full length repeats identified in owls is also high, we were unable to examine recent expansion in Strigiformes in detail due to the lack of a dated phylogeny. In addition to our genus scale analyses, we also examined CR1 expansion in parrots (Psittaciformes) overall, perching birds (Passeriformes) and shorebirds (Charadriiformes) since the divergence of

each group, and compared the expansion in kiwis and their closest living relatives (Casuariiformes).

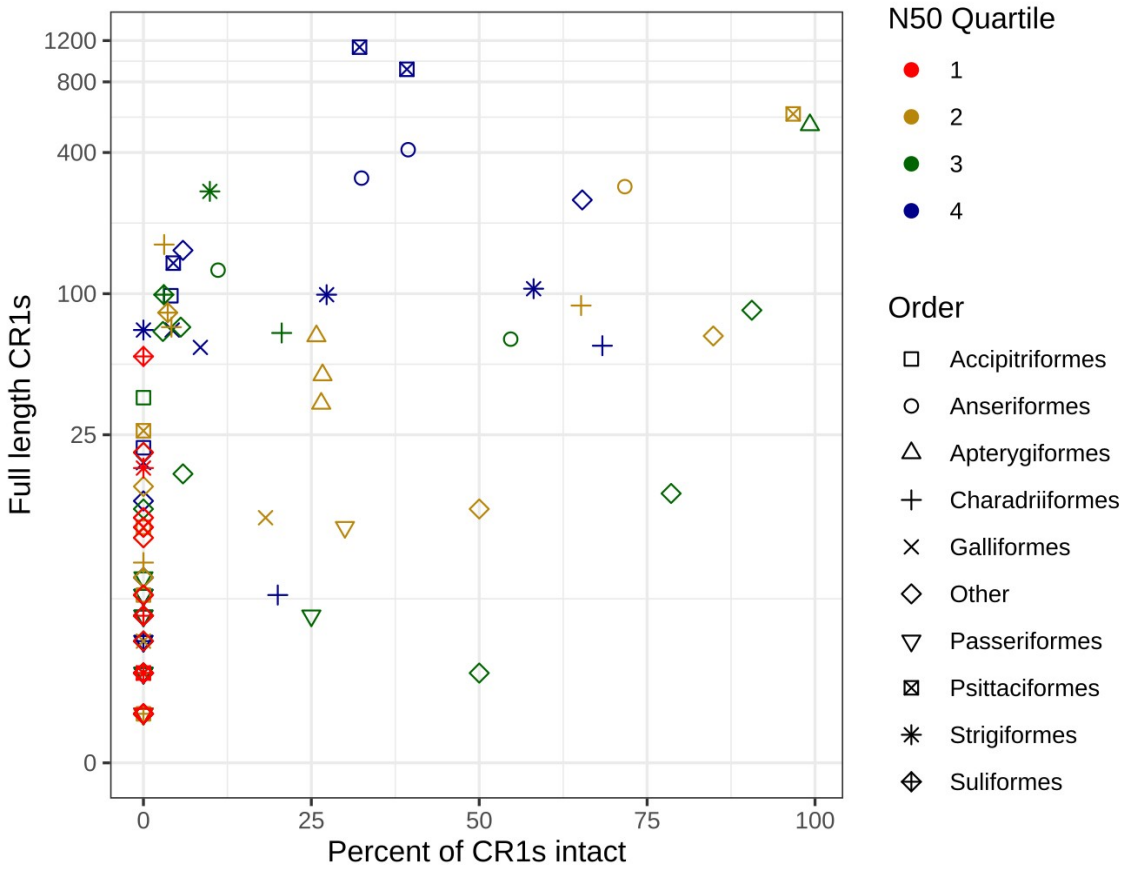
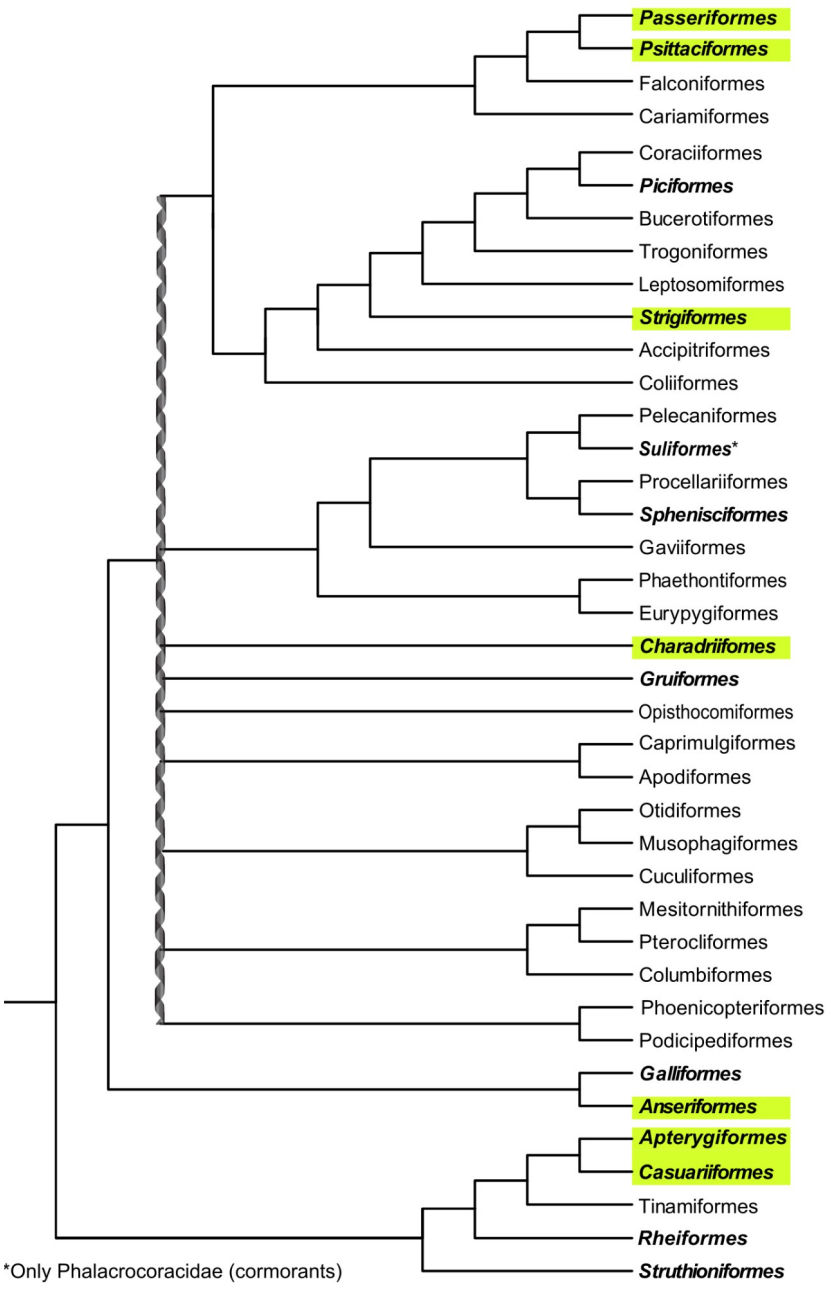


Figure 1: The impact of genome assembly quality on the identification of full length and intact CR1s. CR1s containing both an endonuclease and reverse transcriptase domains were considered full length, and those containing both domains within a single ORF considered intact. Both across all orders and within individual orders, genomes with higher scaffold N50 values (quartiles 2 through 4) had higher numbers of full length CR1s.



Min	Max	Mean	Species
0	10	2.7	23
0	1125	309.4	9
3	6	4.3	3
10	10	10	1
1	1	1	1
5	12	8.5	1
0	0	0	1
9	9	9	1
11	11	11	1
2	273	94.5	6
0	98	23.4	7
1	1	1	1
9	21	14	3
4	99	60	4
2	2	2	1
1	2	1.5	2
2	2	2	1
2	2	2	1
2	2	2	1
1	162	46.9	10
2	153	77.5	2
66	66	66	1
1	1	1	1
85	251	168	2
2	2	2	1
10	10	10	1
6	6	6	1
3	3	3	1
1	1	1	1
13	13	13	1
4	4	4	1
5	5	5	1
0	70	23.4	8
64	411	239.6	5
34	524	167.3	4
15	17	16	2
2	21	9.3	3
69	72	70.5	2
14	14	14	1

*Only Phalacrocoracidae (cormorants)

Figure 2: The number of full length CR1s varies significantly across the diversity of birds sampled. Minimum, maximum and mean number of full length CR1 copies identified in each order of birds, and the number of species surveyed in each order. Largest differences are noticeable between sister clades such as parrots (Psittaciformes) and perching birds (Passeriformes), and landfowl (Galliformes) and waterfowl (Anseriformes). The double helix represents a putative hard polytomy at the root of Neoaves (Suh 2016). Orders bolded contain at least one intact and potentially active

CR1 copy and those highlighted in yellow are the orders examined in detail. For coordinates of full length CR1s within genomes, see SI Data 1. Tree adapted from (Mitchell et al. 2014; Suh 2016).

Order-specific CR1 annotations and a phylogeny of avian CR1s reveal diversity of candidate active CR1s in neognaths

In order to perform comparative analyses of activity within orders, we created order-specific CR1 libraries. Instead of consensus sequences, all full length CR1s identified within an order were clustered and the centroids of the clusters were used as cluster representatives for that avian order. To classify the order-specific centroids, we constructed a CR1 phylogeny from the centroids and full length avian and crocodilian CR1s from Repbase (Figure 3, SI Figure 1, SI Data 2). From this tree, we partitioned CR1s into families to determine if groups of elements have been active in species concurrently. We partitioned the tree by eye based on the phylogenetic position of previously described CR1 families (Vandergon and Reitman 1994; Wicker et al. 2005; Warren et al. 2010; Bao et al. 2015) and long branch lengths rather than a cutoff for divergence, attempting to find the largest monophyletic groups containing as few previously defined CR1 families as possible. We took this “lumping” approach to our classification to avoid paraphyly and excessive splitting, resulting in some previously defined families being grouped together in one family (SI Table 2). For example, all full length CR1s identified in songbirds were highly similar to the previously described CR1-K and CR1-L families and were nested deeply within the larger CR1-J family. As a result, CR1-K, CR1-L and all full length songbird CR1s were reclassified as subfamilies of the larger CR1-J family. Based on the position of well resolved, deep nodes and previously described CR1s in the phylogeny, we defined 7 families of avian CR1s, with a new family, CR1-W, which was restricted to shorebirds. Interestingly, the 3’ microsatellite of the CR1-W family is a 10-mer rather than the octamer found in nearly all amniote CR1s (Suh 2015). With the exception of Palaeognathae (ratites and tinamous), all avian orders that contained large numbers of full length CR1s also contained full length CR1s from multiple CR1 families (Figure 3).

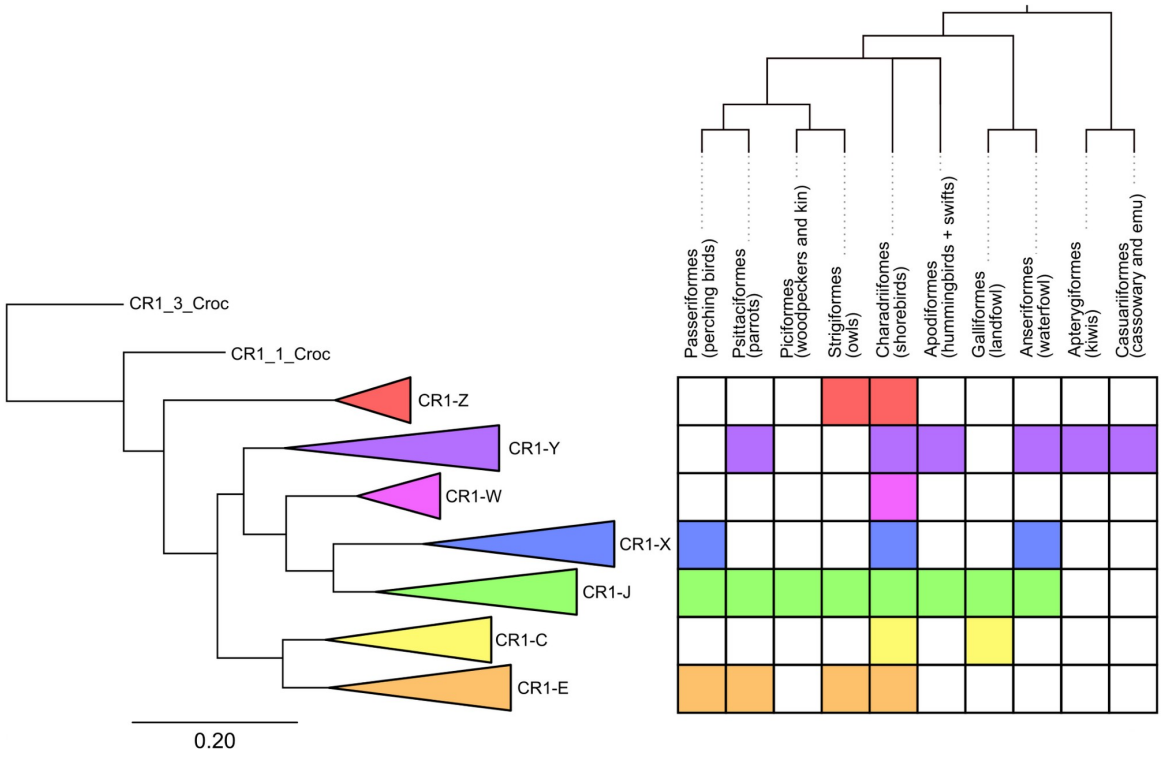


Figure 3: Collapsed tree of full length CR1s and presence of full length copies of CR1 families in selected avian orders. The name of each family is taken from a previously described CR1 present within the family (SI Table 3). The colouring of squares indicates the presence of full length CR1s within the order. All orders shown were chosen due to the presence of high numbers of intact CR1 elements, except for Casuariiformes which are shown due to their recent divergence from Apterygiformes as well as Passeriformes due to their species richness and frequent use as model species (especially zebra finch). The full CR1 tree was constructed using FastTree from a MAFFT alignment of the nucleotide sequences. For the full tree and nucleotide alignment of 1278 CR1s see SI Figure 1 and SI Data 2.

Variable timing of expansion events across avian orders

We used the aforementioned order-specific centroid CR1s and avian and crocodilian Repbase sequences to create order-specific libraries. Throughout the following analysis we ensured CR1 copies identified were 3' anchored, i.e. retain 3' ends with homology to both the hairpin sequences

and microsatellites. We used the order-specific libraries in reciprocal searches to identify and classify 3' anchored CR1s present within all orders in which we had identified full length repeats. Using the classified CR1s we searched for all 3' anchored CR1s (both full length and truncated) and constructed divergence plots to gain a basic understanding of CR1 expansions within each genome (SI Data 3). At high Jukes-Cantor distances, divergence profiles in each order show little difference between species. However, at lower Jukes-Cantor distances divergence, profiles differ significantly between species in some orders. For example, in songbirds at Jukes-Cantor distances higher than 0.1 the overall shape of the divergence plot curves and the proportions of the various CR1 families are nearly identical, while at distances lower than 0.1 higher numbers of the CR1-J family are present in some passerines than others (SI Figure 2a). CR1s most similar to all defined families were present in all orders of Galloanserae and Neoaves examined, with the exception of CR1-X which was restricted to Charadriiformes. Almost all CR1s identified in Palaeognathae genomes were most similar to CR1-Y with a small number of truncated and divergent repeats most similar to crocodilian CR1s (SI Data 3).

Divergence plots may not accurately indicate the timing of repeat insertions as they assume uniform substitution rates across the non-coding portion of the genome. High divergence could be a consequence of either full length CR1s being absent in a genome or the centroid identified by the clustering algorithm being distant from the CR1s present in a genome. To better determine when CR1 families expanded in avian genomes, we first identified regions orthologous to CR1 insertions sized 100-600 bp in related species (see Methods). We compared these orthologous regions and approximated the timing of insertion based on the presence or absence of the CR1 insertion in the other species. In most orders only long term trends could be estimated due to long branch lengths (cf. Figure 2) and high variability of the quality of genome assemblies (cf. Figure 1). Therefore, we focused our presence/absence analyses to reconstruct the timing of CR1 insertions in parrots, waterfowl, perching birds, and kiwis (Figure 4). We also applied the method to owls (SI Figure 3) and shorebirds (Figure 5), however due to the lack of order-specific fossil calibrated phylogenies of owls and long branch lengths of shorebirds, we could not determine how recent the CR1 expansions were.

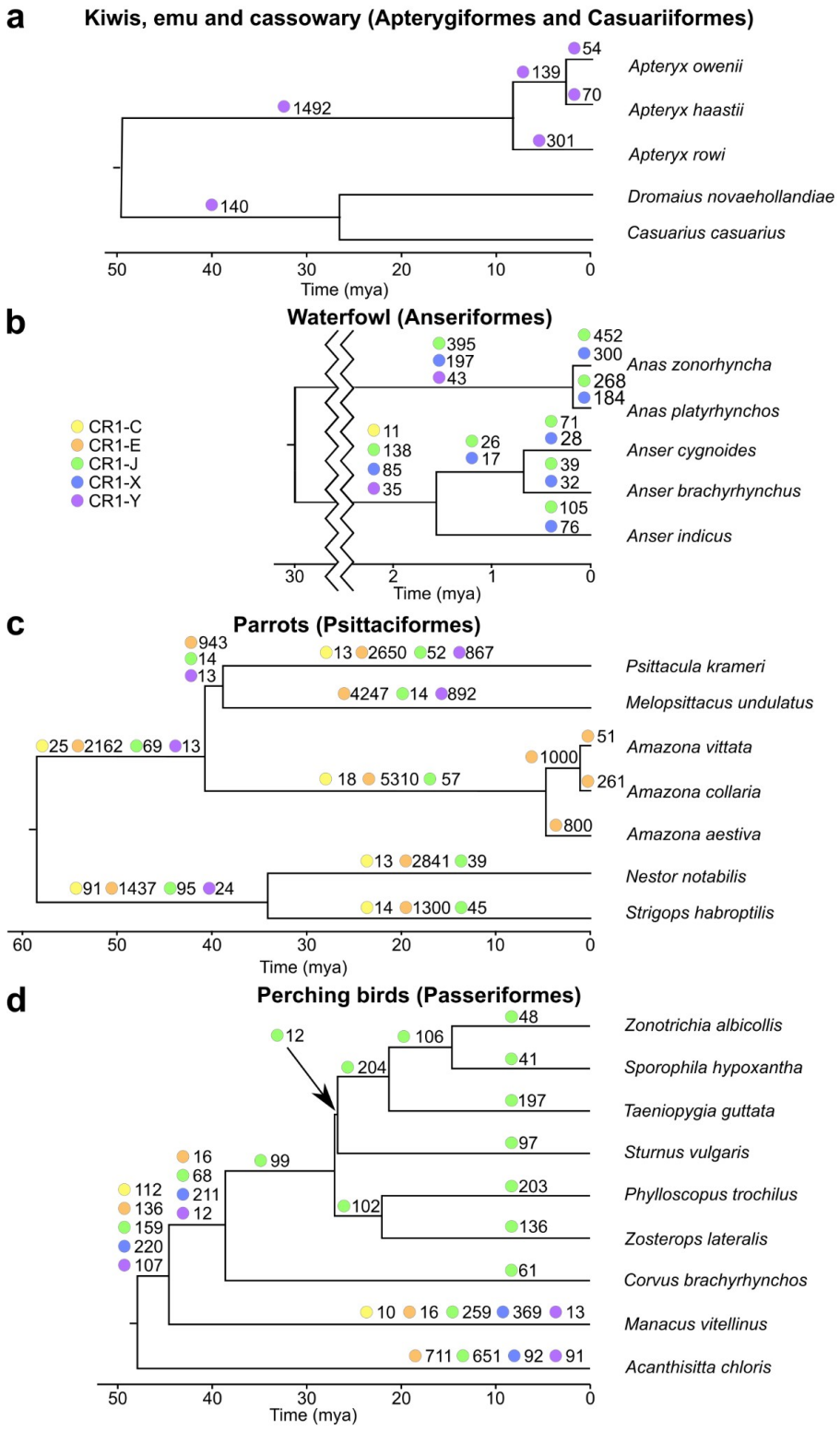


Figure 4: Presence/absence patterns reconstruct the timing of expansions of dominant CR1 families within five selected avian orders. The number next to the coloured circle is the number of CR1 insertions found. Only CR1 families with more than 10 CR1 presence/absence patterns (only

CR1 insertions ranging between 100 and 600 bp were analyzed) are shown, for the complete number of insertions see SI Table 3. Phylogenies adapted from (Mitchell et al., 2014; Oliveros et al., 2019; Silva et al., 2017; Sun et al., 2017).

In analysing the repeat expansion in the kiwi genomes, we used the closest living relatives, the cassowary and emu (Casuariiformes), as outgroups. Following the divergence of kiwis from Casuariiformes, CR1-Y elements expanded, both before and during the recent speciation of kiwis over the last few My. In contrast, there was little CR1 expansion in Casuariiformes, both following their divergence from kiwis, and more recently since their divergence ~28 Mya, with only 1 insertion found in the emu and 3 in the cassowary since they diverged (SI Table 3).

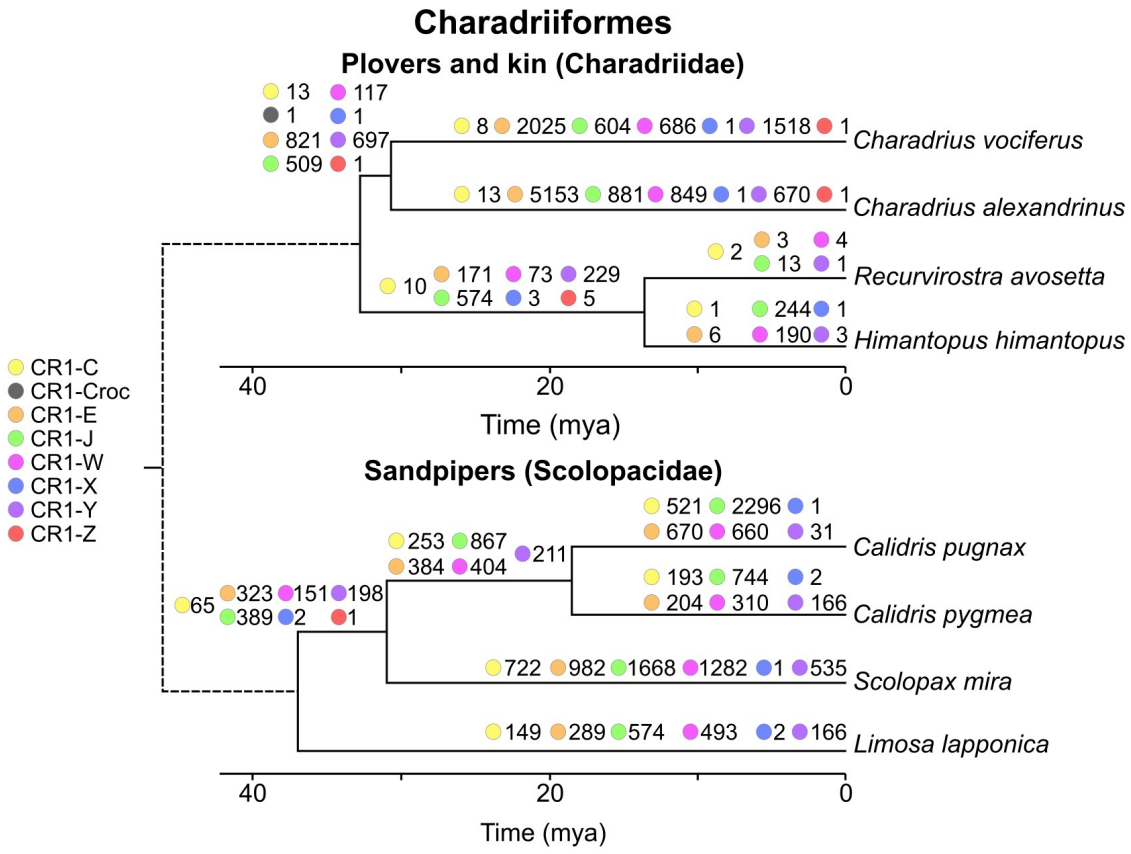
In the waterfowl species examined, both CR1-J and CR1-X families expanded greatly in both ducks and geese during the last 2 million years. Expansion occurred in both examined genera, with greater expansions in the ducks (*Anas*) than the geese (*Anser*). Other CR1 families appear to have been active following the two groups' divergence ~30 Mya, but have not been active since each genus speciated.

Due to the high number of genomes available for passerines, we chose best quality representative genomes from major groups *sensu (Oliveros et al. 2019)*; New Zealand wrens (*Acanthisitta chloris*), Suboscines (*Manacus vitellinus*), Corvides (*Corvus brachyrhynchos*), and Muscicapida (*Sturnus vulgaris*), Sylvida (*Phylloscopus trochilus* and *Zosterops lateralis*) and Passerida (*Taeniopygia guttata*, *Sporophila hypoxantha* and *Zonotrichia albicollis*). Between the divergence of Oscines (songbirds) and Suboscines from New Zealand wrens and the divergence of Oscines, there was a large spike in expansion of multiple families of CR1s, predominantly CR1-X. Since their divergence 30 Mya, only CR1-J remained active in oscines, though the degree of expansion varied between groups.

Of all avian orders examined, we found the highest levels of CR1 expansion in parrots. Because most branch lengths on the species tree were long, the timing of recent expansions could only be

27

282 reconstructed in genus *Amazona*. The species from *Amazona* diverged 5 Mya ago and seem to
283 vary significantly in their level of CR1 expansion. However, genome assembly quality might be a
284 confounder as the number of insertions into a species of *Amazona* was highest in the best quality
285 genome (*Amazona collaria*), and lowest in the worst quality genome (*Amazona vittata*). In all
286 parrots, CR1-E was the predominant expanding CR1 family, however CR1-Y expanded in the
287 *Melopsittacus-Psittacula* lineage, while remaining largely inactive in the other parrot lineages.



288

289 Figure 5: Presence/absence patterns reconstruct the timing of expansions of CR1 families in two
290 lineages of shorebirds (Charadriiformes): plovers and sandpipers. The number next to the coloured
291 circle is the number of CR1 insertions identified and only CR1 insertions between 100 and 600 bp
292 long were analyzed. Divergence dates between plovers and sandpiper clades may differ due to
293 the source phylogenies (Barth et al. 2013; Paton et al. 2003; Baker et al. 2007) being constructed
294 using different approaches.

295

28 14

Multiple expansions of multiple families of CR1s have occurred in the two shorebird lineages examined; plovers (Charadriidae) and sandpipers (Scolopacidae) (Figure 5). The diversity of CR1 families that remained active through time was higher than in the other orders investigated, particularly in sandpipers, with four CR1 families showing significant expansion in *Calidris pugnax* and five in *Calidris pygmaea,* since their divergence. In all other orders examined in detail, CR1 expansions over similar time periods have been dominated by only one or two families, with insertions of fewer than 10 CR1s from non-dominant families (SI Table 2). Unfortunately, due to long branch lengths more precise timing of these expansions is not possible.

Finally, CR1s continuously expanded in true owls since divergence from barn owls, with almost all resolved insertions being CR1-E-like (SI Figure 3). However, due to the lack of a genus-level timed phylogeny, the precise timing of these expansions cannot be determined.

Combined, our CR1 presence/absence analyses demonstrate that the various CR1 families have expanded at different rates both within and across avian orders. These differences are considerable, ranging from an apparent absence of CR1 expansion in the emu and cassowary to slow, continued expansion of a single CR1 family in songbirds, to recent rapid expansions of one or two CR1 families in kiwis, Amazon parrots and waterfowl, as well as a wide variety of CR1 families expanding concurrently in sandpipers.

Discussion

Genome assembly quality impacts repeat identification

The quality of a genome assembly has a large impact on the number of CR1s identified within it, both full length and 5'-truncated. This is made clear when comparing the number of insertions identified within species in recently diverged genera. The three *Amazona* parrot species diverged approximately ~2 Mya (Silva et al. 2017) and the scaffold N50s of *A. vittata, A. aestiva* and *A. collaria* are 0.18, 1.3 and 13 Mbp respectively. No full length CR1s were identified in *A. vittata,* and only 10 in *A. aestiva,* while 1125 were identified in *A. collaria*. Similarly, in *Amazona* the total number of truncated insertions identified increased significantly with higher scaffold N50s. In contrast the three species of kiwi compared, diverged ~7 Mya and have similar N50s (between 1.3 and 1.7 Mbp). This pattern of higher quality genome assemblies leading to higher numbers of both full length and intact CR1s being identified is consistent across most orders examined, and is particularly true of the lowest N50 quartile (Figure 1). The lower number of repeats identified in lower quality assemblies is likely due to the sequencing technology used. Repeats are notoriously hard to assemble and are often collapsed, particularly when using short read Illumina sequencing, leading to fragmented assemblies (Alkan et al. 2011; Treangen and Salzberg 2011). The majority of the genomes we have used are of this data type. The recent sequencing of avian genomes using multiplatform approaches have resolved gaps present in short read assemblies, finding these gaps to be rich in interspersed, simple and tandem repeats (Peona et al. 2021; Li et al. 2021). Of particular note (Li et al. 2021) resolved gaps in the assembly of *Anas platyrhynchos* which we analyzed here using long read sequencing, and found the gaps to be dominated by the two CR1 families that have recently expanded in waterfowl (Anseriformes): CR1-J and CR1-X. Species with low quality assemblies may have full length repeats present in their genome, yet the sequencing technology used prevents the assembly of the repeats and hence detection. Thus TE activity may be even more widespread in birds than we estimate here.

The origin and evolution of avian CR1s

Avian CR1s are monophyletic in regards to other major CR1 lineages found in amniotes (Suh et al. 2014). For comparison, crocodilians contain some CR1 families more similar to those found in testudines and squamates than others in crocodilians. By searching for truncated copies of previously described CR1s in addition to our order-specific CR1s, we were able to uncover how CR1s have evolved in avian genomes as birds have diverged. CR1-Y is the only family with full length CR1s present in Paleognathae, Galloanserae and Neoaves. The omnipresence of CR1-Y indicates it was present in the ancestor of all birds. A small number of highly divergent truncated copies of CR1s most similar to CR1-Z are found in ratites and CR1-J in tinamous (SI Figure 2b). This is potentially indicative of an ancestral presence of CR1-J and CR1-Z in the common ancestor of all birds, or misclassification owing to the high divergence of these CR1 fragments. As mentioned above, we took a lumping approach to classification to CR1 classification to avoid paraphyly, thereby collapsing highly similar families elsewhere considered as separate families. As CR1-C, CR1-E, and CR1-X are present in both Galloanserae and Neoaves but absent from Palaeognathae, we conclude these 4 families likely originated following the divergence of neognaths from paleognaths, but prior to the divergence of Neoaves and Galloanserae. In addition to having a 10 bp microsatellite instead of the typical 8 bp microsatellite, CR1-W is peculiar as it is unique to Charadriiformes but sister to CR1-J and CR1-X (Figure 3). This implies an origin in the neognath ancestor, followed by retention and activity in measurable numbers only in Charadriiformes.

A wide variety of CR1 families has expanded in all orders of neognaths, with many potential expansion events within the past 10 My present in many lineages. As mentioned in the results, it is not possible to conclude that insertions are ancient based on divergence plots alone. Some species with low quality genome assemblies, such as *A. vittata*, contained very few full length repeats compared to relatives (SI Figure 4). As a result of full length repeats not being assembled, the divergence of most or all truncated insertions identified in *A. vittata* would likely be calculated using CR1 centroids identified in *A. collaria*, leading to higher divergence values than those identified in *A. collaria*, and in turn an incorrect assumption of less recent expansion in *A. vittata*

than *A. collaria*. In addition to fewer full length repeats being assembled, fewer truncated repeats also appear to have been assembled in poorer quality genomes.

CR1 family expansions within orders

Across all sampled neognaths, recent expansions appear to be largely restricted to one or two families of CR1. Our presence/absence analyses found this to be the case in waterfowl, parrots, songbirds and owls, with shorebirds and the early passerine divergences the only exceptions. Similarly, based on the phylogeny of full length elements, most orders only retain full length CR1s from two or three families, while shorebirds retain full length CR1s from across all seven families. Our presence/absence analysis revealed likely concurrent expansions of at least four CR1 families in two families of shorebirds: sandpipers of genus *Calidris* and plovers of genus *Charadrius*. In both genera four families of CR1s have significantly expanded since their divergence including the order-specific CR1-W (Figure 5). While in both genera one family accounts for 40 to 50% of insertions, the other three families have hundreds of insertions each. This is highly different to the pattern seen in songbirds and waterfowl which, over a similar time period, have single digit insertions of non-dominant CR1 families (SI Table 3).

This increase of CR1 diversity in shorebirds could be due to some CR1 families in shorebirds having 3' inverted repeat and microsatellite motifs which differ from the typical structure (Suh 2015) (SI Fig). For example, the CR1-W family has an extended 10 bp microsatellite (5'-AAATTCYGTG-3') rather than the 8 bp microsatellite (5'-ATTCTRTG-3') seen in nearly all other avian CR1s. When transcribed the 3' structure upstream of the microsatellite is hypothesized to form a stable hairpin which acts as a recognition site for the cis-encoded reverse transcriptase (Suh 2015; Suh et al. 2017; Luan et al. 1993). The recently active CR1s we identified in other avian orders have 3' microsatellites and hairpins which closely resemble those previously described. While the changes seen in shorebirds are minor we speculate they could impact CR1 mobilisation, allowing for more families to remain active than the typical one or two.

Rates of CR1 expansion can vary significantly within orders

Based on the presence/absence of CR1 insertions and divergence plots, rates of CR1 expansion within lineages appear to vary even across rather short evolutionary timescales. The expansion of CR1-Y in kiwis appears to be a recent large burst of expansion and accumulation, while since Passeriformes diverged CR1-J appear to have continued to expand slowly in all families, however the number of new insertions seen in the American crow is much lower than that seen in the other oscine songbird species surveyed. The expansion of CR1-Y seen in the *Psittacula-Melopsittacus* lineage of parrots, following their divergence from the lineage leading to *Amazona,* appears to result from an increase in expansion, with little expansion in the period prior to divergence and none observed in other lineages of parrots. CR1s appear to have been highly active in all parrots examined since their divergence, however due to the less dense sampling it is not clear if this has been continuous expansion as in songbirds or a burst of activity like that in kiwis. Finally, in sandpipers CR1s have continued to expand in both species of *Calidris* since divergence, however the much lower number of new insertions in *C. pygmaea* suggests the rate of expansion differs significantly between the two species.

All full length CR1s identified in ratites were CR1-Y, and almost all truncated copies found in ratites were most similar to either CR1-Y, or crocodilian CR1s typically not found in birds (Suh et al. 2014). This retention of ancient CR1s and the presence of full length CR1s in species such as the southern cassowary (*Casuarius casuarius*) and emu (*Dromaius novaehollandiae*), yet without recent expansion, reflects the much lower substitution and deletion rates in ratites compared to Neoaves (Zhang et al. 2014; Kapusta et al. 2017). These crocodilian-like CR1s in ratites may be truncated copies of CR1s that were active in the common ancestor of crocodilians and birds (Suh et al. 2014) while we hypothesise that these have long since disappeared in Neoaves due to their higher deletion and substitution rates (Kapusta et al. 2017; Zhang et al. 2014).

Co-occurrence of CR1 expansion with speciation

The four genera containing recent CR1 expansions we have examined co-occur with rapid speciation events. Of particular note, kiwis rapidly speciated into 5 distinct species composed of at least 16 distinct lineages arising due to significant population bottlenecks caused by Pleistocene

glacial expansions (Weir et al. 2016). We speculate that the smaller population sizes might have allowed for CR1s to expand as a result of increased genetic drift (Szitenberg et al. 2016). While we do not see CR1 expansion occurring alongside speciation in passerines, ERVs, which are rare in other birds, have expanded throughout their diversification (Boman et al. 2019; Warren et al. 2010). Investigating the potentially ongoing expansion of CR1s and its relationship to speciation in ducks, geese, and Amazon parrots will require a larger number of genomes from within the same and sister genera to be sequenced, especially in waterfowl due to the high rates of hybridisation even between long diverged species (Ottenburghs et al. 2015).

Comparison to mammals

As mentioned in the introduction, many parallels have been drawn between LINEs in birds and mammals, most notably the expansion of LINEs in both clades being balanced by a loss through purifying selection (Kapusta et al. 2017). Here we have found additional trends in birds previously noted in mammals. The TE expansion during periods of speciation seen in *Amazona*, *Apteryx* and *Anas* has previously been observed across mammals (Ricci et al. 2018). Similarly, the dominance of one or two CR1 families seen in most orders of birds resembles the activity of L1s in mammals (Ivancevic et al. 2016), however the general persistence of activity of individual CR1 families seems to be more diverse (Kriegs et al. 2007; Suh et al. 2011).

Conclusion: the avian genome is more dynamic than meets the eye

While early comparisons of avian genomes were restricted to the chicken and zebra finch, where high level comparisons of synteny and karyotype led to the conclusion that bird genomes were largely stable compared to mammals (Ellegren 2010), the discovery of many intrachromosomal rearrangements across birds (Hooper and Price 2017; Skinner and Griffin 2012; Zhang et al. 2014; Farre et al. 2016) and interchromosomal recombination in falcons, parrots and sandpipers (O'Connor et al. 2018; Coelho et al. 2019; Pinheiro et al. 2021) has shown that at a finer resolution for comparison, the avian genome is rather dynamic. The highly variable rate of TE expansion we have observed across birds extends knowledge from avian orders with "unusual" repeat landscapes, i.e., Piciformes (Manthey et al. 2018) and Passeriformes (Warren et al. 2010), and

provides further evidence that the genome evolution of bird orders and species within orders differs significantly, even though synteny is often conserved. In our comprehensive characterization of CR1 diversity across 117 bird genome assemblies, we have identified significant variation in CR1 expansion rates, both within genera such as *Calidris* and between closely related orders such as kiwis and the cassowary and emu. As the diversity and quality of avian genomes sequenced continues to grow and whole genome alignment methods improve (Feng et al. 2020; Rhie et al. 2020), further analysis of genome stability based on repeat expansions at the family and genus level will become possible. While the chicken and zebra finch are useful model species, models do not necessarily represent diversity of evolutionary trajectories in nature.

Methods and Materials

Identification and curation of potentially divergent CR1s

To identify potentially divergent CR1s we processed 117 bird genomes downloaded from Genbank (Benson et al. 2015) with CARP (Zeng et al. 2018); see SI Table for species names and assembly versions. We used RPSTBLASTN (Altschul et al. 1997) with the CDD library (Marchler-Bauer et al. 2017) to identify protein domains present in the consensus sequences from CARP. Consensuses which contained both an endonuclease and a reverse transcriptase domain were classified as potential CR1s. Using CENSOR (Kohany et al. 2006) we confirmed these sequences to be CR1s, removing others, more similar to different families of LINEs, such as AviRTEs, as necessary.

Confirmed CR1 CARP consensus sequences were manually curated through a “search, extend, align, trim” method as described in (Galbraith et al. 2020) to ensure that the 3’ hairpin and microsatellite were intact. Briefly, this curation method involves searching for sequences highly similar to the consensus with BLASTN 2.7.1+ (Zhang et al. 2000), extending the coordinates of the sequences found by flanks of 600 bp, aligning these sequences using MAFFT v7.453 (Katoh and Standley 2013) and trimming the discordant regions manually in Geneious Prime v2020.1. The final consensus sequences were generated in Geneious Prime from the trimmed multiple sequence alignments by majority rule.

Identification of more divergent and low copy CR1s

To identify more divergent or low copy number CR1s which CARP may have failed to identify, we performed an iterative search of all 117 genomes. Beginning with a library of all avian CR1s in Repbase (Bao et al. 2015) (see SI Table 2 for CR1 names and species names) and manually curated CARP sequences we searched the genomes using BLASTN (-task dc-megablast -max_target_seqs <number of scaffolds in respective genome>), selecting those over 2700 bp and retaining 3' hairpin and microsatellite sequences. Using RPSTBLASTN we then identified the full length CR1s (those containing both endonuclease (EN) and reverse transcriptase (RT) domains) and combined them with the previously generated consensus sequences. We clustered these combined sequences using VSEARCH 2.7.1 (Rognes et al. 2016) (--cluster_fast --id 0.9) and combined the cluster centroids with the Repbase CR1s to use as queries for the subsequent search iteration. This process was repeated until the number of CR1s identified did not increase compared to the previous round. From the output of the final round, order-specific clusters of CR1s were constructed and cluster centroids identified.

Tree construction

To construct a tree of CR1s, the centroids of all order-specific CR1s were combined with all full length avian and two crocodilian CR1s from Repbase and globally aligned using MAFFT (--thread 12 --localpair). We used FastTree 2.1.11 with default nucleotide parameters (Price et al. 2010) to infer a maximum likelihood phylogenetic tree from this alignment, and rooted the tree using the crocodilian CR1s. The crocodilian CR1s were used as an outgroup as all avian CR1s are nested within crocodilian CR1s (Suh et al. 2015). This tree was split into different families of CR1 by eye based on the presence of long branches from high confidence nodes and the position of the previously described CR1 families from Repbase. To avoid excessive splitting and paraphyly of previously described families a lumping approach was taken resulting in some previously distinct families of CR1 from Repbase being treated as members of families they were nested within (SI Table 3).

Identification and classification of CR1s within species

To identify, classify and quantify divergence of all 3' anchored CR1s present within species, order-specific libraries were constructed from the order-specific clusters and the full length avian and crocodilian Repbase CR1s. 3' anchored sequences CR1s were defined as CR1s retaining the 3' hairpin and microsatellite sequences. Using these libraries as queries we identified 3' anchored sequences CR1s present in assemblies using BLASTN. The identified CR1s were then classified using a reciprocal BLASTN search against the original query library.

Determination of presence/absence in related species

To reconstruct the timing of CR1 expansions we selected the identified 3' anchored CR1 copies of 100 and 600 bp length in a species of interest and at least 600 bp from the end of a contig, extending the coordinates of the sequences by 600 bp to include the flanking region and extracting the corresponding sequences. If the flanking regions contained more than 25% unresolved nucleotides ('N' nucleotides) they were discarded.

Using BLASTN we identified homologous regions in species belonging to the same order as the species being analysed, and through the following process of elimination identified the regions orthologous to CR1 insertions and their flanks in the related species. At each step of this process of elimination, if an initial query could not be satisfactorily resolved, we classified it as unscorable (unresolved) to reduce the chance of falsely classifying deletions or segmental duplications as new insertion events. First, we classified all hits containing the entire repeat and at least 150 bp of each flank as shared orthologous insertions. Following this, we discarded all hits with outer coordinates less than a set distance (150 bp) from the boundary of the flanks and CR1s to remove hits to paralogous CR1s insertions. This distance was chosen by testing the effect of a range of distances from 300 bp through to 50 bp in increments of 50 bp on a random selection of CR1s first identified in *Anser cygnoides* and *Corvus brachyrhynchos* and searched for in other species within the same order. Requiring outer coordinates to higher values resulted in higher numbers orthologous regions not being resolved, likely due to insertions or deletions within flanks since divergence. Allowing for boundaries of 50 or 100 bp resulted in many CR1s having multiple potential orthologous regions at 3' flanks, many of were false hits, only showed homology to the target site duplication and

additional copies of the 3' microsatellite sequence. Thus 150 bp was chosen, as it was the shortest possible distance at which a portion of the flanking sequence was always present.

Based on the start and stop coordinates of the remaining hits, we determined the orientation the hit was in and discarded any queries without two hits in the same orientation. In addition, any queries with more than one hit to either strand was discarded. From the remaining data we determined the distance between the two flanks. If the two flanks were within 16 bp of each other in the sister species and the distance between the flanks was near the same length of the query CR1, the insertion was classified as having occurred since divergence. If the distance between the ends of the flanks in both the original species and sister species were similar, the insertion was classified as shared. For a pictorial description of this process including the parameters used, see SI Figure 5. This process was conducted for other species in the same order as the original species. Finally, we determined the timing of each CR1 insertion event by reconciling the presence/absence of each CR1 insertion across sampled species with the most parsimonious placement on the species tree (SI Figure 6).

Acknowledgments

We thank Valentina Peona, Jesper Boman, Julie Blommaert and Alastair Ludington for comments on an earlier version of this manuscript.

References

Alkan C, Sajjadian S, Eichler EE. 2011. Limitations of next-generation genome sequence assembly. *Nat Methods* **8**: 61–65.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Bailey JA, Liu G, Eichler EE. 2003. An Alu transposition model for the origin and expansion of human segmental duplications. *Am J Hum Genet* **73**: 823–834.

Baker AJ, Haddrath O, McPherson JD, Cloutier A. 2014. Genomic support for a moa-tinamou clade and adaptive morphological convergence in flightless ratites. *Mol Biol Evol* **31**: 1686–1696.

Baker AJ, Pereira SL, Paton TA. 2007. Phylogenetic relationships and divergence times of

Charadriiformes genera: multigene evidence for the Cretaceous origin of at least 14 clades of shorebirds. *Biol Lett* **3**: 205–209.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**: 11.

Barth JMI, Matschiner M, Robertson BC. 2013. Phylogenetic position and subspecies divergence of the endangered New Zealand Dotterel (Charadrius obscurus). *PLoS One* **8**: e78068.

Barth NKH, Li L, Taher L. 2020. Independent Transposon Exaptation Is a Widespread Mechanism of Redundant Enhancer Evolution in the Mammalian Genome. *Genome Biol Evol* **12**: 1–17.

Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2015. GenBank. *Nucleic Acids Res* **43**: D30–5.

Boman J, Frankl-Vilches C, da Silva Dos Santos M, de Oliveira EHC, Gahr M, Suh A. 2019. The Genome of Blue-Capped Cordon-Bleu Uncovers Hidden Diversity of LTR Retrotransposons in Zebra Finch. *Genes* **10**. http://dx.doi.org/10.3390/genes10040301.

Bradbury JW, Balsby TJS. 2016. The functions of vocal learning in parrots. *Behav Ecol Sociobiol* **70**: 293–312.

Burt DW, Bruley C, Dunn IC, Jones CT, Ramage A, Law AS, Morrice DR, Paton IR, Smith J, Windsor D, et al. 1999. The dynamics of chromosome evolution in birds and mammals. *Nature* **402**: 411–413.

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

Cloutier A, Sackton TB, Grayson P, Clamp M, Baker AJ, Edwards SV. 2019. Whole-Genome Analyses Resolve the Phylogeny of Flightless Birds (Palaeognathae) in the Presence of an Empirical Anomaly Zone. *Syst Biol* **68**: 937–955.

Coelho LA, Musher LJ, Cracraft J. 2019. A Multireference-Based Whole Genome Assembly for the Obligate Ant-Following Antbird, Rhegmatorhina melanosticta (Thamnophilidae). *Diversity* **11**: 144.

Cornetti L, Valente LM, Dunning LT. 2015. The genome of the “great speciator” provides insights into bird diversification. *Genome Biol*. https://academic.oup.com/gbe/article-abstract/7/9/2680/592400.

Damas J, Kim J, Farré M, Griffin DK, Larkin DM. 2018. Reconstruction of avian ancestral karyotypes reveals differences in the evolutionary history of macro- and microchromosomes. *Genome Biol* **19**: 155.

Ellegren H. 2010. Evolutionary stasis: the stable chromosomes of birds. *Trends Ecol Evol* **25**: 283–291.

Ericson PGP, Anderson CL, Britton T, Elzanowski A, Johansson US, Källersjö M, Ohlson JI, Parsons TJ, Zuccon D, Mayr G. 2006. Diversification of Neoaves: integration of molecular sequence data and fossils. *Biol Lett* **2**: 543–547.

Farre M, Narayan J, Slavov GT, Damas J. 2016. Novel insights into chromosome evolution in birds, archosaurs, and reptiles. *Genome Biol*. https://academic.oup.com/gbe/article-abstract/8/8/2442/2198198.

Feng S, Stiller J, Deng Y, Armstrong J, Fang Q, Reeve AH, Xie D, Chen G, Guo C, Faircloth BC, et al. 2020. Dense sampling of bird diversity increases power of comparative genomics. *Nature* **587**: 252–257.

Galbraith JD, Ludington AJ, Suh A. 2020. New Environment, New Invaders—Repeated Horizontal Transfer of LINEs to Sea Snakes. *Genome Biol*. https://academic.oup.com/gbe/article-abstract/12/12/2370/5918459.

Green RE, Braun EL, Armstrong J, Earl D, Nguyen N, Hickey G, Vandewege MW, St John JA, Capella-Gutiérrez S, Castoe TA, et al. 2014. Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. *Science* **346**: 1254449.

Gregory TR, Nicol JA, Tamm H, Kullman B, Kullman K, Leitch IJ, Murray BG, Kapraun DF, Greilhuber J, Bennett MD. 2007. Eukaryotic genome size databases. *Nucleic Acids Res* **35**: D332–8.

Haddrath O, Baker AJ. 2012. Multiple nuclear genes and retroposons support vicariance and dispersal of the palaeognaths, and an Early Cretaceous origin of modern birds. *Proc Biol Sci* **279**: 4617–4625.

Hooper DM, Price TD. 2017. Chromosomal inversion differences correlate with range overlap in passerine birds. *Nat Ecol Evol* **1**: 1526–1534.

Hughes AL, Hughes MK. 1995. Small genomes for better flyers. *Nature* **377**: 391.

International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**: 695–716.

Ivancevic AM, Kortschak RD, Bertozzi T, Adelson DL. 2016. LINEs between Species: Evolutionary Dynamics of LINE-1 Retrotransposons across the Eukaryotic Tree of Life. *Genome Biol Evol* **8**: 3301–3322.

Jaiswal SK, Gupta A, Saxena R, Prasoodanan VPK, Sharma AK, Mittal P, Roy A, Shafer ABA, Vijay N, Sharma VK. 2018. Genome Sequence of Peacock Reveals the Peculiar Case of a Glittering Bird. *Front Genet* **9**: 392.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**: 1320–1331.

Kaiser VB, van Tuinen M, Ellegren H. 2007. Insertion events of CR1 retrotransposable elements elucidate the phylogenetic branching order in galliform birds. *Mol Biol Evol* **24**: 338–347.

Kapusta A, Suh A, Feschotte C. 2017. Dynamics of genome size evolution in birds and mammals. *Proc Natl Acad Sci U S A* **114**: E1460–E1469.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772–780.

Kohany O, Gentles AJ, Hankus L, Jurka J. 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* **7**: 474.

Kretschmer R, Furo I de O, Gomes AJB, Kiazim LG, Gunski RJ, Del Valle Garnero A, Pereira JC, Ferguson-Smith MA, Corrêa de Oliveira EH, Griffin DK, et al. 2020a. A Comprehensive Cytogenetic Analysis of Several Members of the Family Columbidae (Aves, Columbiformes). *Genes* **11**. http://dx.doi.org/10.3390/genes11060632.

Kretschmer R, Gunski RJ, Garnero ADV, de Freitas TRO, Toma GA, Cioffi M de B, Oliveira EHC de, O'Connor RE, Griffin DK. 2020b. Chromosomal Analysis in Crotophaga ani (Aves, Cuculiformes) Reveals Extensive Genomic Reorganization and an Unusual Z-Autosome Robertsonian Translocation. *Cells* **10**. http://dx.doi.org/10.3390/cells10010004.

Kriegs JO, Matzke A, Churakov G, Kuritzin A, Mayr G, Brosius J, Schmitz J. 2007. Waves of genomic hitchhikers shed light on the evolution of gamebirds (Aves: Galliformes). *BMC Evol Biol* **7**: 190.

Laine VN, Gossmann TI, Schachtschneider KM, Garroway CJ, Madsen O, Verhoeven KJF, de Jager V, Megens H-J, Warren WC, Minx P, et al. 2016. Evolutionary signals of selection on cognition from the great tit genome and methylome. *Nat Commun* **7**: 10474.

Lee J, Han K, Meyer TJ, Kim H-S, Batzer MA. 2008. Chromosomal inversions between human and chimpanzee lineages caused by retrotransposons. *PLoS One* **3**: e4047.

Li J, Zhang J, Liu J, Zhou Y, Cai C, Xu L, Dai X, Feng S, Guo C, Rao J, et al. 2021. A new duck genome reveals conserved and convergently evolved chromosome architectures of birds and mammals. *Gigascience* **10**. http://dx.doi.org/10.1093/gigascience/giaa142.

Lim JK, Simmons MJ. 1994. Gross chromosome rearrangements mediated by transposable elements in Drosophila melanogaster. *Bioessays* **16**: 269–275.

Luan DD, Korman MH, Jakubczak JL, Eickbush TH. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* **72**: 595–605.

Manthey JD, Moyle RG, 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.

Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, et al. 2017. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res* **45**: D200–D203.

Matzke A, Churakov G, Berkes P, Arms EM, Kelsey D, Brosius J, Kriegs JO, Schmitz J. 2012. Retroposon insertion patterns of neoavian birds: strong evidence for an extensive incomplete lineage sorting era. *Mol Biol Evol* **29**: 1497–1501.

Mitchell KJ, Llamas B, Soubrier J, Rawlence NJ, Worthy TH, Wood J, Lee MSY, Cooper A. 2014. Ancient DNA reveals elephant birds and kiwi are sister taxa and clarifies ratite bird evolution. *Science* **344**: 898–900.

Nieder A, Wagener L, Rinnert P. 2020. A neural correlate of sensory consciousness in a corvid bird. *Science* **369**: 1626–1629.

O'Connor RE, Farré M, Joseph S, Damas J, Kiazim L, Jennings R, Bennett S, Slack EA, Allanson E, Larkin DM, et al. 2018. Chromosome-level assembly reveals extensive rearrangement in saker falcon and budgerigar, but not ostrich, genomes. *Genome Biol* **19**: 171.

Oliveros CH, Field DJ, Ksepka DT, Barker FK, Aleixo A, Andersen MJ, Alström P, Benz BW, Braun EL, Braun MJ, et al. 2019. Earth history and the passerine superradiation. *Proc Natl Acad Sci U S A* **116**: 7916–7925.

Organ CL, Shedlock AM, Meade A, Pagel M, Edwards SV. 2007. Origin of avian genome size and structure in non-avian dinosaurs. *Nature* **446**: 180–184.

Ottenburghs J, Ydenberg RC, Van Hooft P, Van Wieren SE, Prins HHT. 2015. The Avian Hybrids Project: gathering the scientific literature on avian hybridization. *Ibis* **157**: 892–894.

Paton TA, Baker AJ, Groth JG, Barrowclough GF. 2003. RAG-1 sequences resolve phylogenetic relationships within Charadriiform birds. *Mol Phylogenet Evol* **29**: 268–278.

Peona V, Blom MPK, Xu L, Burri R, Sullivan S, Bunikis I, Liachko I, Haryoko T, Jønsson KA, Zhou

Q, et al. 2021. Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. *Mol Ecol Resour* **21**: 263–286.

Petkov CI, Jarvis ED. 2012. Birds, primates, and spoken language origins: behavioral phenotypes and neurobiological substrates. *Front Evol Neurosci* **4**: 12.

Pfenning AR, Hara E, Whitney O, Rivas MV, Wang R, Roulhac PL, Howard JT, Wirthlin M, Lovell PV, Ganapathy G, et al. 2014. Convergent transcriptional specializations in the brains of humans and song-learning birds. *Science* **346**: 1256846.

Pinheiro MLS, Nagamachi CY, Ribas TFA, Diniz CG, O´Brien PCM, Ferguson-Smith MA, Yang F, Pieczarka JC. 2021. Chromosomal painting of the sandpiper (Actitis macularius) detects several fissions for the Scolopacidae family (Charadriiformes). *BMC Ecology and Evolution* **21**: 8.

Platt RN 2nd, Blanco-Berdugo L, Ray DA. 2016. Accurate Transposable Element Annotation Is Vital When Analyzing New Genome Assemblies. *Genome Biol Evol* **8**: 403–410.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**: e9490.

Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Gedman GL, et al. 2020. Towards complete and error-free genome assemblies of all vertebrate species. *bioRxiv* 2020.05.22.110833. https://www.biorxiv.org/content/10.1101/2020.05.22.110833v1.full-text (Accessed March 31, 2021).

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.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**: e2584.

Salter JF, Oliveros CH, Hosner PA, Manthey JD. 2020. Extensive paraphyly in the typical owl family (Strigidae). *Auk*. https://academic.oup.com/auk/article-abstract/137/1/ukz070/5673551.

Shetty S, Griffin DK, Graves JA. 1999. Comparative painting reveals strong chromosome homology over 80 million years of bird evolution. *Chromosome Res* **7**: 289–295.

Silva T, Guzmán A, Urantówka AD, Mackiewicz P. 2017. A new parrot taxon from the Yucatán Peninsula, Mexico-its position within genus Amazona based on morphology and molecular phylogeny. *PeerJ* **5**: e3475.

Skinner BM, Griffin DK. 2012. Intrachromosomal rearrangements in avian genome evolution: evidence for regions prone to breakpoints. *Heredity* **108**: 37–41.

St John J, Cotter J-P, Quinn TW. 2005. A recent chicken repeat 1 retrotransposition confirms the Coscoroba-Cape Barren goose clade. *Mol Phylogenet Evol* **37**: 83–90.

Suh A. 2016. The phylogenomic forest of bird trees contains a hard polytomy at the root of Neoaves. *Zool Scr* **45**: 50–62.

Suh A. 2015. The Specific Requirements for CR1 Retrotransposition Explain the Scarcity of Retrogenes in Birds. *J Mol Evol* **81**: 18–20.

Suh A, Bachg S, Donnellan S, Joseph L, Brosius J, Kriegs JO, Schmitz J. 2017. De-novo emergence of SINE retroposons during the early evolution of passerine birds. *Mob DNA* **8**: 21.

Suh A, Churakov G, Ramakodi MP, Platt RN 2nd, Jurka J, Kojima KK, Caballero J, Smit AF, Vliet

KA, Hoffmann FG, et al. 2014. Multiple lineages of ancient CR1 retroposons shaped the early genome evolution of amniotes. *Genome Biol Evol* **7**: 205–217.

Suh A, Kriegs JO, Donnellan S, Brosius J, Schmitz J. 2012. A universal method for the study of CR1 retroposons in nonmodel bird genomes. *Mol Biol Evol* **29**: 2899–2903.

Suh A, Paus M, Kiefmann M, Churakov G, Franke FA, Brosius J, Kriegs JO, Schmitz J. 2011. Mesozoic retroposons reveal parrots as the closest living relatives of passerine birds. *Nat Commun* **2**: 443.

Suh A, Smeds L, Ellegren H. 2018. Abundant recent activity of retrovirus-like retrotransposons within and among flycatcher species implies a rich source of structural variation in songbird genomes. *Mol Ecol* **27**: 99–111.

Suh A, Smeds L, Ellegren H. 2015. The Dynamics of Incomplete Lineage Sorting across the Ancient Adaptive Radiation of Neoavian Birds. *PLoS Biol* **13**: e1002224.

Sun Z, Pan T, Hu C, Sun L, Ding H, Wang H, Zhang C, Jin H, Chang Q, Kan X, et al. 2017. Rapid and recent diversification patterns in Anseriformes birds: Inferred from molecular phylogeny and diversification analyses. *PLoS One* **12**: e0184529.

Szitenberg A, Cha S, Opperman CH, Bird DM, Blaxter ML, Lunt DH. 2016. Genetic Drift, Not Life History or RNAi, Determine Long-Term Evolution of Transposable Elements. *Genome Biol Evol* **8**: 2964–2978.

Treangen TJ, Salzberg SL. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* **13**: 36–46.

Treplin S, Tiedemann R. 2007. Specific chicken repeat 1 (CR1) retrotransposon insertion suggests phylogenetic affinity of rockfowls (genus Picathartes) to crows and ravens (Corvidae). *Mol Phylogenet Evol* **43**: 328–337.

Underwood CJ, Choi K. 2019. Heterogeneous transposable elements as silencers, enhancers and targets of meiotic recombination. *Chromosoma* **128**: 279–296.

Vandergon TL, Reitman M. 1994. Evolution of chicken repeat 1 (CR1) elements: evidence for ancient subfamilies and multiple progenitors. *Mol Biol Evol* **11**: 886–898.

Wang D, Qu Z, Yang L, Zhang Q, Liu Z-H, Do T, Adelson DL, Wang Z-Y, Searle I, Zhu J-K. 2017. Transposable elements (TEs) contribute to stress-related long intergenic noncoding RNAs in plants. *Plant J* **90**: 133–146.

Warren IA, Naville M, Chalopin D, Levin P, Berger CS, Galiana D, Volff J-N. 2015. Evolutionary impact of transposable elements on genomic diversity and lineage-specific innovation in vertebrates. *Chromosome Res* **23**: 505–531.

Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Künstner A, Searle S, White S, Vilella AJ, Fairley S, et al. 2010. The genome of a songbird. *Nature* **464**: 757–762.

Watanabe M, Nikaido M, Tsuda TT, Inoko H, Mindell DP, Murata K, Okada N. 2006. The rise and fall of the CR1 subfamily in the lineage leading to penguins. *Gene* **365**: 57–66.

Weir JT, Haddrath O, Robertson HA, Colbourne RM, Baker AJ. 2016. Explosive ice age diversification of kiwi. *Proc Natl Acad Sci U S A* **113**: E5580–7.

Weissensteiner MH, Bunikis I, Catalán A, Francoijs K-J, Knief U, Heim W, Peona V, Pophaly SD, Sedlazeck FJ, Suh A, et al. 2020. Discovery and population genomics of structural variation in a songbird genus. *Nat Commun* **11**: 3403.

Wicker T, Robertson JS, Schulze SR, Feltus FA, Magrini V, Morrison JA, Mardis ER, Wilson RK, Peterson DG, Paterson AH, et al. 2005. The repetitive landscape of the chicken genome. *Genome Res* **15**: 126–136.

Wiens JJ. 2015. Explaining large-scale patterns of vertebrate diversity. *Biol Lett* **11**. http://dx.doi.org/10.1098/rsbl.2015.0506.

Wright NA, Ryan Gregory T, Witt CC. 2014. Metabolic "engines" of flight drive genome size reduction in birds. *Proceedings of the Royal Society B: Biological Sciences* **281**: 20132780. http://dx.doi.org/10.1098/rspb.2013.2780.

Zeng L, Kortschak RD, Raison JM, Bertozzi T, Adelson DL. 2018. Superior ab initio identification, annotation and characterisation of TEs and segmental duplications from genome assemblies. *PLoS One* **13**: e0193588.

Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, Storz JF, Antunes A, Greenwold MJ, Meredith RW, et al. 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**: 1311–1320.

Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. *J Comput Biol* **7**: 203–214.

Zhou Y, Mishra B. 2005. Quantifying the mechanisms for segmental duplications in mammalian genomes by statistical analysis and modeling. *Proc Natl Acad Sci U S A* **102**: 4051–4056.

SI Information

Figures

SI Figure 1. Phylogenetic tree of newly identified full length CR1s and full length avian CR1s from Repbase. The full length CR1s used are the centroids of order specific clusters constructed using VSEARCH at 90% identity. Phylogeny constructed using FastTree from a MAFFT alignment of the nucleotide sequences.

SI Figure 2. Scaled divergence of 3' anchored CR1s identified in a) selected passerines and b) selected paleognaths. CR1s were initially identified using a reciprocal BLAST search based on libraries consisting of RepBase avian and crocodilian repeats and the centroids of full length sequences identified within the order clustered in VSEARCH.

SI Figure 3. Number of high confidence insertions of dominant CR1 families in owls approximated by presence/absence patterns of orthologous CR1 insertions between 100 and 600 bp in length. CR1 subfamilies are labeled by colour (see legend). Phylogeny adapted from (Salter et al. 2020).

SI Figure 4. Scaled divergence of 3' anchored CR1s identified in species of Amazon parrot (*Amazona*). CR1s were initially identified using a reciprocal BLAST search based on a consisting of RepBase avian and crocodilian repeats and the centroids of full length sequences identified within parrots clustered in VSEARCH.

SI Figure 5. Presence/absence workflow. 3' anchored CR1 insertions in a genome between 100 and 600 bp (1) were identified with BLASTN and had coordinates extended to include 600 bp of flanking sequence at both the 5' and 3' ends (2). The resulting 1300-1800 bp long sequences were searched for in a related genome using BLASTN. Hits containing the entire insertion and at least 150 bp of each flank were treated as ancestral insertions (3). Hits to insertion not containing any flanking region, with hits to the flanking sequence on differing strands or multiple hits to a single flanking sequence far from each other were treated as unresolvable and discarded. Insertions having at least 150 bp of each flank in close proximity and one flank containing at least 90 bp of

the insertion were treated as ancestral insertions of which part was deleted in the species being searched (4). Sequences remaining were either flanks in close proximity or flanks plus a portion of the CR1 insertion. The distance between the flanks potentially containing part of the insertion was calculated in both species, qdist in the query species and sdist in the related species (5). If qdist was greater or equal to the length of the original CR1 insertion (olen) minus the length of 3x the 3' microsatellite monomer and sdist was within the length of 2x the 3' microsatellite monomer the insertion was treated as since divergence (6). If qdist was within the length of 2x the 3' microsatellite monomer and the sdist was greater than 90 bp the insertion was treated as ancestral (7). Any insertions not fitting these criteria were treated as unresolvable and discarded. This strict process was calibrated through adjusting variables and viewing resulting pairwise alignments between regions identified as orthologous, using the presence of target site duplications in the query species and if part of the CR1 insertion was present in the related species to determine if insertions had truly occurred in an orthologous region, erring on the side of discarding new insertions over misclassifying partially deleted ancestral insertions as new insertions.

SI Figure 6. Presence/absence resolution - Example of the method we used to resolve the presence/absence, and hence insertion timing, of each CR1 in a species (species a), two related species (species b and c) and an outgroup (species d). The CR1 insertion in question is represented in green, the flanking regions in black and the branches labelled 1-3. The branch in bold italics is the branch on which the insertion occurred. If an CR1 was present in species a through c we considered the repeat to have been inserted at branch 1 (i), if in a and b at branch 2 (ii) and if in species a alone to be since the divergence from the immediate sister species and on branch 3 (iii). If present in all three species and the outgroup species examined we consider the repeat to be ancestral (iv). If a CR1 was absent from an immediate sister species but present in the more distant related species we considered this to be a result of deletion in the immediate sister species (v). Finally, if the orthologous region was present in a species or group of species but could not be resolved in the immediate sister species we considered the timing of insertion to be unresolvable (vi).

Tables

SI Table 1. Genome assemblies used throughout this analysis. All genomes were downloaded from GenBank.

SI Table 2 - Reclassification of previously described full length avian CR1s based on their position within our CR1 phylogeny (SI Figure 1; same color coding).

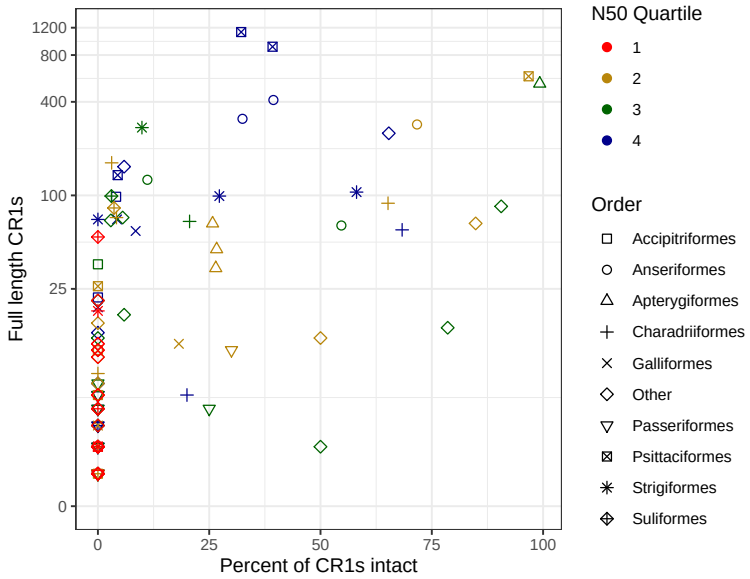
SI Table 3. Resolution of presence or absence of orthologous CR1 insertions between 100 and 600 bp in related species in waterfowl, shorebirds, perching birds, parrots, owls, and kiwis + cassowary + emu genomes. Cells highlighted in yellow are the values used to construct Figures 4 and 5 and SI Figure 3.

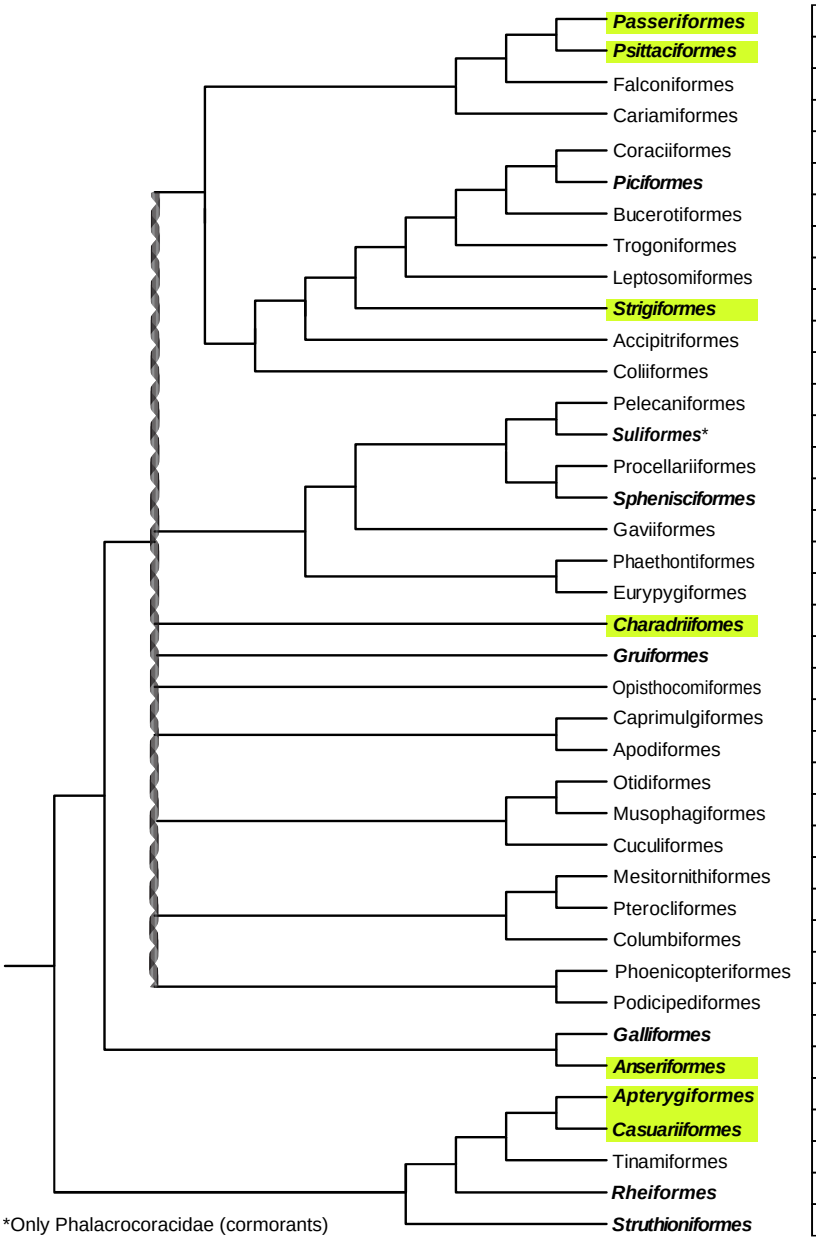
Data

SI Data 1 - Coordinates of full length CR1s identified in each genome in BED format. For the appropriate genome version see SI Table 1.

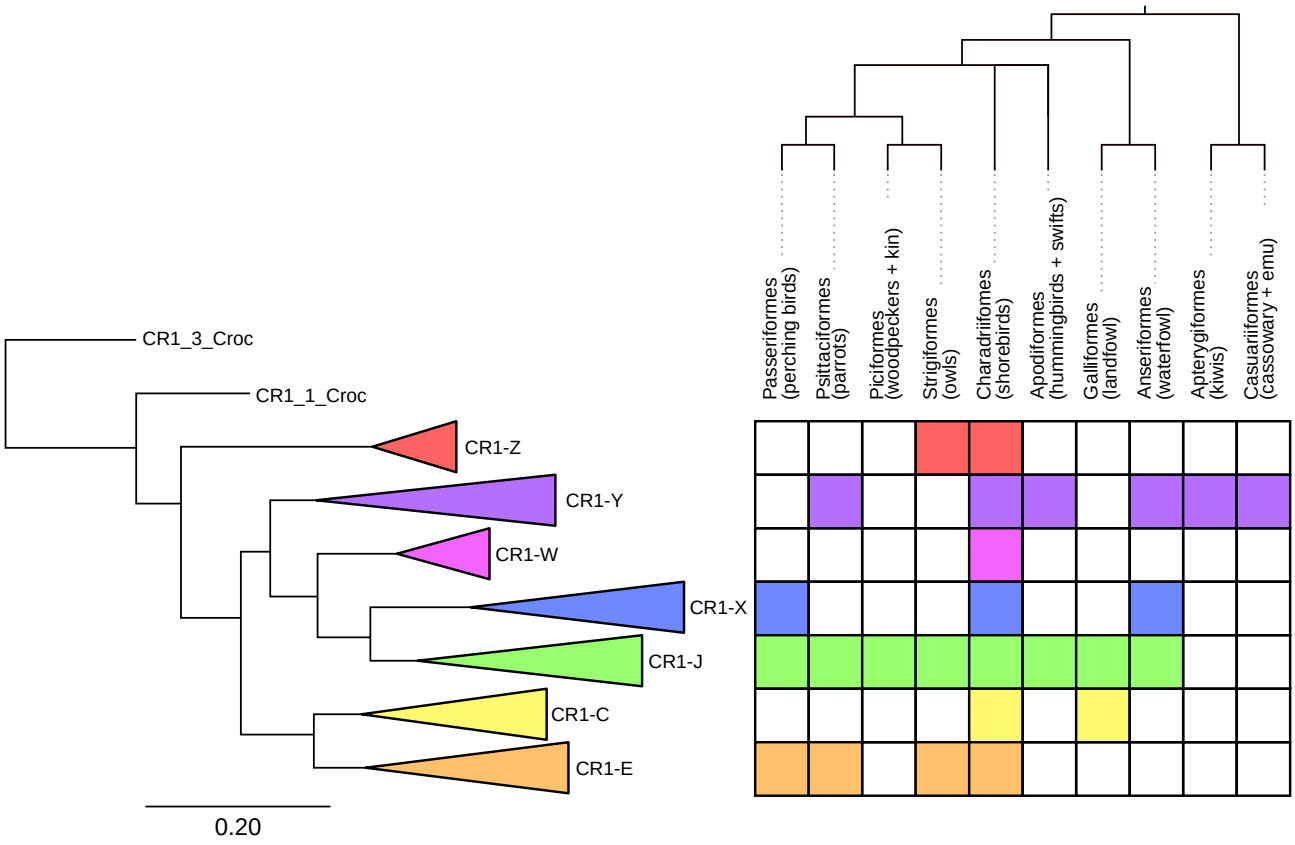
SI Data 2 - Multiple sequence alignment used to create the CR1 phylogeny (SI Figure 1) and Newick tree of said phylogeny.

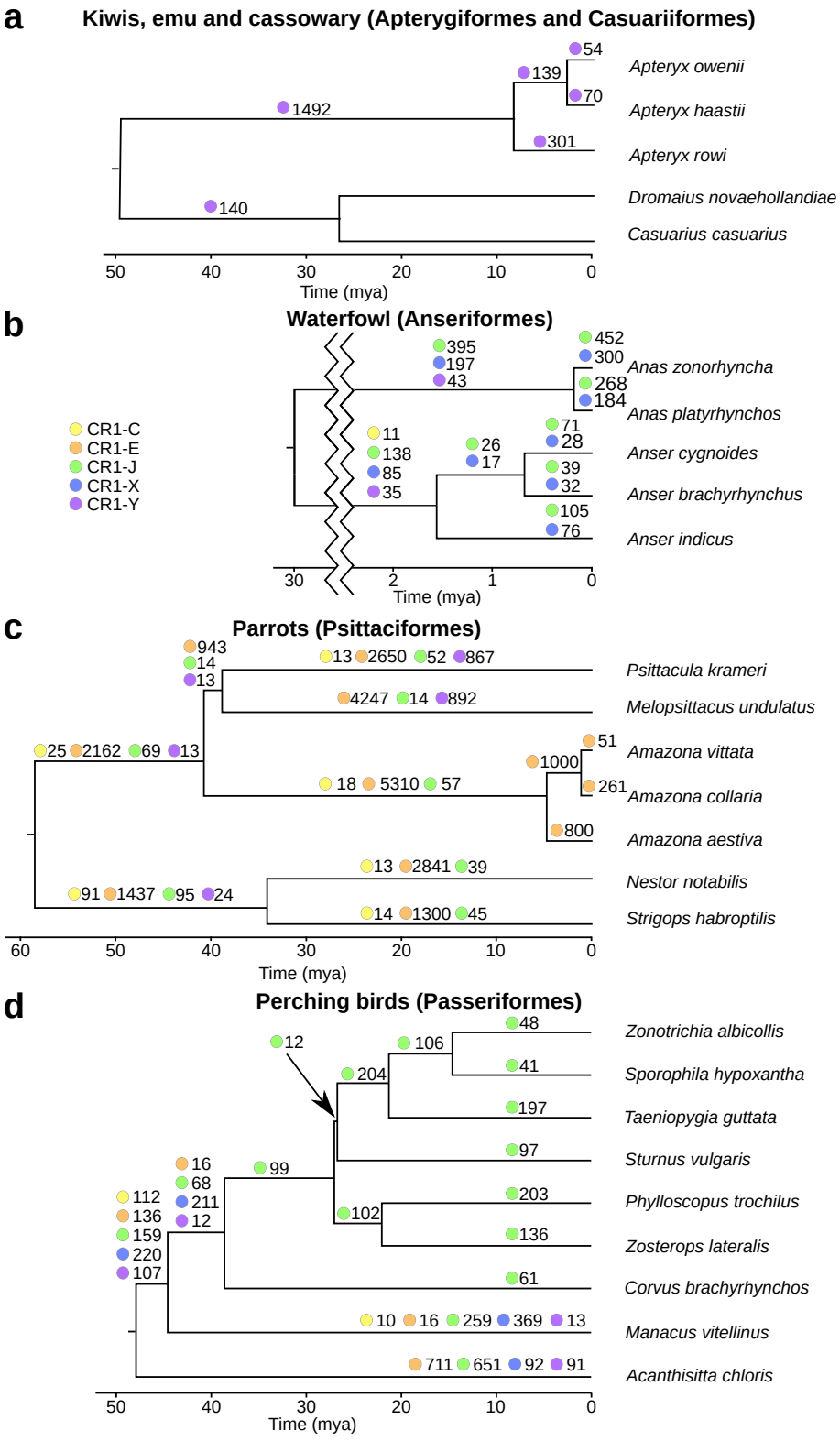
SI Data 3. Divergence plots of 3' anchored CR1s identified in each species of bird belonging to orders in which we detected full length CR1s. CR1s were identified using a reciprocal BLAST search based on libraries consisting of Repbase avian and crocodilian repeats and the centroids of full length sequences identified within the order clustered in VSEARCH. Jukes-Cantor distance was calculated from the reciprocal BLAST search output.

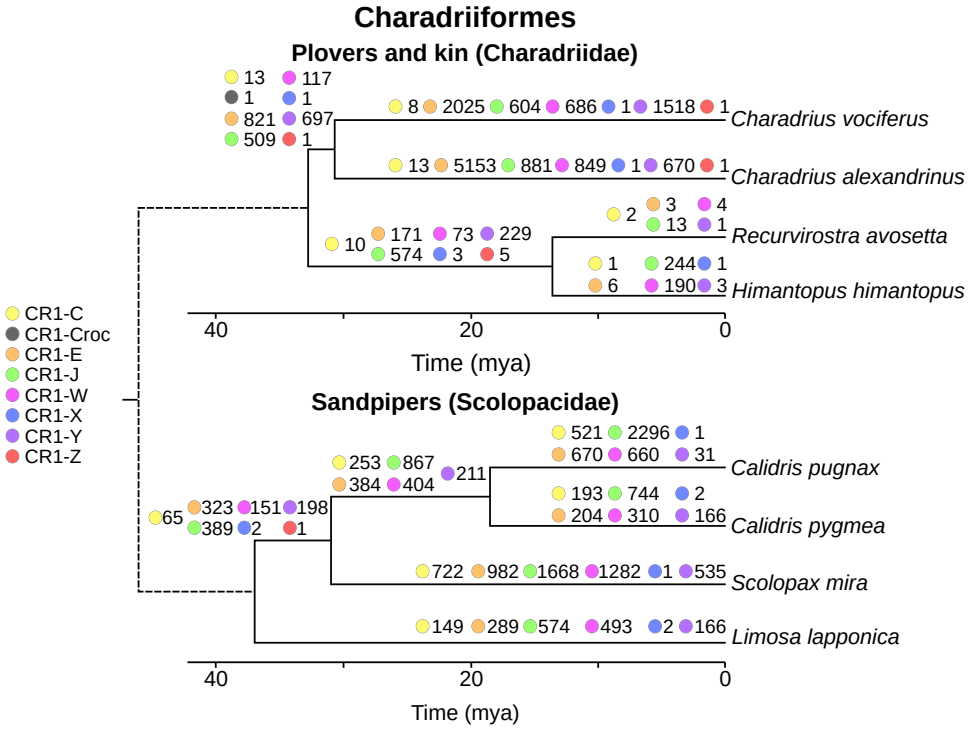

Full length CR1s
1200
800
400
100
25
0
0
25
50
75
100
Percent of CR1s intact
N50 Quartile
1
2
3
4
Order
Accipitriformes
Anseriformes
Apterygiformes
Charadriiformes
Galliformes
Other
Passeriformes
Psittaciformes
Strigiformes
Suliformes



	Min	Max	Mean	Species
Passeriformes	0	10	2.7	23
Psittaciformes	0	1125	309.4	9
Falconiformes	3	6	4.3	3
Cariamiformes	10	10	10	1
Coraciiformes	1	1	1	1
Piciformes	5	12	8.5	1
Bucerotiformes	0	0	0	1
Trogoniformes	9	9	9	1
Leptosomiformes	11	11	11	1
Strigiformes	2	273	94.5	6
Accipitriformes	0	98	23.4	7
Coliiformes	1	1	1	1
Pelecaniformes	9	21	14	3
Suliformes*	4	99	60	4
Procellariiformes	2	2	2	1
Sphenisciformes	1	2	1.5	2
Gaviiformes	2	2	2	1
Phaethontiformes	2	2	2	1
Eurypygiformes	2	2	2	1
Charadriiformes	1	162	46.9	10
Gruiformes	2	153	77.5	2
Opisthocomiformes	66	66	66	1
Caprimulgiformes	1	1	1	1
Apodiformes	85	251	168	2
Otidiformes	2	2	2	1
Musophagiformes	10	10	10	1
Cuculiformes	6	6	6	1
Mesitornithiformes	3	3	3	1
Pterocliformes	1	1	1	1
Columbiformes	13	13	13	1
Phoenicopteriformes	4	4	4	1
Podicipediformes	5	5	5	1
Galliformes	0	70	23.4	8
Anseriformes	64	411	239.6	5
Apterygiformes	34	524	167.3	4
Casuariiformes	15	17	16	2
Tinamiformes	2	21	9.3	3
Rheiformes	69	72	70.5	2
Struthioniformes	14	14	14	1


CR1_3_Croc
CR1_1_Croc
CR1-Z
CR1-Y
CR1-W
CR1-X
CR1-J
CR1-C
CR1-E
0.20
Passeriformes (perching birds)
Psittaciformes (parrots)
Piciformes (woodpeckers + kin)
Strigiformes (owls)
Charadriiformes (shorebirds)
Apodiformes (hummingbirds + swifts)
Galliformes (landfowl)
Anseriformes (waterfowl)
Apterygiformes (kiwis)
Casuariiformes (cassowary + emu)


a
Kiwis, emu and cassowary (Apterygiformes and Casuariiformes)
54
139
70
1492
301
140
Apteryx owenii
Apteryx haastii
Apteryx rowi
Dromaius novaehollandiae
Casuarius casuarius
50 40 30 20 10 0
Time (mya)
b
Waterfowl (Anseriformes)
CR1-C
CR1-E
CR1-J
CR1-X
CR1-Y
395
197
43
452
300
268
184
11
138
85
35
26
17
71
28
39
32
105
76
Anas zonorhyncha
Anas platyrhynchos
Anser cygnoides
Anser brachyrhynchus
Anser indicus
30 2 1 0
Time (mya)
c
Parrots (Psittaciformes)
943
14
13
13 2650 52 867
4247 14 892
25 2162 69 13
18 5310 57
51
1000
261
800
13 2841 39
91 1437 95 24
14 1300 45
Psittacula krameri
Melopsittacus undulatus
Amazona vittata
Amazona collaria
Amazona aestiva
Nestor notabilis
Strigops habroptilis
60 50 40 30 20 10 0
Time (mya)
d
Perching birds (Passeriformes)
12
106
48
204
41
197
97
16
68
211
12
99
112
136
159
220
107
203
102
136
61
10 16 259 369 13
711 651 92 91
Zonotrichia albicollis
Sporophila hypoxantha
Taeniopygia guttata
Sturnus vulgaris
Phylloscopus trochilus
Zosterops lateralis
Corvus brachyrhynchos
Manacus vitellinus
Acanthisitta chloris
50 40 30 20 10 0
Time (mya)


Charadriiformes
Plovers and kin (Charadriidae)
13 117
1 1
821 697
509 1
8 2025 604 686 1 1518 1
Charadrius vociferus
13 5153 881 849 1 670 1
Charadrius alexandrinus
2 3 4
13 1
10 171 73 229
574 3 5
Recurvirostra avosetta
1 244 1
6 190 3
Himantopus himantopus
CR1-C
CR1-Croc
CR1-E
CR1-J
CR1-W
CR1-X
CR1-Y
CR1-Z
40
20
0
Time (mya)
Sandpipers (Scolopacidae)
521 2296 1
670 660 31
253 867
384 404
211
Calidris pugnax
193 744 2
204 310 166
Calidris pygmea
65 323 151 198
389 2 1
722 982 1668 1282 1 535
Scolopax mira
149 289 574 493 2 166
Limosa lapponica
40
20
0
Time (mya)

New Environment, New Invaders — Repeated Horizontal Transfer of LINEs to Sea Snakes

"The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka' but 'That's funny." – Isaac Asimov.

During my investigation into CR1s in birds, due to an interest in both TEs and reptiles, I was approached to annotate the mobilome of the olive sea snake (*Aipysurus laevis*) genome. Sea snakes belong to the family Hydrophiinae, a highly diverse group of elapids also encompassing sea kraits and terrestrial snakes native to Australia and New Guinea. Since their split from terrestrial Hydrophiinae 13-15 Mya sea snakes have adapted rapidly to become fully marine, developing paddle shaped and photosensitive tails. Like in birds, little research has examined the evolution of TEs in squamates outside of the model species, *Anolis carolinensis*. Additionally, little research has investigated the role TEs may play in species' adaptation to novel environments. As a member of a diverse clade of snakes which had recently adapted to a novel environment, *Aipysurus laevis* provided an ideal model to investigate the evolution of the mobilome following massive habitat change.

All supplementary data for this chapter can be found at
github.com/jamesdgalbraith/thesis_supplementary_material/tree/main/Chapter_2

Statement of Authorship

Title of Paper	New Environment, New Invaders—Repeated Horizontal Transfer of LINEs to Sea Snakes
Publication Status	☒ Published ☐Accepted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style ☐ Submitted for Publication
Publication Details	James D. Galbraith, Alastair J. Ludington, Alexander Suh, Kate L. Sanders, David L. Adelson (2021), New Environment, New Invaders—Repeated Horizontal Transfer of LINEs to Sea Snakes, *Genome Biology and Evolution*, Volume 12, Issue 12, December 2020, Pages 2370–2383.

Principal Author

Name of Principal Author (Candidate)	James Galbraith		
Contribution to the Paper	Designed and performed analysis, interpreted results and wrote manuscript		
Overall percentage (%)	85%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	24/06/2021

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Alastair Ludington		
Contribution to the Paper	Assisted in analysing the results and writing the manuscript		
Signature		Date	22/06/2021

Name of Co-Author	Kate Sanders		
Contribution to the Paper	Assisted in analysing the results and writing the manuscript		
Signature		Date	24/06/2021

Please cut and paste additional co-author panels here as required.

Statement of Authorship

Name of Co-Author	Alexander Suh		
Contribution to the Paper	Supervised the development of the work and assisted in analysing the results and writing the manuscript		
Signature		Date	2021-06-21

Name of Co-Author	David Adelson		
Contribution to the Paper	Supervised the development of the work and assisted in analysing the results and writing the manuscript		
Signature		Date	2021-06-21

GBE

New Environment, New Invaders—Repeated Horizontal Transfer of LINEs to Sea Snakes

James D. Galbraith[1], Alastair J. Ludington[1], Alexander Suh[2,3,4], Kate L. Sanders[1], and David L. Adelson [1,*]

[1]School of Biological Sciences, University of Adelaide, Australia

[2]Department of Ecology and Genetics—Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Sweden

[3]Department of Organismal Biology—Systematic Biology, Evolutionary Biology Centre, Uppsala University, Sweden

[4]School of Biological Sciences, University of East Anglia, Norwich, United Kingdom

*Corresponding author: E-mail: david.adelson@adelaide.edu.au.

Accepted: 29 September 2020

Abstract

Although numerous studies have found horizontal transposon transfer (HTT) to be widespread across metazoans, few have focused on HTT in marine ecosystems. To investigate potential recent HTTs into marine species, we searched for novel repetitive elements in sea snakes, a group of elapids which transitioned to a marine habitat at most 18 Ma. Our analysis uncovered repeated HTTs into sea snakes following their marine transition. The seven subfamilies of horizontally transferred LINE retrotransposons we identified in the olive sea snake (*Aipysurus laevis*) are transcribed, and hence are likely still active and expanding across the genome. A search of 600 metazoan genomes found all seven were absent from other amniotes, including terrestrial elapids, with the most similar LINEs present in fish and marine invertebrates. The one exception was a similar LINE found in sea kraits, a lineage of amphibious elapids which independently transitioned to a marine environment 25 Ma. Our finding of repeated horizontal transfer events into marine snakes greatly expands past findings that the marine environment promotes the transfer of transposons. Transposons are drivers of evolution as sources of genomic sequence and hence genomic novelty. We identified 13 candidate genes for HTT-induced adaptive change based on internal or neighboring HTT LINE insertions. One of these, ADCY4, is of particular interest as a part of the KEGG adaptation pathway "Circadian Entrainment." This provides evidence of the ecological interactions between species influencing evolution of metazoans not only through specific selection pressures, but also by contributing novel genomic material.

Key words: horizontal transfer, transposable element, Serpentes.

Significance

Recent research has found horizontal transfer (HT) of transposons between marine animals. We analyzed the olive sea snake (*Aipysurus laevis*) genome, uncovering HT of six novel retrotransposons into sea snakes since their marine transition within the last 18 Ma. All six are absent from terrestrial animals and are most similar to retrotransposons found in fish, corals, and the independently marine sea kraits. All six retrotransposons are likely still active and expanding across the genome in *A. laevis*. Our findings suggest the marine environment is ideal for the HT of transposons; and provide evidence that changing environments can influence evolution not only through novel selective pressures, but also by contributing novel genomic material.

Introduction

Transposons are a major component of metazoan genomes, making up between 24% and 56% of squamate genomes (Pasquesi et al. 2018). Transposons are split into two classes: Class I containing LINEs (long interspersed elements) and LTR (long terminal repeat) retrotransposons; and Class II

containing DNA transposons (Wicker et al. 2007). Although all three groups of transposons are present in squamates, recent activity is dominated by LINEs including CR1s, RTE-BovBs, Rex1, and L2s (Pasquesi et al. 2018). Although transposons are normally vertically transmitted (parent to offspring) there have been many instances of horizontal transposon transfer (HTT) observed between distantly related species. HTT of DNA transposons and LTR retrotransposons appears to be more common, yet many examples of HTT of non-LTR retrotransposons (LINEs) have been described (Peccoud et al. 2018). These include transfers of RTE-BovBs between multiple distant lineages (Ivancevic et al. 2018), of AviRTEs between birds and parasitic nematodes (Suh et al. 2016), and of Rex1 elements between teleost fish (Volff et al. 2000; Zhang et al. 2020). As transposons proliferate throughout a genome they can contribute novel coding sequences, alter gene regulatory networks, modify coding regions, and lead to gene copy number variation (Rebollo et al. 2012; Chuong et al. 2017; Cerbin and Jiang 2018; Schrader and Schmitz 2019). Within a lifetime most insertions will be neutral and some may be deleterious; however, on an evolutionary time scale, some TE insertions constitute a key source of genomic innovation as organisms adapt to new and changing environments (Casacuberta and González 2013; Salces-Ortiz et al. 2020). Previous studies in *Drosophila* found HTT to increase following colonization of new habitats due to exposure to new species (Biémont et al. 1999; Vieira et al. 2002).

Hydrophiinae (Elapidae) is a prolific radiation of more than 100 terrestrial snakes plus ~70 aquatic species. The aquatic species form two separate lineages which independently transitioned to a marine habitat: the fully marine sea snakes and the amphibious sea kraits (*Laticauda*) (Lee et al. 2016). Sea snakes are phylogenetically nested inside the terrestrial hydrophiine radiation and appeared ~6–18 Ma, whereas sea kraits form the sister lineage to all other Hydrophiinae and diverged 25 Ma (Sanders et al. 2008; Lee et al. 2016). Sea snakes include >60 species in two major clades, *Hydrophis* and *Aipysurus-Emydocephalus*, which shared a semi-aquatic common ancestor ~6–18 Ma and exhibit highly contrasting evolutionary histories since their transitions from terrestrial to marine habits (Sanders et al. 2013; Lee et al. 2016; Nitschke et al. 2018). Both of these lineages have independently developed adaptations to the aquatic environment including a lingual notch allowing for full closure of the mouth when underwater, and tail paddles for efficient underwater movement (Lillywhite 2014). However, the *Aipysurus-Emydocephalus* lineage has continued to evolve at the same rate as terrestrial lineages of Hydrophiinae, diverging into nine species, whereas the *Hydrophis* lineage has rapidly radiated into 48 species (Sanders et al. 2010).

Following major ecological transitions, such as sea snakes' transition from a terrestrial to a marine habitat, organisms must adapt to their new environment, with transposons potentially being a key genomic source for genomic adaptations (Schlötterer et al. 2015; Marques et al. 2018). Peng et al. (2020) found expansions of LTR retrotransposons in Shaw's sea snake (*Hydrophis curtus*) to be linked to its adaptation to the marine environment. This was based on overrepresentation of GO terms of genes near inserted LTR retrotransposons and found potential links to locomotory behavior, eye pigmentation, cellular hypotonic response, positive regulation of wound healing, and olfactory bulb interneuron development. Here we analyzed transposons in three sea snake genomes and one sea krait genome, where the marine environment appears to have fostered the repeated, independent acquisition of these transposons through HTT. The repeated HTT suggests that direct effects of the environment on genome structure may be an important but overlooked driver of evolutionary change during major ecological transitions.

Results

Annotation of Sea Snake Transposons

We performed ab initio repeat annotation of the olive sea snake (*Aipysurus laevis*) genome (Ludington et al., dx.doi.org/10.5281/zenodo.3975254) using CARP (Zeng et al. 2018) and RepeatModeler (Smit and Hubley 2017) to characterize repetitive content. Most repetitive sequences identified by both CARP and RepeatModeler were not well classified because both software tools rely on homology to reference sequences from Repbase (Bao et al. 2015), a database of repeats from highly studied species that are evolutionarily distant to Hydrophiinae. The reliance on sequence homology alone for genome-wide repeat annotation of newly sequenced species often results in the incorrect and misannotation of repeats (Platt et al. 2016). We used a structural homology approach based on the presence of a variety of protein domains in these poorly annotated repeats to identify subfamilies of LINEs, Penelope and LTR retrotransposons, endogenous retroviruses, and DNA transposons. Consensus sequences containing the characteristic protein domains and, if appropriate, TIRs or LTRs were considered as full length and confidently assigned to the lowest Transposable Element (TE) taxonomy level possible. For example, sequences identified as containing 90% of a reverse transcriptase domain and 90% of an endonuclease domain were classified as LINEs.

To identify potential HTT events which may have occurred since the transition of elapids to a marine habitat, we looked for transposons identified in *A. laevis* that were not present in genome assemblies of its closest sequenced terrestrial relatives, *Notechis scutatus* (tiger snake) and *Pseudonaja textilis* (eastern brown snake). All the TE subfamilies characterized in the *A. laevis* genome were found to be present in *P. textilis* and *N. scutatus* with the exception of five LINE subfamilies discussed below (see fig. 1). These subfamilies were further classified based on CENSOR (Kohany et al. 2006) searches against Repbase (Bao et al. 2015) using the online interface.

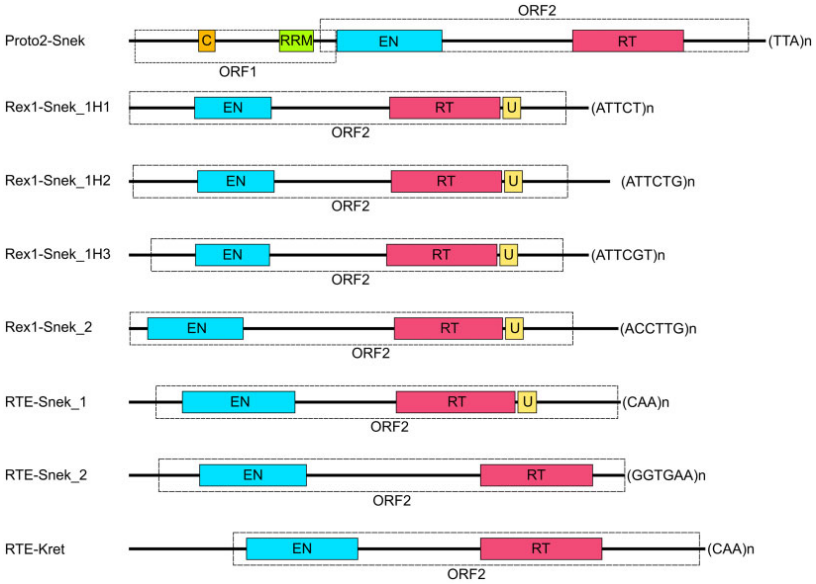


Fig. 1.—Structure of the seven HTT *Aipysurus* and one *Laticauda* LINE subfamilies. Cyan represents endonuclease (EN), red reverse transcriptase (RT), orange coiled coil (CC), green RNA-recognition motif (RRM), and yellow domain of unknown function 1891 (U). Protein domains were identified using RPSBLAST (Marchler-Bauer et al. 2017) and HHpred (Zimmermann et al. 2018) searches against CDD and Pfam (Finn et al. 2016; Marchler-Bauer et al. 2017) databases and the coiled-coil domain was identified using PCOILS (Gruber et al. 2006).

Consensus sequences containing the characteristic protein domains were confidently assigned to the lowest TE taxonomy level possible.

In *A. laevis* two of the five LINEs subfamilies, Rex1-Snek_1 (five full-length copies found) and Rex1-Snek_2 (three full-length copies found) belong to the CR1/Jockey superfamily but share less than 100-bp nucleotide sequence homology. Manual curation (see Methods) of a multiple sequence alignment of the five full-length copies identified by CARP revealed Rex1-Snek_1 to be three subfamilies; henceforth named Rex1-Snek_1H1, Rex1-Snek_1H2 and Rex1-Snek_1H3. Rex1-Snek_1H2 and Rex1-Snek_1H3 have 90% and 89% pairwise identity with Rex1-Snek_1H1, respectively. The other three subfamilies, RTE-Snek_1 (three full-length sequences found), RTE-Snek_2 (one full-length sequence found), and Proto2-Snek (one full-length sequence found) belong to the RTE superfamily but have no significant nucleotide sequence homology based on BLASTN searches using default parameters. In addition to the full-length sequences, we identified hundreds of highly similar copies with 5′ truncation patterns characteristic of recently active LINEs (fig. 2, supplementary tables 1 and 2, Supplementary material online). Specifically, coverage plots of the RTE-Snek_1, RTE-Snek_2, and Proto2-Snek families are typical of LINEs, with a clear pattern of 5′-truncated insertions (Luan et al. 1993). All seven LINE subfamilies were most similar to Repbase TE reference sequences from a marine annelid worm, a marine crustacean, and teleost fishes (Bao et al. 2015) (see table 1, supplementary dataset 1, Supplementary material online).

The absence of these recently active LINE subfamilies from terrestrial snakes that shared a common ancestor with sea snakes within the last approximately 18 Ma, combined with the finding that they were most similar to LINEs from distantly related aquatic organisms, suggested HTT as the most plausible explanation. There are three diagnostic features of HTT: 1) the sporadic presence of a TE family within a set of closely related species, 2) a higher than expected degree of sequence identity in long diverged species, and 3) discordant topologies for the phylogenies of transposons and their host species (Silva et al. 2004).

Presence/Absence in Closely Related Species

As mentioned above, the seven LINE subfamilies were absent from the closest terrestrial relatives of *A. laevis*. To test if the subfamilies have a sporadic distribution in closer relatives, we performed reciprocal BLASTN searches for their presence in two closely related sea snake genome assemblies, *Hydrophis melanocephalus* (black-headed sea snake) and *Emydocephalus ijimae* (Ijima's turtle-headed sea snake); the two closest (available) terrestrial species, *N. scutatus* and *P. textilis*; an independently aquatic species, *Laticauda colubrina* (yellow-lipped sea krait); and a distant terrestrial relative, *Ophiophagus hannah* (king cobra). The reciprocal search for

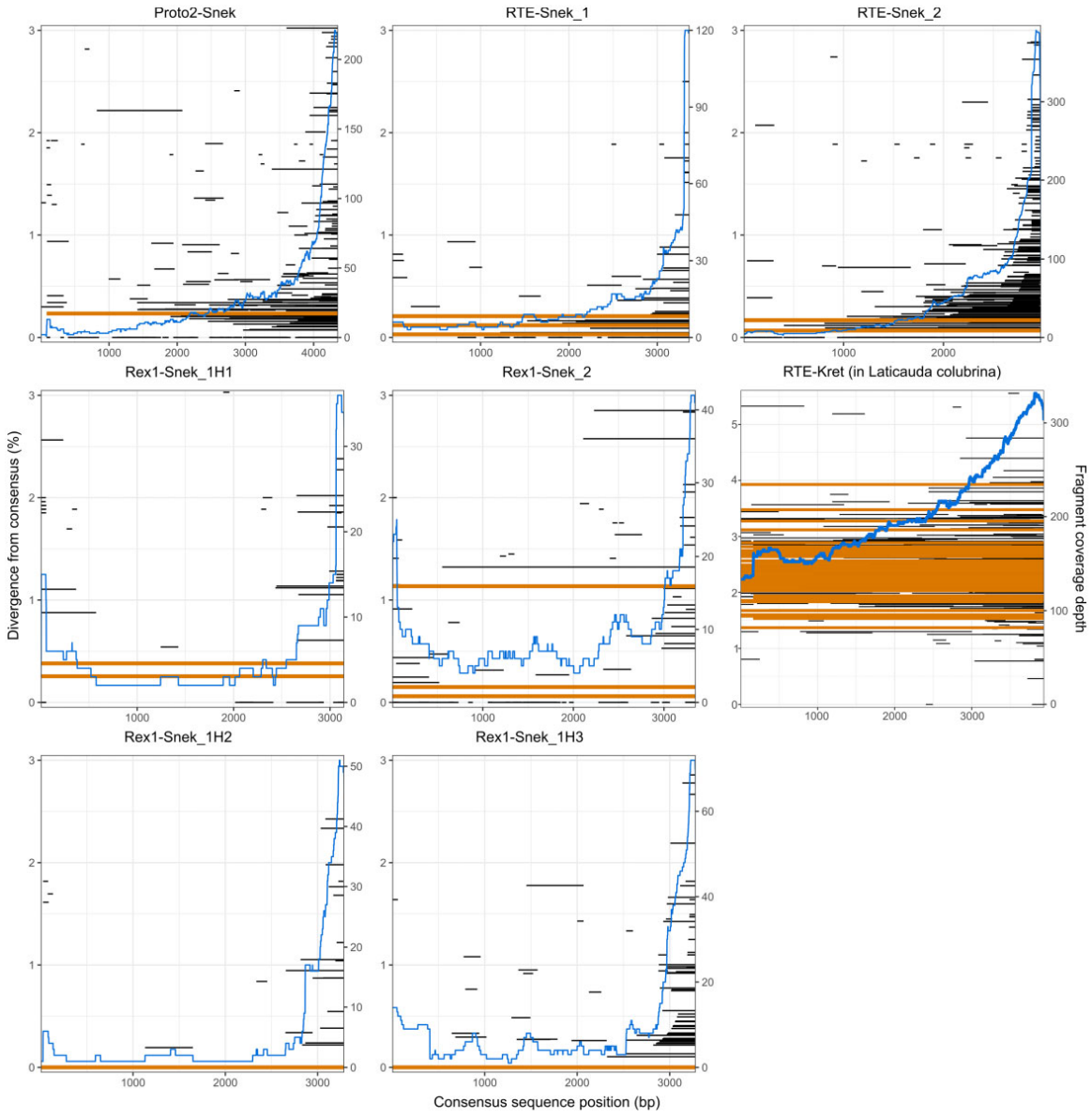


Fig. 2.—Coverage and divergence from consensus of the seven horizontally transferred LINE subfamilies identified in the *Aipysurus laevis* genome and the one identified in *Laticauda colubrina*. LINE fragments were identified with BLASTN (Altschul et al. 1990; Camacho et al. 2009) and plotted using ggplot2 (Wickham 2011) using the consensus2genome script (https://github.com/clemgoub/consensus2genome, last accessed September 16, 2020). The blue line represents the depth of coverage of fragments aligned to the subfamily consensus sequence (shown on right-hand *y* axis). Each horizontal line represents the divergence of a fragment and its position mapped to the repeat consensus (position shown on *x* axis); orange shows full-length repeats and black shows repeat fragments. The divergence from consensus of the repeats is shown on the left-hand *y* axis.

RTE-Snek_1 revealed a similar yet distinct RTE subfamily present in *L. colubrina*, henceforth referred to as RTE-Kret. From these searches, we found RTE-Snek_1 was restricted to *A. laevis* and RTE-Kret to be restricted to *L. colubrina*. In addition to being present in *A. laevis*, Proto2-Snek was also present in *E. ijimae*; Rex1-Snek_1H1, Rex1-Snek_2, and RTE-Snek_2 in *E. ijimae* and *H. melanocephalus*; and Rex1-SnekH2 and Rex1-SnekH3 in *H. melanocephalus*. This reciprocal

Table 1
Most Similar Repbase and Curated Repeats for Each LINE Subfamily in Species Outside of Closely Related Snakes

Repeat (query)	Species (target repeat)	Percent identity	Hit length (bp)
Most similar Repbase sequences			
Rex-Snek_1H1	*Petromyzon marinus* (Rex1-1_PM)	67.5	1,359
Rex-Snek_1H2	*Petromyzon marinus* (Rex1-1_PM)	66.7	1,359
Rex-Snek_1H3	*Petromyzon marinus* (Rex1-1_PM)	64.2	2,796
Rex-Snek_2	*Cyprinus carpio* (Rex1-1_CCa)	75.9	2,795
RTE-Snek_1	*Petromyzon marinus* (RTE-2_PM)	62.9	3,100
RTE-Snek_2	*Chrysemys picta* (RTE-9_CPB)	65.3	2,926
Proto2-Snek	*Oryzias latipes* (Proto2-1_OL)	65.6	666
RTE-Kret	*Petromyzon marinus* (RTE-2_PM)	63.6	3,102
Most similar curated repeats			
Rex-Snek_1H1	*Oryzias latipes*	85.0	2,987
Rex-Snek_1H2	*Oryzias latipes*	82.2	2,973
Rex-Snek_1H3	*Oryzias latipes*	81.6	2,960
Rex-Snek_2	*Miichthys miiuy*	78.7	2,594
RTE-Snek_1	*Laticauda colubrina* (RTE-Kret)	84.9	3,252
RTE-Snek_2	*Hippocampus comes*	74.4	3,184
Proto2-Snek	*Epinephelus lanceolatus*	75.4	3,299
RTE-Kret	*Aipysurus laevis* (RTE-Snek_1)	84.9	3.252

NOTE.—Repbase was searched using the seven consensus *Aipysurus laevis* LINEs using relaxed BLASTN parameters (see Materials and Methods). A database of our curated repeats from all searched species (see Materials and Methods) was searched using the seven consensus *A. laevis* repeats using default BLASTN parameters.

search confirmed all seven subfamilies were absent from both terrestrial (*N. scutatus*, *P. textilis*, and *O. hannah*) and aquatic (*L. colubrina*) outgroups, and RTE-Kret was restricted to *L. colubrina* (fig. 3, supplementary figs. 1–8, Supplementary material online).

We used two approaches to estimate the number and timing of HTT events into sea snakes. Based on the presence or absence of the seven *A. laevis* LINEs in *O. hannah*, *L. colubrina*, *P. textilis*, *N. scutatus*, *H. melanocephalus*, *E. ijimae*, and *A. laevis*, we conservatively estimated nine HTT events into sea snakes dated using the species divergence times from Sanders et al. (2008, 2009, 2013) and Lee et al. (2016) (fig. 3, supplementary table 2, Supplementary material online). Due to the lack of fragments of Rex1-Snek_1H2 and Rex1-Snek_1H3 in *Emydocephalus* (supplementary figs. 10 and 11, Supplementary material online), these two subfamilies were likely transferred independently into *Aipysurus* and *Hydrophis*. In addition, we calculated the timing of HTT into the *Aipysurus* lineage using the average substitutions per site of each LINE subfamily and an estimated genome-wide substitution rate. The insertion time based on substitution rate (supplementary table 2, Supplementary material online) suggests that the HTTs postdate the divergence of *Aipysurus* and *Emydocephalus*. Taking the high standard deviation into account, the timing of HTT events estimated by both methods overlapped with the exception of the transfer of RTE-Snek_2 (supplementary table S2, Supplementary material online).

As an independent verification of presence/absence and to look for potential current activity of the LINEs, we searched assembled transcriptomes of a variety of tissues from three sea snakes—*A. laevis*, *A. tenuis*, and *Hydrophis major* from Crowe-Riddell (2019) (see supplementary dataset 2, Supplementary material online). We identified high-identity transcripts (>95% identity) of all Rex1-Snek1H1, Rex1-Snek1H2, Rex1-Snek1H3, Rex1-Snek_2, and RTE-Snek_2 in at least one tissue of *A. laevis*, *A. tenuis*, and *H. major*. High-identity transcripts of RTE-Snek_1 and Proto2-Snek were present in *A. laevis* and *A. tenuis*, yet absent from all *H. major* tissues, with one small fragment of an RTE-Snek_1-like transcript present in an *H. major* testis transcriptome. The presence of transcripts of all seven LINE subfamilies both confirmed the presence/absence pattern of the specific subfamilies in *A. laevis* and indicates potential ongoing retrotransposition of these elements.

Verification of HTT and Search for HTT Donor Species

Although the absence of the marine-specific TEs in close terrestrial species supported HTT to sea snakes, we needed to rule out the possibility that those TEs were lost from those terrestrial species. In order to confirm HTT versus loss of TEs, we searched for all seven LINE subfamilies in 630 metazoan genomes using BLASTN with relaxed parameters (see Materials and Methods). Our search identified homologous, yet divergent Rex1s in fish and squamates, Proto2s in fish, and RTEs widespread across a variety of marine organisms including fish, echinoderms, corals, and sea kraits (see fig. 4, supplementary dataset 4, Supplementary material online). Using these hits as seeds, we curated consensus repeats of each LINE subfamily in the species they were identified in.

We then aligned our original LINE sequences against a database containing both our curated repeats and Repbase

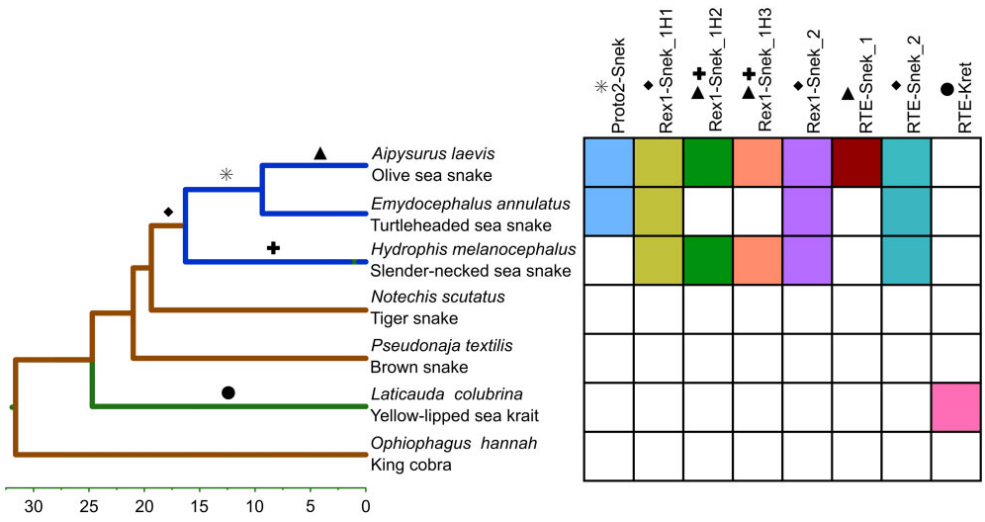


Fig. 3.—Presence of the eight HTT LINE subfamilies across the phylogeny of elapid snakes (adapted from Lee et al. [2016]). Color of lineage represents habitat—marine species are blue, terrestrial brown, and amphibious green. Each symbol represents the likely timing of horizontal transfers, for example, the square indicates the likely transfer date of both Rex1-Snek_1H1 and Rex1-Snek_2. Presence/absence determined using reciprocal BLASTN search (Altschul et al. 1990; Camacho et al. 2009) using default parameters.

repeats. All seven of our original LINE subfamilies were most similar to curated LINEs found in marine species (table 1) with pairwise identity for all closest hits between 75–85%. Rex1-Snek_1H1, Rex1-SnekH2, Rex1-SnekH3, and Rex1-Snek_2 were most similar to Rex1s curated from a variety of fish genomes. Proto2-Snek was most similar to a Proto2 from the European carp (*Cyprinus carpio*) genome and RTE-Snek_1 most similar to RTE-Kret from *L. colubrina*. If the LINE subfamilies were present in sea snakes yet absent from terrestrial and amphibious elapids due to repeated losses, we would expect to find highly similar LINEs to still be present in other squamates. However, we failed to identify highly similar repeats in any squamates; therefore, the most parsimonious explanation supports HTT and rules out loss. We used the results of this search in an attempt to identify the likely donor or vector species by looking for species hosting our HTT LINEs with a comparable degree of sequence divergence to that observed in *A. laevis*. However, none of the cross-species alignments were greater than 87% nucleotide sequence identity and therefore did not show comparable sequence divergence which would be required to identify potential donor species (table 1).

Discordant Phylogenies of RTEs and of Rex1s Compared with Host Species

As extreme discordance between repeat and species phylogenies would further support HTT, we compared the respective tree topology of all RTEs, Proto2s, and Rex1s, using both Repbase sequences and our curated sequences, to the species tree topology. As illustrated in figure 5, the species and repeat phylogenies of all seven sea snake LINE subfamilies and the *L. colubrina* RTE are highly discordant, evidenced by their clustering with teleost fishes. This confirms likely HTT events from marine organisms into sea snakes and sea kraits, and further refutes independent losses from terrestrial Australian elapids.

Insertions in and Near Coding Regions

To identify any insertions of these LINEs in *A. laevis* which may have the potential to alter gene expression or protein structure, we identified all insertions in or near regions annotated as genes, in particular exons and untranslated regions (UTRs) (supplementary table 1, Supplementary material online). Intersects of gene and repeat annotation intervals in the *A. laevis* assembly initially revealed 23 insertions of HTT LINEs in or near genes: 19 insertions in 5′ UTRs or within 5,000 bp upstream, 1 into a coding exon and 3 into 3′ UTRs.

To test for potential assembly errors that might have yielded erroneous insertions near genes, we searched for the flanking regions of the 23 insertions in the closely related *E. ijimae* and *H. melanocephalus*. Eight of the 23 insertions were disregarded as the likely result of assembly errors in *A. laevis*, as their flanking sequences were in the middle of two different contigs in both *E. ijimae* and *H. melanocephalus*. The flanking regions of the remaining 15 insertions were contiguous in *E. ijimae* and *H. melanocephalus*. We report these 15 insertions in table 2. We consider the insertion of RTE_Snek_2 into the 3′ UTR of the Adenylate Cyclase Type 4 (*ADCY4*) gene as the most interesting of these, as it is the only gene out the 15 that is present in a KEGG environmental adaptation pathway (circadian entrainment). However,

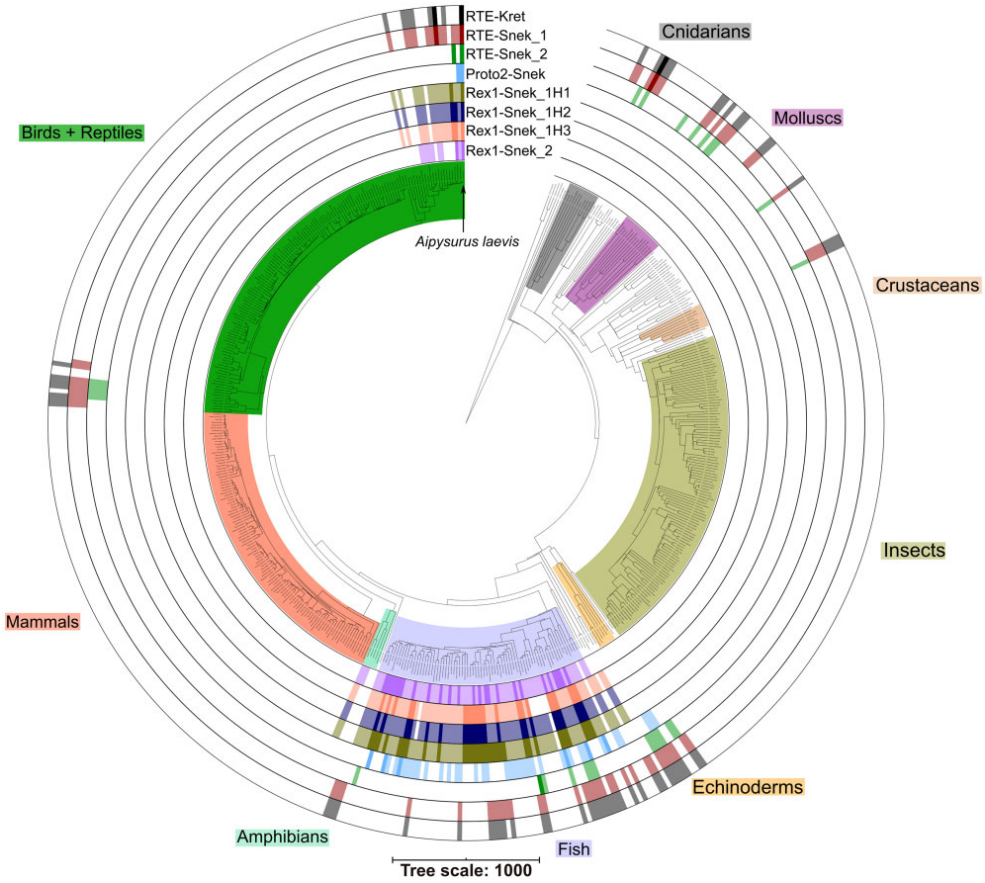


Fig. 4.—Presence of the seven *Aipysurus* and one *Laticauda* HTT LINE subfamilies across 540 Metazoa. In each ring, darker shading represents the presence of at least one sequence over 1,000 bp in length showing 75% or higher pairwise identity to the LINE, lighter shading represents the presence of more than one sequence over 1,000 bp with less than 75% pairwise identity, and white represents the complete absence of similar sequences. Presence of LINEs identified using BLASTN with custom parameters (see Materials and Methods) (Altschul et al. 1990; Camacho et al. 2009) and plotted in iToL (Letunic and Bork 2019). Species tree generated using TimeTree (Hedges et al. 2006), manually edited to correct elapid phylogeny to fit (Lee et al. 2016). Interactive tree available at https://itol.embl.de/shared/jamesdgalbraith (last accessed September 16, 2020).

testing the adaptive significance of these insertions will have to await improvement of the genome assembly and population genetic data for *A. laevis*. We note that many of these genes are likely to have pleiotropic effects as regulators of transcription or protein turnover, thus complicating future assessments of their adaptive significance. However, changes in pleiotropic genes have the potential to amplify adaptive changes in other loci (Østman et al. 2012).

Discussion

We have identified seven LINE subfamilies present in sea snakes and one present in sea kraits, yet absent from their terrestrial relatives. The two competing hypotheses for this presence/absence pattern are loss from the terrestrial species or HTT to the marine species. If the seven subfamilies were lost from the terrestrial species, we would expect to see similar subfamilies still present in other squamates. Our search of 630 additional metazoans revealed the seven subfamilies to be absent not just from other squamates, but from all other tetrapods. For the majority of the seven subfamilies, the most similar LINE was present in a teleost fish, indicating either that the LINEs were repeatedly lost from all other tetrapods following their divergence from teleost fish 400 Ma, or the subfamilies were horizontally transferred into sea snakes and sea kraits following their divergence from terrestrial relatives.

Based on the observed patchy phylogenetic distribution, the high similarity of HTT TEs to those from distantly related marine species, and the discordance of the species and LINE phylogenies (figs. 3 and 5), the most parsimonious explanation is that the seven LINEs identified in *A. laevis* and one identified in *L. colubrina* were horizontally transferred from

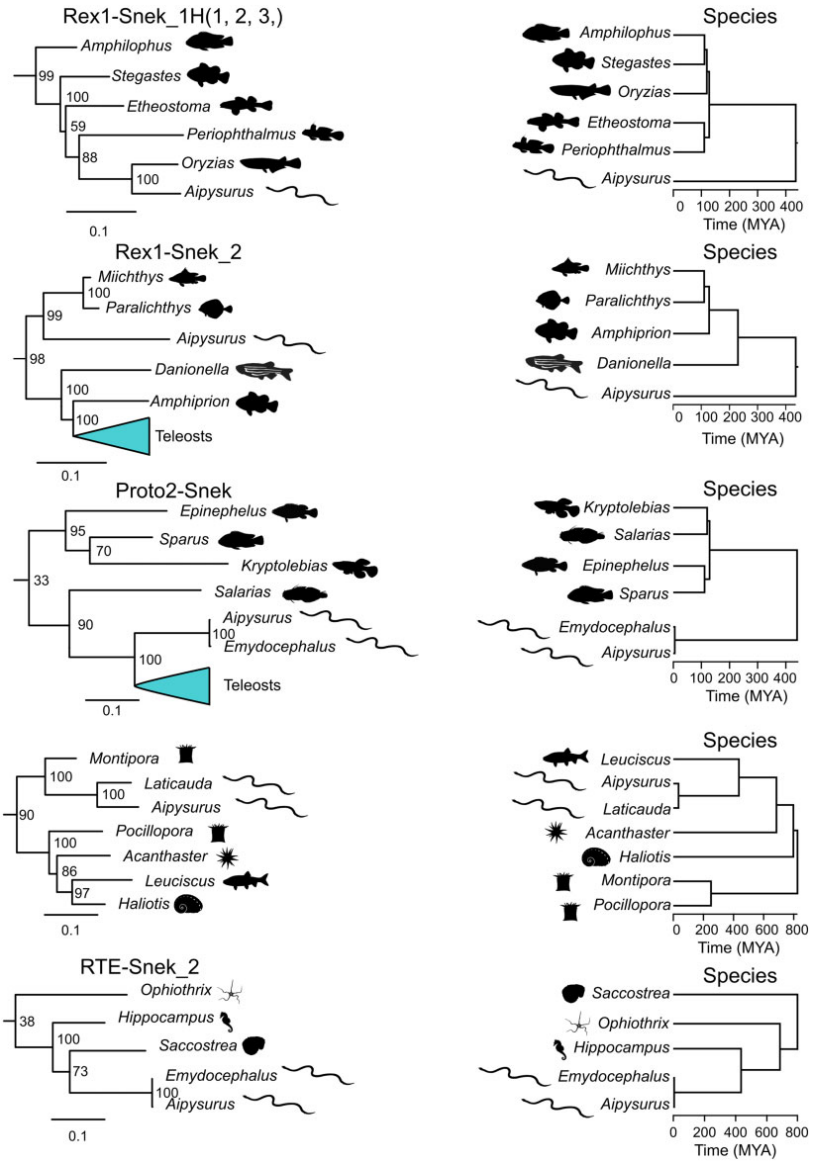


Fig. 5.—Excerpts from the phylogenies of all intact curated and Repbase RTEs and all intact curated and Repbase Rex1s compared with host species phylogeny. The blue triangles on the left represent condensed large subtrees of LINE sequences. TE phylogeny scale bar represents substitutions per site. The numbers next to each node in the repeat trees are the support value. Extracts from larger phylogenies constructed using RAxML (Stamatakis 2014) based on MAFFT (Katoh and Standley 2013) nucleotide alignments trimmed with Gblocks (Talavera and Castresana 2007) (for full phylogenies see supplementary Appendix, figs. S1 and S2, Supplementary material online). Species trees constructed with TimeTree (Hedges et al. 2006).

marine species following the transition of the ancestors of these snakes to a marine habitat. Additionally, the estimated timing of transfer supports independent transfers of both Rex1-Snek_1H2 and Rex1-Snek_1H3 into the *Aipysurus* and *Hydrophis* lineages (supplementary table 2, Supplementary material online). Although all seven LINE subfamilies are currently expressed in *A. laevis* based on transcriptome data, the number of large, near-identical fragments of RTE-Snek_1, RTE-Snek_2, and Proto2-Snek found within the *A. laevis* genome is larger than for the Rex1s. This indicates potentially greater replication of RTE-Snek_1, RTE-Snek_2, and Proto2-Snek since the HTT events in the past 3–17 Myr (Sanders et al 2008, 2012, 2013; Lee et al. 2016).

Table 2
HTT LINEs Inserted into Exons, UTRs, or within 5,000 bp Upstream of 5′ UTRs of Genes within the *A. laevis* Assembly and Transcriptome

Gene	LINE	Distance to 5′ UTR (bp)	Insertion size (bp)
Acetyl-CoA Acyltransferase 1 (*ARIH1*)	Proto2-Snek	223	85
KN Motif And Ankyrin Repeat Domains 4 (*KANK4*)	Proto2-Snek	4,987	161
Potassium Calcium-Activated Channel Subfamily N Member 4 (*KCNN4*)	Proto2-Snek	3,746	98
Outer Mitochondrial Membrane Lipid Metabolism Regulator OPA3 (*OPA3*)	Rex1-Snek_2	3,149	81
Rabaptin, RAB GTPase Binding Effector Protein 1 (*RABEP1*)	Proto2-Snek	1,389	99
Valosin Containing Protein Lysine Methyltransferase (*VCPKMT*)	Rex1-Snek_1H1	512	76
Cdc42 effector protein 4 (*CDC42EP4*)	RTE-Snek_2	1,475	422
Gamma-aminobutyric acid receptor subunit alpha-3 (*GABRA3*)	RTE-Snek_2	4,247	95
Leucine-zipper-like transcriptional regulator 1 (*LZTR1*)	RTE-Snek_2	2,066	421
Polyadenylate-binding protein 2 (*PABPN1*)	RTE-Snek_2	145	431
Parvalbumin alpha (*PVALB*)	RTE-Snek_2	4,152	52
Deaminated glutathione amidase (*NIT1*)	RTE-Snek_2	In coding exon	228
Adenylate cyclase type 4 (*ADCY4*)	RTE-Snek_2	In 3′ UTR and transcript	130
CAP-Gly Domain Containing Linker Protein Family Member 4 (*CLIP4*)	RTE-Snek_1	In transcript	—
BLOC-1 Related Complex Subunit 8 (*BORCS8*)	Rex1-Snek_1H3	In transcript	—

As all seven of the HTT LINE subfamilies are most similar to LINEs found in distantly related marine metazoans, we hypothesize that the donor species for each is likely a marine fish or invertebrate. However, the degree of sequence divergence between the LINE from *L. colubrina* and the seven LINEs from *A. laevis* from the most similar LINEs from aquatic species means we cannot identify a specific donor species. Likely donors and vectors of HTT are pathogens, predators, prey, parasites, and epibionts (Gilbert and Feschotte 2018). Sea snake diets vary greatly; some species are generalists that eat a wide variety of fish and occasionally crustaceans, cephalopods, and mollusks, whereas others specialize on burrowing eel-like or goby-like fish or feed exclusively on fish eggs (Sherratt et al. 2018). Parasites of sea snakes include isopods, nematodes, tapeworms, and flatworms, whereas epibionts include various, hydrozoans, polychaetes, decapods, gastropods, bivalves, and Bryozoa (Saravanakumar 2012; Gillett 2017). As very few species with ranges overlapping those of *Laticauda* and *Aipysurus* have been sequenced, and the range of *Aipysurus* spans highly biodiverse habitats, it is unlikely we will further narrow the donor of any of these eight LINE subfamilies without significant additional genome sequence data from Indo-West Pacific tropical marine species.

Although we were unable to identify specific donor species, our finding of HTT between marine species is in line with multiple past studies that reported HTT within and across marine phyla. HTT is prolific and particularly well described in aquatic microbial communities (reviewed in-depth in Sobecky and Hazen [2009]). HTT of LINEs, LTR retrotransposons and DNA transposons has been reported in marine metazoans, with past studies describing the transfer of Rex1s and Rex3s between teleost fishes (Volff et al. 2000; Carducci et al. 2018), *Steamer*-like LTR retrotransposons both within and across phyla (Metzger et al. 2018), L1 and BovB LINEs within and across phyla (Ivancevic et al. 2018), Mariner DNA transposons between diverse crustaceans (Casse et al. 2006), and a wide variety of TEs between tetrapods and teleost fish (Zhang et al. 2020). What sets our findings apart is that HTTs in this report have occurred multiple times as a result of the recent terrestrial to marine transition of the *Aipysurus*/*Hydrophis* common ancestor. The transfer of all seven LINEs occurred <18 Ma from aquatic animal donor species that diverged from snakes >400 Ma (Broughton et al. 2013; Hughes et al. 2018). As illustrated in figure 3, the varying presence/absence of the seven LINEs across the three species of sea snakes is indicative of nine independent HTT events as opposed to a single event. The recent timing of HTT into marine squamates is not specific to sea snakes, as we found transfer of an RTE-Kret to the sea kraits which underwent an independent transition to the marine habitat. These repeated invasions suggest aquatic environments potentially foster HTT, with more examples likely to be revealed by additional genome sequences from marine species.

The likely ongoing replication of all seven *A. laevis* HTT LINEs, as evidenced by both the presence of insertions and transcripts with near 100% identity, continues to contribute genetic material to the evolution of *Aipysurus*. Previous investigators have reported entire genes, exons, regulatory sequences, and noncoding RNAs in vertebrates derived from transposons, as well as TE insertions leading to genomic rearrangement (reviewed in-depth in Warren et al. [2015]). For snakes, Peng et al. (2020) described the expansion of LTR elements across *H. curtus* leading to adaptive changes in the marine environment. Similarly, the insertion of CR1 fragments near phospholipase A2 venom genes in vipers led to nonallelic homologous recombination, in turn causing

duplication of these genes (Fujimi et al. 2002). Rapid genomic innovation would have been necessary for *Aipysurus* to adapt to the marine environment, with the independent evolution of paddle-like tails, salt excretion glands, and dermal photoreception following their divergence from their most recent common ancestor with *Hydrophis* (Brischoux et al. 2012; Sanders et al. 2012; Crowe-Riddell et al. 2019). Other adaptations are likely to have occurred or are occurring for sea snakes to conform to their marine habitat, as evolutionary transitions from terrestrial to marine habits entail massive phenotypic changes spanning metabolic, sensory, locomotor, and communication-related traits. Our finding that 15 genes, most with likely pleiotropic effects, contain HTT insertions and thus may have altered expression will require further investigation. One of these genes, ADCY4 is particularly interesting as it is part of the circadian entrainment pathway. Transition to a marine environment is likely to require altered sensitivity of the circadian entrainment pathway to environmental cues of light intensity and wavelength. Future research to examine the association between these HTT-derived sequences and adaptation will require investigation of differential regulation of these genes between terrestrial and marine snakes in a variety of tissues as well as improvement of the *A. laevis* genome assembly and collection of population genomic data.

Conclusions

Our findings reveal repeated HTT of LINEs into fully marine and amphibious lineages of marine elapids as a result of their transition from a terrestrial environment. The HTT LINE insertions near genes and continued expression of all seven HTT LINE subfamilies is indicative of possible ongoing impact on the adaptive evolution of *Aipysurus*. Taken together, our results support a likely role for habitat transitions as direct contributors to the evolution of metazoan genomes, rather than solely acting through selection from altered environmental conditions.

Materials and Methods

Outline of Methods

Our study aimed primarily to identify TE subfamilies present in sea snakes yet absent from close terrestrial relatives, determine if their absence was due to TE loss or HTT, and if due to HTT find the potential donor or vector species. Our secondary aim was to determine if HTT subfamilies likely remain active in sea snakes based on transcriptomic data. Our final aim was to check if any HTT TE subfamilies discovered may have impacted the evolution of sea snakes since their divergence from terrestrial snakes by identifying insertions near/in genes and if these genes had roles in pathways important in adaptation to the marine habitat.

Identification and Classification of Repetitive Sequences in *A. laevis*

We identified repetitive sequences present in the Ludington et al. (dx.doi.org/10.5281/zenodo.3975254) *A. laevis* assembly using CARP (Zeng et al. 2018). Using RPSTBLASTN 2.7.1+ (Marchler-Bauer and Bryant 2004) and a custom library of position-specific scoring matrices from the CDD and Pfam databases (Finn et al. 2016; Marchler-Bauer et al. 2017), we identified protein domains present in all consensus sequences over 800 bp in length found by CARP. Sequences were classified as potential LINEs, LTR retroelements and various DNA transposons based on the presence of relevant protein domains following the Wicker et al. (2007) classification. For example, we treated consensus sequences containing over 80% of both an exo-endonuclease domain and a reverse transcriptase domain as potential LINEs. For a full breakdown of protein domains used to classify retroelements, see supplementary table 3, Supplementary material online. We used CENSOR 4.2.29 (Kohany et al. 2006) to further classify the consensus sequences. To reduce redundancy, we aligned all potential TEs to all other potential TEs using BLASTN 2.7.1+ (Altschul et al. 1990; Camacho et al. 2009) with default parameters and removed consensus sequences with both 94% or higher pairwise identity to, and 50% or higher coverage by longer consensus sequences.

Search for Ab Initio Annotated TEs in Close Terrestrial Relatives

To determine if the TEs subfamilies discovered were present in closely related species, we used megablast 2.7.1+ (Altschul et al. 1990; Camacho et al. 2009) to perform a nucleotide search for the consensus sequences of each subfamily in the genomes of two closely related terrestrial elapids (*N. scutatus* and *P. textilis*) (provided by Richard Edwards), and a more distantly related semi-marine elapid (*L. colubrina*) (Kishida et al. 2019). We treated all CARP sequences which were found by megablast in both *N. scutatus* and *P. textilis* as ancestrally shared, and all others as potential HTT candidates (all were LINEs). After discovering a highly similar subfamily was present in *L. colubrina* but absent from the two terrestrial snakes (RTE-Kret), we manually curated it using a "search, extend, align, trim" method adapted from Platt et al. (2016) and Suh et al. (2018) (see supplementary Methods, Supplementary material online and description below).

Curation of TEs Absent from Close Terrestrial Relatives

To create a better consensus for each LINE subfamily, we manually curated new consensus sequences using a "search, extend, align, trim" method (explained in greater detail in supplementary Methods, Supplementary material online, script at https://github.com/jamesdgalbraith/HT_Workflow/blob/master/PresenceAbsence/extendAlignSoloRstudio.R,

last accessed September 16, 2020). We used megablast 2.7.1+ (Altschul et al. 1990; Camacho et al. 2009) to search for the consensus sequence of a subfamily within the *A. laevis* genome. We selected the 25 best hits over 1,000 bp based on bitscore and extended the coordinates of these sequences by 1,000 bp at each end of the hit. We constructed multiple sequence alignments (MSAs) of the extended sequences using MAFFT v7.310 (Katoh and Standley 2013). Where multiple full-length sequences showing significant lack of homology were present, the LINE subfamily was split into multiple subfamilies (see supplementary fig. 9, Supplementary material online). Finally, we manually edited the extended sequences in Geneious Prime 2020.0.2 to remove nonhomologous regions and created a new consensus sequence. If only one full-length copy of a subfamily was present in the genome, it was used instead of a consensus sequence. We used PCOILs (Gruber et al. 2006) and HHpred (Zimmermann et al. 2018) searches of the translated Open Reading Frames (ORFs) against the CDD and Pfam databases (Finn et al. 2016; Marchler-Bauer et al. 2017) to identify any additional protein domains or structures present in the seven LINEs.

Search for HTT Candidate LINEs in the Genomes and Transcriptomes of Other Sea Snakes

Similar to the search of closely related terrestrial species, we used megablast to perform reciprocal searches for the consensus sequences of the seven *Aipysurus* LINE subfamilies in the genomes of *H. melanocephalus* and *Emydocephalus annulatus* (Kishida 2019), and assembled transcriptomes from various tissues of *A. laevis*, *A. tenuis*, and *H. major* from Crowe-Riddell et al. (2019).

Estimating Timing of HTT Events by Substitution Rate

We estimated the timing of the seven HTT events using a custom R script (https://github.com/jamesdgalbraith/HT_Workflow/blob/master/Divergence/insertion_time_calculator.R, last accessed September 16, 2020). We identified all copies of the seven *A. laevis* HTT subfamilies in the *A. laevis* assembly using megablast. A reciprocal megablast search using the identified copies was carried out against the seven *A. laevis* HTT subfamily consensus sequences to identify the most similar sequence based on pairwise identity. Using the reciprocal megablast search output, we calculated the mean substitutions per site for each HTT subfamily. Finally, using an elapid whole-genome substitution rate estimate from Ludington and Sanders (under review by *Molecular Ecology*) of 1.25 × 10e−08 per site per generation and a generation time of 10 years, we calculated the HTT event timing of each subfamily (supplementary table 2, Supplementary material online).

Search for and Curation of Similar TEs in Other Metazoan Genomes

To identify other species containing the seven *Aipysurus* and one *Laticauda* LINE subfamilies, we used the HTT LINE consensus sequences for BLASTN searches in of over 630 metazoan genomes downloaded from GenBank (Benson et al. 2017) using relaxed parameters (-evalue 0.00002 -reward 3 -penalty -4 -xdrop_ungap 80 -xdrop_gap 130 -xdrop_gap_final 150 -word_size 10 -dust yes -gapopen 30 -gapextend 6). We treated species containing a hit of at least 1,000 bp as potentially containing a similar LINE subfamily. From the BLASTN hits from these species, we attempted to manually curate subfamilies using a variant of the "search, extend, align, trim" method described in the supplementary Methods, Supplementary material online. If only one copy of the LINE subfamily was present in a genome assembly we did not include that species in the list of species containing similar LINEs in order to reduce false positives. We used a consensus sequence derived from the initial hits within the species as the query for the BLASTN search of the genome, and extended hits by 3,000 bp in the 5′ and 3′ directions. As illustrated in supplementary figure 9, Supplementary material online, if an MSA appeared to contain multiple LINE subfamilies, as judged by lack of sequence homology or gaps, it was split and consensuses were constructed for each individual family. As homologous, yet highly diverged, Rex1 and RTE subfamilies were identified in other elapids we used the same "search, extend, align, trim" method to curate the most similar repeats in the *A. laevis* assembly, using the consensus from *N. scutatus* as the initial query. All subfamilies identified in *N. scutatus* had highly similar homolog in *A. laevis*.

Characterizing Divergence Patterns in the HT Repeats across Hydrophiinae

To identify fragments of the seven *Aipysurus* and one *Laticauda* HTT LINE subfamilies and determine their divergence from the consensus sequences, we performed a reciprocal best hit search using BLASTN 2.7.1+ (Altschul et al. 1990; Camacho et al. 2009) on the *A. laevis*, *E. ijimae*, *Hydrophis cyanocinctus*, *H. melanocephalus*, *N. scutatus*, *P. textilis*, *L. colubrina*, and *O. hannah* assemblies. HTT consensus sequences were used as the initial query, with resulting hits then used as queries against a database containing the original consensus sequences.

Repeat Phylogeny Construction

For constructing repeat phylogenies, we created two libraries; one containing all Rex1s we curated and Rex1s derived from Repbase; and another containing all RTEs, we curated and all RTE-like (Proto2, RTE, and BovB) sequences from Repbase. In addition, each library contained an outgroup LINE based on

the Eickbush and Malik (Eickbush and Malik 2002) phylogeny of LINEs. We removed all sequences not containing at least 80% of both the endonuclease and reverse transcriptase domains from each library based on RPSTBLASTN (Marchler-Bauer and Bryant 2004) searches against the NCBI CDD (Marchler-Bauer et al. 2017).

We created nucleotide MSAs of each library of LINEs using MAFFT v7.310 (Katoh and Standley 2013) and removed poorly aligned regions using Gblocks (Talavera and Castresana 2007) allowing smaller final blocks, gap positions within the final blocks and less strict flanking positions. Finally, we constructed phylogenies from the trimmed MSA using RAxML (Stamatakis 2014) with 20 maximum likelihood trees and 500 bootstraps.

Species Phylogeny Construction

We used TimeTree (Hedges et al. 2006) to infer species phylogenies presented in figure 4. In cases in which a species of interest was not present in the TimeTree database, where possible we used an appropriate species from the same clade in its place and corrected the species names on the resulting tree.

Repeat Insertions Near and in Genes

Using the plyranges (Lee et al. 2019) and GenomicRanges R packages (Lawrence et al. 2013) (53, 54), we identified any insertions of the HTT LINEs into coding exons, UTRs and upstream of 5′ UTRs for gene annotations from Ludington et al. (https://dx.doi.org/10.5281/zenodo.3975254, last accessed September 16, 2020) (https://github.com/jamesdgalbraith/HT_Workflow/blob/master/GeneInteraction/overlapSearch.R, last accessed September 16, 2020).

To confirm that insertions were assembled correctly, we used BLASTN to search for the repeats extended by 2,000 bp in each direction in the *E. ijimae* and *H. melanocephalus* assemblies. We selected the best hits from each species based on query coverage and percent identity. Using MAFFT v7.310 (Katoh and Standley 2013), we constructed MSAs of each extended repeat and the corresponding regions from the two other assemblies (https://github.com/jamesdgalbraith/HT_Workflow/blob/master/GeneInteraction/insertionConfirmation.R, last accessed September 16, 2020). By manually viewing the resulting alignment in Geneious and the raw BLASTN output, we determined if the repeat insertions were assembled correctly. To confirm the insertion of RTE-Snek_2 identified in the 3′ UTR of *ADCY4*, we perform megablast searches of the *A. laevis* transcriptome from Ludington et al. (https://dx.doi.org/10.5281/zenodo.3993854, last accessed September 16, 2020).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank Richard Edwards (University of New South Wales) for providing the *Notechis scutatus* and *Pseudonaja textilis* genome assemblies; Dan Kortschak (University of Adelaide), Jesper Boman (Uppsala University), Valentina Peona (Uppsala University), Atma Ivancevic (University of Colorado), and Ed Chuong (University of Colorado) for helpful suggestions and discussions. We also thank the editor and anonymous reviewers for their comments on and suggestions for the manuscript. K.L.S was funded by the Australian Research Council (FT130101965).

Author Contributions

J.D.G., A.L., A.S., and D.L.A. designed research; K.L.S. and A.L. provided olive sea snake genome assembly; A.L. provided olive sea snake genome transcriptome; J.D.G. and A.L. performed research; and J.D.G., K.L.S., and D.L.A. wrote the paper with input from A.L. and A.S.

Data Availability

All scripts are available at hhttps://github.com/jamesdgalbraith/HT_Workflow. Repeat sequences and phylogenies are in supplementary datasets 5–8.

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215(3):403–410.
Bao W, Kojima KK, Kohany O. 2015. Repbase update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 6:11.
Benson DA, et al. 2017. GenBank. Nucleic Acids Res. 45(D1):D37–D42.
Biémont C, Vieira C, Borie N, Lepetit D. 1999. Transposable elements and genome evolution: the case of *Drosophila simulans*. Genetica 107(1/3):113–120.
Brischoux F, Tingley R, Shine R, Lillywhite HB. 2012. Salinity influences the distribution of marine snakes: implications for evolutionary transitions to marine life. Ecography 35(11):994–1003.
Broughton RE, Betancur-R R, Li C, Arratia G, Ortí G. 2013. Multi-locus phylogenetic analysis reveals the pattern and tempo of bony fish evolution. PLoS Curr. 5:ecurrents.tol.2ca8041495ffafd0c92756e75247483e.
Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10(1):421.
Carducci F, et al. 2018. Rex retroelements and teleost genomes: an overview. Int J Mol Sci. 19(11):3653.
Casacuberta E, González J. 2013. The impact of transposable elements in environmental adaptation. Mol Ecol. 22(6):1503–1517.
Casse N, et al. 2006. Species sympatry and horizontal transfers of Mariner transposons in marine crustacean genomes. Mol Phylogenet Evol. 40(2):609–619.
Cerbin S, Jiang N. 2018. Duplication of host genes by transposable elements. Curr Opin Genet Dev. 49:63–69.
Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet. 18(2):71–86.
Crowe-Riddell JM, et al. 2019. Phototactic tails: evolution and molecular basis of a novel sensory trait in sea snakes. Mol Ecol. 28(8):2013–2028.
Eickbush TH, Malik HS. 2002. Origins and evolution of retrotransposons. In: Craig NL, Craigie R, Gellert M, Lambowitz AM, editors.

Mobile DNA II. Washington (DC): American Society of Microbiology Press. p. 1111–1144.
Finn RD, et al. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44(D1):D279–D285.
Fujimi TJ, Tsuchiya T, Tamiya T. 2002. A comparative analysis of invaded sequences from group IA phospholipase A2 genes provides evidence about the divergence period of genes groups and snake families. Toxicon 40(7):873–884.
Gilbert C, Feschotte C. 2018. Horizontal acquisition of transposable elements and viral sequences: patterns and consequences. Curr Opin Genet Dev. 49:15–24.
Gruber M, Söding J, Lupas AN. 2006. Comparative analysis of coiled-coil prediction methods. J Struct Biol. 155(2):140–145.
Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22(23):2971–2972.
Hughes LC, et al. 2018. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. Proc Natl Acad Sci U S A. 115(24):6249–6254.
Ivancevic AM, Kortschak RD, Bertozzi T, Adelson DL. 2018. Horizontal transfer of BovB and L1 retrotransposons in eukaryotes. Genome Biol. 19(1):85.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30(4):772–780.
Kishida T, et al. 2019. Loss of olfaction in sea snakes provides new perspectives on the aquatic adaptation of amniotes. Proc R Soc B. 286(1910):20191828.
Kohany O, Gentles AJ, Hankus L, Jurka J. 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 7:474.
Lawrence M, et al. 2013. Software for computing and annotating genomic ranges. PLoS Comput Biol. 9:e1003118.
Lee MSY, Sanders KL, King B, Palci A. 2016. Diversification rates and phenotypic evolution in venomous snakes (Elapidae). R Soc Open Sci. 3(1):150277.
Lee S, Cook D, Lawrence M. 2019. plyranges: a grammar of genomic data transformation. Genome Biol. 20(1):4.
Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 47(W1):W256–W259.
Lillywhite HB. 2014. How snakes work: structure, function and behavior of the world's snakes. Oxford: Oxford University Press.
Luan DD, Korman MH, Jakubczak JL, Eickbush TH. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell 72(4):595–605.
Marchler-Bauer A, Bryant SH. 2004. CD-Search: protein domain annotations on the fly. Nucleic Acids Res. 32:W327–W331.
Marchler-Bauer A, et al. 2017. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. 45(D1):D200–D203.
Marques DA, Jones FC, Di Palma F, Kingsley DM, Reimchen TE. 2018. Experimental evidence for rapid genomic adaptation to a new niche in an adaptive radiation. Nat Ecol Evol. 2(7):1128–1138.
Metzger MJ, Paynter AN, Siddall ME, Goff SP. 2018. Horizontal transfer of retrotransposons between bivalves and other aquatic species of multiple phyla. Proc Natl Acad Sci U S A. 115(18):E4227–E4235.
Nitschke CR, Hourston M, Udyawer V, Sanders KL. 2018. Rates of population differentiation and speciation are decoupled in sea snakes. Biol Lett. 14(10):20180563.
Østman B, Hintze A, Adami C. 2012. Impact of epistasis and pleiotropy on evolutionary adaptation. Proc R Soc B. 279(1727):247–256.
Pasquesi GIM, et al. 2018. Squamate reptiles challenge paradigms of genomic repeat element evolution set by birds and mammals. Nat Commun. 9(1):2774.
Peccoud J, Cordaux R, Gilbert C. 2018. Analyzing horizontal transfer of transposable elements on a large scale: challenges and prospects. BioEssays 40(2):1700177.
Peng C, et al. 2020. The genome of Shaw's sea snake (*Hydrophis curtus*) reveals secondary adaptation to its marine environment. Mol Biol Evol. 37:1744–1760.
Platt RN, Blanco-Berdugo L, Ray DA. 2016. Accurate transposable element annotation is vital when analyzing new genome assemblies. Genome Biol Evol. 8(2):403–410.
Rebollo R, Romanish MT, Mager DL. 2012. Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annu Rev Genet. 46(1):21–42.
Salces-Ortiz J, Vargas-Chavez C, Guio L, Rech GE, González J. 2020. Transposable elements contribute to the genomic response to insecticides in *Drosophila melanogaster*. Phil Trans R Soc B. 375(1795):20190341.
Sanders KL, Lee MSY, Leys R, Foster R, Scott Keogh J. 2008. Molecular phylogeny and divergence dates for Australasian elapids and sea snakes (hydrophiinae): evidence from seven genes for rapid evolutionary radiations. J Evol Biol. 21(3):682–695.
Sanders KL, Lee MSY, Mumpuni Bertozzi T, Rasmussen AR. 2013. Multilocus phylogeny and recent rapid radiation of the viviparous sea snakes (Elapidae:Hydrophiinae). Mol Phylogenet Evol. 66:575–591.
Sanders KL, Mumpuni Lee, MSY. 2010. Uncoupling ecological innovation and speciation in sea snakes (Elapidae, Hydrophiinae, Hydrophiini). J Evol Biol. 23(12):2685–2693.
Sanders KL, Rasmussen AR, Elmberg J. 2012. Independent innovation in the evolution of paddle-shaped tails in viviparous sea snakes (Elapidae:Hydrophiinae). Integr Comp Biol. 52(2):311–320.
Saravanakumar A, Balasubramanian T, Raja K, Trilles J-P. 2012. A massive infestation of sea snakes by cymothoid isopods. Parasitol Res. 110(6):2529–2531.
Schlötterer C, Kofler R, Versace E, Tobler R, Franssen SU. 2015. Combining experimental evolution with next-generation sequencing: a powerful tool to study adaptation from standing genetic variation. Heredity 114(5):431–440.
Schrader L, Schmitz J. 2019. The impact of transposable elements in adaptive evolution. Mol Ecol. 28(6):1537–1549.
Sherratt E, Rasmussen AR, Sanders KL. 2018. Trophic specialization drives morphological evolution in sea snakes. R Soc Open Sci. 5(3):172141.
Silva JC, Loreto EL, Clark JB. 2004. Factors that affect the horizontal transfer of transposable elements. Curr Issues Mol Biol. 6(1):57–72.
Smit A, Hubley R. RepeatModeler version open-1.11 (downloaded 2017). Available from: http://www.repeatmasker.org/RepeatModeler/
Sobecky PA, Hazen TH. 2009. Horizontal Gene Transfer and Mobile Genetic Elements in Marine Systems. In: Gogarten MB, Gogarten JP, Olendzenski LC, editors. Horizontal gene transfer: genomes in flux. Methods in molecular biology. Totowa (NJ): Humana Press. p. 435–453. Available from: 10.1007/978-1-60327-853-9_25
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313.
Suh A, Smeds L, Ellegren H. 2018. Abundant recent activity of retrovirus-like retrotransposons within and among flycatcher species implies a rich source of structural variation in songbird genomes. Mol Ecol. 27(1):99–111.
Suh A, et al. 2016. Ancient horizontal transfers of retrotransposons between birds and ancestors of human pathogenic nematodes. Nat Commun. 7:11396.
Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. 56(4):564–577.

Vieira C, Nardon C, Arpin C, Lepetit D, Biémont C. 2002. Evolution of genome size in Drosophila. Is the invader's genome being invaded by transposable elements? Mol Biol Evol. 19(7):1154–1161.
Volff J-N, Körting C, Schartl M. 2000. Multiple lineages of the non-LTR retrotransposon Rex1 with varying success in invading fish genomes. Mol Biol Evol. 17(11):1673–1684.
Warren IA, et al. 2015. Evolutionary impact of transposable elements on genomic diversity and lineage-specific innovation in vertebrates. Chromosome Res. 23(3):505–531.
Wicker T, et al. 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 8(12):973–982.
Wickham H. 2011. ggplot2. Wiley Interdiscip. WIREs Comp Stat. 3(2):180–185.
Zeng L, Kortschak RD, Raison JM, Bertozzi T, Adelson DL. 2018. Superior *ab initio* identification, annotation and characterisation of TEs and segmental duplications from genome assemblies. PLoS One 13:e0193588.
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):1362.
Zimmermann L, et al. 2018. A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. J Mol Biol. 430(15):2237–2243.

Associate editor: Gonzalez Josefa

Horizontal transfer and subsequent explosive expansion of a DNA transposon in sea kraits (*Laticauda*)

"... no occurrence is sole and solitary, but is merely a repetition of a thing which has happened before, and perhaps often." - Mark Twain

While annotating the olive sea snake genome and identifying repeated horizontal transfer of retrotransposons into sea snakes, I identified one retrotransposon which had also been horizontally transferred into a sea krait. Within Hydrophiinae, sea kraits are the basal lineage to sea snakes and the terrestrial hydrophiines. While the genome of the yellow-lipped sea krait (*Laticauda colubrina*) had been sequenced, little attention was paid to the TEs present. Since sea kraits are a sister lineage to sea snakes that also transitioned to a marine habitat, independently of sea snakes, my discovery of a potential HTT event within sea kraits highlighted the importance of annotating TEs across the entire hydrophiinae lineage. In accordance with this goal, I performed an in-depth investigation of the mobilomes of sea kraits.

All supplementary data for this chapter can be found at
github.com/jamesdgalbraith/thesis_supplementary_material/tree/main/Chapter_3

Statement of Authorship

Title of Paper	Horizontal transfer and subsequent explosive expansion of a DNA transposon in sea kraits (Laticauda)
Publication Status	☐ Published ☐Accepted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style ☒ Submitted for Publication
Publication Details	James D. Galbraith, Alastair J. Ludington, Kate L. Sanders, Alexander Suh, David L. Adelson (2021), Horizontal transfer and subsequent explosive expansion of a DNA transposon in sea kraits (Laticauda), Submitted as a Research Article to Open Biology

Principal Author

Name of Principal Author (Candidate)	James D. Galbraith		
Contribution to the Paper	Designed and performed analysis, interpreted results and wrote manuscript		
Overall percentage (%)	85%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	24/06/2021

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Alastair J. Ludington		
Contribution to the Paper	Assisted in analysing the results and writing the manuscript		
Signature		Date	22/06/2021

Name of Co-Author	Kate L. Sanders		
Contribution to the Paper	Assisted in analysing the results and writing the manuscript		
Signature		Date	24/06/2021

Please cut and paste additional co-author panels here as required.

Statement of Authorship

Name of Co-Author	Alexander Suh		
Contribution to the Paper	Supervised the development of the work and assisted in analysing the results and writing the manuscript		
Signature		Date	2021-06-21

Name of Co-Author	David L. Adelson		
Contribution to the Paper	Supervised the development of the work and assisted in analysing the results and writing the manuscript		
Signature		Date	2021-06-22

Horizontal transfer and subsequent explosive expansion of a DNA transposon in sea kraits (*Laticauda*)

James D. Galbraith[1], Alastair J. Ludington[1], Kate L. Sanders[1], Alexander Suh*[2,3], David L. Adelson*[1]

1) School of Biological Sciences, University of Adelaide, Adelaide, SA 5005, Australia

2) School of Biological Sciences, University of East Anglia, Norwich Research Park, NR4 7TU, Norwich, United Kingdom

3) Department of Organismal Biology - Systematic Biology, Evolutionary Biology Centre, Uppsala University, SE-752 36 Uppsala, Sweden

* Joint corresponding authors, equally contributed to the paper.

Email: david.adelson@adelaide.edu.au

James: https://orcid.org/0000-0002-1871-2108

Alastair: https://orcid.org/0000-0003-3994-6023

Alexander: https://orcid.org/0000-0002-8979-9992

Kate: https://orcid.org/0000-0002-9581-268X

David: https://orcid.org/0000-0003-2404-5636

Classification

Biological Sciences: Evolution

Keywords

horizontal transfer, transposable element, Serpentes

Author Contributions

J.D.G., A.S and D.L.A. designed research; D.L.A. and A.S supervised research; K.L.S. and A.L. provided olive sea snake genome assembly; J.D.G. performed research; and J.D.G., K.L.S. and D.L.A. wrote the paper with input from A.L. and A.S.

Article type: Research Articles

Abstract

Transposable elements (TEs) are self replicating genetic sequences and are often described as important “drivers of evolution”. This driving force is because TEs promote genomic novelty by enabling rearrangement, and through exaptation as coding and regulatory elements. However, most TE insertions will be neutral or harmful, therefore host genomes have evolved machinery to supress TE expansion. Through horizontal transposon transfer (HTT) TEs can colonise new genomes, and since new hosts may not be able to shut them down, these TEs may proliferate rapidly. Here we describe HTT of the *Harbinger-Snek* DNA transposon into sea kraits (*Laticauda*), and its subsequent explosive expansion within *Laticauda* genomes. This HTT occurred following the divergence of *Laticauda* from terrestrial Australian elapids ~15-25 Mya. This has resulted in numerous insertions into introns and regulatory regions, with some insertions into exons which appear to have altered UTRs or added sequence to coding exons. *Harbinger-Snek* has rapidly expanded to make up 8-12% of *Laticauda* spp. genomes; this is the fastest known expansion of TEs in amniotes following HTT. Genomic changes caused by this rapid expansion may have contributed to adaptation to the amphibious-marine habitat.

Introduction

Transposable elements (TE) are selfish genetic elements that mobilize themselves across the genome. A substantial proportion of eukaryotic genomes is composed of TEs, with most reptilian and mammalian genomes comprising between 30 and 60%. As TEs proliferate within a genome, most insertions will be either neutral or deleterious [1]. However, over evolutionary timescales the movement of TEs can enable major adaptive change; being exapted as coding and regulatory sequences, and by promoting both inter- and intra-chromosomal rearrangements such as segmental duplications, inversions and deletions through non-allelic homologous recombination [2,3].

TE expansion can also be harmful, driving eukaryotes to evolve various defence and regulatory mechanisms. Genomic shocks can disrupt this regulation, allowing TEs to expand [4]. One example of a shock is horizontal transposon transfer (HTT), in which a TE jumps from one species to another. While the exact mechanisms of HTT are unknown, many instances across eukaryotes

have been reported [5–9]. Following HTT the expansion of new TEs is quickly slowed or halted due to the potentially deleterious effects they can cause [1,10], and any continued expansion will likely be slow. For example, following ancient HTT events the BovB retrotransposon has taken 32-39 My and 79-94 My for these elements to colonise between 6 and 18% of ruminant and Afrotheria genomes, respectively [6,11,12]. However rapid expansion of TEs following HT has previously been noted in *Myotis* bats, where *hAT* transposons expanded to cover 3.3% of the genome over the space of 15 Mya [13–15].

Here we report the HT of a *Harbinger* DNA transposon, *Harbinger-Snek*, into *Laticauda*, a genus of marine snakes which diverged from terrestrial Australian snakes 15-25 Mya [16–18]. Surprisingly, none of the available terrestrial animal genomes contained any trace of Harbinger-Snek, with highly similar sequences instead identified in sea urchins. Since diverging from terrestrial snakes *Laticauda* transitioned to amphibious-marine habits, foraging on coral reefs and returning to land only to digest prey, mate and lay eggs [19]. Due to the absence of *Harbinger-Snek*-like sequences from terrestrial species and highly similar sequences present in marine species, we propose *Harbinger-Snek* was horizontally transferred to *Laticauda* from a marine donor genome by habitat transition. Furthermore, since this initial HTT event, *Harbinger-Snek* has expanded rapidly within the genomes of *Laticauda* and now accounts for 8% of the *L. laticaudata* assembly and 12% of the *L. colubrina* assembly.

Methods

All scripts/code used at: https://github.com/jamesdgalbraith/Laticauda_HT

Ab initio repeat annotation of elapids

Using RepeatModeler2 [20] we performed *ab initio* annotation of the four Austro-Melanisian elapid genomes: *Laticauda colubrina* [21], *Notechis scutatus, Pseudonaja textilis,* and *Aipysurus laevis* [22]. To improve the RepeatModeler2 libraries we manually classified consensus sequences over 200 bp using a BLAST, extend, align and trim method, described by Galbraith et al. [23].

Identification of horizontal transfer and potential source/vectors

To identify any TEs restricted to a single lineage of elapid, we searched for all TEs identified by

RepeatModeler2 using BLASTN (-task dc-megablast) [24] in the three other assemblies, as well assemblies of the Asian elapids *Naja naja* [25] and *Ophiophagus hannah* [26]. TEs present in high numbers in a species, but not present in the other elapids, were considered potential HTT. This yielded a high copy number of *Harbinger* elements in *L. colubrina*. To rule out contamination, we searched for this element in a *L. laticaudata* genome assembly from GenBank. Using RPSBLAST [27] and the Pfam database [28] we identified *Harbinger* copies with intact protein-coding domains. To identify potential source or vector species, we searched all metazoan RefSeq genomes with a contig N50 of at least 10 kbp with BLASTN (-penalty -5 -reward 4 -out -word_size 11 -gapopen 12 -gapextend 8) . In species containing similar elements, we created consensus sequences using the aforementioned BLAST, extend, align and trim method. As we had identified similar *Harbinger* elements in fish, bivalves and echinoderms from RefSeq, we repeated this process for all GenBank assemblies of other species from these clades with a contig N50 of at least 10 kbp.

We identified transposase domains present in curated *Harbinger* sequences and all autonomous *Harbinger* elements available from Repbase [29] using RPSBLAST [27] and the Pfam database [28] . Using MAFFT (--localpair) [30] we created a protein multiple sequence alignment (MSA) of identified transposase domains. After trimming the MSA with Gblocks [31] we constructed a phylogenetic tree using FastTree [32] and from this tree chose an appropriate outgroup to use with curated elements. We subsequently constructed a protein MSA of the curated transposases using MAFFT, trimmed the MSA with Gblocks and created a phylogeny using IQ-TREE 2 (-m MFP -B 1000), which selected TVMe+I+G4 as the best model [33–35]. For comparison we also created phylogenies using the same MSA with MrBayes and RAxML [36,37]. To compare the repeat and species phylogenies, we created a species tree of major sampled animal taxa using TimeTree [38].

Potential interaction of *Harbinger-Snek* with genes

Using the improved RepeatModeler2 libraries and the Repbase (-lepidosaur) library, we used RepeatMasker [39] to annotate the two species of *Laticauda*. Using Liftoff [40] we transferred the *No. scutatus* gene annotation from RefSeq [41] to the *L. colubrina* and *L. laticaudata* genome assemblies. To identify *Harbingers* in genes, exons and regulatory regions we intersected the RepeatMasker intervals and transferred gene intervals using plyranges [42]. To test for potential

effects of these insertions on biological processes and molecular functions in *Laticauda* we ran PANTHER overrepresentation tests [43] of each using *Anolis carolinensis* as reference with genes annotated in *Laticauda* as a filter.

Continued expression of *Harbinger-Snek*

To test if *Harbinger-Snek* is expressed in *L. laticaudata* we aligned raw RNA-seq reads from four tissues to the *L. laticaudata* genome from Kishida et al. [21] (BioProject PRJDB7257) using STAR [44] and examined the location of intact *Harbinger-Snek* TEs in IGV [45]and exons in which we had identified *Harbinger* insertions.

Results and discussion

***Harbinger-Snek* is unlike transposons seen in terrestrial elapid snakes**

Our *ab initio* repeat annotation revealed a novel *Harbinger* DNA transposon in *L. colubrina*, *Harbinger-Snek*. Using BLASTN we found *Harbinger-Snek* present in both *L. colubrina* and *L. laticaudata*, but failed to identify any similar sequences in terrestrial relatives. *Harbingers* are a superfamily of transposons encoding two proteins, a transposase and a Myb-like DNA-binding protein [46]. While both are necessary for transposition [47], we identified multi-copy variants of *Harbinger-Snek* which encoded only one of the two proteins. These variants likely result from large deletions, and may be non-autonomous. In addition, we identified many short non-autonomous variants which retain the same TSDs and terminal motifs, yet encode no proteins.

Harbinger-Snek* was horizontally transferred to *Laticauda

Harbingers have previously been reported in a wide variety of aquatic vertebrates including fish, crocodilians and testudines, but not in terrestrial vertebrates [29]. Our repeat annotation of the *Laticauda, Aipysurus*, *Naja, Notechis* and *Pseudonaja* assemblies confirmed *Harbinger-Snek* is unique to the two *Laticauda* species examined and is the dominant transposable element in both species (Table 1). This absence from relatives suggested *Harbinger-Snek* was horizontally transferred into the ancestral *Laticauda* genome and our search of over 600 metazoan genome assemblies identified similar sequences only in echinoderms, bivalves and teleosts.

The nucleotide sequences most similar to *Harbinger-Snek* were identified in the purple sea urchin, *Strongylocentrotus purpuratus*, and were ~90% identical to the transposase coding region and ~88% identical to the DNA-binding protein. Based on a) high numbers of *Harbinger-Snek* in both species of *Laticauda* sampled and b) similar sequences only present in in marine species, we conclude that *Harbinger-Snek* was likely horizontally transferred to *Laticauda* following their divergence from terrestrial snakes 15-25 Mya, and prior to the crown group divergence of the eight recognised species in *Laticauda* (spanned by *L. colubrina and L. laticaudata*) ~15 Mya [16].

Our phylogenetic analysis (Figure 1) of similar *Harbinger* transposase sequences placed *Harbinger-Snek* in a strongly supported cluster with *Harbingers* found in two sea urchins, *S. purpuratus* and *Hemicentrotus pulcherrimus* (order Echinoida). Interestingly, neither Echinoida assembly contained more than 10 *Harbinger-Snek*-like transposons, none of which encode both proteins. *H. pulcherrimus Harbinger-Snek*-like transposons only contained the transposase, while the *S. purpuratus* assembly contained *Harbinger-Snek*-like transposons encoding either the transposase or the DNA binding protein. In addition, the species that cluster together elsewhere on the tree are not closely related, for example, the sister cluster to the *Laticauda*-Echinoidea cluster contains a variety of fish and bivalve species. The mismatch of the species tree and the transposase tree suggests horizontal transfer of *Harbinger-Snek*-like transposons may be widespread among these marine organisms.

Harbinger-Snek expanded rapidly in _Laticauda_ and is now much less active

Both the RepeatMasker annotation and BLASTN searches reveal a massive expansion in both *Laticauda* species, making up 8% of the *L. laticaudata* assembly and 12% of the larger *L. colubrina* assembly (Table 1, Figure 2). To become established within a host genome following horizontal transfer, TEs must rapidly proliferate, or be lost due to genetic drift or negative selection [48]. To our knowledge the largest previously described expansion of DNA transposons in amniotes following HT is that of *hAT*s in the bat *Myotis lucifugus* [13–15]. Following HT ~30 Mya, *hAT* transposons quickly expanded over 15 My at an estimated rate of ~0.7 Mbp/My and currently make up ~3.3% of the *M. lucifugus* genome. Using the upper bound of *Harbinger-Snek*'s transfer of 25

My (directly after their divergence from terrestrial Australian snakes), we calculate *Harbinger-Snek* to have expanded in *L. colubrina* at a rate of 11.3 Mbp/My and in *L. laticauda* a rate of 8.12 Mby/My. Therefore, our finding is the largest described expansion of a TE in an amniote following HTT.

Mass expansion of existing TEs during speciation has previously been seen in many groups including primates [49], woodpeckers [50] and salmonids [51]. By making the genome more dynamic these expansions fostered rapid adaptations. The sharp peak in the divergence profile (Figure 2) indicates *Harbinger-Snek's* expansion was rapid, and the small number of near-identical copies suggests expansion has slowed massively, especially in *L. laticaudata*. Many more copies of *Harbinger-Snek* able to transpose are present in the *L. colubrina* assembly than the *L. laticaudata* assembly, with only 1 fully intact copy in *L. laticaudata*, but 269 in *L. colubrina*. Our alignment of *L. laticaudata* RNA-seq data from four tissues (vomeronasal organ, nasal cavity, tongue and liver) to the *L. laticaudata* genome revealed reads mapping across both coding regions of the intact copy of *Harbinger-Snek*. Therefore, *Harbinger-Snek* and its non-autonomous derivatives may still be transposing in *L. laticaudata*.

In addition to containing many more intact copies of the full element, *Laticauda colubrina* also contains a higher number of the aforementioned "solo-ORF" variants than *L. laticaudata*, with 2263 intact transposase only variants compared to 35, and 452 intact DNA binding protein only variants compared to 6. Based on this stark contrast, since divergence ~15 Mya [16] either *L. colubrina* has maintained a higher rate of *Harbinger-Snek* expansion or *L. laticaudata* has had a higher rate of *Harbinger-Snek* loss; or more likely, a combination of these two effects.

The accordion model - the expansion of *Harbinger-Snek* has been balanced by loss in *L. laticaudata*

The peak in *Harbinger-Snek* expansion in *L. colubrina* is both higher and more recent than *L. laticaudata* (Figure 2). In addition *L. laticaudata* has a much lower overall *Harbinger-Snek* content and genome size (Table 1). Past observations in birds, mammals and squamates found increases in genome size due to transposon expansion are balanced by loss due to deletions through non-allelic homologous recombination (NAHR) [52,53]. We expect that the mass expansion of

Harbinger-Snek in *Laticauda* has generated many near identical sites in the genome, in turn promoting NAHR. In spite of the explosive expansion of *Harbinger-Snek* in *L. laticaudata*, the genome size and total TE content is very similar to that of the terrestrial *Pseudonaja* and *Notechis* (Table 1). This retention of a similar genome size is not seen in *L. colubrina*, the genome assembly of which is 20% larger than the terrestrial species. However, the overall TE content of the *L. colubrina* genome remains similar to that of *L. laticaudata* and the terrestrial species, with the expansion of TEs only contributing half of the total increase in genome size. This is consistent with the aforementioned balancing of TE expansion by deletions.

Expansion of *Harbinger-Snek* has potentially impacted gene function

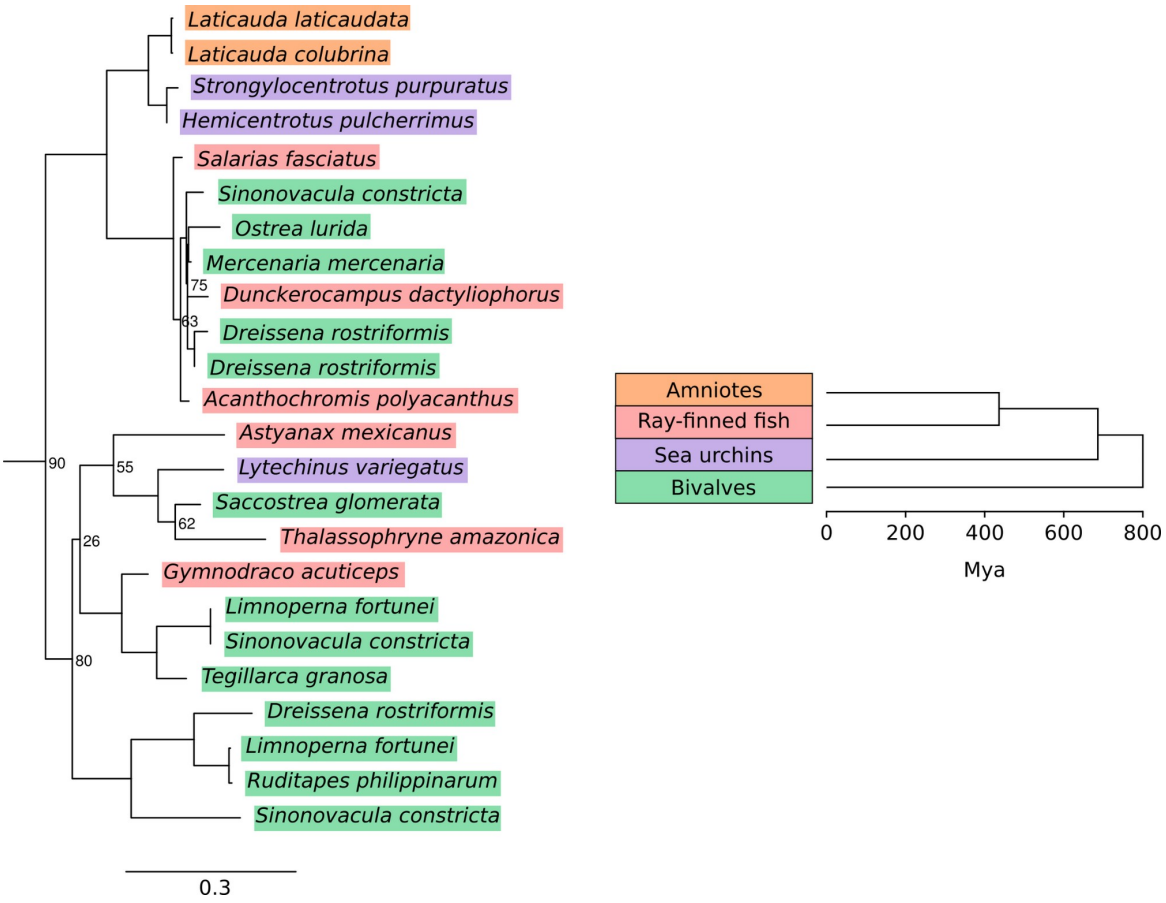
In both species of *Laticauda* many insertions of *Harbinger-Snek* overlap with or are contained within exons, regulatory regions and introns. Insertions overlapped with the exons of 56 genes in *L. colubrina* and 31 in *L. laticaudata*, 17 of which are shared (SI Table 1). By manually inspecting transcripts mapped to the *L. laticaudata* genome we determined 8 3' UTRs and 2 coding exons predicted by Liftoff now contain *Harbinger-Snek* insertions which contribute to mRNA (SI Table 1). These genes have a wide range of functions, many of which could be significant in the context of adaptation. We also identified insertions into 1685 and 888 potentially regulatory regions (within 5 kbp of the 5' UTR in genes) and into introns of 4141 and 1440 genes in *L. colubrina* and *L. laticauda* respectively. PANTHER over/under-representation tests of these in gene and regulatory region insertions identified a number of pathways of potential adaptive significance (SI Tables 2-5). Therefore, *Harbinger-Snek* is a prime candidate in the search for genomic changes responsible for *Laticauda's* adaptation to a marine environment through altered gene expression.

Conclusion

In this report, we describe the rapid expansions of *Harbinger-Snek* TEs in *Laticauda* spp., to our knowledge, the fastest expansion of a DNA transposon in amniotes reported to date. The large number of insertions of *Harbinger-Snek* into exons and regulatory regions may have contributed to speciation and adaptation to a new habitat; this suggests a number of future lines of investigation. As the HTT was prior to the divergence of 8 *Laticauda* species, *Harbinger-Snek* presents a unique

235 opportunity to reconstruct subsequent molecular evolution and determine the impact of HTT on the adaptation of *Laticauda* to the amphibious-marine habitat.

Figures/Tables



240 Figure 1. **The absence of *Harbinger-Snek* from terrestrial vertebrates and its highest similarity to *Harbingers* present in sea urchins support its horizontal transfer to *Laticauda* since transitioning to a marine habitat.** Nodes without support values have support of 95% or higher. The distribution of species across this tree suggests *Harbinger-Snek*-like transposons were horizontally transferred into a wide variety of species. This figure is an extract of a maximum 245 likelihood phylogeny constructed from the aligned nucleotide sequences of the transposases present in curated elements using IQ-TREE 2 [33], for the full tree see SI Figure 1. We also reconstructed trees with similar topologies using RAxML and MrBayes (see methods). Species phylogeny constructed with TimeTree [38].

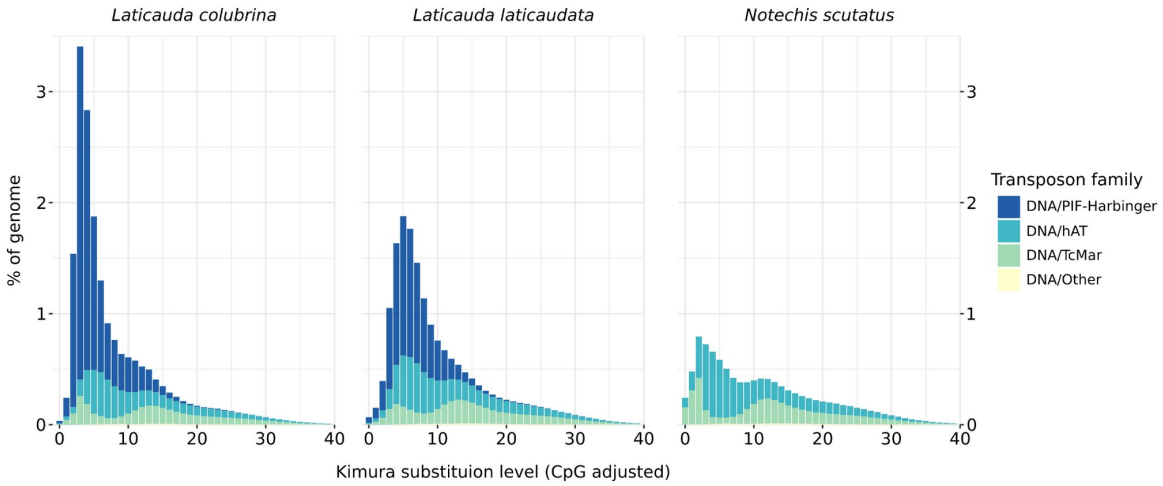


Figure 2. **Rapid, recent expansion of *Harbinger-Snek* PIF-Harbinger transposons.** Horizontal transfer of this transposon into the *Laticauda* ancestor has occurred within the past 15-25 My [16] . Due to expansions since then, these transposons have become the dominant DNA transposon in *Laticauda* genomes, in contrast to the genomes of their closest terrestrial relatives such as *Notechis scutatus* (diverged ~15-25 Mya). Repeat content calculated with RepeatMasker [39].

	Terrestrial	*L. colubrina*		*L. laticaudata*	
Retrotransposons			Diff. Mbp (%)		Diff. Mbp (%)
SINEs (Mbp)	25.81	24.31	-1.27 (-0.06%)	24.57	-1.00 (-0.06%)
Penelopes (Mbp)	33.19	42.34	+9.20 (0.45%)	45.28	+12.15 (0.78%)
LINEs (Mbp)	277.65	262.89	-9.33 (-0.46%)	235.46	-36.76 (-2.36%)
LTR elements (Mbp)	175.52	202.06	+27 (1.33%)	131.33	-43.73 (-2.81%)
DNA transposons					
hATs (Mbp)	88.63	79.33	-6.92 (-0.34%)	77.62	-8.63 (-0.55%)
Tc1/Mariners (Mbp)	61.56	57.80	-1.11 (-0.05%)	55.43	-3.48 (-0.22%)
Harbinger (Mbp)	0.44	229.84	+229.42 (11.33%)	126.84	+126.42 (8.11%)
Rolling-circles (Mbp)	3.24	3.09	-0.13 (-0.01%)	3.01	-0.20 (-0.01%)
Unclassified (Mbp)	165.40	140.72	-20.15 (-1.00%)	134.11	-26.77 (-1.72%)
Total TEs (Mbp)	798.05	999.63	+217.30 (10.73%)	788.05	5.72 (0.37%)
Assembly size (Mbp)	1,665.53	2,024.69	+396.91 (19.60%)	1,558.71	-69.01 (-4.43%)

Table 1: **The expansion of *Harbinger* elements in *Laticauda* spp.** This expansion, along with that of LTR elements, in *L. colubrina* has contributed to *L. colubrina having* a larger genome than terrestrial species. This gain in *L. laticaudata* appears to have been offset to some degree by loss from other TE families. Mbp or percentage difference in assembly repeat content between *Laticauda* and the average of the terrestrial *Notechis scutatus* and *Pseudonaja textilis*. Repeat content was annotated using RepeatMasker [39] using a combined Repbase [29] and curated RepeatModeler2 [20] library.

Supplementary Information

SI Table 1 - *Laticauda colubrina and Laticauda laticaudata* genes with *Harbinger-Snek* insertions into or overlapping open reading frames, and any noticeable effects on insertion noted from transcript data. Gene coordinates predicted with Liftoff [40] using the RefSeq *Notechis scutatus* assembly and gene annotation as reference. Repeat annotation performed with RepeatMasker [39] using a custom repeat library (see Methods). Intersect performed using BEDTools [54]. Transcripts

mapped to the genome assembly using STAR [44] and viewed in IGV [45].

SI Table 2 - Biological processes with an over/under-representation of *Harbinger-Snek* insertions into *Laticauda colubrina* genes. Representation test performed using PANTHER [43]. Gene coordinates predicted with Liftoff [40] using the RefSeq *Notechis scutatus* assembly and gene annotation as reference. Repeat annotation performed with RepeatMasker [39] using a custom repeat library (see Methods). Intersect performed using plyranges [42].

SI Table 3 - Molecular functions with an over/under-representation of *Harbinger-Snek* insertions into *Laticauda colubrina* genes. Representation test performed using PANTHER [43]. Gene coordinates predicted with Liftoff [40] using the RefSeq *Notechis scutatus* assembly and gene annotation as reference. Repeat annotation performed with RepeatMasker [39] using a custom repeat library (see Methods). Intersect performed using plyranges [42].

SI Table 4 - Biological processes with an over/under-representation of *Harbinger-Snek* insertions into potential regulatory regions of *Laticauda colubrina* genes. Representation test performed using PANTHER [43]. Gene coordinates predicted with Liftoff [40] using the RefSeq *Notechis scutatus* assembly and gene annotation as reference. Repeat annotation performed with RepeatMasker [39] using a custom repeat library (see Methods). Intersect performed using plyranges [42].

SI Table 5 - Molecular functions with an over/under-representation of *Harbinger-Snek* insertions into potential regulatory regions of *Laticauda colubrina* genes. Representation test performed using PANTHER [43]. Gene coordinates predicted with Liftoff [40] using the RefSeq *Notechis scutatus* assembly and gene annotation as reference. Repeat annotation performed with RepeatMasker [39] using a custom repeat library (see Methods). Intersect performed using plyranges [42].

SI Table 6 - Latin species names and versions of all public genomes used. All were downloaded from RefSeq [41] when available, else from GenBank [55].

References

1. Cosby RL, Chang N-C, Feschotte C. 2019 Host-transposon interactions: conflict, cooperation, and cooption. *Genes Dev.* **33**, 1098–1116.

2. Bourque G. 2009 Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Curr. Opin. Genet. Dev.* **19**, 607–612.

3. Warren IA, Naville M, Chalopin D, Levin P, Berger CS, Galiana D, Volff J-N. 2015 Evolutionary impact of transposable elements on genomic diversity and lineage-specific innovation in vertebrates. *Chromosome Res.* **23**, 505–531.

4. 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.

5. El Baidouri M *et al.* 2014 Widespread and frequent horizontal transfers of transposable elements in plants. *Genome Res.* **24**, 831–838.

6. Ivancevic AM, Kortschak RD, Bertozzi T, Adelson DL. 2018 Horizontal transfer of BovB and L1 retrotransposons in eukaryotes. *Genome Biol.* **19**, 85.

7. 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.

8. Reiss D, Mialdea G, Miele V, de Vienne DM, Peccoud J, Gilbert C, Duret L, Charlat S. 2019 Global survey of mobile DNA horizontal transfer in arthropods reveals Lepidoptera as a prime hotspot. *PLoS Genet.* **15**, e1007965.

9. 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**, 1362.

10. Gilbert C, Feschotte C. 2018 Horizontal acquisition of transposable elements and viral sequences: patterns and consequences. *Curr. Opin. Genet. Dev.* **49**, 15–24.

11. Foley NM, Springer MS, Teeling EC. 2016 Mammal madness: is the mammal tree of life not yet resolved? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **371**. (doi:10.1098/rstb.2015.0140)

12. Chen L *et al.* 2019 Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. *Science* **364**. (doi:10.1126/science.aav6202)

13. Ray DA, Feschotte C, Pagan HJT, Smith JD, Pritham EJ, Arensburger P, Atkinson PW, Craig NL. 2008 Multiple waves of recent DNA transposon activity in the bat, Myotis lucifugus. *Genome Res.* **18**, 717–728.

14. Pace JK 2nd, Gilbert C, Clark MS, 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.

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

16. Sanders KL, Lee MSY, Leys R, Foster R, Keogh JS. 2008 Molecular phylogeny and divergence dates for Australasian elapids and sea snakes (hydrophiinae): evidence from seven genes for rapid evolutionary radiations. *J. Evol. Biol.* **21**, 682–695.

17. Sanders KL, Mumpuni, Lee MSY. 2010 Uncoupling ecological innovation and speciation in sea snakes (Elapidae, Hydrophiinae, Hydrophiini). *J. Evol. Biol.* **23**, 2685–2693.

18. Lee MSY, Sanders KL, King B, Palci A. 2016 Diversification rates and phenotypic evolution in venomous snakes (Elapidae). *R Soc Open Sci* **3**, 150277.

19. Mirtschin P, Rasmussen A, Weinstein S. 2017 Australia's Dangerous Snakes. (doi:10.1071/9780643106741)

20. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020 RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9451–9457.

21. Kishida T, Go Y, Tatsumoto S, Tatsumi K, Kuraku S, Toda M. 2019 Loss of olfaction in sea snakes provides new perspectives on the aquatic adaptation of amniotes. *Proc. Biol. Sci.* **286**, 20191828.

22. Ludington AJ, Sanders KL. 2021 Demographic analyses of marine and terrestrial snakes (Elapidae) using whole genome sequences. *Mol. Ecol.* **30**, 545–554.

23. Galbraith JD, Ludington AJ, Suh A. 2020 New Environment, New Invaders—Repeated Horizontal Transfer of LINEs to Sea Snakes. *Genome Biol. Evol.*

24. Zhang Z, Schwartz S, Wagner L, Miller W. 2000 A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7**, 203–214.

25. Suryamohan K *et al.* 2020 The Indian cobra reference genome and transcriptome enables comprehensive identification of venom toxins. *Nat. Genet.* **52**, 106–117.

26. Vonk FJ *et al.* 2013 The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 20651–20656.

27. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.

28. Mistry J *et al.* 2021 Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419.

29. Bao W, Kojima KK, Kohany O. 2015 Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11.

30. Katoh K, Standley DM. 2013 MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780.

31. Castresana J. 2000 Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552.

32. Price MN, Dehal PS, Arkin AP. 2010 FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490.

33. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020 IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534.

34. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589.

35. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018 UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522.

36. Huelsenbeck JP, Ronquist F. 2001 MRBAYES: Bayesian inference of phylogenetic trees.

Bioinformatics **17**, 754–755.

37. Stamatakis A. 2014 RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313.

38. Kumar S, Stecher G, Suleski M, Hedges SB. 2017 TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol. Biol. Evol.* **34**, 1812–1819.

39. SMIT, A. 2004 Repeat-Masker Open-3.0. *http://www.repeatmasker.org*

40. Shumate A, Salzberg SL. 2020 Liftoff: accurate mapping of gene annotations. *Bioinformatics* (doi:10.1093/bioinformatics/btaa1016)

41. O'Leary NA *et al.* 2016 Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–45.

42. Lee S, Cook D, Lawrence M. 2019 plyranges: a grammar of genomic data transformation. *Genome Biol.* **20**, 4.

43. Mi H, Ebert D, Muruganujan A, Mills C, Albou L-P, Mushayamaha T, Thomas PD. 2021 PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res.* **49**, D394–D403.

44. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013 STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21.

45. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011 Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26.

46. Kapitonov VV, Jurka J. 2004 Harbinger transposons and an ancient HARBI1 gene derived from a transposase. *DNA Cell Biol.* **23**, 311–324.

47. Sinzelle L, Kapitonov VV, Grzela DP, Jursch T, Jurka J, Izsvák Z, Ivics Z. 2008 Transposition of a reconstructed Harbinger element in human cells and functional homology with two transposon-derived cellular genes. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 4715–4720.

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

49. Pace JK 2nd, Feschotte C. 2007 The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. *Genome Res.* **17**, 422–432.

50. Manthey JD, Moyle RG, 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.

51. de Boer JG, Yazawa R, Davidson WS, Koop BF. 2007 Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. *BMC Genomics* **8**, 422.

52. Kapusta A, Suh A, Feschotte C. 2017 Dynamics of genome size evolution in birds and mammals. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E1460–E1469.

53. Pasquesi GIM *et al.* 2018 Squamate reptiles challenge paradigms of genomic repeat element evolution set by birds and mammals. *Nat. Commun.* **9**, 2774.

54. Quinlan AR, Hall IM. 2010 BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842.

55. Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2015 GenBank. *Nucleic Acids Res.* **43**, D30–5.

Horizontal transfer and southern migration: the tale of Hydrophiinae's marine journey.

"Every great story seems to begin with a snake." - Nicolas Cage

Following my discovery of multiple, independent horizontal transfers into two distinct lineages of marine hydrophiines, two questions became apparent. Is horizontal transfer common in aquatic environments? Or are squamates, more specifically hydrophiines, highly susceptible to horizontal transfer? In favour of the first question, recent studies have identified repeated horizontal transfer in fish and bivalves. Alternatively, in favour of the second question, many past studies into horizontal transfer have found DNA transposons and retrotransposons horizontally transferred into other reptiles such as pythons and various lizards. To gain insight into these questions I performed comprehensive analyses of the mobilome of two terrestrial hydrophiines, a sea snake and a sea krait, using two Asian elapid as outgroups. By doing so I was able to investigate how the TE landscape of hydrophiines had evolved since their divergence from Asian elapids 25-30 Mya. This would allow me to determine if HTT had occurred in ancestral lineages and ascertain if the variability in TEs seen across squamates is present within a single family.

All supplementary data for this chapter can be found at
github.com/jamesdgalbraith/thesis_supplementary_material/tree/main/Chapter_4

Statement of Authorship

Title of Paper	Horizontal transfer and southern migration: the tale of Austro-Melanesian elapids' marine journey.
Publication Status	☐ Published ☐Accepted for Publication ☒ Unpublished and Unsubmitted work written in manuscript style ☐ Submitted for Publication
Publication Details	James D. Galbraith, Alastair J. Ludington, Richard J. Edwards, Kate L. Sanders, Alexander Suh, David L. Adelson (2021) Horizontal transfer and southern migration: the tale of Austro-Melanesian elapids' marine journey. Prepared as a submission as a Research Article to Proceedings of the Royal Society B.

Principal Author

Name of Principal Author (Candidate)	James D. Galbraith		
Contribution to the Paper	Designed and performed analysis, interpreted results and wrote manuscript		
Overall percentage (%)	75%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	24/06/2021

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Alastair J. Ludington		
Contribution to the Paper	Assisted in analysing the results and writing the manuscript		
Signature		Date	22/06/2021

Name of Co-Author	Kate L. Sanders		
Contribution to the Paper	Assisted in analysing the results and writing the manuscript		
Signature		Date	24/06/2021

Please cut and paste additional co-author panels here as required.

Statement of Authorship

Name of Co-Author	Alexander Suh
Contribution to the Paper	Supervised the development of the work and assisted in analysing the results and writing the manuscript
Signature	Date: 2021-06-21

Name of Co-Author	David L. Adelson
Contribution to the Paper	Supervised the development of the work and assisted in analysing the results and writing the manuscript
Signature	Date: 2021-06-22

Name of Co-Author	RICHARD EDWARDS
Contribution to the Paper	ASSEMBLY OF TIGER SNAKE AND EASTERN BROWN SNAKE GENOMES
Signature	Date: 21/6/21

Horizontal transfer and southern migration: the tale of Hydrophiinae's marine journey.

James D. Galbraith[1], Alastair J. Ludington[1], Richard J. Edwards[2], Kate L. Sanders[1], Alexander Suh*[3,4], David L. Adelson*[1]

1) School of Biological Sciences, University of Adelaide, Adelaide, SA 5005, Australia

2) School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW 2052, Australia

3) School of Biological Sciences, University of East Anglia, Norwich Research Park, NR4 7TU, Norwich, United Kingdom

4) Department of Organismal Biology - Systematic Biology, Evolutionary Biology Centre, Uppsala University, SE-752 36 Uppsala, Sweden

* David L. Adelson and Alexander Suh corresponding authors and contributed equally to this work.

Email: david.adelson@adelaide.edu.au

James: https://orcid.org/0000-0002-1871-2108

Alastair: https://orcid.org/0000-0003-3994-6023

Alexander: https://orcid.org/0000-0002-8979-9992

Kate: https://orcid.org/0000-0002-9581-268X

David: https://orcid.org/0000-0003-2404-5636

Classification

Biological Sciences: Evolution

Keywords

transposable element, comparative genomics, Serpentes

Author Contributions

J.D.G., A.S and D.L.A. designed research; D.L.A. and A.S supervised research; R.J.E. provided olive sea snake genome assembly; J.D.G. and A.L. performed research; and J.D.G., A.S. and D.L.A. wrote the paper with input from A.J.L. and K.L.S.

Abstract

Transposable elements (TEs), also known as jumping genes, are sequences able to move or copy themselves within a genome. As TEs move throughout genomes they can be exapted as coding and regulatory elements, or can promote genetic rearrangement. In so doing TEs act as a source of genetic novelty, hence understanding TE evolution within lineages is key in understanding adaptation to their environment. Studies into the TE content of lineages of mammals such as bats have uncovered horizontal transposon transfer (HTT) into these lineages, with squamates often also containing the same TEs. Despite the repeated finding of HTT into squamates, little comparative research has examined the evolution of TEs within squamates. The few broad scale studies in Squamata which have been conducted found both the diversity and total number of TEs differs significantly across the entire order. Here we examine a diverse family of Australo-Melanesian snakes (Hydrophiinae) to examine if this pattern of variable TE content and activity holds true on a smaller scale. Hydrophiinae diverged from Asian elapids ~15-25 Mya and have since rapidly diversified into 6 amphibious, ~60 marine and ~100 terrestrial species which fill a broad range of ecological niches. We find TE diversity and expansion differs between hydrophiines and their Asian relatives and identify multiple HTTs into Hydrophiinae, including three transferred into the ancestral hydrophiine likely from marine species. These HTT events provide the first tangible evidence that Hydrophiinae reached Australia from Asia via a marine route.

Introduction

Elapids are a diverse group of venomous snakes found across Africa, Asia, the Americas and Australia. Following their divergence from Asian elapids ~30 Mya, the Australo-Melanesian elapids (Hydrophiinae) have rapidly diversified into more than 160 species including ~100 terrestrial snakes, ~60 fully marine sea snakes, and 6 amphibious sea kraits [1]. Both the terrestrial and fully marine hydrophiines have adapted to a wide range of habitats and niches. Terrestrial Hydrophiinae are found across Australia, for example the eastern brown snake *Pseudonaja textilis* in open habitats, the tiger snake (*Notechis scutatus*) in subtropical and temperate habitats, and the inland taipan (*Oxyuranus microlepidotus*) to inland arid habitats [2]. Since transitioning to a marine habitat, many sea snakes have specialised to feed on a single prey such as fish eggs, catfish, eels or burrowing gobies, while others such as *Aipysurus laevis* are generalists [3,4]. Sea kraits (*Laticauda*) are amphibious and have specialised to hunt various fish including eels and anguilliform-like fish at sea, while digesting prey, mating and shedding on land [5]. Since transitioning to marine environments, both sea snakes and sea kraits have been the recipients of multiple independent horizontal

transposon transfer (HTT) events, which may have had adaptive potential [6,7].

Transposable elements (TEs) are mobile genetic elements that can move or copy themselves across the genome, and account for a large portion of most vertebrate genomes [8,9]. Though often given short shrift in genome analyses, TEs are important agents of genome evolution and generate genomic diversity [10,11]. For example, the envelope gene of endogenous retroviruses was exapted by both mammals and viviparous lizards to function in placental development [12]. In addition, unequal crossing over caused by CR1 retrotransposons led to the duplication, and hence diversification, of PLA_2 venom genes in pit vipers [13]. Transposable elements (TEs) are classified into one of two major classes based on their structure and replication method [14]. DNA transposons (Class II) proliferate through a “cut and paste” method, possess terminal inverted repeats and are further split based on the transposase sequence used in replication. Retrotransposons (Class I) are split into LTR retrotransposons and non-LTR retrotransposons, which proliferate through ``copy and paste” methods. Both subclasses of retrotransposons are split into numerous superfamilies based on both coding and structural features [15–17].Within the diverse lineages of higher vertebrates, the evolution of TEs is well described in eutherian mammals and birds. The total repetitive content of both bird and mammal genomes is consistently at 7-10% and 30-50% respectively. Similarly, most lineages of both birds and eutherian mammals are dominated by a single superfamily of non-LTR retrotransposons (CR1s and L1s respectively) and a single superfamily of LTR retrotransposons (endogenous retroviruses in both) [8,18]. Some lineages of birds and mammals contain horizontally transferred retrotransposons which have variably been successful (AviRTE and RTE-BovB respectively) [19,20].

In stark contrast to mammals and birds, squamates have highly variable mobilomes, both in terms of the diversity of their TE superfamilies and the level activity of said superfamilies within each genome [21]. While these broad comparisons have found significant variation in TEs between distant squamate lineages, none have examined how TEs have evolved within a single family of squamates. The one in depth study into the mobilome of snakes found the Burmese python genome is approximately ~21% TE and appears to have low TE expansion, while that of a pit viper is ~45% TE due to the expansion of numerous TE superfamilies and microsatellites since their divergence ~90 Mya [22,23] . Unfortunately it is unclear whether similar expansions have occurred within other lineages of venomous snakes. Here we examine the TE landscape of the family Hydrophiinae, and in doing so discover horizontal transfer events into the ancestral hydrophiine, sea kraits and sea snakes.

Methods

Ab initio TE annotation of the elapid genomes

We used RepeatModeler2 [24] to perform *ab initio* TE annotation of the genome assemblies of 4 hydrophiines (*Aipysurus laevis*, *Notechis scutatus*, *Pseudonaja textilis* and *Laticauda colubrina*) and 2 Asian elapids (*Naja naja* and *Ophiophagus hannah*). We manually curated the subfamilies of TEs identified by RepeatModeler (rm-families) to ensure they encompassed the full TE, were properly classified and that each species' library was non-redundant.

We first purged redundant rm-families from each species library based on pairwise identity to and coverage by other rm-families within the library. Using BLAST [25] we calculated the similarity between all rm-families. Any rm-family with over 75% of its length aligning to a larger rm-family at 90% pairwise identity or higher was removed from the library. We then searched for each non-redundant rm-family within their source genome with BLASTN (-task dc-megablast) and selected the best 30 hits based on bitscore. In order to ensure we could retrieve full length TE insertions, we extended the flanks of each hit by 4000 bp. Using BLASTN (-task dc-megablast) we pairwise aligned each of the 30 extended sequences to others, trimming trailing portions of flanks which did not align to flanks of the other 29 sequences. Following this, we constructed a multiple sequence alignment (MSA) of the 30 trimmed sequences with MAFFT [26] (--localpair). Finally we trimmed each MSA at the TE target site duplications (TSDs) and constructed a consensus from the multiple sequence alignments using Geneious Prime 2021.1.1 (www.geneious.com) which we henceforth refer to as a mc-subfamily (manually curated subfamily).

To classify the mc-subfamilies we searched for intact protein domains in the consensus sequences using RPSBLAST [27] and the CDD library [28] and identified homology to previously described TEs in Repbase using CENSOR online [29] . Using this data in conjunction with the classification set out in Wicker (2007) [14], we classified previously unclassified mc-subfamilies where possible and corrected the classification of mc-subfamilies where necessary. Where possible we used the criteria of Feschotte and Pritham (2007) [17] to identify unclassified DNA transposons using TSDs and terminal inverted repeats. Finally, we removed any genes from the mc-subfamily libraries based on searches using online NCBI BLASTN and BLASTX searches against the nt/nr and UniProt libraries respectively [30,31]. Any mc-subfamilies unable to be classified were labelled as "Unknown".

TE annotation of the elapid genomes

We constructed a custom library for TE annotation of the elapid genome assemblies by combining the mc-subfamilies from the six assemblies with previously described lepidosaur TEs identified using RepeatMasker's "queryRepeatDatabase.pl" utility. Using RepeatMasker, we generated repeat annotations of all six elapid genome assemblies.

Estimating ancestral TE similarity

To estimate the sequence conservation of ancestral TEs, and hence categorise recently expanding TEs as either ancestral or horizontally transferred, we identified orthologous TE insertions and their flanks present in both the *Notechis scutatus* and *Naja naja* genome assemblies. From the *Notechis* repeat annotation, we took a random sample of 5000 TEs over 500 bp in length and extended each flank by 1000 bp. Using BLASTN (-task dc-megablast) we searched for the TEs and their flanks in the *Naja* assembly and selected all hits containing at least 250 bp of both the TE and the flank. Sequences with more than one hit containing flanks were treated as potential segmental duplications. We also removed any potential segmental duplications from the results. We then used the orthologous sequences to estimate the expected range in similarity between TEs present in the most recent common ancestor of Australian and Asian elapids. Based on this information, TEs with 95% or higher pairwise identity to the mc-subfamily used to identify them were treated as likely inserted in hydrophiine genomes since their divergence from Asian elapids. In addition, mc-subfamilies which we had identified as recently expanding in hydrophiines but were not found at 80% or higher pairwise identity in other serpentine genomes, were identified as candidates for horizontal transfer.

Identifying recent TE expansions

In each of the four hydrophiines, using the RepeatMasker output we identified mc-subfamilies comprising at least 100 total kbp having 95% or higher pairwise identity to the mc-subfamily. We treated these mc-subfamilies as having expanded since Hydrophiinae's divergence from Asian elapids. We reduced any redundancy between recently expanding mc-subfamilies by clustering Using CD-HIT-EST (-c 0.95 -n 6 -d 50) [32]. Using BWA [33], we mapped raw transcriptome reads of eye tissue taken from each of the hydrophiines [34] for back to these mc-subfamilies. Retrotransposons with RNA-seq reads mapping across their whole length and DNA transposons with RNA-seq reads mapping to their coding regions were treated as expressed and therefore currently expanding.

Continued expansion or horizontal transfer

Using BLASTN (-task dc-megablast), we searched for homologs of recently expanding mc-subfamilies in a

range of snake genomes including Asian elapids, colubrids, vipers and a python. We classified mc-subfamilies having copies of 80% or higher pairwise identity to the query sequence in other snakes as ancestral. All hydrophiine mc-subfamilies we were unable to find in other snakes were treated as candidates for horizontal transfer. We searched for the horizontal transfer candidates in approximately 600 additional metazoan genomes using BLASTN (-task dc-megablast). We classified all mc-subfamilies present in non-serpentine genomes at 80% or higher pairwise identity and absent from other serpentine genomes at 80% or higher pairwise identity as horizontally transferred into hydrophiines.

Results and Discussion

Genome quality affects repeat annotation

Previous studies have highlighted the importance of genome assembly quality in repeat annotation, with higher sequencing depths and long read technologies critical for resolving TEs [35,36]. Our repeat analysis reveals significant variation in total TE content between genome assemblies (Table 1, Figure 1), however some of this variation is likely due to large differences in assembly quality rather than differential TE expansions or contractions in certain lineages. Most notably, the TE content of the *Ophiophagus* assembly is significantly lower than that of that of the other species (~36% compared to ~46%). The TE content of the *Aipysurus* assembly is also notably lower, however to a lesser extent (41% compared to ~46%). The *Naja*, *Laticauda*, *Notechis*, and *Pseudonaja* assemblies are much higher quality assemblies than the *Ophiophagus* and *Aipysurus* assemblies, having longer contigs and scaffolds (SI Table 1). This discrepancy is because the *Ophiophagus* and *Aipysurus* genomes are both assembled solely from short read data with a low sequencing depth (28x and 30x respectively). In stark contrast the *Naja* genome was assembled from a combination of long read (PacBio and Oxford Nanopore) and short read (Illumina) data, scaffolded using Chicago and further improved using Hi-C and optical mapping (Bionano) technologies. In the middle ground, the *Laticauda*, *Notechis* and *Pseudonaja* assemblies utilized a combination of 10X Chromium linked read and short read technologies. Many of the recently expanded TEs in the *Ophiophagus* and *Aipysurus* genomes likely collapsed during assembly because of their very high sequence similarity. Therefore, the apparent lack of recent activity in *Ophiophagus* and *Aipysurus* is a likely artefact of assembly quality. As the total TE content annotated in the *Naja*, *Laticauda*, *Notechis* and *Pseudonaja* is comparable at 46-48% of the genome and the four genomes are of comparable quality, the majority of the following analyses focuses on these four species.

	Asian elapids		Hydrophiinae			
	Naja naja	***Ophiophagus hannah***	***Laticauda colubrina***	***Pseudonaja textilis***	***Notechis scutatus***	***Aipysurus laevis***
Class I (Retroelements)	30.58	19.73	23	27.81	27.29	25.91
Penelope	1.72	1.37	2.08	2.06	1.98	1.64
LINE/CR1	4.37	4.12	4.12	5.54	6.28	4.67
LINE/L1	3.24	2.21	2.94	4.04	3.4	2.72
LINE/L2	7.14	3.39	1.05	1.41	1.4	1.27
LINE/Rex-Babar	1.17	1.11	1.08	1.44	1.42	1.26
LINE/RTE	1.4	1.52	1.07	1.5	1.43	1.25
LINE/Other	0.62	0.69	0.54	0.66	0.64	0.59
SINE	0.37	0.43	0.3	0.4	0.39	0.36
LTR/Copia	0.71	0.47	1.02	1.27	1.65	0.89
LTR/DIRS	0.81	0.61	0.86	1.04	1.19	1.57
LTR/ERV	0.6	0.42	1.23	1.38	1.69	1.16
LTR/Gypsy	7.01	1.97	5.6	5.61	4.38	7.22
LTR/Other	1.42	1.42	1.11	1.46	1.44	1.31
Class II (DNA Transposons)	7.81	7.45	18.08	9.02	9.2	7.95
DNA/hAT	4.25	4.13	3.78	5.09	5.13	4.47
DNA/PIF-Harbinger	0.02	0.03	11.19	0.02	0.02	0.02
DNA/Tc1-Mariner	3.13	2.88	2.79	3.47	3.63	3.08
DNA/Other	0.22	0.24	0.17	0.24	0.22	0.2
RC/Helitron	0.19	0.17	0.15	0.2	0.2	0.18
Unknown	7.93	8.57	6.46	9.18	9.26	7.89
Total interspersed repeats	46.32	35.75	47.54	46.01	45.75	41.75
Satellite	0.17	0.18	0.13	0.17	0.17	0.18
Simple repeat	1.31	1.37	1.28	1.66	1.67	1.28

Table 1: **Repeat composition of hydrophiine and Asian elapid genome assemblies.** Variation in assembly repeat content varies both within hydrophiines and between hydrophiines and Asian elapids. Genome assemblies annotated with RepeatMasker [37] and a custom library of curated RepeatModeler2 libraries [24] and previously described lepidosaur TEs from the Repbase RepeatMasker library [38].

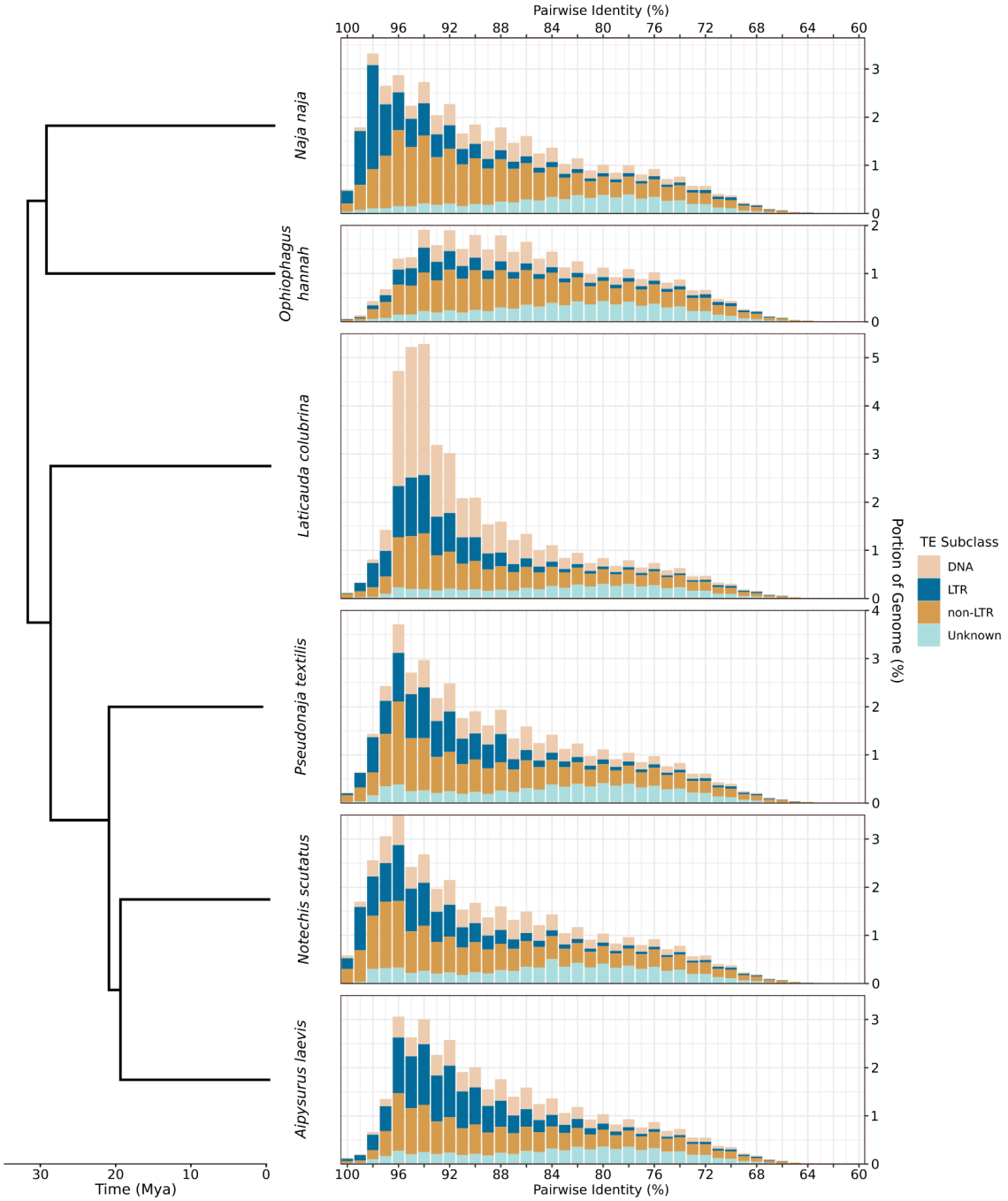


Figure 1. **Overall TE divergence profile of four hydrophiines and two Asian elapids.** Ancestral TE expansion is similar across hydrophiines and Asian elapids while recent expansion varies between species. Due to much lower genome assembly quality resulting in collapsed TEs, little recent expansion in the *Aipysurus laevis* and *Ophiophagus hannah* genomes was detected. TEs were identified using RepeatMasker [37] and a custom repeat library (see methods).

Recent insertions vs ancestral insertions

Recent TE insertions are likely to have diverged only slightly from the sequences RepeatMasker used to identify them, while ancestral insertions will likely be highly divergent. Based on this assumption, we discerned between recent and ancestral insertions using the pairwise identity of TE insertions to the mc-subfamily used to identify them. To estimate the expected divergence of ancestral TE insertions from consensus sequences compared to new insertions we searched for orthologues of 5000 randomly selected *Notechis* TE insertions and their flanks in the *Naja* assembly (Figure 2). From the 5000 TEs we were able to identify 2192 orthologues in *Naja naja*. As expected the median pairwise % identity of the ancestral TEs to the curated consensus sequences was notably lower compared to that of all TEs, at 90.5% compared to 93.6%. Similarly the 95% quantile was lower, at 96.2% compared to 98.4%. Based on these results we treated TEs with a percent identity of 96 % or higher to consensus sequences as having likely been inserted since divergence from Asian elapids.

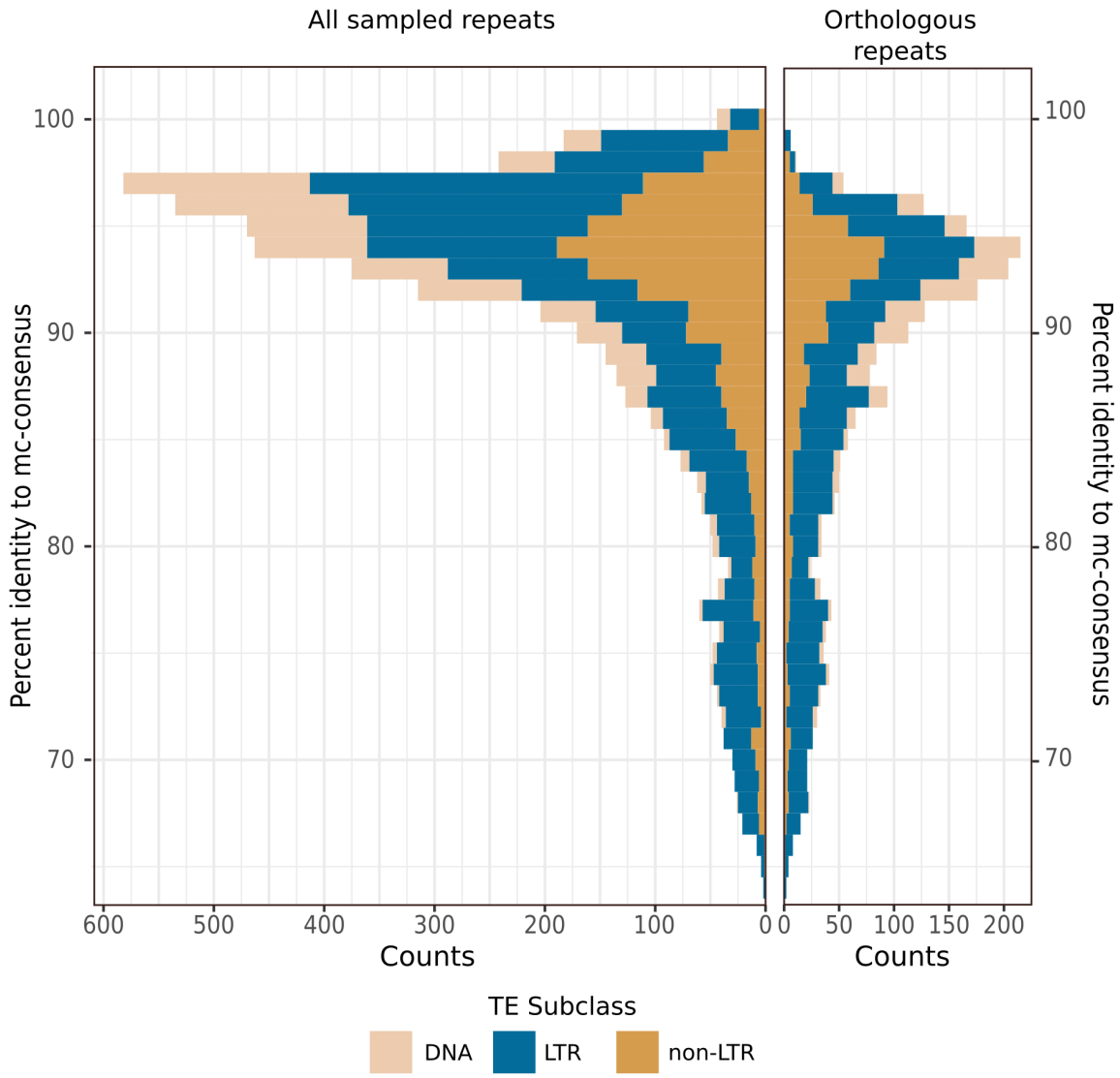


Figure 2: **The similarity of 5000 randomly selected TE insertions in *Notechis scutatus* to the consensus sequence used to identify them compared to that of the subset of 2192 having orthologues in *Naja naja*.** The similarity of ancestral insertions from mc-consensuses used to identify them was notably lower than that of TEs likely inserted since the species diverged. TEs were initially identified in *Notechis* using RepeatMasker [37]. The presence of orthologues in *Naja* was determined using BLASTN (-task dc-megablast) [25].

Recent expansion of specific superfamilies

By comparing TE divergence profiles of the various assemblies, we can gain an overall picture of how TE superfamilies have expanded since the split of Hydrophiinae from Asian elapids (Figures 3-5). Large

expansions of Gypsy retrotransposons are apparent in both the *Naja* and hydrophiine assemblies, however Copia, DIRS and ERVs appear inactive in *Naja* while expanding in hydrophiines. The divergence profile of DNA transposons suggests *Tc1-Mariner* and *hAT* transposons to have either been expanding at a similar rate in all species and/or result from ancestral expansion, with the exception of the explosive expansion of *PIF-Harbinger* transposons in *Laticauda* (see [7]). The greatest variation was seen within non-LTR retrotransposons, with L2s highly active in *Naja* yet completely inactive in hydrophiines. Instead, multiple other LINE superfamilies expanded, in particular CR1s and L1s.. This difference in TE expansion between snake lineages is similar to what was reported by Castoe et. al (2011) [23] , Yin et. al (2016) [39] and Pasquesi et. al (2018) [21], except here we see variation within a family of snakes, not just between families of snakes.

Without highly contiguous assemblies of all species it is difficult to rigorously identify recent or ongoing TE expansions. However, by using transcription as a proxy for transposition we identified currently expressed TE families in present day species as candidates for being active and potentially expanding. To achieve this, we first identified TE subfamilies in each species with over 100 kbp of copies with over 95% pairwise identity to the consensus sequences used to identify them; treating these subfamilies as potentially expanding. By mapping raw transcriptome reads back to these consensuses, we were able to identify expressed TE subfamilies. In all four species, diverse TEs were expressed including subfamilies of Copia, ERV, DIRS, Gypsy, Penelope, CR1, L1s, *Rex1*, RTE, *hAT* and *Tc1-Mariner*.

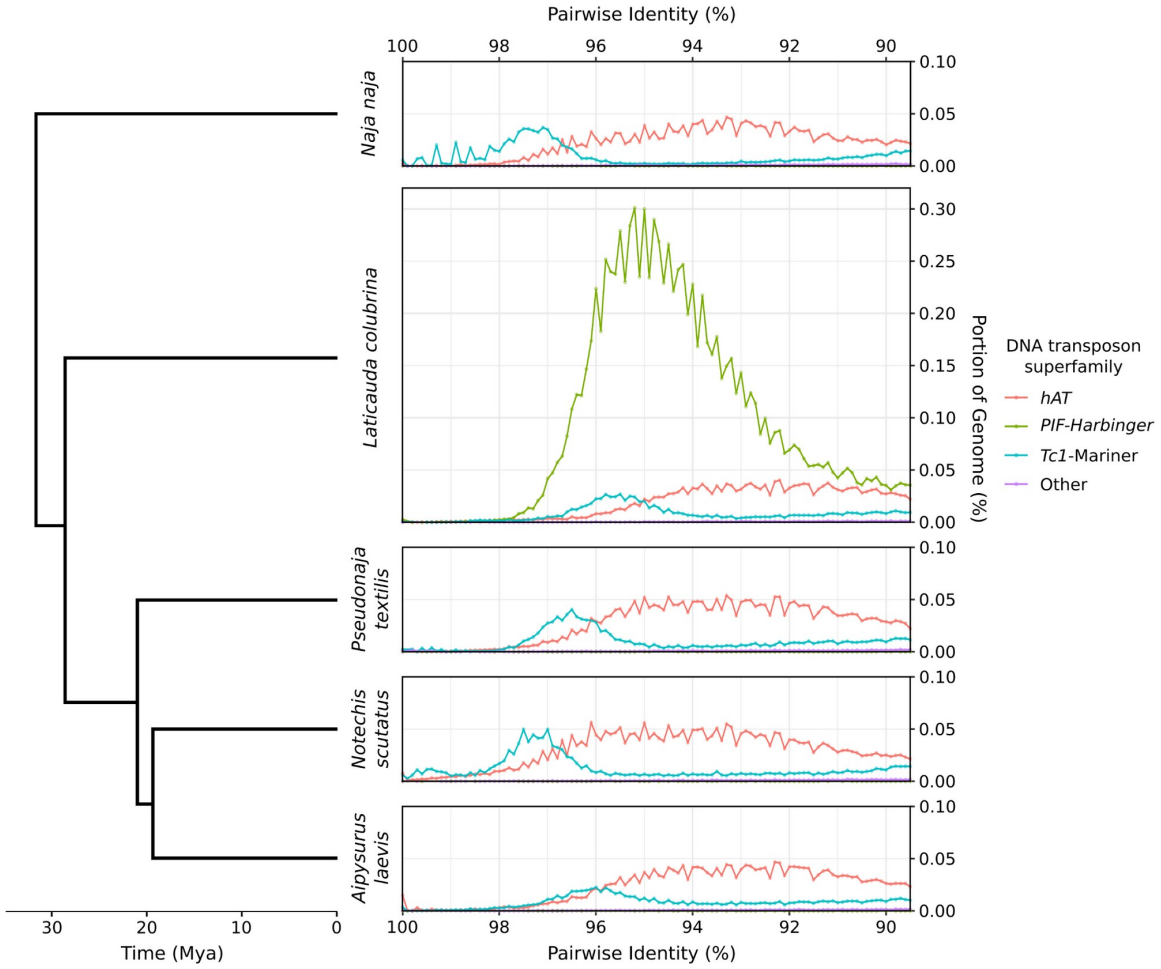


Figure 3 - **Recent DNA transposon divergence profile of four hydrophiines and the Indian cobra.** Across elapids, *hAT* transposons appear inactive, *Tc1-Mariner* transposons appear to have expanded in multiple lineages and *PIF-Harbinger* transposons appear to have expanded rapidly in *Laticauda* following horizontal transfer from a marine species. TEs were identified using RepeatMasker [37] and a custom repeat library (see methods).

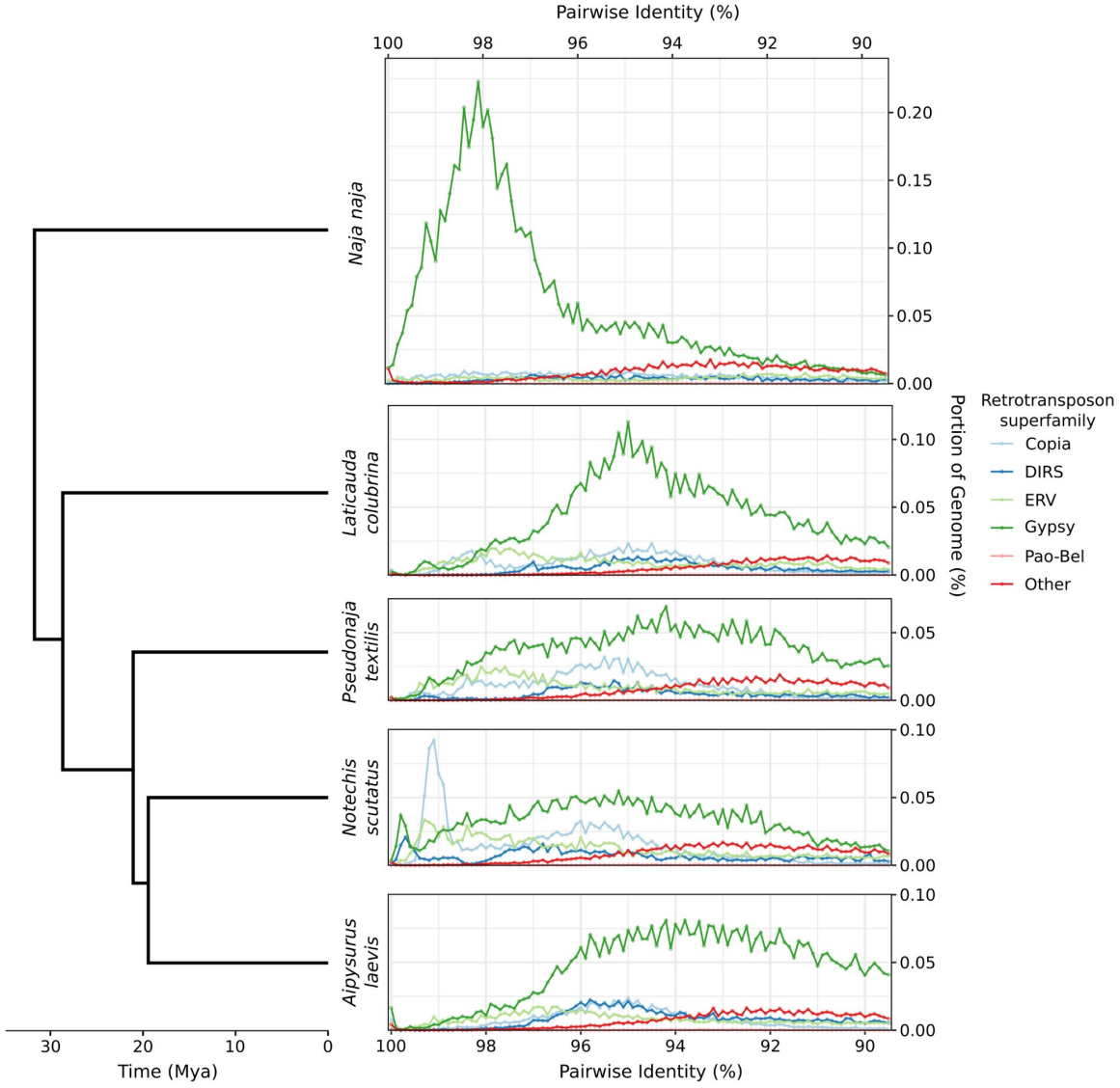


Figure 4 - **Recent LTR retrotransposon divergence profile of four hydrophiines and the Indian cobra.** Gypsy elements are the dominant superfamily in all five species of elapids; ERVs, Copia and DIRS elements have expanded in all Hydrophiinae but have been near inactive in the cobra outgroup. TEs were identified using RepeatMasker [37] and a custom repeat library (see methods).

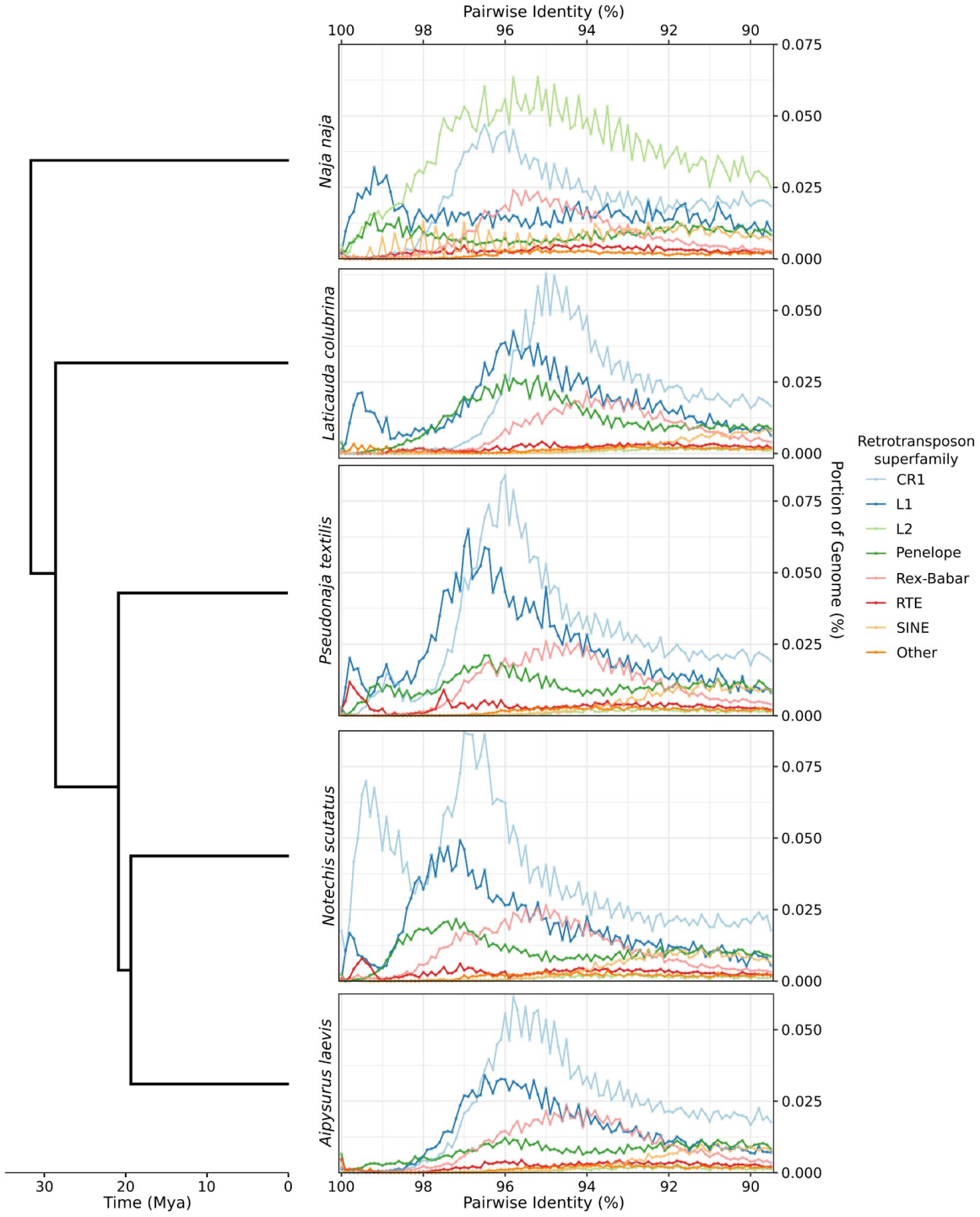


Figure 5 - **Recent non-LTR retrotransposon divergence profile of four hydrophiines and the Indian cobra.** A wide diversity of non-LTR retrotransposons have expanded in all five genomes, with CR1s and L1s being most active in hydrophiines and L2s most active in the cobra outgroup. TEs were identified using RepeatMasker [37] and a custom repeat library (see methods).

Continued expansion or horizontal transfer

The TE subfamilies which we have identified as recently expanded within Hydrophiinae could be ancestral and continuously expanding since diverging from Asian elapids or have been horizontally transferred from long diverged species. Differentiating between ancestral and horizontally transferred TEs is difficult and must meet strict conditions. Horizontally transferred sequences are defined as having a patchy phylogenetic distribution and higher similarity to sequences in another species than would be expected based on divergence time. To identify any TEs which may have been horizontally transferred into Hydrophiinae we conservatively estimated the expected minimum similarity of TEs present in both hydrophiines and Asian elapids using the 2192 orthologous sequences identified in *Notechis* and *Naja* to be 80% (Figure 6). Based on this, any vertically inherited TE subfamily classified as recently expanding in hydrophiines will likely have copies of 80% or higher similarity present in Asian elapids.

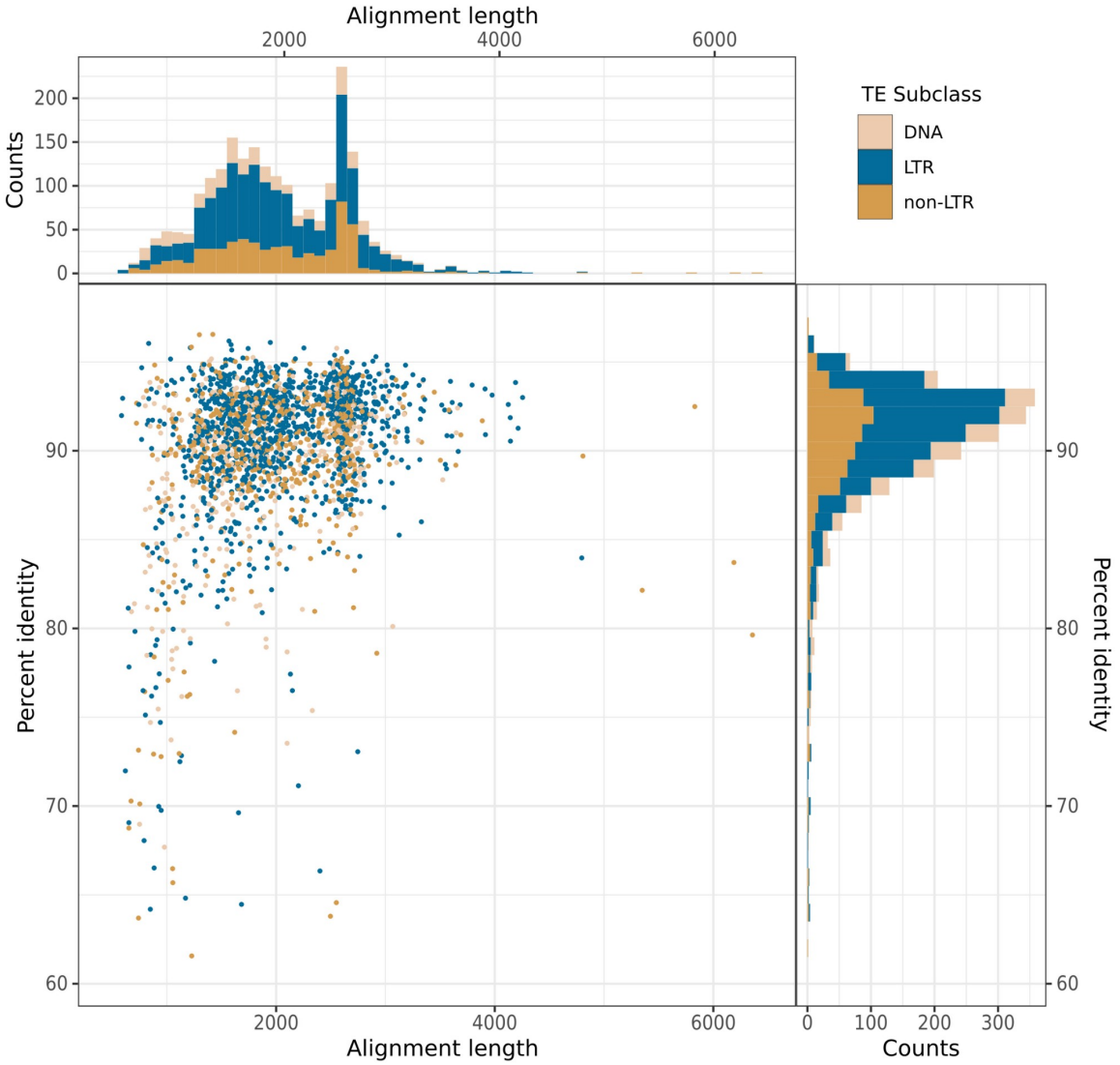


Figure 6. **Similarity (% pairwise identity) of orthologous TEs in *Notechis scutatus* and *Naja naja* genomes.** TE initially identified in *Notechis scutatus* using RepeatMasker, orthologues identified in and pairwise identity calculated for *Naja naja* using BLASTN (-task dc-megablast) [25].

To determine whether any recently expanding TE subfamilies were horizontally transferred into hydrophiines following their divergence from Asian elapids, we searched for them in the genomes of *Naja*, *Ophiophagus* and an additional 8 non-elapid snakes. Some recently expanding subfamilies absent from *Naja* and *Ophiophagus* were present in non-elapid snakes at 80% or higher identity. To be conservative we treated these TEs as ancestral, likely being lost from Asian elapids. The remaining TE subfamilies, those present in hydrophiines but absent from other snakes, were treated as horizontal transfer candidates. To confirm these candidate TEs were horizontally transferred into hydrophiines we searched for them in over 600 metazoan

genomes. This search revealed at least eleven autonomous TEs present in non-serpentine genomes at 80% or higher identity and are therefore likely to have been horizontally transferred into hydrophiines. Of these eleven, three were transferred into the ancestral hydrophiine, five into sea kraits, one into sea snakes and one into the common ancestor of terrestrial hydrophiines and sea snakes (Figure 7).

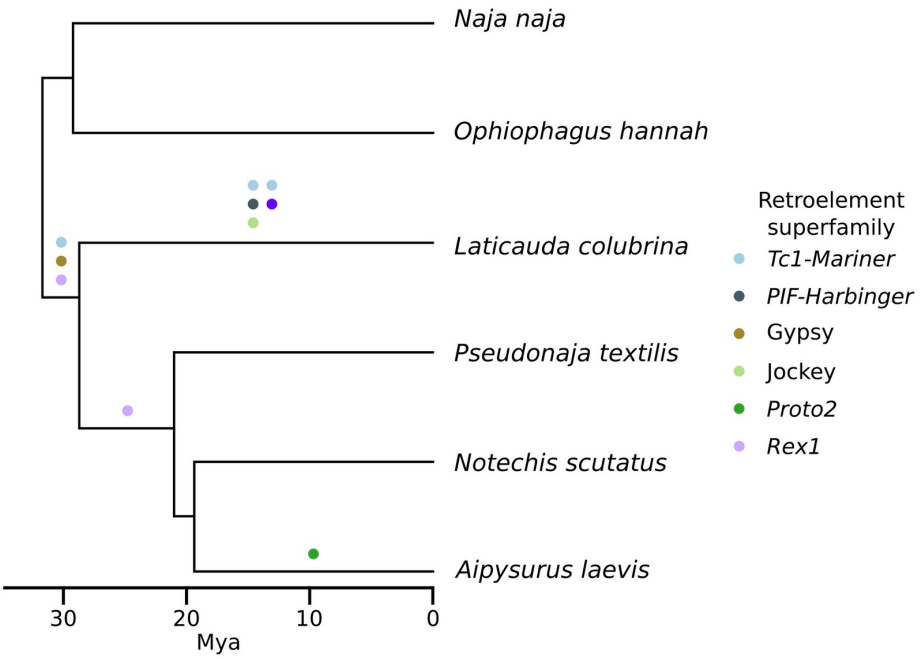


Figure 7: **Horizontal transfer of TEs into hydrophiines since their divergence from Asian elapids.** At least eleven autonomous TEs have been horizontally transferred into hydrophiines, with most likely from marine organisms. We have previously described the horizontal transfer of the *Proto2* to *Aipysurus laevis* in Galbraith et al. (2020) [6] and the *PIF-Harbinger* to *Laticauda* in Galbraith et al. (2021) [7]

We have previously described 2 of the 11 HT events in detail, that of *Proto2-Snek* to *Aipysurus* and *Harbinger-Snek* to*Laticauda*, both of which were likely transferred from a marine species (see [6,7]). Three of the four newly identified HT events identified in *Laticauda* were probably also from an aquatic species, because similar sequences are only found in marine or amphibious species. Therefore, the transfer of these elements likely occurred following the transition of each group to a marine habitat. The exception is a *Tc1-Mariner* which is most similar to sequences identified in hemipterans, a beetle and a spider, however as *Laticauda* is amphibious this is perhaps not surprising.

The *Rex1* transferred to the common ancestor of terrestrial hydrophiines and sea snakes was only identified in the central bearded dragon (*Pogona vitticeps*), an agamid lizard native to the inland woodlands and

shrublands of eastern and central Australia [40]. As this TE is restricted to another species of Australian squamate, this HTT appears to have occurred after hydrophiines reached Australia.

The most interesting of the horizontally transferred TEs are the *Tc1-Mariner*, Gypsy and *Rex1* which were horizontally transferred into the ancestral hydrophiine following its divergence from Asian elapids. Those three are most similar to sequences identified in marine species, either fish or tunicates. Marine elapids (sea kraits and sea snakes) and terrestrial Australian elapids were originally considered two distinct lineages [41–43], however recent adoption of molecular phylogenomics has resolved Hydrophiinae as a single lineage, with sea kraits as a deep-branch and sea snakes nested within terrestrial Australian snakes [1,44,45]. Fossil evidence combined with an understanding of plate tectonics has revealed Hydrophiinae, like many other lineages of Australian reptiles, likely colonised Australia via hopping between islands formed in the Late Oligocene-Early Miocene by the collision of the Australian and Eurasian plates [46–50]. Alternatively, it has also been proposed the common ancestor of Hydrophiinae may have been a semi-marine “proto-*Laticauda*”, which colonised Australia in the Late Oligocene directly from Asia [51]. The horizontal transfer of three TEs into the ancestral hydrophiine likely from a marine organism provides tangible support for the hypothesis that the ancestral hydrophiine was a semi-marine or marine snake.

Conclusion

In our survey of elapid genomes, we have found that TE diversity and their level of expansion varies significantly within a single family of squamates, similar to the variation previously seen across all squamates or within long diverged snakes. This diversity and variation is much greater than what has been reported for mammals and birds. Our finding of HTT into lineages of hydrophiine exposed to novel environments indicates that environment may play a large role in HTT through exposure to new TEs. Additionally, the HTT of three TEs found solely in marine organisms into the ancestral hydrophiine provides evidence that terrestrial Australian elapids are derived from a marine or amphibious ancestor.

As long read genome sequencing becomes feasible for more species, genome assembly quality will continue to increase and multiple genomes of non-model species will be able to be sequenced. Using these higher quality genomes, we will be able to better understand HTT and the role TEs play in adaptive evolution. Due to their rapid adaptation to a wide range of environments and multiple HTT events into different lineages, Hydrophiinae provide the ideal system for such studies.

Acknowledgements

We thank James Nankivell and Vicki Thomson for their discussions on Hydrophiinae's phylogeny and dispersal to Australia.

Supplementary Information

SI Table 1

Genome assembly statistics. Repeat composition was calculated using RepeatMasker [37] and a custom library of curated RepeatModeler libraries and previously described lepidosaur TEs from the Repbase RepeatMasker library [38].

SI Table 2

A list of the metazoan genome assemblies searched for HTT. All genomes were downloaded from NCBI RefSeq and GenBank [52,53].

SI Table 3

Horizontally transferred TEs and the number of hits in species they were identified in. TEs were identified using BLASTN (-task dc-megablast) [25].

References

1. Sanders KL, Lee MSY, Leys R, Foster R, Keogh JS. 2008 Molecular phylogeny and divergence dates for Australasian elapids and sea snakes (hydrophiinae): evidence from seven genes for rapid evolutionary radiations. *J. Evol. Biol.* **21**, 682–695. (doi:10.1111/j.1420-9101.2008.01525.x)
2. Mirtschin P, Rasmussen A, Weinstein S. 2017 *Australia's dangerous snakes : identification, biology and envenoming*. Clayton South, VIC, Australia: CSIRO Publishing. (doi:10.1071/9780643106741)
3. Glodek GS, Voris HK. 1982 Marine Snake Diets: Prey Composition, Diversity and Overlap. *Copeia* **1982**, 661–666. (doi:10.2307/1444667)
4. Voris HK, Voris HH. 1983 Feeding strategies in marine snakes: an analysis of evolutionary, morphological, behavioral and ecological relationships. *Am. Zool.* **23**, 411–425.
5. Brischoux F, Bonnet X. 2009 Life history of sea kraits in New Caledonia. *Zoologia Neocaledonica* **7**, 37–51.
6. Galbraith JD, Ludington AJ, Suh A. 2020 New Environment, New Invaders—Repeated Horizontal Transfer of LINEs to Sea Snakes. *Genome Biol. Evol.*
7. Galbraith JD, Ludington AJ, Sanders KL, Suh A, Adelson DL. 2021 Horizontal transfer and subsequent explosive expansion of a DNA transposon in sea kraits (Laticauda). *bioRxiv*. , 2021.06.13.448261. (doi:10.1101/2021.06.13.448261)

8. 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. (doi:10.1093/gbe/evv005)

9. Sotero-Caio CG, Platt RN 2nd, Suh A, Ray DA. 2017 Evolution and Diversity of Transposable Elements in Vertebrate Genomes. *Genome Biol. Evol.* **9**, 161–177. (doi:10.1093/gbe/evw264)

10. Volff J-N. 2006 Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *Bioessays* **28**, 913–922. (doi:10.1002/bies.20452)

11. Kazazian HH Jr. 2004 Mobile elements: drivers of genome evolution. *Science* **303**, 1626–1632. (doi:10.1126/science.1089670)

12. Cornelis G *et al.* 2017 An endogenous retroviral envelope syncytin and its cognate receptor identified in the viviparous placental Mabuya lizard. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E10991–E11000. (doi:10.1073/pnas.1714590114)

13. Ikeda N, Chijiwa T, Matsubara K, Oda-Ueda N, Hattori S, Matsuda Y, Ohno M. 2010 Unique structural characteristics and evolution of a cluster of venom phospholipase A2 isozyme genes of Protobothrops flavoviridis snake. *Gene* **461**, 15–25. (doi:10.1016/j.gene.2010.04.001)

14. Wicker T *et al.* 2007 A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982. (doi:10.1038/nrg2165)

15. Jurka J, Kapitonov VV, Kohany O, Jurka MV. 2007 Repetitive sequences in complex genomes: structure and evolution. *Annu. Rev. Genomics Hum. Genet.* **8**, 241–259. (doi:10.1146/annurev.genom.8.080706.092416)

16. Kapitonov VV, Tempel S, Jurka J. 2009 Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. *Gene* **448**, 207–213. (doi:10.1016/j.gene.2009.07.019)

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

18. Kapusta A, Suh A, Feschotte C. 2017 Dynamics of genome size evolution in birds and mammals. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E1460–E1469. (doi:10.1073/pnas.1616702114)

19. Ivancevic AM, Kortschak RD, Bertozzi T, Adelson DL. 2018 Horizontal transfer of BovB and L1 retrotransposons in eukaryotes. *Genome Biol.* **19**, 85. (doi:10.1186/s13059-018-1456-7)

20. Suh A, Paus M, Kiefmann M, Churakov G, Franke FA, Brosius J, Kriegs JO, Schmitz J. 2011 Mesozoic retroposons reveal parrots as the closest living relatives of passerine birds. *Nat. Commun.* **2**, 443. (doi:10.1038/ncomms1448)

21. Pasquesi GIM *et al.* 2018 Squamate reptiles challenge paradigms of genomic repeat element evolution set by birds and mammals. *Nat. Commun.* **9**, 2774. (doi:10.1038/s41467-018-05279-1)

22. Kumar S, Stecher G, Suleski M, Hedges SB. 2017 TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol. Biol. Evol.* **34**, 1812–1819. (doi:10.1093/molbev/msx116)

23. Castoe TA *et al.* 2011 Discovery of highly divergent repeat landscapes in snake genomes using high-throughput sequencing. *Genome Biol. Evol.* **3**, 641–653. (doi:10.1093/gbe/evr043)

24. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020 RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9451–9457. (doi:10.1073/pnas.1921046117)

25. Zhang Z, Schwartz S, Wagner L, Miller W. 2000 A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7**, 203–214. (doi:10.1089/10665270050081478)

26. Katoh K, Standley DM. 2013 MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780. (doi:10.1093/molbev/mst010)

27. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997 Gapped BLAST

and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402. (doi:10.1093/nar/25.17.3389)

28. Marchler-Bauer A *et al.* 2017 CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* **45**, D200–D203. (doi:10.1093/nar/gkw1129)

29. Kohany O, Gentles AJ, Hankus L, Jurka J. 2006 Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* **7**, 474. (doi:10.1186/1471-2105-7-474)

30. NCBI Resource Coordinators. 2018 Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **46**, D8–D13.

31. The UniProt Consortium. 2021 UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489.

32. Huang Y, Niu B, Gao Y, Fu L, Li W. 2010 CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**, 680–682. (doi:10.1093/bioinformatics/btq003)

33. Li H. 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]*.

34. Simões BF *et al.* 2020 Spectral Diversification and Trans-Species Allelic Polymorphism during the Land-to-Sea Transition in Snakes. *Curr. Biol.* **30**, 2608–2615.e4. (doi:10.1016/j.cub.2020.04.061)

35. Rhie A *et al.* 2021 Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746. (doi:10.1038/s41586-021-03451-0)

36. Peona V *et al.* 2021 Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. *Mol. Ecol. Resour.* **21**, 263–286. (doi:10.1111/1755-0998.13252)

37. Smit AF. 2004 Repeat-Masker Open-3.0. *http://www.repeatmasker.org*

38. Bao W, Kojima KK, Kohany O. 2015 Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11. (doi:10.1186/s13100-015-0041-9)

39. Yin W *et al.* 2016 Evolutionary trajectories of snake genes and genomes revealed by comparative analyses of five-pacer viper. *Nat. Commun.* **7**, 13107. (doi:10.1038/ncomms13107)

40. Melville J, Wilson SK. 2019 *Dragon Lizards of Australia: Evolution, Ecology and a Comprehensive Field Guide*. Museums Victoria Publishing. See https://play.google.com/store/books/details?id=nWJ_xQEACAAJ.

41. Smith MA. 1926 *Monograph of the sea-snakes (Hydrophiidae)*. Printed by order of the Trustees of the British museum.

42. Hoffstetter R. 1939 Contribution à l'étude des Elapidae actuels et fossiles et de l'ostéologie des ophidiens. *Publications du musée des Confluences*

43. Storr GM. 1964 Some aspects of the geography of Australian reptiles. *Senckenb. Biol.* **45**, 577–589.

44. Keogh JS. 1998 Molecular phylogeny of elapid snakes and a consideration of their biogeographic history. *Biol. J. Linn. Soc. Lond.*

45. Sanders KL, Lee MSY. 2008 Molecular evidence for a rapid late-Miocene radiation of Australasian venomous snakes (Elapidae, Colubroidea). *Mol. Phylogenet. Evol.* **46**, 1165–1173. (doi:10.1016/j.ympev.2007.11.013)

46. Hall R. 2013 The palaeogeography of Sundaland and Wallacea since the Late Jurassic. *J. Limnol.* **72**, e1.

47. Yang T, Gurnis M, Zahirovic S. 2016 Mantle-induced subsidence and compression in SE Asia since the early Miocene. *Geophys. Res. Lett.* **43**, 1901–1909. (doi:10.1002/2016gl068050)

48. Esquerré D, Donnellan S, Brennan IG, Lemmon AR, Moriarty Lemmon E, Zaher H, Grazziotin FG,

Keogh JS. 2020 Phylogenomics, Biogeography, and Morphometrics Reveal Rapid Phenotypic Evolution in Pythons After Crossing Wallace's Line. *Syst. Biol.* **69**, 1039–1051. (doi:10.1093/sysbio/syaa024)

49. Heinicke MP, Greenbaum E, Jackman TR, Bauer AM. 2011 Phylogeny of a trans-Wallacean radiation (Squamata, Gekkonidae, Gehyra) supports a single early colonization of Australia. *Zool. Scr.* **40**, 584–602. (doi:10.1111/j.1463-6409.2011.00495.x)

50. Scanlon JD, Lee MSY, Archer M. 2003 Mid-Tertiary elapid snakes (Squamata, Colubroidea) from Riversleigh, northern Australia: early steps in a continent-wide adaptive radiation. *Geobios Mem. Spec.* **36**, 573–601. (doi:10.1016/S0016-6995(03)00056-1)

51. Heatwole H, Grech A, Marsh H. 2017 Paleoclimatology, Paleogeography, and the Evolution and Distribution of Sea Kraits (Serpentes; Elapidae; Laticauda). *Herpetological Monographs.* **31**, 1–17. (doi:10.1655/HERPMONOGRAPHS-D-16-00003)

52. O'Leary NA *et al.* 2016 Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–45. (doi:10.1093/nar/gkv1189)

53. Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2015 GenBank. *Nucleic Acids Res.* **43**, D30–5. (doi:10.1093/nar/gku1216)

Conclusions and Future Directions

"I was just so interested in what I was doing I could hardly wait to get up in the morning and get at it. One of my friends, a geneticist, said I was a child, because only children can't wait to get up in the morning to get at what they want to do." - Barbara McClintock

Over the past half century, studies in model species have made clear that TEs are not simply junk DNA, but rather a rich source of genetic novelty. Due to the rise of cost-effective whole-genome sequencing, studies over the past decade have begun to branch into non-model species and examine variation in TEs and the adaptive effects TEs have had across the tree of life. This thesis has made it clear that to truly understand the adaptive potential of TEs, we need to continue to examine their evolution in non-model species.

One of the biggest challenges in comparative genomics is the high variability of genome assembly quality. Due to the high cost of long read sequencing and optical mapping, many assemblies still rely solely on short read sequencing. As short reads cannot span full length copies of most autonomous TEs, many TE sequences are collapsed during genome assembly. Throughout, this thesis shows that a reliance on low-depth short read sequencing results in less contiguous genome assemblies and an inability to identify recent repeat expansions. Consequently, many recent insertions, which are most likely to have adaptive effects, may not be correctly assembled.

The most widely used TE annotation software packages have largely been developed for the annotation of traditional model species such as humans, mice, maize and rice. As such, while these packages are reliable for related species, such as other primates, rodents or grasses, it proves problematic for long-diverged species, such as hydrophiines. Additionally, non-model species may contain novel multicopy genes, which repeat annotation software will annotate as unknown TEs. As the chicken is an important agricultural species and the zebra finch is a model organism for neurological research, this did not pose a problem in the avian research. However, hydrophiine snakes are long diverged from any model organism and contain many novel multicopy gene families including various toxins and vomeronasal genes. Repeat annotation software will correctly identify these multicopy genes as repetitive DNA, but rather than being classified as genes these sequences are treated as unclassifiable TEs.

Fortunately, future research will be able to overcome these issues due to the decreasing cost of high-quality genome sequencing and improved repeat annotation packages. The advent of cheap long read sequencing nullifies the issue of collapsed reads, allowing for better ab initio repeat annotation. Improved scaffolding through the use of Hi-C and optical mapping results in more contiguous genome assemblies, allowing for chromosomal rearrangements enabled by TEs to be identified. Additionally, more contiguous genomes

will allow for a better understanding of the position of TEs relative to coding and regulatory regions.

Going forward from this thesis there are three key discoveries I have made which warrant further investigation. Firstly, we have identified recent and likely on-going repeat expansion in parrots, one of the most diverse avian orders. Sequencing of species such as *Amazona collaria* at a population level will allow for a greater understanding of repeat expansion, loss and fixation, and if recent TE insertions have had adaptive effects or impacted speciation. Secondly, we have identified potentially adaptive insertions into the regulatory regions of multiple genes in both *Laticauda* and *Aipysurus*. Population level genome and transcriptome sequencing could determine if these insertions have been fixed, if they have impacted gene expression, and hence had adaptive effects. Finally, while we have determined the HTTs into marine hydrophiines and the ancestral hydrophiine were likely from marine organisms, we have not identified potential vectors. Whole genome sequencing of prey along with parasitic organisms such as cestodes, nematodes and barnacles could elucidate mechanisms underlying HTT.

Together this thesis provides the groundwork for research into adaptive change caused by transposable elements. Hydrophiines have rapidly adapted to a wide variety of terrestrial and marine habitats, and during this time have seen large fluctuation in TE content due both to horizontal transfer and expansion of existing TE families. As such, Hydrophiinae is an ideal model family for studying both HTT and the adaptive effects of TEs during adaptation to novel environments.