Evolutionary Genomics of an Adaptive Radiation: Viviparous Sea Snakes (Elapidae)

A dissertation by

Alastair John Ludington

A thesis submitted for the degree of Doctor of Philosophy

The University of Adelaide
School of Biological Sciences
Department of Ecology and Evolutionary Biology

September 2023

THE UNIVERSITY of ADELAIDE

Abstract

Viviparous sea snakes (Hydrophiinae; Elapidae) underwent a dramatic and remarkably recent transition from land to sea in the last 10 million years, yet they remain a conspicuous gap in molecular studies of speciation and adaptation in vertebrates. My thesis addresses this knowledge gap by generating new high-quality genomic resources for sea snakes and analysing these data, alongside existing datasets, to understand the demographic and genetic basis of ecological transitions and rapid radiation within elapid snakes.

Marine systems are often suggested to support large, highly connected populations, but quantitative comparisons with terrestrial systems have been lacking. Further, it has been shown that co-occurring species often have congruent responses to past ecological events. Genomes of organisms contain a record of their evolutionary and ecological past, and as such can shed light on species' responses to past biotic and abiotic factors. Using coalescent methods and whole genome data, I compared the effective population sizes between seven terrestrial and marine elapids. The results of this analysis found that, contrary to expectation, effective population sizes in marine species were not larger, instead showing highly variable patterns that did not clearly correspond with major ecological divisions. This work provided one of the first preliminary studies comparing terrestrial and marine elapids, providing the basis for future demographic studies looking to understand the role of ecological and biogeographic factors in shaping population dynamics.

Understanding the genetic basis of the terrestrial-to-marine transition in sea snakes remains a prominent gap in studies of marine adaptation in vertebrates. Using newly generated chromosome-scale reference genomes, representing the most complete genomes available for Elapidae, I performed comparative genomic analyses between six *Hydrophis* sea snakes and seven terrestrial snakes. Through phylogenomic analyses, we highlight the possibility of reticulate speciation at the root of *Hydrophis*, while whole genome alignments elucidate extensive chromosomal synteny within *Hydrophis*, identifying structural variations that will be key candidates for future studies. Finally, I used a strict screen for positive selection to identify adaptive candidate genes in *Hydrophis,* relative to the terrestrial background, which showed over-representation for hypoxia adaptation, sensory perception, and morphological development.

Finally, genome scans were used to explore the divergence profiles between three ecologically distinct and coexisting species that diverged near-simultaneously 3-8 million years ago. Our results support early admixture between the three species, in addition to shared evolutionary processes that likely shaped their conserved genomic landscapes. Numerous islands of divergence were identified, whose profiles best fit a scenario of long-term linked background and/or positive selection across a stable genomic landscape. Several low divergence balanced loci were also identified, with one notable locus housing numerous chemosensory receptor genes, which play a crucial role in mediating reproductive and feeding behaviours in sea snakes.

This thesis provides novel insights into the demographic and genetic factors that have facilitated the terrestrial-to-marine transition in elapid snakes, and the rapid radiation in *Hydrophis* snakes. This work has generated numerous high-quality genomic

resources, including whole genome sequencing and chromosome-scale reference genomes, that will prove to be powerful resources for future studies investigating these diverse organisms.

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 the thesis resides with the copyright holder(s) of those works.

I 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.

Acknowledgements

I'd like to start by thanking my primary supervisor, Kate Sanders. You've been nothing but encouraging and enthusiastic and I can't thank you enough for all the opportunities you've afforded me over the years. It's been an absolute privilege to both learn and work alongside side you. I must also thank my co-supervisor, Jimmy Breen. I couldn't have asked for a better mentor and friend to show me the ropes and introduce me to the world of bioinformatics and genomics.

To the Sanders Lab, both past and present, you have been one of the best groups to work with. You were nothing but encouraging and I hope to continue enticing you to the world of genomics. I'd also like to acknowledge the past members of the Adelaide Bioinformatics Hub, and specifically Stevie Pederson, who took a chance on a young honours kid. Additionally, to my friend Chris Ward, who has been a constant source of knowledge and advice over the course of this project.

I would especially like to recognise my parents, Vinka and Philip, and my sister Eleanor. You always valued effort over outcome and encouraged us to follow our interests. It's safe to say that I wouldn't have achieved anywhere near as much as I have without your endless support and advice. To my partner Carlee, I am immensely grateful. You have been extraordinarily patient over the last 3 years and your support has been unwavering. Maybe now I'll have time to walk you through that 'introduction to computational genetics' textbook that you printed off however many years ago…

Lastly, I'd like to dedicate this thesis to my close friend Davis who passed away in the final year of this work. You will always be remembered.

Chapter 1:
Introduction

1.1 The speciation continuum

The question of how species originate and adapt to novel environments has been, and still is, one of the cornerstones of evolutionary biology research. An early consensus was that speciation represented complete reproductive isolation, whereby genetic and phenotypic distinctiveness between populations could be maintained even when in geographical proximity (Seehausen et al., 2014). To establish such reproductive isolation, a geographic model of speciation was proposed by Mayr (1942), whereby an extrinsic geographic barrier precludes gene-flow between subpopulations. Under such a model, the build-up of genomic incompatibilities via genetic drift and natural selection is a function of time, as there are no evolutionary forces acting to enforce reproductive compatibility between the incipient species (Turelli et al., 2001). Speciation in allopatry proved popular throughout the 20th century, as it had an intrinsic logic to how divergence would develop; both pre- and postzygotic isolating mechanisms would arise as an inevitable by-product of being separated (Turelli et al., 2001). However, biological systems are rarely so simple, with later theoretical and empirical work exploring and demonstrating that speciation without geographic barriers is plausible and has seemingly occurred in nature (Bolnick & Fitzpatrick, 2007; Seehausen et al., 2014). As Gavrilets (2003) so precisely put it, “sympatric speciation is the emergence of new species from a population where mating is random with respect to the birthplace of the mating partners” (p. 2198). The obvious difficulty facing speciation in sympatry is the antagonism between the development and establishment of genetic divergence and the homogenising effects of gene flow (Feder et al., 2012). It’s widely believed that strong

ecological and/or sexual selection may play an important role early in sympatric speciation to help overcome the effect of gene-flow (Seehausen et al., 2014). As such, the dynamics between selection and genomic architecture are likely to have significantly different roles, and leave different fingerprints on the genome, in speciation in sympatry compared to allopatry (Feder et al., 2012). This led to speciation beginning to be viewed as a continuum, with allopatry residing at one extreme and sympatry at the other (Butlin et al., 2008). Given the time it takes for speciation to occur, it becomes less and less plausible that either extreme occurs in isolation, a finding that is routinely being supported by empirical studies (e.g. Lewanski et al., 2022; Nichols et al., 2015; Schield et al., 2019; Yamasaki et al., 2020).

Marine taxa are one such system that challenge the extremes of the traditional geographic models, presented as the “marine speciation paradox” (Bierne et al., 2003). Marine systems are highly biodiverse, with many marine organisms having high dispersal capabilities and large contact zones (De Jode et al., 2023), making it difficult to explain speciation with a purely allopatric model. Alternatively, recent studies modelling demographic histories of marine species have routinely supported secondary contact, or isolation and migration models (De Jode et al., 2023). Consequently, these results suggest that perhaps marine environments, and the organisms that live in them, are not as connected as previously thought, and instead may undergo speciation in parapatry, or through bouts of allopatry with periods of sympatry. Recent work has supported the presence of barriers in marine systems, such as oceanographic currents (Rossi et al., 2021) and density differences between water masses (Hudson et al., 2020). In summary, it is evident that the relative influence of drift, selection and gene

flow during speciation will vary among systems depending on geographic and ecological context.

1.2 Adaptive radiations: A special case of speciation

Adaptive radiations are a specific case of speciation. They represent a response to natural selection and ecological opportunity, resulting in the multiplication of species and associated adaptations (Glor, 2010). The best opportunities to understand the processes that shape adaptive radiations are provided by recent radiations of sym- or parapatric species whose co-existence is facilitated by some level of resource specialisation (Glor, 2010). Adaptive radiations are believed to start with ecological opportunity; whether through dispersal to a new environment, evolution of a key innovation or escape from an antagonist (Simpson, 1944; Yoder et al., 2010). Ecological release then follows, characterised by a shift in selection pressures on one or more ecological traits (Lister, 1976; Roughgarden, 1972). This can occur even in the face of potentially strong directional selection that may be acting on key traits for adaptation to a new environment or relating to the evolution of a key innovation (Yoder et al., 2010). A characteristic of ecological opportunity is an expansion in population size, potentially driven by an initial reduction in interspecific competition in species-poor habitats (MacArthur et al., 1972). Such an increase in population size is typically correlated with an increase in genetic diversity (Hague & Routman, 2016; Kimura, 1983)**,** but also an increase in intraspecific competition for resources. Consequently, likely in part to a wider range and variety of resources being available, but also due to increases in intraspecific competition, and a shift in the selective landscape, this may result in

divergent selection to use a wider range of resources and habitat (Bolnick, 2001; Svanbäck & Bolnick, 2005). Given time, speciation can occur, translating ecological opportunity into increased species richness (and in some cases elevated rates of lineage diversification) associated with adaptive radiation (Gavrilets & Losos, 2009; Rabosky & Glor, 2010; Yoder et al., 2010). Speciation in this instance is expected to consolidate transient increases of trait variation brought about through either relaxed selection or increased divergent or disruptive selection against intermediate phenotypes, owing to increased intraspecific competition (Yoder et al., 2010)**.** Consequently, speciation can continually increase diversity with each new ecological opportunity through continued adaptive radiations (Coyne et al., 2004). Further, ecological opportunity can drive speciation, especially if the ecological opportunity contributes to reproductive isolation.

1.3 The genetics of rapid adaptation

Adaptive genotypes will be positively selected for in the face of novel stressors, with the historical consensus that new beneficial mutations drive adaptation. When looking at models of adaptation, such as Fisher's "geometric" model and Gillespies "mutational landscapes" model, adaptation results from new, beneficial mutations of variable effect increasing fitness (Fisher, 1930; Gillespie, 1983, 1984). Under such models, a species facing a rapidly changing environment would have to wait for the appearance of the desired allele before adaptation could proceed (Orr & Betancourt, 2001). Significant negative pressures also face new mutations, such as Haldane's sieve (Turner, 1981), in which there is a strong fixation bias against recessive adaptive mutations due to weak

selection, and genetic drift, the stochastic loss of alleles from populations (Kimura, 1983).

An alternate avenue for adaptive alleles to arise is from pre-existing variation within the genome, termed “standing genetic variation”. Such variation is expected to be maintained within populations either through gene-flow between populations experiencing different environmental conditions, or through hybridisation with other species. Alternatively, neutral, or deleterious alleles may be hidden from selection as they do not have any significant effect on phenotype in the ancestral environment (Barrett & Schluter, 2008). This kind of cryptic variation may only manifest as a new phenotype in the event of a significant environmental change, or through the introduction of novel alleles (Gibson & Dworkin, 2004). Compared to adaptive *de novo* mutations, standing variation is expected to reach fixation faster due to the initial higher frequencies of alleles within the population (Hermisson & Pennings, 2005), can result in more rapid evolution in novel environments as the variation is available immediately (Barrett & Schluter, 2008; Zheng et al., 2019), and may have been “pre-tested” by selection in past environments (Rieseberg et al., 2003).

Elucidating the genetic mechanisms that drive adaptation is complex, with both *de novo* mutations and standing genetic variation being viable avenues to a similar destination. However, their respective involvement and contribution to adaptation may be suited to specific evolutionary scenarios, each leaving behind a unique fingerprint in the underlying genome. Consequently, through thorough genomic investigations we can better understand the processes that drive adaptation.

1.4 Genomic landscapes of speciation and divergence

Barrier loci is a broad term that relates to any region in the genome that acts as a barrier to gene-flow between populations and contributes to reproductive isolation (Ravinet et al., 2017). Features that constitute barrier loci are broad, ranging from single loci under divergent selection to features that reduce recombination boundaries like chromosomal rearrangements (Wolf & Ellegren, 2017). In situations where a locally beneficial allele arises, it can undergo a local selective sweep through divergent positive selection, increasing the allele frequency within the population (Ravinet et al., 2017). The barrier effect is the reduction in effective migration around the sweep relative to the remaining genome, which remains porous to gene flow (Feder et al., 2012). Barrier loci have received extensive attention, as these high divergence regions are assumed to harbour functionally relevant genomic elements necessary for speciation, even while gene flow homogenises the surrounding genome (Feder et al., 2012; Ravinet et al., 2017).

The advent of next generation sequencing has made conducting genomic scans of divergence more accessible than ever. Despite this, interpreting divergence profiles has proven to be a difficult task, complicating efforts to identify locally adapted alleles and barriers to gene flow during speciation (Cruickshank & Hahn, 2014; Noor & Bennett, 2009). Complications arise due to alternate evolutionary processes leaving similar genetic traces in the genome. Relative and absolute measures of divergence are commonly used in genome scans to tease apart evolutionary signals. *Fst*, a statistic based on Wright's *Fst* (Wright, 1931, 1943), measures the normalised difference in allele frequencies between populations. It is a common statistic used to identify

divergent regions between species but is a relative measure of divergence due to its dependence on the level of within species polymorphisms. Unfortunately, factors such as genetic drift, gene-flow, recombination, demographic history, and genomic architecture can all influence *Fst* signals, among other divergence statistics (Ravinet et al., 2017; Wolf & Ellegren, 2017). For example, regions of low recombination, which often have reduced diversity due to linked selection, or loci that have experienced background selection, will consistently report higher *Fst* values due to the reduced genetic diversity across these regions, not because of genuine sequence divergence between two species (Cruickshank & Hahn, 2014; Noor & Bennett, 2009). Consequently, absolute measures of divergence are often used in conjunction with relative measures, such as *Dxy* (Nei & Li, 1979), which is independent of within species diversity levels. In combination, *Fst* and *Dxy* become a powerful tool for disentangling divergence patterns. However, *Dxy* is not fool proof. The statistic captures the number of sequence changes at a locus since the most recent common ancestor (Cruickshank & Hahn, 2014). As the most recent common ancestor is likely to have existed in the ancestral population before the species separation, *Dxy* measures not only the differences that have occurred since the split, but also differences that were present at the time of the split. This characteristic means that variation in *Dxy* may be due to divergence since speciation but may also reflect levels of ancestral polymorphisms (Cruickshank & Hahn, 2014).

Demographic factors also complicate the interpretation of differentiation landscapes. Speciation in sympatry is commonly characterised by a heterogeneous divergence profile (Seehausen et al., 2014), but such a profile can also arise through

incomplete lineage sorting (Noor & Bennett, 2009). Furthermore, incomplete lineage sorting affects a greater proportion of the genome when speciation events occur close in time, and the ancestral population size is large (Barton, 2006), an obvious challenge for the study of adaptive radiations. Other demographic factors that make interpreting divergence landscapes difficult are pronounced changes in effective population size (Ne), as bottlenecks can shift the mean and variance of baseline genomic differentiation, making the detection of regions affected by divergent selection difficult (Ferchaud & Hansen, 2016). Consequently, understanding the genomic landscapes of divergence and adaptation are fraught with pitfalls that can result in incorrect interpretations of evolutionary processes. However, by using complementary divergence statistics, and modelling demographic histories, we can start to piece together these complex evolutionary events.

1.5 Viviparous sea snakes as a study system

Elapid snakes, in particular sea snakes, lend themselves to being an excellent study system for investigating macro evolutionary transitions and speciation dynamics. Viviparous sea snakes have a biologically interesting history, descending from Australo-Papuan terrestrial hydrophiines some 9-18mya (Sanders et al., 2008; Sanders & Lee, 2008), experiencing a total environment change from terrestrial to marine. Despite their relatively recent transition to marine life, there over 60 known fully marine snakes (Lee et al., 2016; Sanders et al., 2008), which have overtime adapted marine-specific phenotypes such as elongated bodies and paddle-shaped tails that enable propulsive thrust (Brischoux & Shine, 2011; Sanders et al., 2012), shifted spectral sensitivity

towards longer wavelengths (Simões et al., 2020), and a sublingual salt-secreting gland (Heatwole, 1999). Evidently, the macroevolutionary transition from land to sea has resulted in significant innovation, as indicated by the extensive number of marine sea snake species, along with considerable changes in genome-wide selective pressures. Variation within the group is a further indicator of the strong, yet distinct, evolutionary forces at play. For example, when comparing *Hydrophis* and *Aipysurus*, the two main fully marine genera within Hydrophiinae, *Hydrophis* have an approximately three-fold higher rate of speciation (Lee et al., 2016), and as such have undergone rapid speciation in a way that *Aipysurus* has not. Sea snakes have experienced rapid evolution across a range of complex biological processes, all of which has happened within a single study taxon. The genomic changes that have accompanied the transition of reptiles from terrestrial to marine environments and the subsequent rapid adaptive radiation of *Hydrophis* are largely unknown, yet their phenotypic and adaptive effects are vast and unique. As such, sea snakes provide a unique opportunity to advance our understanding of the genomic basis of aquatic adaptation and rapid radiation in vertebrates, along with speciation in a novel marine system.

1.6 Summary

This thesis used bioinformatics, population genomics and phylogenomics to advance knowledge of the evolution of Australia's prolific venomous snakes (Hydrophiinae; Elapidae). New genomes were generated to investigate the demographic (Aim1) and genomic (Aim 2) basis of the terrestrial-to-marine transition in viviparous sea snakes and their close terrestrial relatives. Population resequencing was then used to examine

the sequence level patterns and processes underpinning the exceptionally rapid *Hydrophis* radiation of fully marine sea snakes (Aim 3). The results of this research represent a significant advance in our understanding of the population history, gene selection and the genomic landscape of a powerful new system for evolutionary studies. This work further provides substantial high-quality genomes for sea snakes that will be important resources for diverse future studies of vertebrate genomics and evolution.

1.7 Thesis outline

Aim 1: Investigate the demographic histories, via effective population size (*Ne),* of terrestrial and marine elapids, comparing the magnitude and timing of historical fluctuations in *Ne* between the two groups.

Demographic histories of co-occurring species are expected to show similar fluctuations in response to shared biotic and abiotic factors. Further, it's believed that given the perceived lack of geographic barriers, marine organisms are expected to have overall larger effective population sizes than their terrestrial counterparts. In chapter 2 I use whole genome data and coalescent methods to explore the effective population sizes of seven elapid snakes: three fully marine sea snakes, a sea krait and three terrestrial snakes. The findings indicate variable population histories between all snakes, with no obvious congruent patterns in effective population sizes within marine or terrestrial snakes.

Aim 2: Generate new genomic resources for the evolutionarily important and rapidly radiating *Hydrophis* sea snakes, and use them to investigate their evolutionary relationships, along with the genetic basis of the land-to-sea transition.

Sea snakes underwent a significant transition from land-to-sea, with *Hydrophis* being an excellent candidate species of an adaptive radiation into a new environment. However, until recently, there has been a deficit of genomic resources for *Hydrophis*, while molecular studies targeting the underlying genetics of the land-to-sea transition were severely lacking. In chapter 3, we present four new reference genomes for four different *Hydrophis* species, three of which are at chromosome-scale. Using these new resources, and publicly available data, we performed synteny analyses between six *Hydrophis* snakes and a terrestrial outgroup *Thamnophis elegans*, detailing extensive synteny across all chromosomes and a relatively stable karyotype within *Hydrophis* (2n = 32). Phylogenomic analyses found evidence of reticulate evolution and concurrent speciation within the *Hydrophis* snakes. Finally, using a strict screen for positive selection amongst protein coding genes, we found a set of 1,402 genes enriched for functional pathways relating to marine adaptation, with the vast majority of the marine-specific positively selected genes showing an intensification of selection relative to their terrestrial counterparts.

Aim 3: Characterise the genetic divergence landscape between three ecologically distinct, but coexisting, *Hydrophis* species.

Hydrophis sea snakes comprise a recent and exceptionally diverse radiation of fully marine reptiles. Chapter 4 presents the first detailed investigation into the divergence profiles between three ecologically distinct species that diverged near-simultaneously 3-8 million years ago. We used whole genome resequencing data to perform genome wide scans of divergence using between and within species diversity statistics. We additionally conducted formal tests of introgression using genome-wide *D* statistics and sliding window *f-statistics*. We found evidence of early gene-flow between the incipient species, with evidence of shared evolutionary factors shaping a relatively stable divergence landscape between all three species. Whole genome scans identified numerous high- and low-divergence peaks, showing patterns of background/positive selection and balancing selection, respectively. Notably, a balanced region on chromosome one was found to contain multiple chemosensory receptor genes, which are critical in mediating reproductive and feeding behaviours.

References

Barrett, R., & Schluter, D. (2008). Adaptation from standing genetic variation. *Trends in Ecology & Evolution*, *23*(1), 38–44. https://doi.org/10.1016/j.tree.2007.09.008

Barton, N. H. (2006). Evolutionary Biology: How Did the Human Species Form? *Current Biology*, *16*(16), R647–R650. https://doi.org/10.1016/j.cub.2006.07.032

Bierne, N., Bonhomme, F., & David, P. (2003). Habitat preference and the marine-speciation paradox. *Proceedings of the Royal Society of London. Series B:*

Biological Sciences, *270*(1522), 1399–1406. https://doi.org/10.1098/rspb.2003.2404

Bolnick, D. I. (2001). Intraspecific competition favours niche width expansion in Drosophila melanogaster. *Nature*, *410*(6827), 463–466. https://doi.org/10.1038/35068555

Bolnick, D. I., & Fitzpatrick, B. M. (2007). Sympatric Speciation: Models and Empirical Evidence. *Annual Review of Ecology, Evolution, and Systematics*, *38*(1), 459–487. https://doi.org/10.1146/annurev.ecolsys.38.091206.095804

Brischoux, F., & Shine, R. (2011). Morphological adaptations to marine life in snakes. *Journal of Morphology*, *272*(5), 566–572. https://doi.org/10.1002/jmor.10933

Butlin, R. K., Galindo, J., & Grahame, J. W. (2008). Sympatric, parapatric or allopatric: The most important way to classify speciation? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1506), 2997–3007. https://doi.org/10.1098/rstb.2008.0076

Coyne, J. A., Coyne, H. A., & Orr, H. A. (2004). *Speciation*. Oxford University Press, Incorporated.

Cruickshank, T. E., & Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, *23*(13). https://doi.org/10.1111/mec.12796

De Jode, A., Le Moan, A., Johannesson, K., Faria, R., Stankowski, S., Westram, A. M., Butlin, R. K., Rafajlović, M., & Fraïsse, C. (2023). Ten years of demographic modelling of divergence and speciation in the sea. *Evolutionary Applications*, *16*(2), 542–559. https://doi.org/10.1111/eva.13428

Feder, J. L., Egan, S. P., & Nosil, P. (2012). The genomics of speciation-with-gene-flow. *Trends in Genetics*, *28*(7), 342–350. https://doi.org/10.1016/j.tig.2012.03.009

Ferchaud, A.-L., & Hansen, M. M. (2016). The impact of selection, gene flow and demographic history on heterogeneous genomic divergence: Three-spine sticklebacks in divergent environments. *Molecular Ecology*, *25*(1), 238–259. https://doi.org/10.1111/mec.13399

Fisher, Ronald Aylmer, Sir. (1930). *The genetical theory of natural selection*. Oxford, Clarendon Press.

Gavrilets, S. (2003). Models of speciation: What have we learned in 40 years? *Evol*, *57*(10), 2197. https://doi.org/10.1554/02-727

Gavrilets, S., & Losos, J. B. (2009). Adaptive Radiation: Contrasting Theory with Data. *Science*, *323*(5915), 732–737. https://doi.org/10.1126/science.1157966

Gibson, G., & Dworkin, I. (2004). Uncovering cryptic genetic variation. *Nature Reviews Genetics*, *5*(9), 681–690. https://doi.org/10.1038/nrg1426

Gillespie, J. H. (1983). A simple stochastic gene substitution model. *Theoretical Population Biology*, *23*(2), 202–215. https://doi.org/10.1016/0040-5809(83)90014-X

Gillespie, J. H. (1984). Molecular Evolution over the Mutational. *Evolution*, *38*(5), 1116–1129. https://doi.org/10.1111/j.1558-5646.1984.tb00380.x

Glor, R. E. (2010). Phylogenetic Insights on Adaptive Radiation. *Annual Review of Ecology, Evolution, and Systematics*, *41*(1), 251–270. https://doi.org/10.1146/annurev.ecolsys.39.110707.173447

Hague, M. T. J., & Routman, E. J. (2016). Does population size affect genetic diversity?

A test with sympatric lizard species. *Heredity*, *116*(1), 92–98. https://doi.org/10.1038/hdy.2015.76

Heatwole, H. (1999). *Sea Snakes*. UNSW Press.

Hermisson, J., & Pennings, P. S. (2005). Soft Sweeps: Molecular Population Genetics of Adaptation From Standing Genetic Variation. *Genetics*, *169*(4), 2335–2352. https://doi.org/10.1534/genetics.104.036947

Hudson, J., Johannesson, K., McQuaid, C. D., & Rius, M. (2020). Secondary contacts and genetic admixture shape colonization by an amphiatlantic epibenthic invertebrate. *Evolutionary Applications*, *13*(3), 600–612. https://doi.org/10.1111/eva.12893

Kimura, M. (1983). *The neutral theory of molecular evolution* (1. paperback ed., re-issued). Cambridge Univ. Press.

Lee, M. S. Y., Sanders, K. L., King, B., & Palci, A. (2016). Diversification rates and phenotypic evolution in venomous snakes (Elapidae). *R. Soc. Open Sci.*, *3*(1), 150277. https://doi.org/10.1098/rsos.150277

Lewanski, A. L., Golcher-Benavides, J., Rick, J. A., & Wagner, C. E. (2022). Variable hybridization between two Lake Tanganyikan cichlid species in recent secondary contact. *Molecular Ecology*, *31*(19), 5041–5059. https://doi.org/10.1111/mec.16636

Lister', B. C. (1976). The Nature of Niche Expansion in West Indian Anolis Lizards II: Evolutionary Components. *Evolution*, *30*(4), 677–692.

MacArthur, R. H., Diamond, J. M., & Karr, J. R. (1972). Density Compensation in Island Faunas. *Ecology*, *53*(2), 330–342. https://doi.org/10.2307/1934090

Mayr, E. (1942). *Systematics and the Origin of Species*. Columbia University Press.

Nei, M., & Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, *76*(10), 5269–5273. https://doi.org/10.1073/pnas.76.10.5269

Nichols, P., Genner, M. J., Van Oosterhout, C., Smith, A., Parsons, P., Sungani, H., Swanstrom, J., & Joyce, D. A. (2015). Secondary contact seeds phenotypic novelty in cichlid fishes. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1798), 20142272. https://doi.org/10.1098/rspb.2014.2272

Noor, M. A. F., & Bennett, S. M. (2009). Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity*, *103*(6), 439–444. https://doi.org/10.1038/hdy.2009.151

Orr, H. A., & Betancourt, A. J. (2001). Haldane's Sieve and Adaptation From the Standing Genetic Variation. *Genetics*, *157*(2), 875–884. https://doi.org/10.1093/genetics/157.2.875

Rabosky, D. L., & Glor, R. E. (2010). Equilibrium speciation dynamics in a model adaptive radiation of island lizards. *Proceedings of the National Academy of Sciences*, *107*(51), 22178–22183. https://doi.org/10.1073/pnas.1007606107

Ravinet, M., Faria, R., Butlin, R. K., Galindo, J., Bierne, N., Rafajlović, M., Noor, M. a. F., Mehlig, B., & Westram, A. M. (2017). Interpreting the genomic landscape of speciation: A road map for finding barriers to gene flow. *Journal of Evolutionary Biology*, *30*(8), 1450–1477. https://doi.org/10.1111/jeb.13047

Rieseberg, L. H., Raymond, O., Rosenthal, D. M., Lai, Z., Livingstone, K., Nakazato, T., Durphy, J. L., Schwarzbach, A. E., Donovan, L. A., & Lexer, C. (2003). Major

Ecological Transitions in Wild Sunflowers Facilitated by Hybridization. *Science*, *301*(5637), 1211–1216. https://doi.org/10.1126/science.1086949

Rossi, V., Lo, M., Legrand, T., Ser-Giacomi, E., de Jode, A., Thierry de Ville d'Avray, L., Pairaud, I., Faure, V., Fraysse, M., Pinazo, C., & Chenuil, A. (2021). Small-scale connectivity of coralligenous habitats: Insights from a modelling approach within a semi-opened Mediterranean bay. *Vie et Milieu / Life & Environment*. https://hal.science/hal-03131637

Roughgarden, J. (1972). Evolution of Niche Width. *The American Naturalist*, *106*(952), 683–718.

Sanders, K. L., & Lee, M. S. Y. (2008). Molecular evidence for a rapid late-Miocene radiation of Australasian venomous snakes (Elapidae, Colubroidea). *Molecular Phylogenetics and Evolution*, *46*(3), 1165–1173. https://doi.org/10.1016/j.ympev.2007.11.013

Sanders, K. L., Lee, M. S. Y., Leys, R., Foster, R., & Keogh, J. S. (2008). Molecular phylogeny and divergence dates for Australasian elapids and sea snakes (hydrophiinae): Evidence from seven genes for rapid evolutionary radiations. *Journal of Evolutionary Biology*, *21*(3), 682–695. https://doi.org/10.1111/j.1420-9101.2008.01525.x

Sanders, K. L., Rasmussen, A. R., & Elmberg, J. (2012). Independent Innovation in the Evolution of Paddle-Shaped Tails in Viviparous Sea Snakes (Elapidae: Hydrophiinae). *Integrative and Comparative Biology*, *52*(2), 311–320. https://doi.org/10.1093/icb/ics066

Schield, D. R., Perry, B. W., Adams, R. H., Card, D. C., Jezkova, T., Pasquesi, G. I. M.,

Nikolakis, Z. L., Row, K., Meik, J. M., Smith, C. F., Mackessy, S. P., & Castoe, T. A. (2019). Allopatric divergence and secondary contact with gene flow: A recurring theme in rattlesnake speciation. *Biological Journal of the Linnean Society*, *128*(1), 149–169. https://doi.org/10.1093/biolinnean/blz077

Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. A., Peichel, C. L., Saetre, G.-P., Bank, C., Brännström, Å., Brelsford, A., Clarkson, C. S., Eroukhmanoff, F., Feder, J. L., Fischer, M. C., Foote, A. D., Franchini, P., Jiggins, C. D., Jones, F. C., … Widmer, A. (2014). Genomics and the origin of species. *Nat Rev Genet*, *15*(3), 176–192. https://doi.org/10.1038/nrg3644

Simões, B. F., Gower, D. J., Rasmussen, A. R., Sarker, M. A. R., Fry, G. C., Casewell, N. R., Harrison, R. A., Hart, N. S., Partridge, J. C., Hunt, D. M., Chang, B. S., Pisani, D., & Sanders, K. L. (2020). Spectral Diversification and Trans-Species Allelic Polymorphism during the Land-to-Sea Transition in Snakes. *Current Biology*, *30*(13), 2608-2615.e4. https://doi.org/10.1016/j.cub.2020.04.061

Simpson, G. G. (1944). *Tempo and mode in evolution.* Columbia Univ. Press.

Svanbäck, R., & Bolnick, D. I. (2005). Intraspecific competition affects the strength of individual specialization: An optimal diet theory method. *Evolutionary Ecology Research*, *7*, 993–1012.

Turelli, M., Barton, N. H., & Coyne, J. A. (2001). Theory and speciation. *Trends in Ecology & Evolution*, *16*(7), 330–343. https://doi.org/10.1016/S0169-5347(01)02177-2

Turner, J. R. G. (1981). Adaptation and Evolution in Heliconius: A Defense of

NeoDarwinism. *Annual Review of Ecology and Systematics*, *12*(1), 99–121. https://doi.org/10.1146/annurev.es.12.110181.000531

Wolf, J. B. W., & Ellegren, H. (2017). Making sense of genomic islands of differentiation in light of speciation. *Nat Rev Genet*, *18*(2), 87–100. https://doi.org/10.1038/nrg.2016.133

Wright, S. (1931). Evolution in Mendelian Populations. *Genetics*, *16*(2), 97–159. https://doi.org/10.1093/genetics/16.2.97

Wright, S. (1943). Isolation by Distance. *Genetics*, *28*(2), 114–138. https://doi.org/10.1093/genetics/28.2.114

Yamasaki, Y. Y., Kakioka, R., Takahashi, H., Toyoda, A., Nagano, A. J., Machida, Y., Møller, P. R., & Kitano, J. (2020). Genome-wide patterns of divergence and introgression after secondary contact between *Pungitius* sticklebacks. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1806), 20190548. https://doi.org/10.1098/rstb.2019.0548

Yoder, J. B., Clancey, E., Des Roches, S., Eastman, J. M., Gentry, L., Godsoe, W., Hagey, T. J., Jochimsen, D., Oswald, B. P., Robertson, J., Sarver, B. a. J., Schenk, J. J., Spear, S. F., & Harmon, L. J. (2010). Ecological opportunity and the origin of adaptive radiations. *Journal of Evolutionary Biology*, *23*(8), 1581–1596. https://doi.org/10.1111/j.1420-9101.2010.02029.x

Zheng, J., Payne, J. L., & Wagner, A. (2019). Cryptic genetic variation accelerates evolution by opening access to diverse adaptive peaks. *Science*, *365*(6451), 347–353. https://doi.org/10.1126/science.aax1837

Chapter 2:
Demographic analyses of marine and terrestrial snakes (Elapidae) using whole genome sequences

Ludington, A. J. & Sanders K. L (2020). Demographic analyses of marine and terrestrial snakes (Elapidae) using whole genome sequences. Molecular Ecology, 30 (2), 545-554

Statement of Authorship

Title of Paper	Demographic analyses of marine and terrestrial snakes (Elapidae) using whole genome sequences
Publication Status	☒ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Published in Molecular Ecology (November 2020). Doi: https://doi.org/10.1111/mec.15726

Principal Author

Name of Principal Author (Candidate)	Alastair John Ludington		
Contribution to the Paper	Conceived the research (with KLS, performed quantitative analyses, interpreted results, wrote the first version of the manuscript and edited subsequent versions.		
Overall percentage (%)	80		
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	28/08/2023

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	Kate Laura Sanders		
Contribution to the Paper	Conceived the research, interpreted results, wrote the first version of the manuscript and edited subsequent versions.		
Signature		Date	31/08/2023

Demographic analyses of marine and terrestrial snakes (Elapidae) using whole genome sequences

Ludington, A.J.[1*], Sanders, K.L.[1*]

[1] School of Biological Science, The University of Adelaide, South Australia 5005, Australia

[*] Correspondence: alastair.ludington@adelaide.edu.au; kate.sanders@adelaide.com.au

Running title: Demographic analyses of land vs sea snakes

Abstract

The question of whether spatial aspects of evolution differ in marine versus terrestrial realms has endured since Ernst Mayr's 1954 essay on marine speciation. Marine systems are often suggested to support larger and more highly connected populations, but quantitative comparisons with terrestrial systems have been lacking. Here, we compared the population histories of marine and terrestrial elapid snakes using the Pairwise Sequentially Markovian Coalescent (PSMC) model to track historical fluctuations in species' effective population sizes (N_e) from individual whole-genome sequences. To do this we generated a draft genome for the olive sea snake (*Aipysurus laevis*) and analysed this alongside six published elapid genomes and their sequence reads (marine species *Hydrophis curtus, H. melanocephalus* and *Laticauda laticaudata*; terrestrial species *Pseudonaja textilis, Naja Naja* and *Notechis scutatus*). Counter to the expectation that marine species should show higher overall N_e and less pronounced fluctuations in N_e, our analyses reveal demographic patterns that are highly variable

among species and do not clearly correspond to major ecological divisions. At deeper time intervals, the four marine elapids appear to have experienced relatively stable N_e, while each terrestrial species shows a prominent upturn in N_e starting at ~4 mya followed by an equally strong decline. However, over the last million years, all seven species show strong and divergent fluctuations. Estimates of N_e in the most recent intervals (~10 kya) are lowest in two of four marine species (*H. melanocephalus* and *Laticauda*), and do not correspond to contemporary range sizes in marine or terrestrial taxa.

Keywords: Elapid snake; effective population size; PSMC; whole genome sequencing; marine; terrestrial.

Introduction

Coalescent methods provide opportunities to track historical fluctuations in effective population sizes (N_e) from individual whole-genome sequences (Li and Durbin, 2011). Such inferences have the potential to contribute insights into species' demographic responses to past biotic pressures, such as predation and disease, and abiotic factors, particularly climatic change. Variation in species' responses to past environments will depend not only on complex interactions among biotic and abiotic processes, but also on species-specific traits such as dispersal propensity and physiological plasticity. Nevertheless, comparative studies of spatial and temporal genetic patterns have shown that co-occurring species often have congruent responses to past ecological and climatic events (e.g. Bowen et al., 2014; Moritz et al., 2009).

Broader differences in population histories have been hypothesised for taxa that occupy different biomes (e.g. Antonelli et al., 2018; Bowen et al., 2016; Lorenzen et al., 2012) and biological realms (e.g. Voskamp et al., 2017). In particular, the terrestrial and marine realms have been suggested to differ in the spatial scales they impose on population size and connectivity (Carr et al., 2003; Grosberg et al., 2012; Lourie and Vincent, 2004; Mayr, 1954). Reports of huge range sizes and weak genetic structure in marine fish and invertebrates support a view that the vast oceans support large populations that are highly connected by dispersal (e.g. Grant, 1998). However, other phylogeographic studies have shown that marine habitats are less contiguous and more ecologically heterogeneous than often assumed, and that many marine taxa have limited dispersive abilities (especially those lacking pelagic larval stages) (Crandall et al., 2008; Lourie et al., 2005). Resolving the perceived dichotomy between marine and terrestrial evolution will require quantitative evidence from suitable case studies (Dawson and Hamner, 2008; Steele, 1985). Front fanged snakes (Elapidae) provide an attractive system in this respect because they comprise a relatively young radiation (their most recent common ancestor is dated at less than 40 million years old), but contain more than 350 terrestrial, amphibious and fully marine taxa (Lee et al., 2016; Sanders et al., 2008). These species are concentrated in the Indo-Pacific region and share many life history attributes relevant to dispersal, including direct development (Shine, 1994).

Our study offers preliminary insights into the evolution of terrestrial versus marine elapids by comparing the population histories of three terrestrial and four marine species with large geographic ranges in the Asian and Australian regions. The India

cobra, *Naja naja*, occupies a wide range of terrestrial habitats throughout much of South Asia. Australian tiger snakes, *Notechis scutatus*, are found in subtropical and temperate habitats throughout most of southern Australia. The eastern brown snake, *Pseudonaja textilis*, occurs mostly in open habitats in eastern and central Australia and southern New Guinea. The blue-banded sea krait, *Laticauda laticaudata*, occurs from the Bay of Bengal to the South West Pacific and is amphibious - hunting in coral reefs and using terrestrial habitats for many other activities. Olive sea snakes, *Aipysurus laevis*, are fully marine and found in varied (reef and soft bottomed) shallow-water habitats from Western Australia to Papua New Guinea and New Caledonia. *Hydrophis curtus* (short sea snake) and *H. melanocephalus* (black-headed sea snake) belong to the rapidly speciating clade of fully marine *Hydrophis* sea snakes. *H. melanocephalus* is found in shallow habitats from Indonesia to Japan; *H. curtus* is widely distributed from the Persian Gulf to Australia.

We inferred past population dynamics for these six elapid species by applying the Pairwise Sequentially Markovian Coalescent (PSMC) model (Li and Durbin, 2011) to quantify changes in effective population size (N_e) from approximately 10 million to 10 thousand years ago. Sequentially Markovian coalescent models reconstruct fluctuations in N_e over time from patterns of heterozygosity along a single genome. N_e is estimated from the rate of coalescence events between alleles in discrete time intervals based on the principle that coalescence is more likely to occur when population size is small (and geographic population substructure is weak) (Mather et al., 2020). If the marine realm supports larger populations that experience higher connectivity (and thus weaker

substructure), we would expect the PSMC plots of marine elapids to show i) higher overall N_e and ii) less pronounced fluctuations in N_e compared to terrestrial species.

Methods

Aipysurus laevis genome sample information and sequencing

Genomic DNA was extracted from the liver tissue of an adult male olive sea snake (*Aipysurus laevis*: NCBI taxonomy ID 8678; specimen ID WAM R22384) collected off Broome in Western Australia according to procedures approved by the University of Adelaide Animal Ethics Committee (S-2015-119) and under a research permit (SF010002) granted by the Department of Parks and Wildlife of Western Australia. Genomic DNA was extracted and purified with a Gentra Puregene extraction kit (Qiagen, Hilden, Germany) following the manufacturer's instructions. Whole genome libraries were prepared with a TruSeq PCR Free Kit (Illumina Inc., San Diego, CA) and sequenced in two lanes with v4 Illumina chemistry in a HiSeq 2000 (Illumina Inc., San Diego, CA) with 125 bp Paired-End Dual indexes. Mate pair libraries with 3 kb, 5 kb and 8 kb insert sizes were constructed using Nextera Mate Pair Library prep kit (Illumina Inc., San Diego, CA) and sequenced in one lane of HiSeq 2000 (Illumina Inc., San Diego, CA). Both whole genome sequencing and mate pair sequencing were done at The University of Queensland Centre for Brain Genomics.

Raw read quality control

The quality of the raw DNA sequence data was assessed using FastQC v0.11.4 (Simon Andrews; https://www.bioinformatics.babraham.ac.uk/projects/fastqc) and ngsReports (Ward et al., 2020) before and after read and quality trimming. Whole genome sequence (WGS) paired-end and mate-pair libraries were then screened against the MiniKraken 8GB database of complete bacterial, archaeal and viral genomes (Wood and Salzberg, 2014) to identify and remove putative sequence contamination. AdapterRemoval v2.2.1 (Schubert et al., 2016) was then used to remove adapters in the paired-end and mate-pair libraries. The data were also quality filtered for ambiguous bases and low-quality sequences using the following parameters: *--trimqualities*, *--trimns*, *--minquality* 10 and *--minlength* 25.

Genome Assembly

Assemblies with K-mer values ranging from 31 to 63 were generated using the SOAPdenovo2 v2.04 assembly software (Luo et al., 2012). Parameters *avg_ins*, *rd_len_cutoff* and *pair_num_cutoff* were changed to 288, 110 and 5 respectively, while all other parameters were left at default settings. SOAPdenovo2 was run with the *all* parameter generating contigs and scaffolds using the paired-end WGS data. The Redundans pipeline (Pryszcz and Gabaldón, 2016) (v0.14a) was then used to conduct a single round of contig reduction, scaffold formation and gap filling. For each assembly, redundant contigs were identified and removed via a *LAST* (http://last.cbrc.jp/) all-by-all sequence alignment, where contigs were classified as redundant if they shared ≥ 95% identity and overlap with other contigs. For redundant clusters the longest, most

representative contig was kept. The non-redundant contigs for each assembly were then scaffolded using SSPACE3 (Boetzer et al., 2011)(v3.0) utilising all mate-pair sequence libraries (3, 5 and 8Kbp), with gaps being filled by GapCloser v1.12 (Luo et al., 2012) using the paired-end sequence data. The assemblies were then assessed on metrics including scaffold contiguity and assembly N50. Using these assembly metrics, the 49-mer assembly was selected for further refinement. PILON (Walker et al., 2014) was then used in diploid mode with no restrictions for two iterations of polish. Corrections included single base differences, indels, local mis-assemblies and gap-filling. For each round of polish the paired-end sequence data were aligned to the genome using BWA-MEM (Li and Durbin, 2009).

Selection of additional elapid genomes

High quality published genomes for six additional elapid species were selected for analysis. Paired end whole genome sequence data sets were downloaded from NCBI for *Notechis scutatus* (PRJ: PRJEB27871), *Pseudonaja textilis* (BP: PRJEB27869), *Naja naja* (PRJ: PRJNA527614)*, Laticauda laticaudata* (PRJ: PRJDB7226), *Hydrophis curtus* (PRJ: PRJNA597425) and *Hydrophis melanocephalus* (PRJ: PRJDB7271), along with their accompanying reference sequences in FASTA format (*N. scutatus* ID:14408; *P. textilis* ID:72610; *N. naja* ID:8395; *L. laticaudata* ID:62470; *H. curtus* DOI: https://doi.org/10.6084/m9.figshare.11391606.v5; *H. melanocephalus* ID:76247). For *N. naja*, only the 19 chromosomes and micro-chromosomes were used for analysis, discarding the 1878 unplaced scaffolds.

Phylogenetic analysis

A SNP alignment was created from variants generated using the BCFtools v1.9 variant pipeline (Li, 2011). BWA-MEM (Li and Durbin, 2009) was used to align whole genome sequence data to the *N. naja* reference genome with the Z-chromosome removed. BCFtools mpileup was used to generate genomic VCF files (gvcf) for each sample, with adjusted (*-C 50*) mapping qualities and filtering (≥ 20) of both mapping and base quality. BCFtools call was then used to call genotypes which were normalised using BCFtools norm, before being merged into a single multisample VCF using BCFtools merge. The merged VCF file was filtered for biallelic SNPs only using BCFtools view (*-m 2 -M 2)*. The biallelic gvcf file was then converted to a standard VCF file using BCFtools convert (--gvcf2vcf) before being filtered for per-sample depth ≥ 10 and variant quality ≥ 20. Low quality SNPs were removed along with any site at which data were missing for one or more of the seven samples (VCFtools --max-missing-count 1). The software vcf2phylip (Ortiz, 2019, p. 2) was then used to convert the filtered multi-sample VCF to a concatenated phylip file, which then had invariant sites removed.

Maximum likelihood (ML) trees were inferred for the concatenated alignment of 1,154,768 SNPs using RAxMl-NG (Kozlov et al., 2019) under the GTGTR4 + gamma model, with the Lewis ascertainment bias correction, and 1000 bootstraps. The Indian cobra, *N. naja*, was specified as an outgroup because there is strong support for the monophyly of hydrophiines (terrestrial Australian elapids and sea snakes) within the terrestrial Afro-Asian elapids (Lee et al., 2016). The best ML tree was visualised and formatted using the ggtree R-package (Yu et al., 2017)

PSMC analysis

Reads can be mapped to related species' genomes in the absence of conspecific reference genomes (e.g. Yim et al., 2014). At the time of our initial analyses, genome assemblies were not available for all of the snakes used in this study. We therefore explored the effect of cross-species alignment by aligning the reads to the then available reference genomes. Reads were aligned using BWA-MEM (Li and Durbin, 2009)**,** and shorter split read alignments were marked as secondary (-M). Average depth statistics were calculated for each alignment using SAMtools depth reporting all sites (-a) and a custom AWK shell script. As reference genomes and their accompanying data became available during the course of the analysis, these data were added to our final PSMC analyses.

As the SAMtools variant calling pipeline is now deprecated (http://www.htslib.org/doc/samtools-mpileup), consensus sequences were generated using the BCFtools v1.9 (Li, 2011) variant calling pipeline. The rationale for using the SAMtools/BCFtools variant calling approach for individual samples has been discussed in Patton et al. (2019) and Nadachowska-Brzyska et al. (2016). To summarise, the BCFtools pipeline uses single samples for SNP calling, does not use population frequencies for variant calling, and does not assume Hardy-Weinberg equilibrium. As we are running PSMC on single samples, this is a valid approach for optimising the number of per-sample variants.

To account for haploid Z-linked variants, which would lead to incorrect variant calls in heterogametic females, scaffolds from all snakes were aligned to the *N. naja* Z-

chromosome using *nucmer* from the MUMmer4 alignment system (Marçais et al., 2018). *N. naja* chromosomes were also aligned to the Z-chromosome to obtain a baseline alignment rate of autosomal sequences to the Z-chromosome. Coordinate files generated by MUMmer4 were filtered for high identity alignments (≥80% identical), and overlapping high-identity alignment fragments were collapsed into contiguous stretches of homology. Scaffolds that had ≥ 50% of their length aligned to the Z-chromosome were marked for removal.

The parameters for calling PSMC variants were based on previous studies (Li and Durbin, 2011; Patton et al., 2019) and the PSMC usage example (https://github.com/lh3/psmc). The arguments -C 50, -q 20 and -Q 25 were passed to BCFtools mpileup to adjust mapping qualities using the BWA-MEM recommended scaling coefficient of 50, to skip alignments with mapping qualities less than 20, and to skip bases with qualities less than 25. Variants were called using BCFtools call with parameter -c, utilising the original variant calling algorithm from SAMtools. SNPs were then filtered for proximity to indels (--SnpGap 10) and depth (1/3 average depth ≤ DP ≤ 2 * average depth) using BCFtools filter. Indels were then excluded from the final output (--exclude-types indels) using BCFtools view.

Filtered variants were then piped to the PSMC accessory function vcf2fq to generate diploid consensus FASTQ files. These were converted to FASTA-like psmcfa files using psmcfa, with the argument -q 20 to remove consensus calls with qualities ≤ 20. Scaffolds marked as putative Z-chromosome sequences were then filtered from the FASTA-like files using seqtk (https://github.com/lh3/seqtk). PSMC was run using default values for upper limit to time to most recent common ancestor (TMRCA) (-t 15) and

theta/rho (-r 4) and a range of atomic time intervals (-p) to assess the effect interval clustering on N_e estimates as per recommendations in Mather et al. (2020). Bootstrap support for estimates of N_e were generated for each PSMC run by splitting the psmcfa input sequences into shorter segments and performing 100 rounds of bootstrapping with replacement. Transformations of the results were performed in R using adapted scaling and plotting scripts from Liu and Hansen, 2017 (https://github.com/a-lud/snakes-demographic-history).

Parameter selection and scaling

To assess the effect of choice of number and length of time intervals (within which N_e is assumed to be constant), four alternative time segment patterns were tested. These were 4+5*3+4 (default), 4+10*3+6+8, 4+25*2+4+6 and 4+30*2+4+6+10. Over-fitting can result in large N_e estimates at ancient and recent time periods and is mitigated by merging neighbouring time segments to have identical coalescent rates (Mather et al., 2020). PSMC plots are scaled using mutation rate and generation time estimates. A per year mutation rate for elapids was estimated using the whole genome percentage sequence divergence of 5% between *H. melanocephalus* and *N. scutatus* calculated using the MUMmer4 utilities nucmer and dnadiff (Marçais et al., 2018), and a fossil-calibrated molecular divergence time for these taxa of 20 million years (Lee et al., 2016). This rate was converted into a rate of $1.25x10^{-08}$ per site per generation, based on a generation time of 10 years. Our generation time estimate is based on an age of sexual maturation of 2-3 years and lifespan of ~20 years (Greer, 1997). Due to the uncertainty in these scaling parameters for elapid snakes (and other non-model

species), we repeated the analysis using mutation rate (7.2 x 10^{-09} per site per generation) and generation time (3 years) estimates for snakes used in a recent study (Pasquesi et al., 2018).

Results

Sequencing and assembly of the *Aipysurus laevis* genome

We generated a draft genome assembly of the olive sea snake, *Aipysurus laevis* (Hydrophiinae). Over 213 Gbp of Illumina paired-end and mate-pair cleaned sequence data were used to assemble the *A. laevis* genome, resulting in a 1.85Gbp (~1.6Gbp without N sequence) genome assembly composed of 196,089 scaffolds with an N50 size of 24,809 Kbp. While quite fragmented, 99.5% of the WGS data aligned to the assembly, of which 88% aligned as proper pairs (Table 1), indicating that the orientation and composition of the assembled data are accurate. The genome size of ~1.6Gbp excluding N sequences was similar to that of other assembled elapid genomes.

Table 1: Alignment statistics of samples mapped to their respective genome assembly. Each sample is aligned using BWA-MEM. Coverage statistics were calculated using SAMtools depth and AWK commands. The alignment statistics were generated from the SAMtools flagstat command. Variant statistics were obtained using BCFtools stats.

Sample	Avg. depth	Breadth of coverage	Aln. (%)	Proper pair (%)	Genome size	No. called sites	No. SNPs
A. laevis	50.8	86.2	99.5	88.0	1,852,689,089	1,350,002,414	2,706,424
H. curtus	86.9	94.5	98.1	89.6	1,629,330,146	1,427,851,482	3,724,931
H. melanocephalus	106.4	89.5	96.8	69.2	1,402,639,853	1,195,169,115	1,632,029
L. laticaudata	120.4	90.8	99.7	83.0	1,558,706,106	1,304,122,637	1,565,188
N. naja	65.6	93.2	96.7	91.8	1,667,692,826	1,427,633,476	5,178,823
N. scutatus	56.9	95.2	98.5	75.1	1,665,525,958	1,015,507,485	4,173,526
P. textilis	60.8	97.5	98.2	76.7	1,590,035,073	1,007,477,368	2,393,916

Sample = Sample aligned to the reference genome

Avg. depth = Average alignment depth

Breadth of coverage = Proportion of bases covered by at least 1x coverage

Aln. (%) = Percentage of reads that aligned to the reference genome

Proper pair (%) = Percentage of reads that aligned in proper pairs

Genome size = Number of bases in the reference genome (including N sequence)

No. called sites = Number of reference positions called by *BCFtools* in the final consensus sequence

No. SNPs = Number of variants called against the reference genome

Phylogenetic analysis

The best-known ML tree (Figure 1) is consistent with previous molecular phylogenies for elapids in topology and relative branch lengths (e.g. Lee et al., 2016; Sanders and Lee,

2008). The sea krait, *L. laticaudata*, is resolved as the sister lineage to the terrestrial and fully marine hydrophiines. The tiger and brown snakes (*N. scutatus* and *P. textilis*) are separated by a short internal branch and form successive sister lineages to the three fully marine sea snakes. The newly sequenced olive sea snake, *A. laevis*, is a relatively distant sister lineage to the recently diverged *H. curtus* and *H. melanocephalus*.

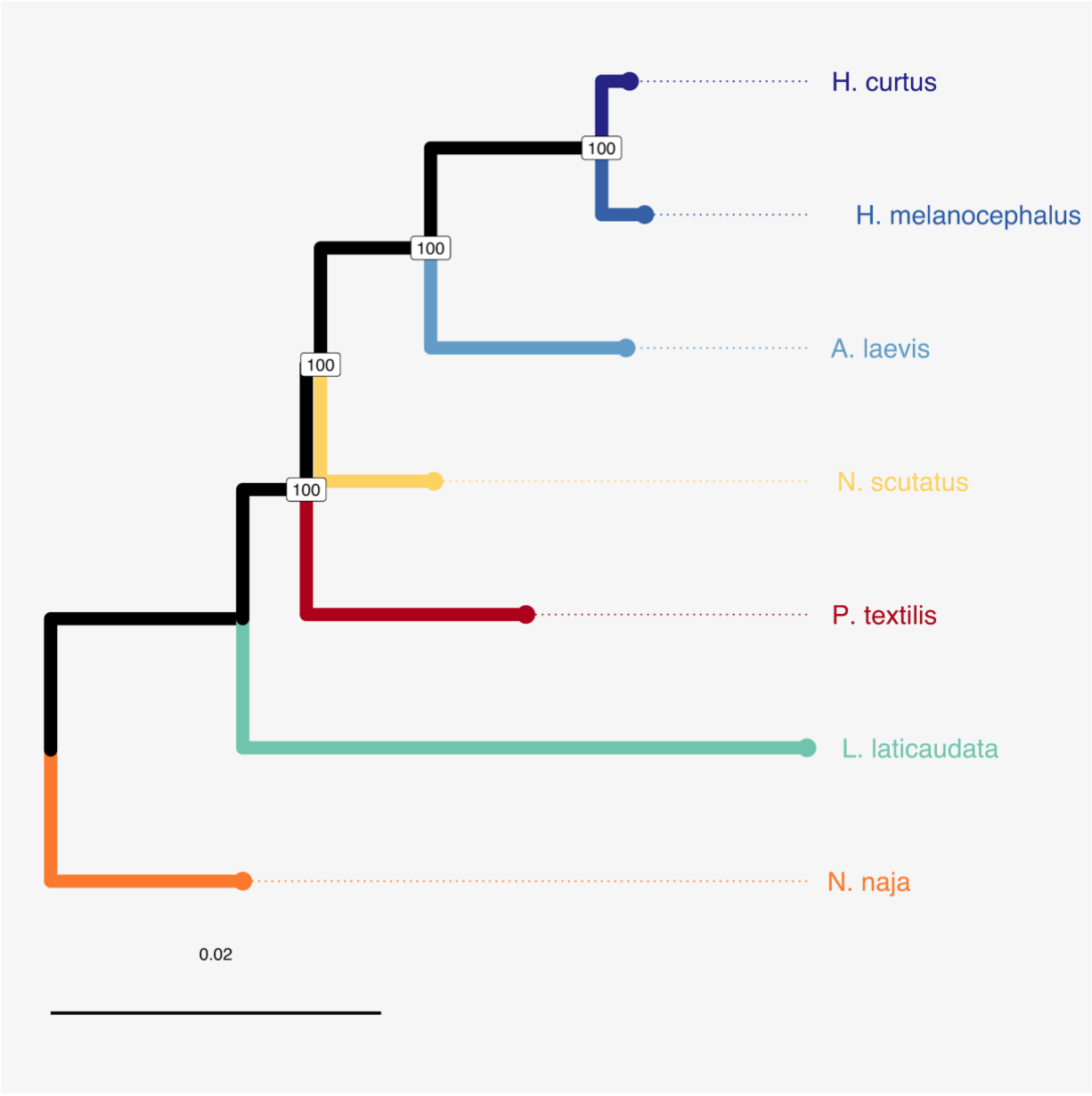


Figure 1: SNP-based maximum likelihood tree of the seven snake genomes used in this study. *N. naja* was specified as the outgroup. Bootstrap support is shown for internal branches (1000 replicates). Scale bar is in substitutions per site. Branches of marine species are shown as cool colours, and terrestrial species are shown as warm colours.

Read alignment to the elapid genomes

The proportion of aligned reads of the six elapid samples against their own reference genomes was high (96.8% - 99.7%), with average alignment depths ranging from 50x to 120x coverage. These values were well above the recommended 18x coverage (Nadachowska-Brzyska et al., 2016). Alignment rates of samples against alternate references (excluding alignments to their own genomes) ranged from 73.5% to 98.1%, with average alignment rates of all samples to each of the genomes ranging between 79.8% and 89.6% (Supplementary information Table S1). The reference coverage was typically lower for samples mapped to alternate references at 37.7x - 99.7x coverage (Supplementary information Table S1), but these values were still above the recommended minimum.

Z-Chromosome filtering

Removal of Z-chromosome sequences from the assembled genomes via *MUMmer4* alignments revealed that autosomal sequences generally share a low proportion of sequence homology (10-25%) to the Z-chromosome (Supplementary information figure S1). Sequences with greater than 50% of their length aligning to the *N. naja* Z-chromosome with high identity were generally shorter than the median scaffold length in

the genome assembly (Supplementary information figures S2 and S3). To be certain that Z-chromosome sequences were removed from all assembled genomes, a stringent filtering threshold was used whereby any sequence that shared greater than 50% of its sequence length with the *N. naja* Z-chromosome at high sequence identity (≥ 80%) was removed from the analysis. Using this strict filtering threshold still retained a majority of each genome's total sequence composition (Supplementary information figure S4). The effect of Z-chromosome variants on PSMC curves appeared to have minimal impact. This is demonstrated for the three female snakes in supplementary information figure S5.

Parameter selection and scaling

The impacts of cross-species alignment and clock selection are presented in supplementary information figure S6. Using the default PSMC clock, N_e dynamics are relatively consistent for each sample when aligned to reference sequences that are not its own. Alignments to the sample's own reference generally yield more pronounced trajectories that deviate from the cross-species results; however the overall shape of the curve is relatively consistent. Conversely, using a more complex clock yields significantly more detailed N_e trajectories at time points ≥ 10 kya, but can result in overfitting of the model and dispersion of bootstrap replicates at time points ≤ 10 kya in cross-species alignments (Supplementary information figure S6). This results in excessive estimates of population size reaching upward of 10^7 - 10^{10}. This observation is consistent with the expectation that few recent coalescence events can be inferred from single genome sequences so that over splitting of time intervals will yield higher variance in the number coalescent events per interval (Mather et al., 2020). The

disproportionately large N_e estimates in recent time periods is observed in all marine snakes aligned to terrestrial snake genomes, and in *N. naja* when aligned to *N. scutatus* and *P. textilis* (Supplementary information figure S6). N_e inferences from samples aligned to their own reference showed low variance in bootstrap replicates and were essentially comparable in trajectory, relative magnitude and timing across the four alternative time interval selections (Supplementary information figure S7). Based on these observations, we focused our discussion on the results of analyses of alignments to each sample's own reference.

The different scaling parameters tested had a strong impact on the magnitude and timeframe of Ne curves **(**Supplementary information figure S8). Combinations of g=3 with $\mu=7.2 \times 10^{-09}$ and g=10 with $\mu=1.25 \times 10^{-08}$ yield time axes up to ~10 mya versus ~19 mya, respectively. We treat the time axes of our Ne plots as highly uncertain and focus our discussion on the results of the runs using the genomic mutation rate calculated for elapids (this study), and a generation time of 10 years based on an age of sexual maturation of 2-3 years and lifespan exceeding 15-20 years (Greer, 1997).

Variation in *Ne* dynamics among elapid snakes

PSMC plots (Fig 2) show highly variable N_e curves among the seven species from approximately 10 million years ago (mya) to 10 thousand years ago (kya) (based on scaling parameter values of g=10 and $u=1.25 \times 10^{-08}$). The curves of the three terrestrial species (warm colours on Figure 2) are relatively stable from 10 mya, then show strong increases in N_e at ~4 mya, and equally strong declines in N_e starting between 400 and 50 kya. The Australian brown snake (*P. textilis*) and Indian cobra (*N. naja*) have

similarly shaped N_e curves (showing bimodal peaks at ~2 mya in both species and at ~200 kya in *P. textilis* and ~300 kya in *N. naja*), but these fluctuations have much higher amplitudes in the brown snake. The initial increase in N_e is gradual in the Australian tiger snake (*N. scutatus*), peaks at ~400 kya, then declines slightly and peaks again at ~50 kya before declining sharply. All four marine species (cool colours on Figure 2) have relatively stable N_e trajectories at deeper time intervals but show dramatic changes in N_e over the last million years. The olive sea snake (*A. laevis*), sea krait (*L. laticaudata*) and short sea snake (*H. curtus*) show gradual declines in N_e, from ~3 mya, but whereas the curve of *L. laticaudata* continues on a mostly downward trajectory up to ~10 kya, *A. laevis* and *H. curtus* undergo increases in N_e from ~500 kya The N_e curve of the black headed sea snake (*H. melanocephalus*) comprises three peaks with increasing amplitude followed by a steep decline in N_e. In recent time intervals, N_e estimates are highest in *A. laevis,* followed by *N. scutatus* and *H. curtus*, and are lowest in *H. melanocephalus* followed by *L. laticaudata*.

All seven species showed highest variance in N_e among bootstrap replicates at recent time intervals (<15 kya), which is consistent with the expectation that few recent coalescence events can be inferred from single genome sequences (Li and Durbin 2011). Estimates of N_e in the most recent intervals (~10 kya) are lowest in *H. melanocephalus* and *L. laticaudata* and highest in *A. laevis*; *N. scutatus* has larger recent N_e than either *P. textilis* or *N. naja*.

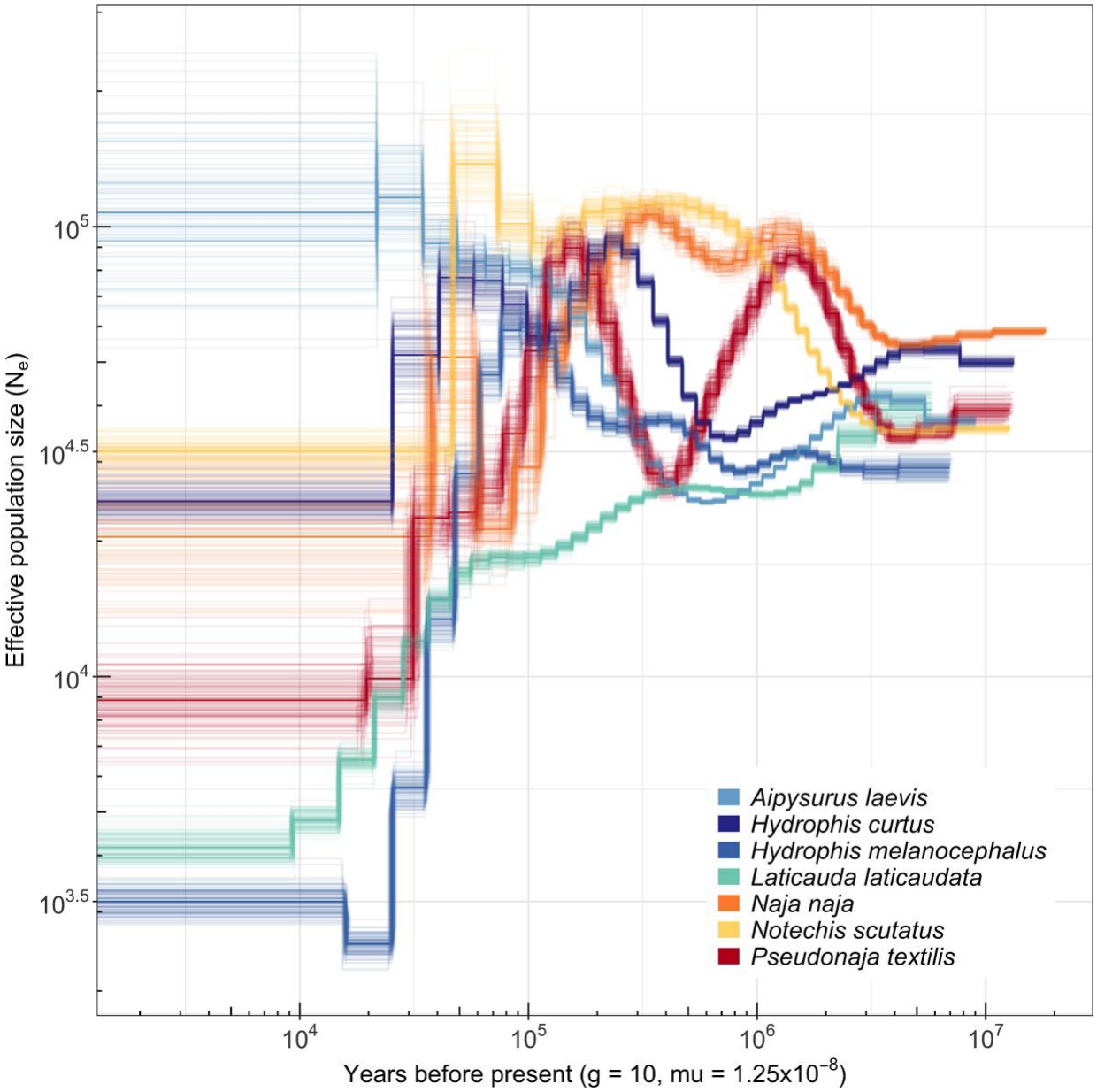


Figure 2: PSMC analyses of marine (cool colours) and terrestrial (warm colours) elapid snakes mapped to their respective genomes. Axes were converted to demographic units (number of individuals, and years) based on approximate generation time and mutation rate estimates of g=10 and μ=$1.25x10^{-08}$.

Discussion

In this paper, we reported a *de novo* genome assembly for the olive sea snake, *Aipysurus laevis* (Elapidae: Hydrophiinae). This genome fills an important gap in the genomic resources available for snakes and was analysed alongside six published genomes to provide new insights into the past population dynamics of marine versus terrestrial elapids. Our comparative study included a terrestrial Asian cobra (*N. naja*) and two Australian terrestrial elapids (*P. textilis* and *N. scutatus*); an amphibious sea krait (*L. laticaudata*); and three fully marine sea snakes (*H. curtus, H. melanocephalus* and *A. laevis*). Historical effective population sizes (N_e) were reconstructed using each elapid genome under the Pairwise Sequentially Markovian Coalescent (PSMC) model.

The data provided sufficient resolution to quantify changes in N_e over an approximate time frame of 10 million years ago (mya) up to 10 thousand years ago (kya). However, some important caveats should be considered when interpreting PSMC inferences. Firstly, among species comparisons of N_e curves are biased by variation in levels of genome coverage, and choice of reference genome for mapping (Nadachowska-Brzyska et al., 2016). Sequences with higher mean coverage will more accurately capture patterns of heterozygosity, resulting in more pronounced signals of change in N_e (Nadachowska-Brzyska et al., 2016; Patton et al., 2019). As divergence between reference and sample increases and alignment rates decline, so does the rate of accurate calls of heterozygous sites, resulting in dampened signals of change in N_e (Cho et al., 2013). In the present study, when samples were aligned to their respective genomes, coverage ranged from ~50x in *A. laevis* to ~120x in *L. laticaudata* (Table 1),

above the recommended minimum from Nadachowska-Brzyska et al. (2016). Mapping to divergent reference genomes decreased the proportion of high-quality alignments, resulting in lower average coverage and fewer called consensus sites (Supplementary information table S1**)**. This effect is illustrated in supplementary information figures S6, which highlights the importance of exploring the effect of reference selection when no genome is available for the species under investigation. To minimise these effects, we used sample data mapped to their respective genome assemblies.

PSMC analyses are scaled using mutation rate and generation time estimates. Choice of values for scaling parameters does not impact the general shape of N_e curves, but smaller generation times yield higher N_e estimates, and mutation rates per generation determine the timescale of reconstructions (supplementary information figure S8). In this study, we calculated a genome wide mutation rate using fossil-calibrated molecular divergence times for elapids and whole genome alignments between *N. scutatus* and *H. melanocephalus*. While there is considerable uncertainty in this rate estimate, the closely related species analysed here are unlikely to vary substantially in their per year mutation rates. The use of a single generation time estimate for all seven species was probably more problematic, but there are currently few data with which to estimate species specific generation times (and thus per generation mutation rates) for scaling PSMC profiles.

To avoid being misled by biases due to scaling and variability of genome coverage, we focus our among species comparisons primarily on the trajectories of change in N_e, and emphasise that the time axes are approximate. The PSMC plot for each species mapped to its own reference genome (Figure 2) shows idiosyncratic N_e

trajectories from ~10 mya to 10 kya. However, each terrestrial species (warm colours on Figure 2) shows a prominent upturn in N_e starting at ~4 mya, and an equally strong decline in N_e in more recent time intervals (from 400-200 kya in the cobra and brown snake, and from 50 kya in the tiger snake). N_e curves are clearly bimodal in the cobra and brown snake and somewhat bimodal in the tiger snake. The four marine lineages (cool colours on Figure 2) have relatively stable N_e at deeper time periods compared to the terrestrial species and have strongly fluctuating N_e over more recent intervals. The three fully marine sea snakes (*H. curtus, H. melanocephalus* and *A. laevis*) show coincident increases in N_e from ~500kya, but their trajectories depart at ~100 kya with *H. melanocephalus* undergoing a sharp decline, *H, curtus* following a bimodal trajectory, and *A. laevis* continuing on an upward trajectory. The curve of the sea krait, *L. laticaudata*, follows a mostly downward trajectory from 5 mya, and this is steepest from ~400 to ~10 kya.

Historical fluctuations in the N_e of marine and terrestrial elapids are not unexpected in light of the climatic oscillations that dramatically impacted land-sea configurations in the Indo-Pacific over the last few million years (Voris, 2000). Mitochondrial phylogeographic data for all six species show signals of geographic genetic structure suggestive of historical isolation followed by expansion and colonisation of new habitats (*P. textilis*: Skinner et al., 2005; *N. scutatus*: Keogh et al., 2005; *N. naja*:Dissanayake et al., 2015; *L. laticaudata*: Tandavanitj et al., 2013; *A. laevis*: Nitschke et al 2018; *H. curtus*: Ukuwela et al., 2014; *H. melanocephalus*: Sanders et al., 2013a). Particularly notable in this respect is the Australian brown snake, *P. textilis*. This species contains three distinct mitochondrial clades on mainland

Australia and appears to have colonised New Guinea at least twice in the Pleistocene via a land bridge that connected the two landmassess during lower sea levels (Skinner et al., 2005; Williams et al., 2008). Our analyses show much stronger fluctuations in N_e in the brown snake compared to the Indian cobra (*N. naja*), although both species have wide continental distributions and use varied habitats. This difference is not merely an artifact of variation in sequence coverage because these are similar in the two species (Table 1).

PSMC curves are expected to track changes in population size based on the principle that coalescence times of alleles will be longer in large populations. However, geographic population substructure will also hinder rates of coalescence, so that N_e trajectories might additionally (or alternatively) represent changes in isolation and migration (Mather et al., 2020; Mazet et al., 2016). The fully marine *Hydrophis* lineage has a complex history of speciation and introgression, comprising a very rapid radiation of at least 50 species with overlapping ranges in the Indo-Pacific (Sanders et al., 2013a). The *H. melanocephalus* genome shows signals of fluctuating N_e during deeper time intervals, but the most dramatic changes in N_e appear to have occured over the last one million years. A previous study placed *H. melanocephalus* in a clade of four nominal species that have overlapping ranges and show a lack mitochondrial monophyly indicative of divergence within the last ~200 thousand years and/or interspecific introgression (Sanders et al., 2013b). Microsatellite data suggest limited recent gene flow among these species, including *H. melanocephalus* and its closest relative, *H. cyanocinctus*. It is possible that reduced gene flow between *H.*

melanocephalus and co-occurring *Hydrophis* species explains the sharp decline in N_e from ~100 kya in *H. melanocephalus.*

If marine populations are larger and more highly connected compared to terrestrial species, we would expect the PSMC plots of marine species to show higher overall N_e and less pronounced fluctuations in N_e (regardless of whether N_e profiles reflect changes in population size or structure). Our analyses of elapid snakes reveal demographic patterns that are highly variable among species and do not clearly correspond to major ecological divisions. Marine elapids appear to have experienced more stable N_e compared to terrestrial species at deeper time intervals, but all seven species show strong and divergent fluctuations over the last million years. Estimates of N_e in the most recent intervals (~10 kya) are lowest in two of the four marine species, and do not correspond to contemporary range sizes. *H. curtus* and *L. laticaudata* occupy much larger geographic areas than *A. laevis* but show much smaller recent N_e. *N. scutatus* has the smallest geographic range among the terrestrial species but has the largest recent N_e. Interspecific differences in the spatial extent of recent population genetic structure provide a plausible explanation for these counterintuitive patterns. However, land- and seascape genomic analyses of the spatial scale and timing of demographic events will be needed to interpret interspecific differences in recent N_e. Our study offers preliminary comparisons of the demographic histories of marine versus terrestrial elapids. As additional elapid genomes become available, coalescent methods will provide valuable opportunities to quantify differences among these taxa and better understand the factors that drive within- versus between-realm differences.

Acknowledgements

KLS was supported by an Australian Research Council Future Fellowship (FT130101965) and Australian Research Council Discovery Grant (DP180101688). We thank the Western Australian Department of Biodiversity, Conservation and Attractions and The University of Adelaide Animal Ethics Committee for permissions to collect the *Aipysurus laevis* specimen used for genome sequencing. The Queensland Brain Institute carried out the sequencing of *A. laevis.* Mick Woddley (Absolute Oceans Charters, Broome) and Jenna Crowe-Riddell are thanked for field assistance, and James Breen and Christopher Ward for helpful discussions.

References

Antonelli, A., Ariza, M., Albert, J., Andermann, T., Azevedo, J., Bacon, C., Faurby, S., Guedes, T., Hoorn, C., Lohmann, L.G., Matos-Maraví, P., Ritter, C.D., Sanmartín, I., Silvestro, D., Tejedor, M., ter Steege, H., Tuomisto, H., Werneck, F.P., Zizka, A., Edwards, S.V., 2018. Conceptual and empirical advances in Neotropical biodiversity research. PeerJ 6, e5644. https://doi.org/10.7717/peerj.5644

Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., Pirovano, W., 2011. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27, 578–579. https://doi.org/10.1093/bioinformatics/btq683

Bowen, B.W., Gaither, M.R., DiBattista, J.D., Iacchei, M., Andrews, K.R., Grant, W.S., Toonen, R.J., Briggs, J.C., 2016. Comparative phylogeography of the ocean planet. Proc. Natl. Acad. Sci. 113, 7962–7969.

https://doi.org/10.1073/pnas.1602404113

Bowen, B.W., Shanker, K., Yasuda, N., Celia, M., Malay, M.C. (Machel) D., von der Heyden, S., Paulay, G., Rocha, L.A., Selkoe, K.A., Barber, P.H., Williams, S.T., Lessios, H.A., Crandall, E.D., Bernardi, G., Meyer, C.P., Carpenter, K.E., Toonen, R.J., 2014. Phylogeography unplugged: comparative surveys in the genomic era. Bull. Mar. Sci. 90, 13–46. https://doi.org/10.5343/bms.2013.1007

Carr, M.H., Neigel, J.E., Estes, J.A., Andelman, S., Warner, R.R., Largier, J.L., 2003. Comparing Marine and Terrestrial Ecosystems: Implications for the Design of Coastal Marine Reserves. Ecol. Appl. 13, 90–107. https://doi.org/10.1890/1051-0761(2003)013[0090:CMATEI]2.0.CO;2

Cho, Y.S., Hu, L., Hou, H., Lee, H., Xu, J., Kwon, S., Oh, S., Kim, H.-M., Jho, S., Kim, Sangsoo, Shin, Y.-A., Kim, B.C., Kim, H., Kim, C., Luo, S.-J., Johnson, W.E., Koepfli, K.-P., Schmidt-Küntzel, A., Turner, J.A., Marker, L., Harper, C., Miller, S.M., Jacobs, W., Bertola, L.D., Kim, T.H., Lee, S., Zhou, Q., Jung, H.-J., Xu, X., Gadhvi, P., Xu, P., Xiong, Y., Luo, Y., Pan, S., Gou, C., Chu, X., Zhang, J., Liu, Sanyang, He, Jing, Chen, Y., Yang, L., Yang, Y., He, Jiaju, Liu, Sha, Wang, Junyi, Kim, C.H., Kwak, H., Kim, J.-S., Hwang, S., Ko, J., Kim, C.-B., Kim, Sangtae, Bayarlkhagva, D., Paek, W.K., Kim, S.-J., O'Brien, S.J., Wang, Jun, Bhak, J., 2013. The tiger genome and comparative analysis with lion and snow leopard genomes. Nat. Commun. 4, 2433. https://doi.org/10.1038/ncomms3433

Crandall, E.D., Jones, M.E., Muñoz, M.M., Akinronbi, B., Erdmann, M.V., Barber, P.H., 2008. Comparative phylogeography of two seastars and their ectosymbionts within the Coral Triangle. Mol. Ecol. 17, 5276–5290.

https://doi.org/10.1111/j.1365-294X.2008.03995.x

Dawson, M.N., Hamner, W.M., 2008. A biophysical perspective on dispersal and the geography of evolution in marine and terrestrial systems. J. R. Soc. Interface 5, 135–150. https://doi.org/10.1098/rsif.2007.1089

Dissanayake, D., Rajapakse, R., Bandara, K., Ukuwela, K., Kularatne, S., Ranasinghe, S., 2015. Phylogeography of the Indian Cobra (Naja naja) reveals genetically divergent populations between the Indian subcontinent and Sri Lanka. Genome 58, 212–213.

Grant, W., 1998. Shallow population histories in deep evolutionary lineages of marine fishes: insights from sardines and anchovies and lessons for conservation. J. Hered. 89, 415–426. https://doi.org/10.1093/jhered/89.5.415

Greer, A.E., 1997. The biology and evolution of Australian snakes.

Grosberg, R.K., Vermeij, G.J., Wainwright, P.C., 2012. Biodiversity in water and on land. Curr. Biol. 22, R900–R903. https://doi.org/10.1016/j.cub.2012.09.050

Keogh, J.S., Scott, I.A.W., Hayes, C., 2005. Rapid and Repeated Origin of Insular Gigantism and Dwarfism in Australian Tiger Snakes. Evolution 59, 226–233. https://doi.org/10.1111/j.0014-3820.2005.tb00909.x

Lee, M.S.Y., Sanders, K.L., King, B., Palci, A., 2016. Diversification rates and phenotypic evolution in venomous snakes (Elapidae). R. Soc. Open Sci. 3, 150277. https://doi.org/10.1098/rsos.150277

Li, H., 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27, 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, H., Durbin, R., 2011. Inference of human population history from individual whole-genome sequences. Nature 475, 493–496. https://doi.org/10.1038/nature10231

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Liu, S., Hansen, M.M., 2017. PSMC (pairwise sequentially Markovian coalescent) analysis of RAD (restriction site associated DNA) sequencing data. Mol. Ecol. Resour. 17, 631–641. https://doi.org/10.1111/1755-0998.12606

Lorenzen, E.D., Heller, R., Siegismund, H.R., 2012. Comparative phylogeography of African savannah ungulates1. Mol. Ecol. 21, 3656–3670. https://doi.org/10.1111/j.1365-294X.2012.05650.x

Lourie, S.A., Green, D.M., Vincent, A.C.J., 2005. Dispersal, habitat differences, and comparative phylogeography of Southeast Asian seahorses (Syngnathidae: Hippocampus). Mol. Ecol. 14, 1073–1094. https://doi.org/10.1111/j.1365-294X.2005.02464.x

Lourie, S.A., Vincent, A.C.J., 2004. Using Biogeography to Help Set Priorities in Marine Conservation. Conserv. Biol. 18, 1004–1020. https://doi.org/10.1111/j.1523-1739.2004.00137.x

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Yunjie, Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Yong, Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D.W., Yiu, S.-M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, Jian, Lam, T.-W., Wang, Jun, 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler.

GigaScience 1, 18. https://doi.org/10.1186/2047-217X-1-18

Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., Zimin, A., 2018. MUMmer4: A fast and versatile genome alignment system. PLOS Comput. Biol. 14, e1005944. https://doi.org/10.1371/journal.pcbi.1005944

Mather, N., Traves, S.M., Ho, S.Y.W., 2020. A practical introduction to sequentially Markovian coalescent methods for estimating demographic history from genomic data. Ecol. Evol. 10, 579–589. https://doi.org/10.1002/ece3.5888

Mayr, E., 1954. Geographic Speciation in Tropical Echinoids. Int. J. Org. Evol. 8, 19.

Mazet, O., Rodríguez, W., Grusea, S., Boitard, S., Chikhi, L., 2016. On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference? Heredity 116, 362–371. https://doi.org/10.1038/hdy.2015.104

Moritz, C., Hoskin, C.J., MacKenzie, J.B., Phillips, B.L., Tonione, M., Silva, N., VanDerWal, J., Williams, S.E., Graham, C.H., 2009. Identification and dynamics of a cryptic suture zone in tropical rainforest. Proc. R. Soc. B Biol. Sci. 276, 1235–1244. https://doi.org/10.1098/rspb.2008.1622

Nadachowska-Brzyska, K., Burri, R., Smeds, L., Ellegren, H., 2016. PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white *Ficedula* flycatchers. Mol. Ecol. 25, 1058–1072. https://doi.org/10.1111/mec.13540

Ortiz, E.M., 2019. vcf2phylip v2.0: convert a VCF matrix into several matrix formats for phylogenetic analysis. Zenodo. https://doi.org/10.5281/zenodo.2540861

Pasquesi, G.I.M., Adams, R.H., Card, D.C., Schield, D.R., Corbin, A.B., Perry, B.W.,

Reyes-Velasco, J., Ruggiero, R.P., Vandewege, M.W., Shortt, J.A., Castoe, T.A., 2018. Squamate reptiles challenge paradigms of genomic repeat element evolution set by birds and mammals. Nat. Commun. 9, 2774. https://doi.org/10.1038/s41467-018-05279-1

Patton, A.H., Margres, M.J., Stahlke, A.R., Hendricks, S., Lewallen, K., Hamede, R.K., Ruiz-Aravena, M., Ryder, O., McCallum, H.I., Jones, M.E., Hohenlohe, P.A., Storfer, A., 2019. Contemporary Demographic Reconstruction Methods Are Robust to Genome Assembly Quality: A Case Study in Tasmanian Devils. Mol. Biol. Evol. 36, 2906–2921. https://doi.org/10.1093/molbev/msz191

Pryszcz, L.P., Gabaldón, T., 2016. Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic Acids Res. 44, e113–e113. https://doi.org/10.1093/nar/gkw294

Sanders, K.L., Lee, M.S.Y., 2008. Molecular evidence for a rapid late-Miocene radiation of Australasian venomous snakes (Elapidae, Colubroidea). Mol. Phylogenet. Evol. 46, 1165–1173. https://doi.org/10.1016/j.ympev.2007.11.013

Sanders, K.L., Lee, M.S.Y., Leys, R., Foster, R., Keogh, J.S., 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. https://doi.org/10.1111/j.1420-9101.2008.01525.x

Sanders, K.L., Lee, M.S.Y., Mumpuni, Bertozzi, T., Rasmussen, A.R., 2013a. Multilocus phylogeny and recent rapid radiation of the viviparous sea snakes (Elapidae: Hydrophiinae). Mol. Phylogenet. Evol. 66, 575–591. https://doi.org/10.1016/j.ympev.2012.09.021

Sanders, K.L., Rasmussen, A.R., Mumpuni, Elmberg, J., Silva, A. de, Guinea, M.L., Lee, M.S.Y., 2013b. Recent rapid speciation and ecomorph divergence in Indo-Australian sea snakes. Mol. Ecol. 22, 2742–2759. https://doi.org/10.1111/mec.12291

Schubert, M., Lindgreen, S., Orlando, L., 2016. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res. Notes 9, 88. https://doi.org/10.1186/s13104-016-1900-2

Shine, R., 1994. Allometric Patterns in the Ecology of Australian Snakes. Copeia 1994, 851. https://doi.org/10.2307/1446709

Skinner, A., Donnellan, S.C., Hutchinson, M.N., Hutchinson, R.G., 2005. A phylogenetic analysis of Pseudonaja (Hydrophiinae, Elapidae, Serpentes) based on mitochondrial DNA sequences. Mol. Phylogenet. Evol. 37, 558–571. https://doi.org/10.1016/j.ympev.2005.06.020

Steele, J.H., 1985. A comparison of terrestrial and marine ecological systems. Nature 313, 355–358. https://doi.org/10.1038/313355a0

Tandavanitj, N., Ota, H., Cheng, Y.-C., Toda, M., 2013. Geographic Genetic Structure in Two Laticaudine Sea Kraits, *Laticauda laticaudata* and *Laticauda semifasciata* (Serpentes: Elapidae), in the Ryukyu-Taiwan Region as Inferred from Mitochondrial Cytochrome *b* Sequences. Zoolog. Sci. 30, 633–641. https://doi.org/10.2108/zsj.30.633

Ukuwela, K.D.B., Silva, A. de, Mumpuni, Fry, B.G., Sanders, K.L., 2014. Multilocus phylogeography of the sea snake Hydrophis curtus reveals historical vicariance and cryptic lineage diversity. Zool. Scr. 43, 472–484.

https://doi.org/10.1111/zsc.12070

Voris, H.K., 2000. Maps of Pleistocene sea levels in Southeast Asia: shorelines, river systems and time durations. J. Biogeogr. 27, 1153–1167. https://doi.org/10.1046/j.1365-2699.2000.00489.x

Voskamp, A., Baker, D.J., Stephens, P.A., Valdes, P.J., Willis, S.G., 2017. Global patterns in the divergence between phylogenetic diversity and species richness in terrestrial birds. J. Biogeogr. 44, 709–721. https://doi.org/10.1111/jbi.12916

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., Earl, A.M., 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS ONE 9, e112963. https://doi.org/10.1371/journal.pone.0112963

Ward, C.M., To, T.-H., Pederson, S.M., 2020. ngsReports: a Bioconductor package for managing FastQC reports and other NGS related log files. Bioinformatics 36, 2587–2588. https://doi.org/10.1093/bioinformatics/btz937

Williams, D., O'Shea, M., DAGUERRE, R., POOK, C., Wüster, W., HAYDEN, C., McVay, J., Paiva, O., Matainaho, L., Winkel, K., Austin, C., 2008. Origin of the eastern brownsnake, Pseudonaja textilis (Duméril, Bibron and Duméril) (Serpentes: Elapidae: Hydrophiinae) in New Guinea: evidence of multiple dispersals from Australia, and comments on the status of Pseudonaja textilis pughi Hoser 2003. Zootaxa 1703, 47–61. https://doi.org/10.11646/zootaxa.1703.1.3

Wood, D.E., Salzberg, S.L., 2014. Kraken: ultrafast metagenomic sequence

classification using exact alignments. Genome Biol. 15, R46. https://doi.org/10.1186/gb-2014-15-3-r46

Yim, H.-S., Cho, Y.S., Guang, X., Kang, S.G., Jeong, J.-Y., Cha, S.-S., Oh, H.-M., Lee, Jae-Hak, Yang, E.C., Kwon, K.K., Kim, Y.J., Kim, T.W., Kim, W., Jeon, J.H., Kim, S.-J., Choi, D.H., Jho, S., Kim, H.-M., Ko, J., Kim, H., Shin, Y.-A., Jung, H.-J., Zheng, Y., Wang, Z., Chen, Y., Chen, M., Jiang, A., Li, E., Zhang, S., Hou, H., Kim, T.H., Yu, L., Liu, S., Ahn, K., Cooper, J., Park, S.-G., Hong, C.P., Jin, W., Kim, H.-S., Park, C., Lee, K., Chun, S., Morin, P.A., O'Brien, S.J., Lee, H., Kimura, J., Moon, D.Y., Manica, A., Edwards, J., Kim, B.C., Kim, S., Wang, J., Bhak, J., Lee, H.S., Lee, Jung-Hyun, 2014. Minke whale genome and aquatic adaptation in cetaceans. Nat. Genet. 46, 88–92. https://doi.org/10.1038/ng.2835

Yu, G., Smith, D.K., Zhu, H., Guan, Y., Lam, T.T.-Y., 2017. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. 8, 28–36. https://doi.org/10.1111/2041-210X.12628

Data Accessibility

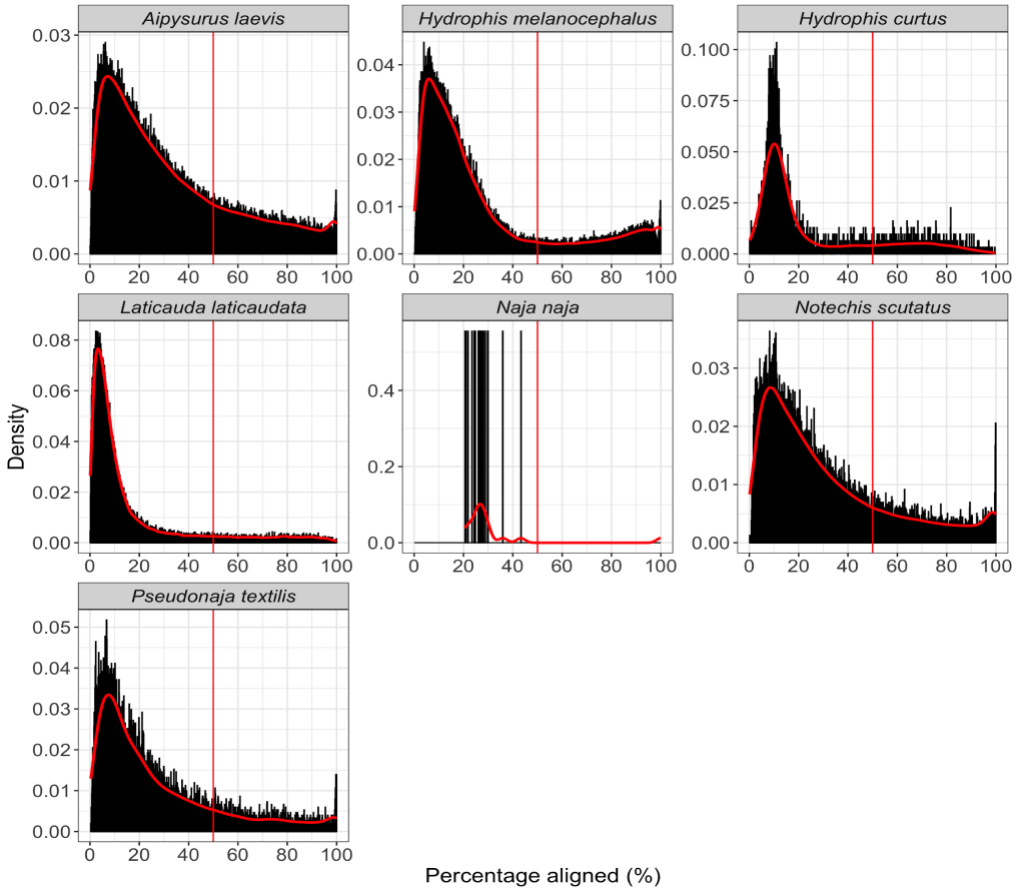
Newly generated sequence data for *Aipysurus laevis* has been submitted to SRA under Bioproject PRJNA669384, sample identifier SAMN16448117**.** The draft *Aipysurus laevis* genome assembly is available at https://doi.org/10.5281/zenodo.3975254.

Bioinformatics code is available at https://github.com/a-lud/snakes-demographic-history.

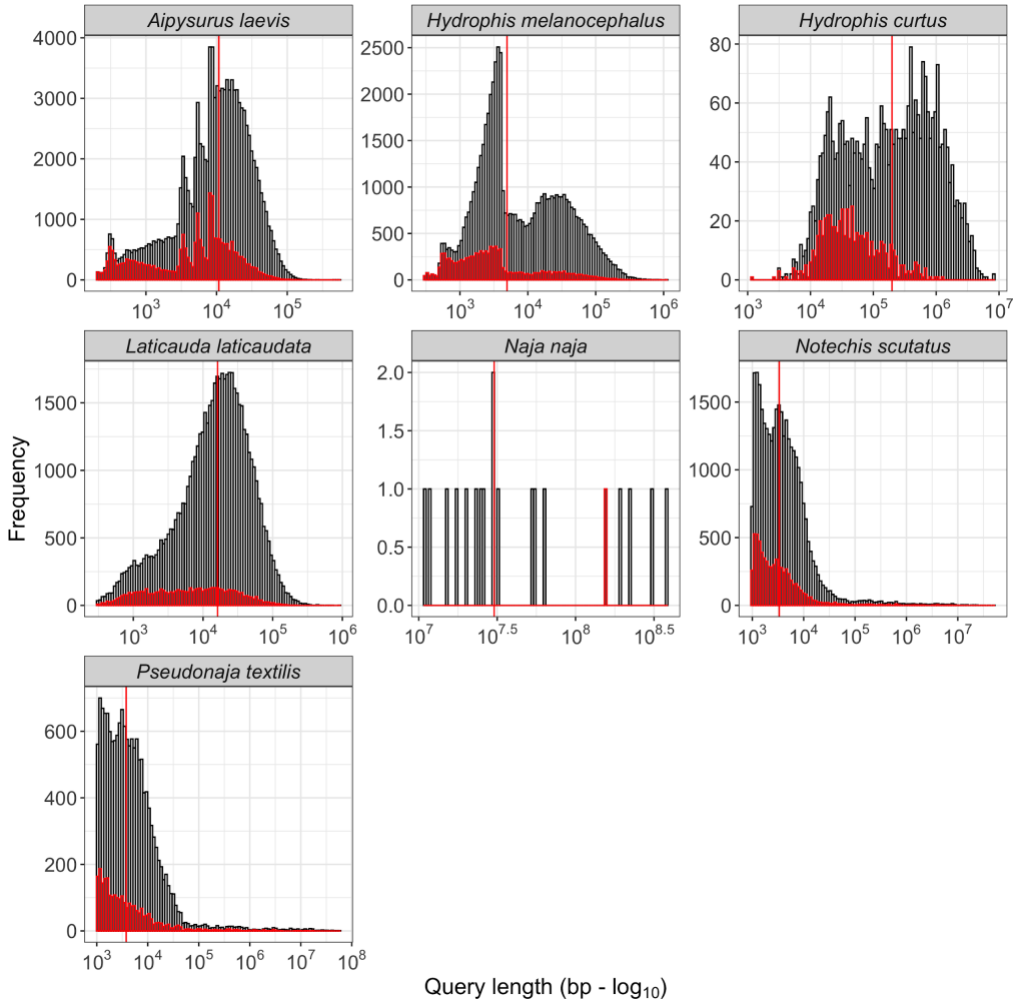
Author Contributions

KLS conceived the study and conducted fieldwork. AJL assembled the *Aipysurus laevis* genome and performed coalescent analysis with input from KLS. KLS and AJL interpreted the results and wrote the manuscript.

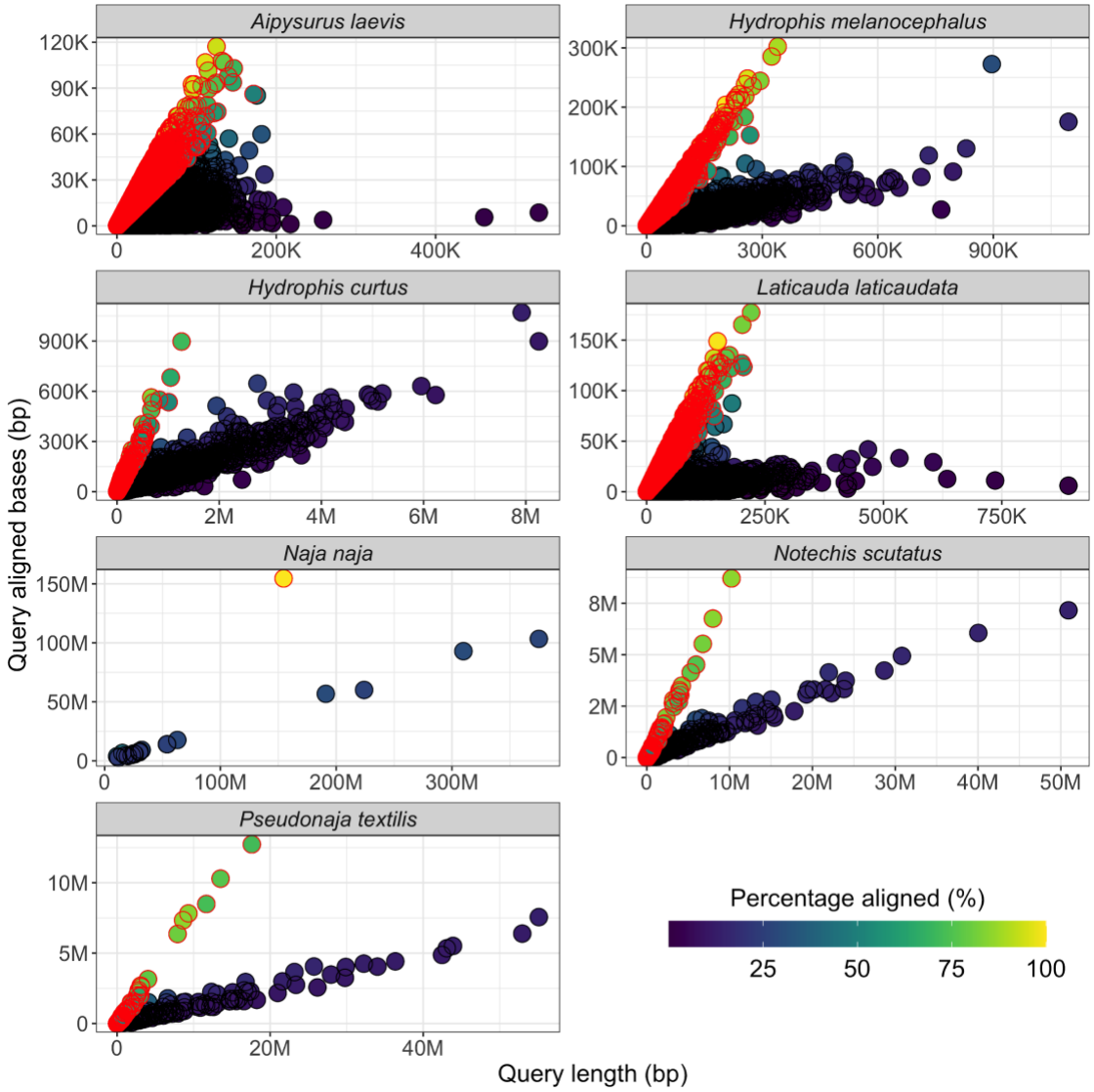
Supplementary figures



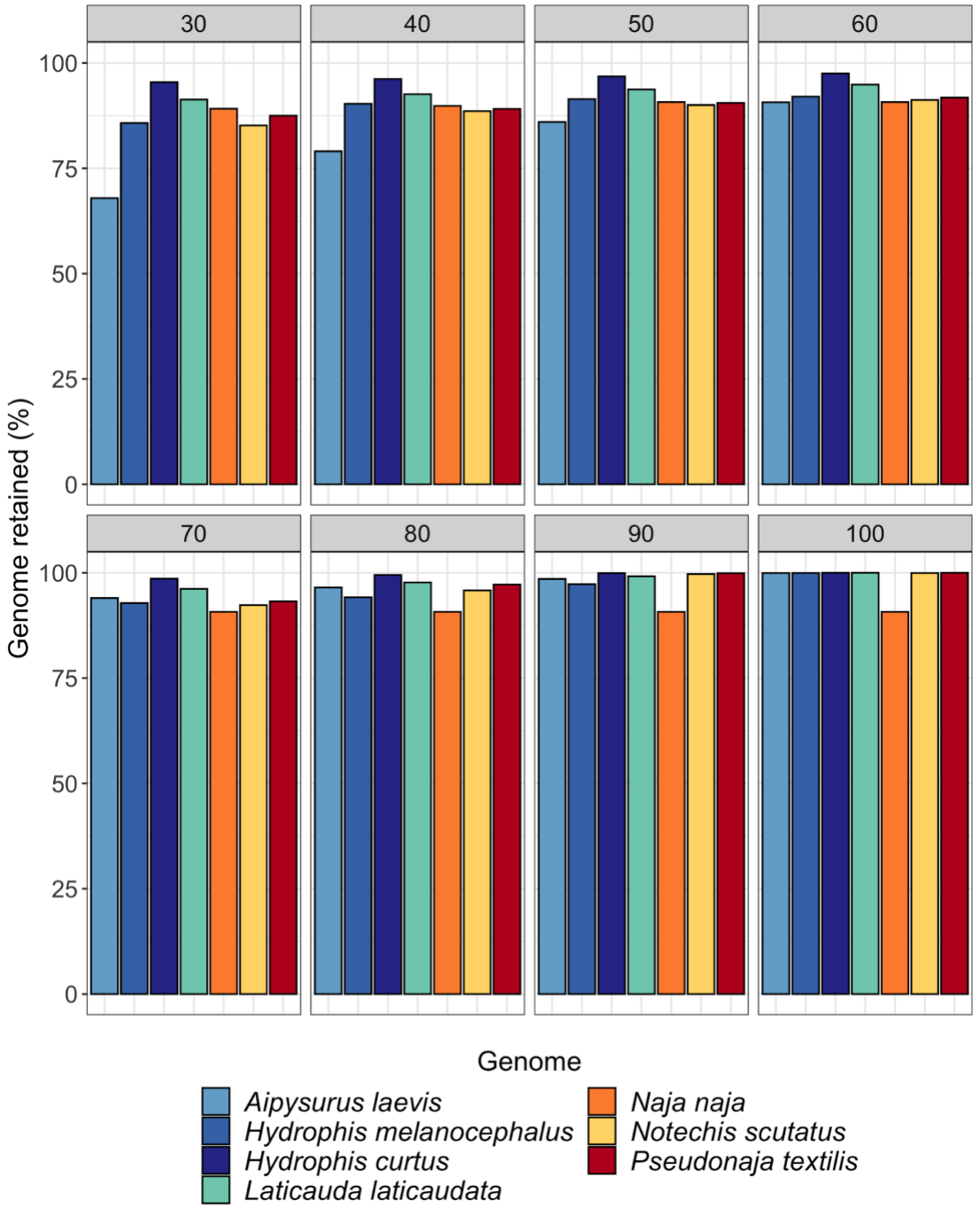
Supplementary figure 1: Density histograms for each snake species showing the proportion of a query sequence (scaffold/chromosome) that mapped to the *N. naja* Z-chromosome. Black vertical bars represent proportions binned into 100 segments, while the red curve represents the smoothed density of the bars. The red vertical line intercepts the x-axis at 50%, where scaffolds/chromosomes to the left of the red vertical line had less than 50% of their total sequence length map to the Z-chromosome. Scaffolds/chromosomes to the right of the red vertical line had greater than 50% of their total length align to the Z-chromosome. Alignment-proportion peaks all reside within the 10-20% regions, with few scaffolds/chromosomes having more than 50% of their length aligning to the Z-chromosome.



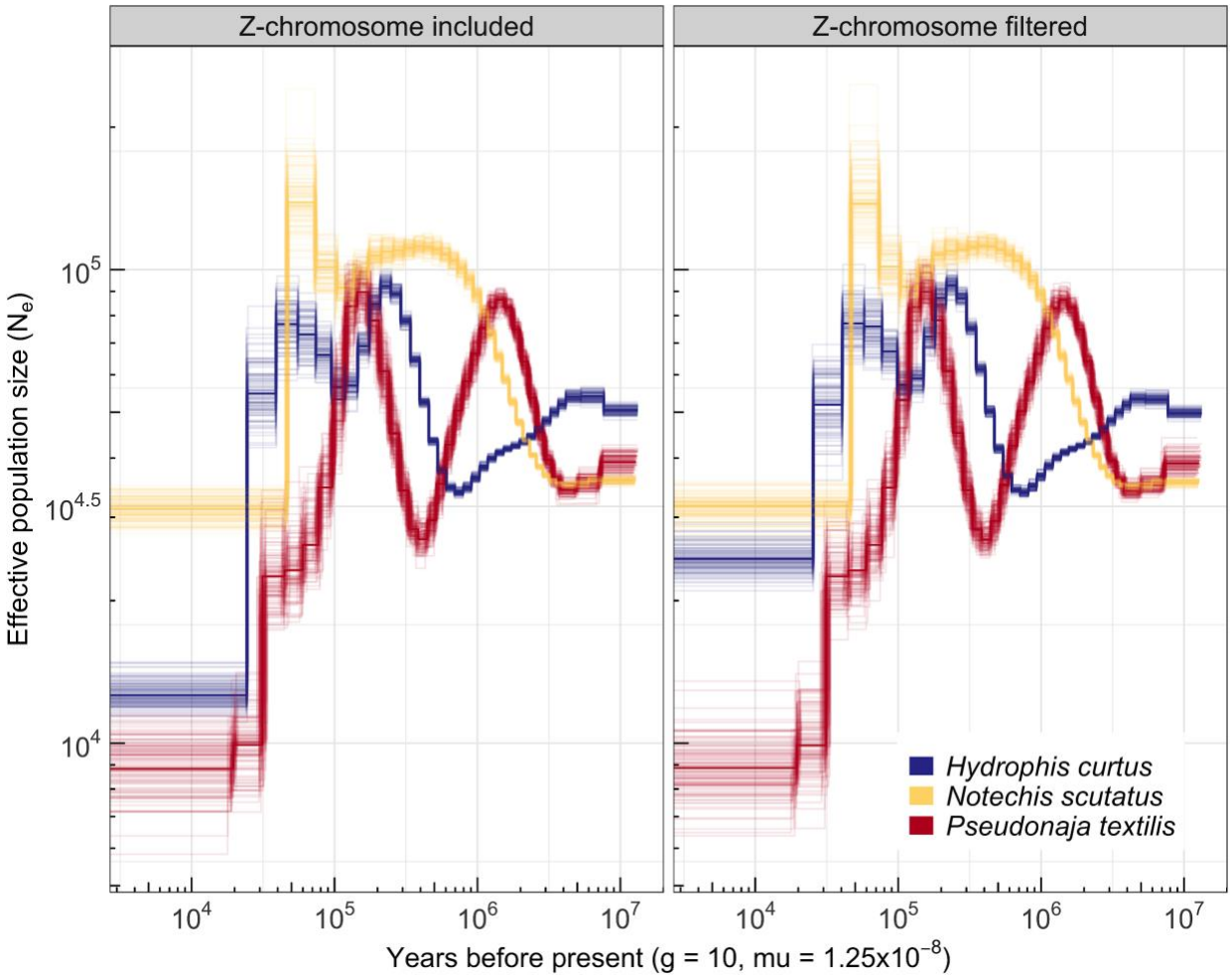
Supplementary figure 2: Histograms of genome scaffold/chromosome lengths (black) and scaffold/chromosomes lengths of scaffolds failing Z-chromosome alignment cut off (red). The black vertical bars along the x-axis (log_{10}) represent length bins for all genomic sequences belonging to each sample. Red bins correspond to scaffolds/chromosomes that had ≥ 50% of their sequence length align to *N. naja*'s Z-chromosome. The vertical red line represents the median sequence length of all genomic sequences belonging to that sample. Sequences with at least 50% of their length aligning to the Z-chromosome are generally smaller than the median sequence length.



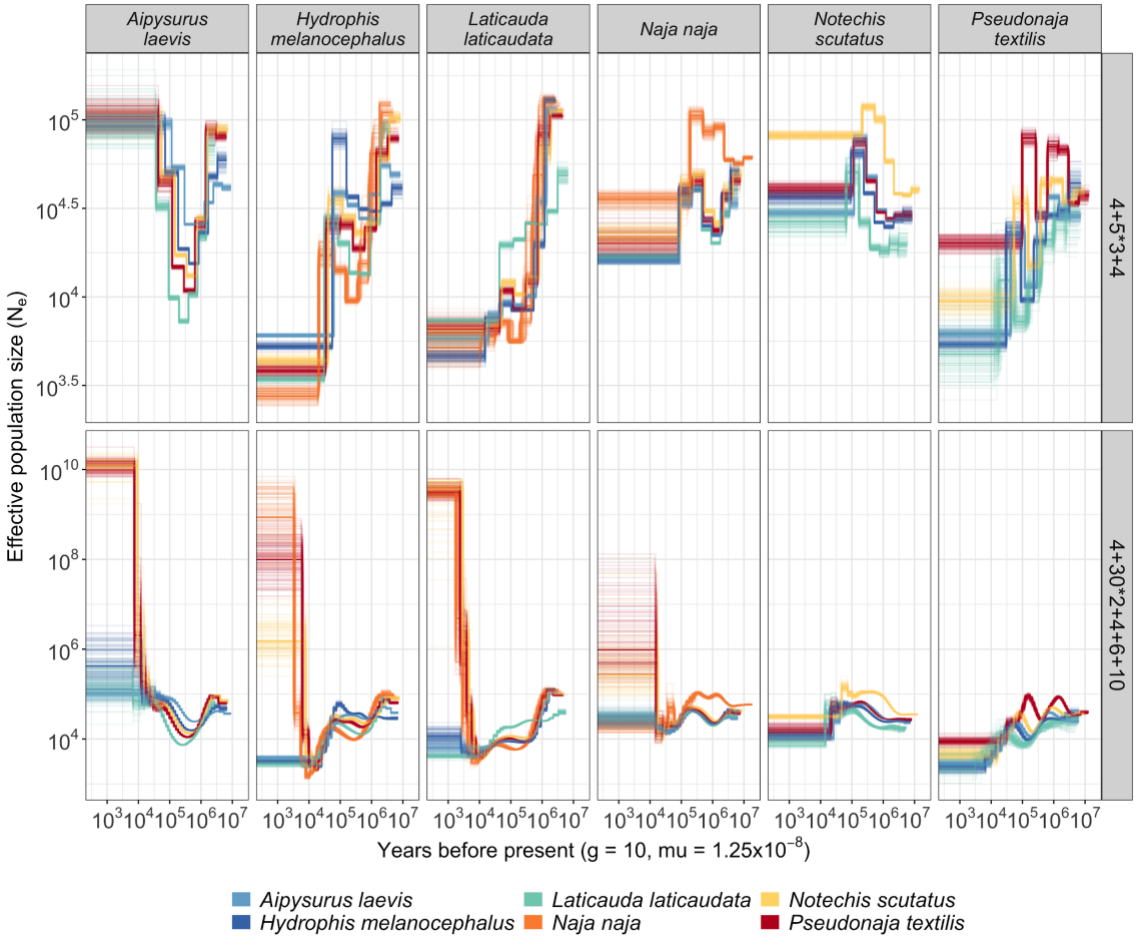
Supplementary figure 3: Scatter plots of scaffolds/chromosome lengths (queries) against the proportion of bases that aligned to the *N. naja* Z-chromosome from each respective query, faceted by sample. The colour filling represents the percentage of the query length that aligned, with cooler tones indicating less alignment and warmer tones more. Points with red outlines are scaffolds/queries where greater than 50% of their sequence length aligned to the *N. naja* Z-chromosome. With the exception of *N. naja's* own Z-chromosome, most query sequences with high alignment proportions were short in length.



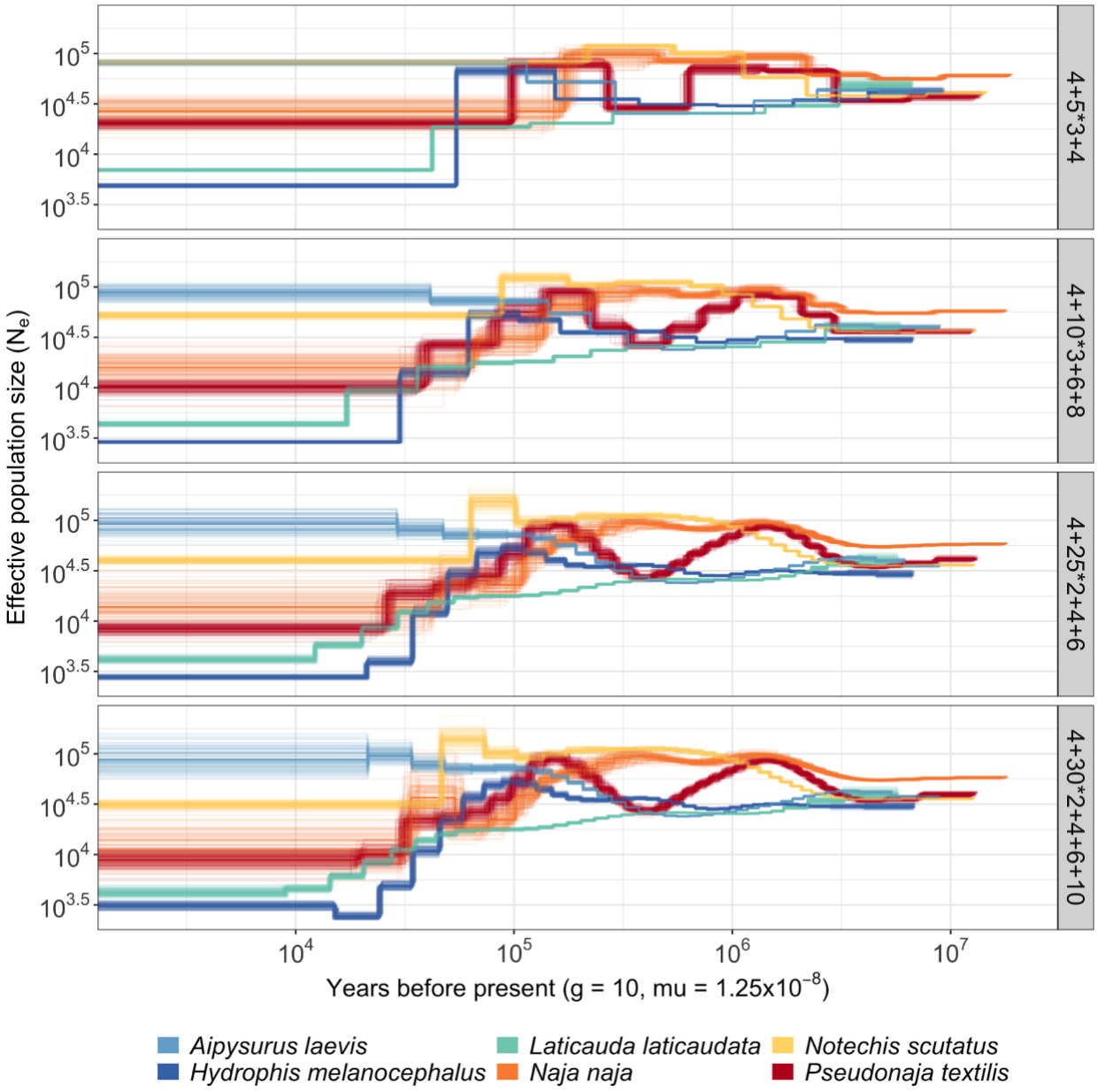
Supplementary figure 4: Bar plot representing how much of the total genome sequence is retained after removing scaffolds/chromosomes from sample genome assemblies that failed the varying proportion-aligned cutoff values (30 - 100%).



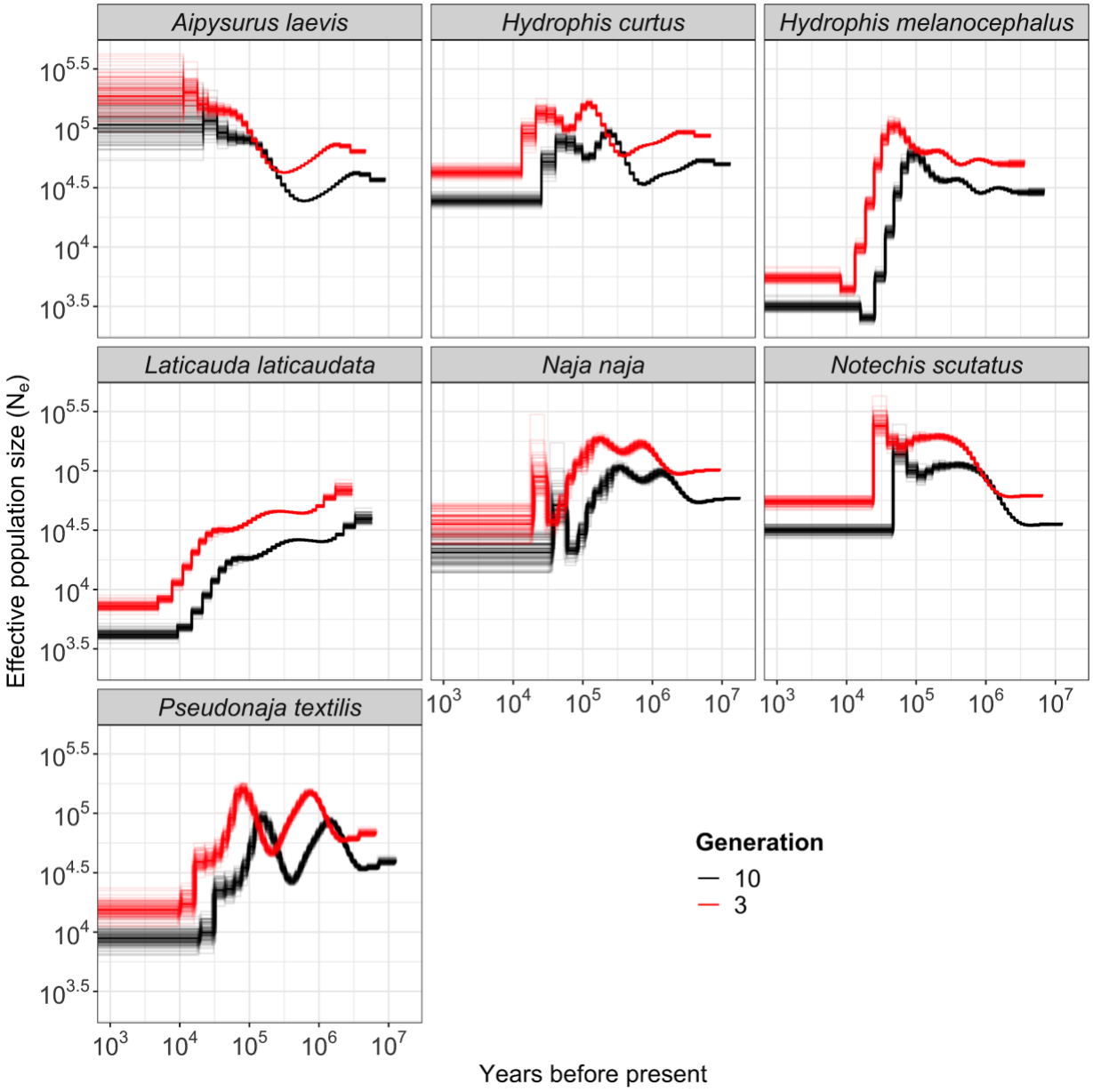
Supplementary figure 5: PSMC curves (clock: 4+30*2+4+6+10) for female samples; *H. curtus, N. scutatus and P. textilis.* After removing putative Z-chromosome sequences from each of the snakes, the overall shape of the curve remains near identical. Only *H. curtus* sees a noticeable change in the most recent time segment trending to $10^{4.5}$ from 10^4.



Supplementary figure 6: Preliminary results demonstrating the effect of reference selection and clock complexity. Columns represent each sample with the coloured stairway plots corresponding to the reference genome the PSMC data were generated from. Marine snake reference genomes are represented by cool colours and terrestrial snake reference genomes are shown using warm colours. Rows in the figure correspond to different atomic clock patterns (top=simple; bottom=more complex). The effect of reference sequence is evident in the separation and shapes of the stairway plots for each individual. Highly inflated N_e estimates are seen in recent time intervals in the cross species alignments, particularly for the marine species in the complex clocks. This may be an artifact of model overfitting in recent intervals (which contain few coalescent events) for samples with reduced coverage due to poor reference mapping.



Supplementary figure 7: Preliminary PSMC plots of samples aligned to their respective reference genomes, scaled using a generation time of 10 and mutation rate of 1.25x10^{-08}, and faceted by atomic time intervals. Plots become increasingly smooth as the number of free interval parameters increase. Data used to generate these plots are pre Z-chromosome filtering.



Supplementary figure 8: Effect of mutation rate and generation time on scaling PSMC plots. Each facet corresponds to a sample aligned to its respective reference genome. The black line assumes a generation time of 10 years and a mutation rate of $1.25x10^{-08}$; the red line is scaled with a generation time of 3 years and a mutation rate of $7.2x10^{-09}$. Smaller generation times and mutation rates shift the curves up and to the left, while larger mutation and generation estimates are lower and extend further back in time.

Supplementary Tables

Supplementary table 1: Alignment statistics from cross species comparisons. Each sample aligned to the respective genomes using *BWA-MEM*. Coverage statistics were calculated using *SAMtools depth* and *AWK* commands. The alignment statistics were generated from the *SAMtools flagstat* command. Variant statistics were obtained using *BCFtools stats*.

Reference	Sample	Avg. depth	Breadth of coverage	Aln. (%)	Proper pair (%)	Genome size	No. called sites	No. SNPs
A. laevis	*L. laticaudata*	79.7	74.4	83.7	64.1	1,852,689,089	582,813,887	13,436,026
A. laevis	*H. melanocephalus*	86.2	83.0	98.1	77.7	1,852,689,089	1,068,071,320	17,826,137
A. laevis	*N. scutatus*	46.8	79.2	96.7	64.1	1,852,689,089	456,570,742	6,327,510
A. laevis	*P. textilis*	46.7	77.0	94.7	63.6	1,852,689,089	347,921,420	5,010,269
A. laevis	*N. naja*	43.1	70.1	79.7	62.2	1,852,689,089	589,711,847	17,481,603
H. melanocephalus	*L. laticaudata*	99.7	82.9	81.7	59.8	1,402,639,853	489,420,169	10,703,484
H. melanocephalus	*N. scutatus*	58.3	86.6	95.0	60.2	1,402,639,853	291,359,925	3,718,275
H. melanocephalus	*P. textilis*	58.6	85.2	92.8	60.4	1,402,639,853	214,090,365	2,783,779
H. melanocephalus	*A. laevis*	56.2	87.4	95.4	73.2	1,402,639,853	916,380,072	13,690,586
H. melanocephalus	*N. naja*	54.6	80.6	76.9	59.5	1,402,639,853	522,050,683	14,798,603
L. laticaudata	*H. melanocephalus*	79.4	78.3	82.5	58.8	1,558,706,106	524,394,830	11,842,036
L. laticaudata	*N. scutatus*	51.2	79.7	93.6	59.7	1,558,706,106	160,758,799	2,198,652
L. laticaudata	*P. textilis*	51.8	79.2	92.1	59.6	1,558,706,106	129,745,006	1,786,622
L. laticaudata	*A. laevis*	44.3	77.9	85.6	65.0	1,558,706,106	395,164,880	7,567,441

L. laticaudata	*N. naja*	50.1	76.0	78.4	60.9	1,558,706,106	531,102,917	15,633,721
N. naja	*L. laticaudata*	77.4	73.9	77.0	56.1	1,667,692,826	348,377,723	8,170,383
N. naja	*H. melanocephalus*	67.3	73.2	73.5	51.1	1,667,692,826	375,913,229	8,794,552
N. naja	*N. scutatus*	44.3	74.4	90.8	56.1	1,667,692,826	115,432,000	1,642,472
N. naja	*P. textilis*	44.7	74.0	89.3	56.0	1,667,692,826	97,424,663	1,407,528
N. naja	*A. laevis*	37.7	72.3	78.9	58.1	1,667,692,826	265,304,497	5,212,802
N. scutatus	*L. laticaudata*	90.2	83.7	83.6	68.3	1,665,525,958	709,582,040	15,312,979
N. scutatus	*H. melanocephalus*	88.9	87.1	92.6	74.7	1,665,525,958	931,738,150	17,328,852
N. scutatus	*P. textilis*	53.5	87.3	95.5	68.0	1,665,525,958	486,976,070	6,327,639
N. scutatus	*A. laevis*	49.2	86.4	93.4	77.5	1,665,525,958	853,353,719	13,967,823
N. scutatus	*N. naja*	49.2	78.9	80.6	64.7	1,665,525,958	673,710,643	18,710,630
P. textilis	*L. laticaudata*	93.4	86.1	83.1	66.7	1,590,035,073	597,962,403	13,137,110
P. textilis	*H. melanocephalus*	87.7	88.5	87.9	69.4	1,590,035,073	791,933,023	15,958,358
P. textilis	*N. scutatus*	55.4	90.5	96.8	66.5	1,590,035,073	410,704,767	5,230,874
P. textilis	*A. laevis*	48.6	87.8	89.3	73.0	1,590,035,073	677,994,807	11,822,402
P. textilis	*N. naja*	50.7	81.1	80.0	63.5	1,590,035,073	609,689,091	17,356,940

Reference = Reference genome

Sample = Sample aligned to the reference genome

Avg. depth = Average alignment depth

Breadth of coverage = Proportion of bases covered by at least 1x coverage

Aln. (%) = Percentage of reads that aligned to the reference genome

Proper pair (%) = Percentage of reads that aligned in proper pairs

Genome size = Number of bases in the reference genome (including N sequence)

No. called sites = Number of reference positions called by BCFtools in the final consensus sequence

No. SNPs = Number of variants called against the reference genome

Chapter 3:
New chromosome-scale genomes provide insights into marine adaptations of sea snakes (*Hydrophis*: Elapidae)

Ludington, A. J., Hammond, J. M., Breen, J, Deveson, I. W., Sanders K. L. (2023). New chromosome-scale genomes provide insights into marine adaptations of sea snakes (*Hydrophis*: Elapidae). **Accepted for publication in BMC Biology.**

Statement of Authorship

Title of Paper	New chromosome-scale genomes provide insights into marine adaptations of sea snakes (Hydrophis: Elapidae)
Publication Status	☐ Published ☐ Accepted for Publication ☒ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Submitted to BMC Biology and is currently under review (as of 28-08-2023).

Principal Author

Name of Principal Author (Candidate)	Alastair John Ludington		
Contribution to the Paper	Conceived the research (with KLS), performed quantitative analyses, interpreted results, wrote the first version of the manuscript and edited subsequent versions.		
Overall percentage (%)	80		
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	28/08/2023

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	Jillian M. Hammond
Contribution to the Paper	Carried out quantitative analysis, interpreted results. Contributed to the writing and editing of the manuscript.

Signature		Date	30/08/2023

Name of Co-Author	James Breen		
Contribution to the Paper	Contributed to the writing and editing of the manuscript.		
Signature		Date	29/08/2023

Name of Co-Author	Ira W. Deveson		
Contribution to the Paper	Contributed to the writing and editing of the manuscript.		
Signature		Date	31/08/2023

Name of Co-Author	Kate Laura Sanders		
Contribution to the Paper	Conceived the research, conducted quantitative analyses, interpreted results, wrote the first version of the manuscript, and edited subsequent versions.		
Signature		Date	30/08/2023

New chromosome-scale genomes provide insights into marine adaptations of sea snakes (*Hydrophis*: Elapidae)

Alastair J. Ludington[1,*], Jillian M. Hammond[2,3], James Breen[4,5], Ira W. Deveson[2,3,6] and Kate L. Sanders[1,7*]

[1] School of Biological sciences, The University of Adelaide, Adelaide, SA 5005, Australia

[2] Genomics Pillar, Garvan Institute of Medical Research, Sydney, NSW 2010, Australia

[3] Centre for Population Genomics, Garvan Institute of Medical Research and Murdoch Children's Research Institute, Australia

[4] Indigenous Genomics, Telethon Kids Institute, Adelaide, Australia

[5] John Curtin School of Medical Research, College of Health & Medicine, Australian National University, Canberra, Australia

[6] Faculty of Medicine, University of New South Wales, Sydney, NSW 2052, Australia

[7]The South Australian Museum, Adelaide, Australia

[*] Corresponding authors: A.J.L: alastair.ludington@adelaide.edu.au; K.L.S: kate.sanders@adelaide.edu.au

Abstract

Background

Sea snakes underwent a complete transition from land to sea within the last ~15 million years, yet they remain a conspicuous gap in molecular studies of marine adaptation in vertebrates.

Results

Here, we generate four new annotated sea snake genomes, three of these at chromosome-scale (*Hydrophis major*, *H. ornatus* and *H. curtus*), and perform detailed comparative genomic analyses of sea snakes and their closest terrestrial relatives. Phylogenomic analyses highlight the possibility of near-simultaneous speciation at the root of *Hydrophis*, and synteny maps show intra-chromosomal variations that will be important targets for future adaptation and speciation genomic studies of this system. We then used a strict screen for positive selection in sea snakes (against a background of seven terrestrial snake genomes) to identify genes over-represented in hypoxia adaptation, sensory perception, immune response, and morphological development.

Conclusions

We provide the best reference genomes currently available for the prolific and medically important elapid snake radiation. Our analyses highlight the phylogenetic complexity and conserved genome structure within *Hydrophis*. Positively selected marine-

associated genes provide promising candidates for future, functional studies linking genetic signatures to the marine phenotypes of sea snakes and other vertebrates.

Keywords: chromosome-scale genome, synteny, positive selection, marine adaptation, sea snake

Background

Major evolutionary transitions, such as from terrestrial to marine habitats, present powerful opportunities to understand genomic mechanisms underlying adaptation. In the genomes of secondarily marine mammals and marine-diving birds, hundreds of genes linked to metabolic and cellular processes, physiology, and functional morphology have been identified as under diversifying or relaxed selection pressures [1–5]. Far fewer studies have examined the specific genomic changes that have accompanied marine transitions in reptiles [6,7] even though snakes and lizards have become important models for studying genome evolution [8].

Reptiles, mammals and birds must have encountered many of the same challenges during their marine transitions, particularly the biomechanical and energetic demands of aquatic locomotion, low levels of oxygen (hypoxia) during extended submergence, maintenance of body water balance in a hyperosmotic environment, dramatically shifted sensory perception, and novel pathogenic environments. Reptiles also have special adaptations to marine life that are not found in mammalian and avian divers, and the many distinct aspects of reptile physiology, particularly ectothermy and flexible vascular circulation [9], might lead to different adaptive solutions. In these

respects, marine reptiles are important taxa for advancing genomic studies of aquatic adaptations in vertebrates.

By far the most specialised and species rich lineage of extant aquatic reptiles are the viviparous sea snakes (Elapidae: Hydrophiinae) [10]. This group comprises more than 60 known species that share a recent marine origin, having descended from Australo-Papuan terrestrial hydrophiines (taipans, death adders, tiger snakes, etc) only ≈9-18 million years ago [6,11,12]. Sea snakes possess a suite of marine-associated characteristics: they have dorsoventrally elongate bodies and paddle-shaped tails that together generate propulsive thrust [13,14], haemostatic nostrils that are sealed underwater by erectile tissue, and a sublingual salt-secreting gland. Specialised respiratory traits allow them to remain active underwater for extended periods, particularly a high degree of cutaneous gas exchange [15–17] that is facilitated by low partial pressure of oxygen in the arterial blood [18,19]. The sensory systems of sea snakes have also diverged from those of their terrestrial counterparts: the visual pigments of many sea snakes have spectral sensitivities that are shifted towards the longer wavelengths that dominate marine environments [20], sea snake mechanosensory scale organs are often large and protruding [21,22], and some species are able to withdraw their vulnerable tail paddle in response to light, a sense shared only with distantly related fish and aquatic amphibians [23].

To date, a lack of high-quality genomic resources for sea snakes and their closest terrestrial relatives has hindered genomic studies of the land-sea transition in elapids. To address this, we assembled and annotated four new sea snake genomes, three of these at chromosome-scale. Analyses of newly generated (*Hydrophis major*, *H.*

ornatus, *H. curtus* (West) and *H. elegans*), and existing (*H. curtus* (East) and *H. cyanocinctus*) data identified many candidate positively selected genes specific to *Hydrophis*, uncovered macro-and micro-chromosomal rearrangements among marine and terrestrial species, and highlighted the phylogenetic challenges of resolving the initial rapid radiation in *Hydrophis*.

Results

Genome assembly and annotation

The differing assembly strategies implemented for *Hydrophis major*, *H. ornatus* and *H. curtus* (West) and *H. elegans* generated highly accurate and contiguous assemblies. The assembly process yielded a primary chromosomal assembly and dual assembly (haplotypes) for *H. major*, two chromosome-level assemblies for *H. ornatus* and *H. curtus* (West), and a highly contiguous contig assembly for *H. elegans* (Fig. 1; Table 1; Additional file 1: Figs. S1 and S2; Additional file 2: tables S2 and S3). The karyotypes of the chromosome scale assemblies were consistent, identifying six macro-chromosomes, nine micro-chromosomes and the Z sex chromosome (Additional file 1: Figs. S2 and S4). Numerous completeness metrics supported the accuracy and quality of the newly assembled reference genomes. The length of the assembled genomes was consistent with their estimated genome sizes (Table1; Additional file 1: Fig. S5; Additional file 2: Table S2) and previously reported *Hydrophis* snakes [6], with chromosome sequences accounting for 90-98.7% of the total sequence length (Additional file 2: Table S2). Genome completeness was assessed using k-mer spectra analysis, with 94.9-97.3% of the sequence read k-mers being accounted for in each

primary assembly (Table 1; Additional file 1: Figs. S6A-C and S7A-F; Additional file 2: Table S4), increasing to 98.9% for the *H. major* dual assembly (Additional file 1: Fig. S6D). The assembly consensus qualities (QV) surpassed 99.9% accuracy (QV30), ranging between QV34.6 - 61.5 (Table 1; Additional file 2: Table S4), while BUSCO completeness scores surpassed 95% for each assembly (Fig. 1; Table 1; Additional file 1: Fig. S8).

De novo gene prediction was performed for *H. major* and the two previously published *H. cyanocinctus* and *H. curtus* reference genomes [6], as their annotations were not available at the time of writing. *H. curtus* is currently recognised as a single widespread species, but contains two deeply divergent (species-level) lineages [24]; the Indian Ocean lineage is represented by the *H. curtus* genome generated here (referred to here as "*H. curtus* (West)") and the Southeast Asian-Australian lineage is represented by the *H. curtus* genome from Li et al. [6] ("*H. curtus* (East)"). A total of 30,425 non-redundant protein-coding genes were identified in *H. major*, 26,730 in *H. curtus* (East) and 27,689 in *H. cyanocinctus*. Lift-over annotations were generated for *H. ornatus*, *H. curtus* (West) and *H. elegans* from *H. major* as RNA-sequencing was not available for these samples, resulting in 27,688-28,381 gene predictions (Table 1; Additional file 2: Table S5). De novo gene annotations reported similar feature characteristics to RefSeq annotated snakes (Additional file 1: Fig. S9), typically reporting slightly lower average numbers of exon/coding sequences of shorter length (Additional file 2: Table S5), a phenomenon that may be explained by the excess of short gene models predicted by the *de novo* pipeline. BUSCO completeness measures were high for all de novo and lift-over gene annotations, surpassing 90% completeness,

except for samples *H. curtus* (East) and *H. cyanocinctus*, which reached 85.4% and 85.6% completeness, respectively (Additional file 1: Fig. S10).

Table 1: Summary statistics for the four newly assembled *Hydrophis* genomes.

	Hydrophis major	***Hydrophis ornatus***	***Hydrophis curtus* (West)**	***Hydrophis elegans***
Sequencing data	HiFi + Hi-C	Nanopore + Hi-C + Short-read	Nanopore + Hi-C + Short-read	Nanopore + Short-read
Status	Chromosome	Chromosome	Chromosome	Contig
Assembly size (Gbp)				
Total	2.17	1.93	1.93	2.01
Chromosome	1.96	1.90	1.88	
Estimated	1.88	1.95	1.93	1.82
Contig number	1,815	1,399	3,208	3,126
Scaffold number	1,320	548	809	3,126
Assembly N50 (Mbp)	268	270	268	17.9
BUSCO (complete %)	96.0	95.9	95.8	95.9
QV	61.5	47.3	46.4	34.6
K-mer completeness (%)	95.5	97.3	95.9	94.9
LAI	23.9	26.2	26.7	25.9
Gene count	30,425	27,688	27,701	28,381
Repeat content (% of genome)	60.3	56.7	56.2	58.0

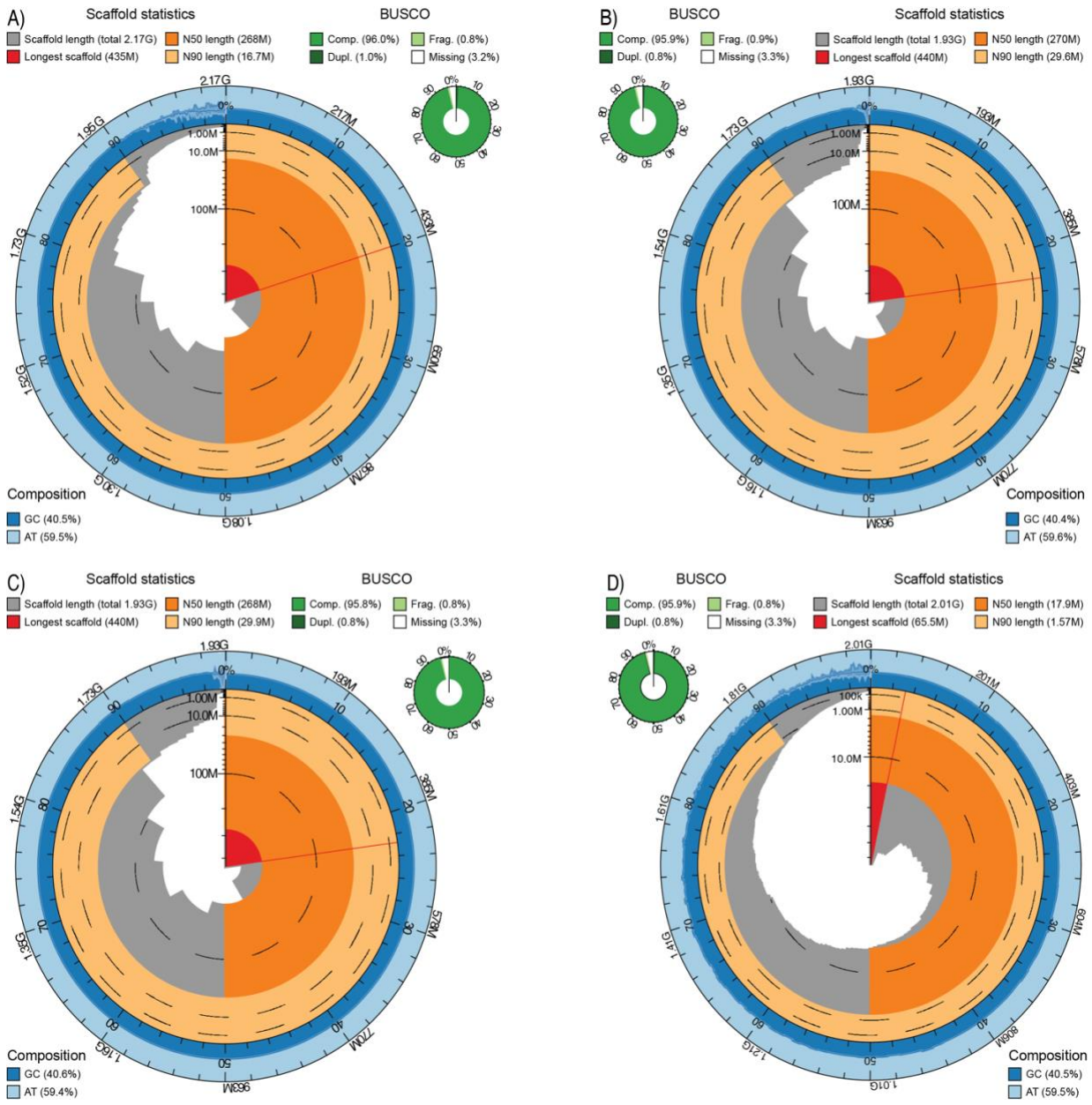


Fig. 1: Snail plots summarising each of the *de novo* assembled *Hydrophis* sea snakes. For each assembly, chromosomes are arranged by length clockwise around the circle, with the largest chromosome/scaffold represented by the red segment and line at the start. The dark and light orange sections represent the N50 and N90 values, respectively, while the dark and light blue rings represent the GC and AT content in the genome. BUSCO scores for each genome are presented in the green ring next to each

snail plot. Facets represent A) *H. major B) H. ornatus C) H. curtus* (West) and D) *H. elegans*.

Repeat annotation and genome size

The repeat content of the newly assembled genomes is approximately 10-15% higher than previously reported for most snakes but consistent with the values reported for *H. curtus* (East) and *H. cyanocinctus* [6]. *Hydrophis major* exhibited the highest proportion of interspersed repeats (60.3% of the total genome length), followed by *H. elegans* (58%), *H. ornatus* (56.7%) and *H. curtus* (West) (56.2%) (Table 1; Additional file 2: Table S6), with all snakes sharing similar proportions of each repeat family (Fig.2A-D). The repeat annotations were validated by running BUSCO on the hard masked genomes of each snake to ensure non-repetitive sequences were not misclassified. This saw complete BUSCOs only drop by approximately 1.3% relative to the unmasked genomes (Additional file 1: Fig. S11). Further, LTR Assembly Index (LAI) scores for all four assemblies exceeded the LAI gold-standard of LAI ≥ 20 [25] (Additional file 1: Figs. S12-15; Additional file 2: Table S7), indicating that not only are the repeat annotations accurate, but that these assessments likely indicate that transposable elements in *Hydrophis* snakes exceed 55% of the total genome sequence.

Repeat histories for the newly assembled *Hydrophis* snakes all show similar Kimura divergence profiles. Across all snakes, MITE elements account for a small percentage of the total genome and fall within the 15-25% divergence range, while LINE elements are slightly more abundant but appear to have been most active in the past (35-50% divergence), with some recent activity indicated by the bands observed between 15-20% (Fig. 2E). The largest repeat signals in all snakes come from DNA and

long terminal repeat (LTR) elements. DNA elements show persistent activity in each of the snakes based on the breadth of their divergence profiles and consistent proportions. LTR elements are also abundant in each snake, however historical activity that was observed in *H. major, H. curtus* (West) and *H. elegans* was not observed within *H. ornatus,* as indicated by the increase of LTR elements beginning at the 40% divergence range. Unique to *H. major* is a recent expansion of 'Unknown' repeats; these account for 3.4% of the total genome sequence (Fig. 2E) and consist of *de novo* modelled repeats along with repeat families from the curated DFAM and RepBase libraries.

In all the assembled sea snakes, LTRs were one of the most abundant repeat elements, accounting for 26.4% of all annotated interspersed repeats in *H. major*, 25.5% in *H. elegans*, 27.6% in *H. ornatus* and 27.3% in *H. curtus* (West) (Fig. 2A-D). Gypsy and Copia LTR elements constitute the predominant signal of LTR expansion in each of the four assemblies, with estimated insertion ages beginning ~12.5 million years ago (mya), with a peak in expansion occurring between 2.5-5 mya (Fig. 2F). The distribution and time of insertion is consistent with previous findings for *H. cyanocinctus* and *H. curtus* (see supplementary Table S4 in [6]). The insertion times and distribution of LTR elements in each of the four snakes are almost identical, which likely indicates shared expansions of these elements in the *Hydrophis* ancestor. However, in more recent time intervals, *H. major, H. curtus* (West) and *H. elegans* continue to share similar counts of Copia elements, while *H. ornatus* has had an expansion of these elements. In the same timeframe, *H. curtus* (West) had a dramatic increase in Gypsy elements that is not observed in any of the other snakes. Given the similarity across the

rest of the distribution, it is likely that these represent recent species-specific expansions.

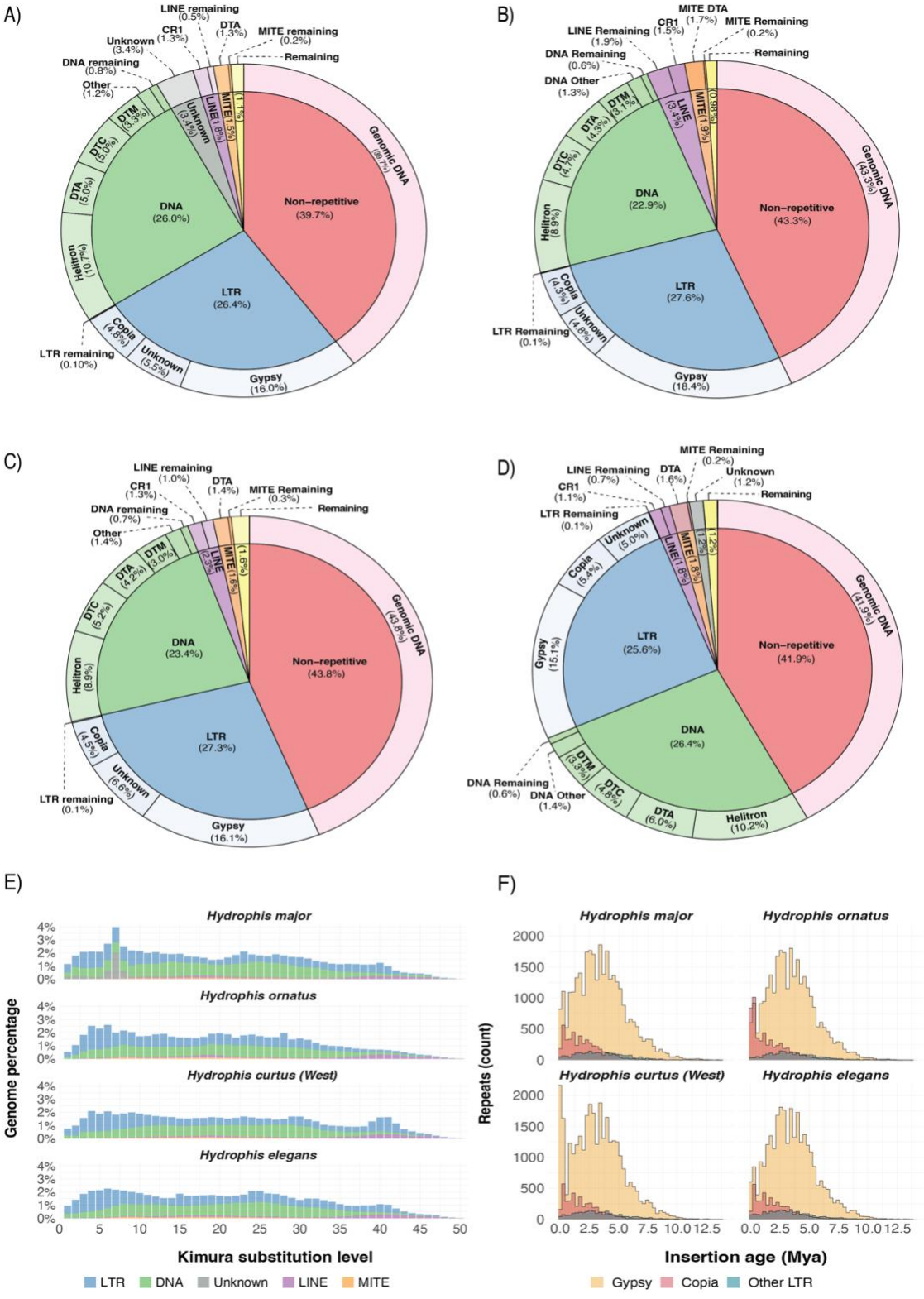


Fig. 2: Summary of repeat annotations in the four *de novo* repeat-annotated sea snakes. A-D) PieDonut repeat-summaries for the four *Hydrophis* snakes: A) *Hydrophis major* B) *Hydrophis ornatus* C) *Hydrophis curtus* (West) D) *Hydrophis elegans*. The inner circle represents the broad classification, with the outer donut ring consisting of the sub-family percentages. E) Distribution of sequence divergence between TEs in each of the four *Hydrophis* assemblies relative to consensus references. The x-axis is the Kimura 2-parameter sequence divergence estimate, while the y-axis is the percentage of each genome that is annotated as TEs. F) Insertion ages of LTR elements in the *Hydrophis* snakes. The x-axis shows the estimated insertion time (mya), estimated from the divergence level and mutation rate, while the y-axis shows the count of TEs inserted at each time interval.

Phylogenomic trees and networks

Our six *Hydrophis* species bridge the backbone of short, unresolved internal branches that characterise molecular phylogenies of the exceptionally rapid *Hydrophis* radiation [11,26] (Fig. 3A). We therefore attempted to resolve their relationships using the newly generated genomes. A combination of phylogenomic tree and network analyses were performed to account for the possibility that the early radiation of *Hydrophis* species does not conform to a bifurcating tree.

Species tree and network analyses both resolved three reciprocally monophyletic lineage pairs: *H. curtus* (East) + *H. curtus* (West); *H. cyanocinctus* + *H. elegans; H. ornatus* + *H. major* (Fig. 3B-C). Internal branches in the species tree are much shorter than the six terminal branches and are mirrored by edges of lower weight in the splits network (Fig. 3D). Species tree and network analyses yielded discordant internal

relationships, with network analyses recovering *H. curtus* (East) + *H. curtus* (West) as sister to the other lineages pairs (Fig. 3C), while the IQTREE/ASTRAL-III species tree placed *H. cyanocinctus* + *H. elegans* as sister to the *H. curtus* lineages and *H. ornatus* + *H. major* with each node present in only ~20-30% of genealogies (Fig. 3B). Consistent with low gene tree concordance factors, network reconstruction using PhyloNet found support for a network over a tree (no reticulations) model, with model fit increasing with the number of reticulations specified (0 to 4) (Additional file 1: Fig. S16; Additional file 2: Table S8). The optimal network contained four reticulations (Fig. 3C) and was consistently recovered in replicate searches. Reticulations were concentrated early in the *Hydrophis* radiation and among the geographically overlapping *H. curtus* (East) and *H. cyanocinctus* lineages.

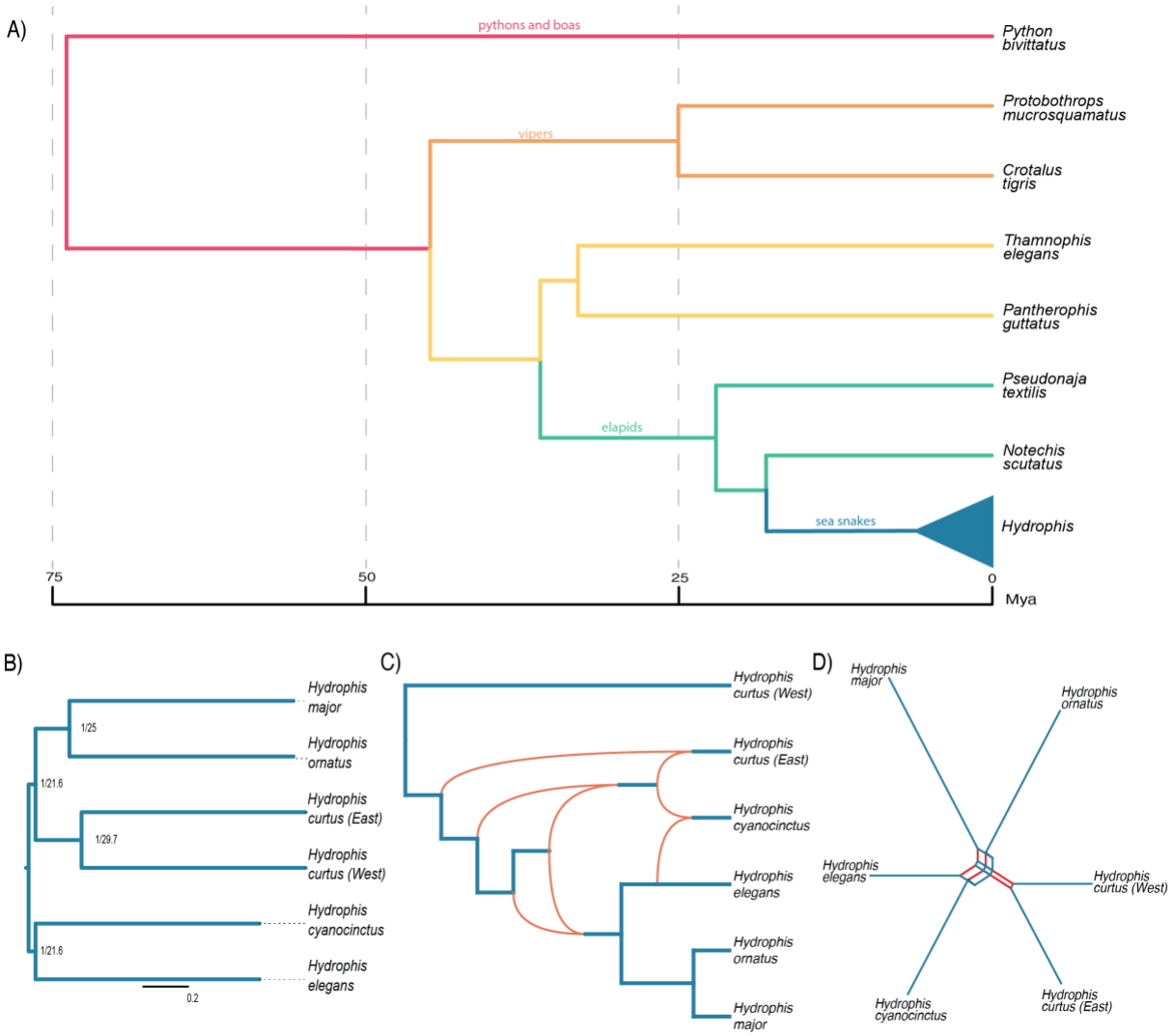


Fig. 3: Species tree inference using *Hydrophis* single copy orthologs and whole genome sequences A) time tree of sampled taxa drawn with Archaeopteryx 0.968 beta BG using relationships and divergence times from Lee et al. [11] and Zaher et al. [27] B) species tree inferred from *Hydrophis* specific single-copy orthologs using IQ-TREE and ASTRAL-III. Node labels are in the form ML support/gene concordance factor C) PhyloNet maximum-likelihood network illustrating potential ILS/Introgression signals (orange arcs) between the six *Hydrophis* snakes D) a SANS serif weakly compatible splits network between the six *Hydrophis* snakes generated from the genome assemblies. Splits between snakes are represented by the red parallel edges.

Genome synteny

Genome-wide synteny analyses between the five chromosome-scale *Hydrophis* snakes, and outgroup *Thamnophis elegans*, was used to investigate karyotype, chromosome synteny, chromosome evolution and structural variation between these snakes. Chromosomal synteny between the *Hydrophis* snakes, both for macro- and micro-chromosomes, is extensive, with the karyotype varying slightly between snakes (Fig. 4; Additional file 1: Figs. S17 and S18). The substantial synteny between the *Hydrophis* snakes is not wholly unexpected based on their recent divergence (Fig. 3). Within the *Hydrophis* snakes, the six macro-chromosomes share broad homology, while the micro-chromosomes share significant homology but appear to be the main source of karyotypic variation. As reported in Li et al. [6], chromosome 14 exists as a microchromosome in *H. cyanocinctus,* as well as in *H. curtus* (West) and *H. ornatus* but is a part of a macro-chromosome 7 and 6 in *H. curtus* (East) and *H. major*, respectively. Chromosome 9 in *H. ornatus* appears to be a chromosome fusion of chromosomes 12 and 14 in *H. major*, or the two respective syntenic chromosomes in the four other *Hydrophis* snakes. Similarly, chromosome 8 in *H. cyanocinctus* looks to be formed from chromosomes 9 and 11 in *H. curtus* (East), with chromosome 11 in *H. curtus* (East) sharing homology to chromosome 15 in *H. cyanocinctus*. Some micro-chromosomes share limited-to-no homology to the other snakes, such as the *H. curtus* (East) chromosome 17, or its chromosome 16, which only shares homology with chromosome 17 in *H. cyanocinctus*. In *H. ornatus* and *H. major*, chromosome 15 shares homology to the Z-chromosome in all the other snakes, while chromosome 18 in *H. cyanocinctus* shares homology with the beginning of chromosome 2 in *H. curtus* (East), potentially

reflecting a misassembly in *H. cyanocinctus*. Relative to the outgroup *T. elegans*, *Hydrophis* snakes share considerable homology, albeit across a vastly different karyotype. The five largest macro-chromosomes in *Hydrophis* all share homology to multiple different chromosomes in *T. elegans*, while the inverse is true for micro-chromosomes in *Hydrophis*, where multiple micro-chromosomes share homology to single chromosomes in *T. elegans* (Fig. 4). This suggests that there has been significant chromosomal evolution through time within elapid snakes, in the form of chromosome fission/fusion events, with *Hydrophis* snakes seemingly having settled on a relatively stable karyotype.

Whilst broad syntenic patterns were observed between the *Hydrophis* genomes, we screened for structural variants (SVs) within *Hydrophis* using whole genome alignments (Additional file 1: Fig. S18). Note that chromosome names used below refer to the chromosome identifiers used in Additional file 1: Fig. S18**.** Approximately 1.5Gbp of sequence was classified as syntenic between all pairs of species alignments, while unaligned regions accounted for anywhere between 130 - 300Mbp of the remaining sequence (Additional file 2: Table S9). Inverted regions were the most substantial SV in all species comparisons, affecting between 106-212Mbp of sequence in each genome comparison. The ends of chromosomes proved to be SV hotspots, with chromosome 1 housing a ~20mb inversion that alternates between all genome comparisons except *H. ornatus* and *H. major*, while the beginning of chromosome 2 has extensive SVs between all snakes. This region of chromosome 2 is dense with repetitive elements (Additional file 1: Fig. S18), perhaps indicating that its disorderly nature may be a result of repeat-activity, whilst also potentially reflecting the limitations of assembly methods in

resolving such regions. Interestingly, between *H. curtus* (East) and *H. cyanocinctus*, small clusters of SVs were consistently observed towards the ends of all macro-chromosomes, a pattern not observed consistently in any of the other genome comparisons. Micro-chromosomes also proved to be littered with SVs, with chromosomes 11 and 12 harbouring significant proportions of SVs along their total length. However, the Z-chromosome is most notable, with nearly every portion of its sequence length being affected by a SV in at least one of the species comparisons.

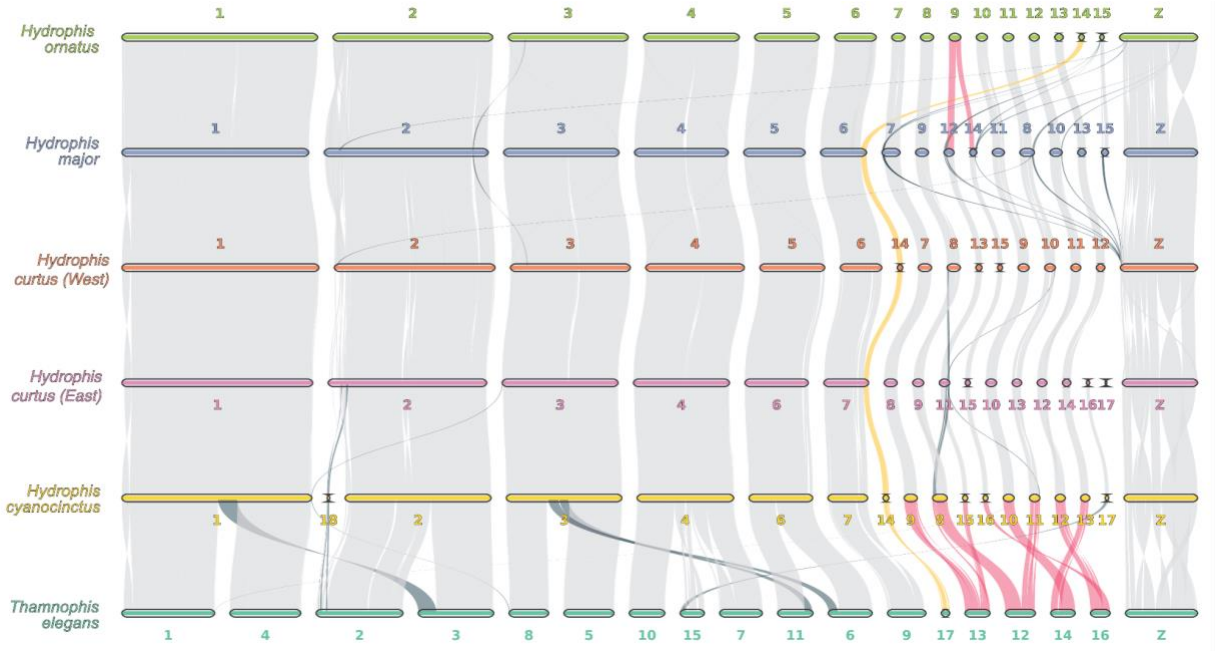


Fig. 4: Synteny between the five-chromosome scale *Hydrophis* sea snakes and *Thamnophis elegans.* Chromosome sequences have been reverse transcribed in some instances to correct for strand variation between assemblies to improve interpretability (see Additional file 1: Fig. S16). Inter-chromosomal rearrangements are highlighted in dark-blue, chromosome fusion/fission events are in red and the chromosome 6 and 14 fusion/fission event which alternates between each of the *Hydrophis* snakes is in yellow.

Gene selection during the marine transition of sea snakes

Using a set of 8,654 single-copy orthologs obtained from thirteen snakes, we aimed to identify candidate genes associated with adaptive marine traits via their signal of positive selection within *Hydrophis* (Additional file 2: tables S10 and S11). To identify marine-specific positively selected genes (PSGs), we used two selection testing methods that are both designed to identify signatures of positive selection that are trait specific and not phylogeny wide. Using the PAML drop-out method [28], we identified 2,670 positively selected genes unique to *Hydrophis* after correcting for multiple testing (Additional file 2: Table S12), while the BUSTED-PH [29] approach reported 1,608 genes as experiencing positive selection specific to *Hydrophis* (Additional file 2: Table S13). The final high-confidence gene-set was obtained by intersecting the significant genes from each method, resulting in 1,402 PSGs that were reported as only under positive selection within *Hydrophis* snakes (Additional file 2: Table S14). While there was considerable overlap in the genes identified by each method, the PAML approach identified an additional 1,268 genes not reported by BUSTED-PH, whereas only 206 genes identified by BUSTED-PH were not also found by the PAML drop-out method (Additional file 1: Fig. S19). Mean ω ratios for single-copy orthologs were typically less than one, as is expected when averaging signatures of selection over branches and sites [30,31], although mean ω values were notably higher for PSGs relative to the non-PSG set (Fig. 5A). This pattern can be explained by the proportion of sites in each rate-class and their respective ω values. Most sites in each ortholog belong to the purifying or nearly neutral ω rate categories ($\omega < 1$ and $\omega \leq 1$), with very few sites in the PSGs assigned to the third ω category ($\omega > 1$) (Fig. 5B; Additional file 1: Fig. S20).

Consequently, the few positively selected sites with large ω values increase the overall average ω in the PSGs, even though the gene-wide average ω remains less than one. Whilst there is evidence that some PSGs may also have experienced positive selection in the background lineages, indicated by the non-zero proportion of sites in the third ω rate category whose average ω is slightly above one (Fig. 5B), the overwhelming indication is that we have accurately identified a core set of genes that have only experienced positive selection in the *Hydrophis* lineages. Exploring the distribution of PSGs across the genome highlighted a relatively even spread, with PSGs residing on all chromosomes, however no PSGs were identified in the tangle of SVs at the beginning of chromosome 2 (Additional file 1: Fig. S18).

In addition to identifying marine-specific PSGs, we also formally tested the strength of natural selection acting on the single-copy orthologs in *Hydrophis* relative to the terrestrial snakes using RELAX [32]. A total of 2,119 single-copy orthologs were reported as significant after correcting for multiple testing *(FDR ≤ 0.01)*, of which 1,677 reported an intensification of selection, 442 reported a relaxation of selection and 6,532 remained insignificant (Fig. 5C; Additional file 2: Table S15). RELAX failed to run for three near-identical single-copy orthologs. Intersecting the RELAX results with the marine-specific PSGs showed that a large majority (969 of 1402) PSGs experienced an intensification of selection, 30 PSGs reported a relaxation of selection and 403 PSGs reported no significant intensification or relaxation of selection (Fig. 5C).

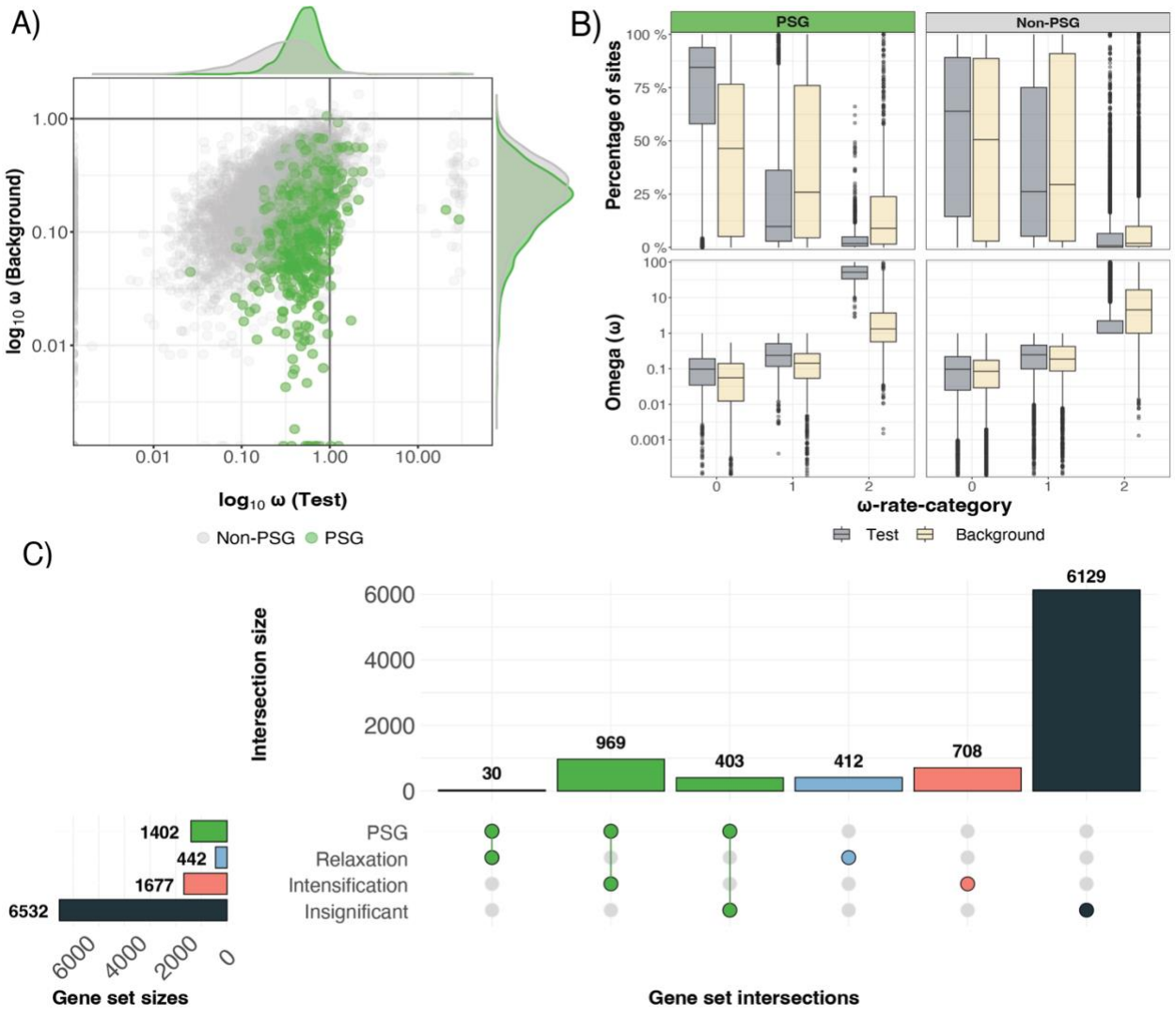


Fig. 5: Exploration of selection testing results, along with the overlap between the marine positively selected genes (PSGs) and *RELAX* results. A) Gene-wide *ω* values were computed during the BUSTED-PH analysis for each gene using the MG94xREV method. The x- and y-axes show the $\log_{10}$ transformed *ω*-values for *Test* and *Background* branches, respectively. B) Summary of the BUSTED-PH unconstrained model results ($\omega \geq 1$). The first facet column represents the marine PSGs (green) and the second facet column represents non-PSGs (grey). The top row shows the percentage of sites falling into each of the three *ω*-rate-classes, while the bottom row

shows the distribution of ω-values in each rate-class. C) UpSet plot visualising intersections between the marine PSGs and each RELAX category (intensification, relaxation, insignificant). The central interaction matrix shows the combination of gene sets, with the top bar plot representing the size of the overlap. The left horizontal bar plot represents the size of the gene sets being compared.

GO term overrepresentation and semantic clustering

To understand the biological significance of the genes exhibiting positive selection in *Hydrophis*, we performed Gene Ontology (GO) under- and over-representation analysis using PANTHER [33,34]. From the 1,302 PSGs used to identify significantly overrepresented GO terms (100 PSGs did not have annotated gene symbols), we identified 120, 26 and 12 GO terms as significantly enriched in biological process (BP), cellular component (CC) and molecular function (MF), respectively, after correcting for multiple testing *(FDR ≤ 0.05)* (Additional file 2: Table S16). Taking the most specific enriched GO terms belonging to each hierarchical cluster resulted in a reduced set of terms belonging to each ontology (30 in BP, 10 in CC and 6 in MF). Of the overrepresented terms, tRNA 5'-end processing (GO:0099116) had the highest fold enrichment, while other terms with increased fold enrichments related to catabolic processes within the cell (GO:0045732: positive regulation of protein catabolic process, GO:0031331: positive regulation of cellular catabolic process, GO:0044265: cellular macromolecule catabolic process, GO:0030162: regulation of proteolysis), regulation of transcription (GO:0010608: post-transcriptional regulation of gene expression, GO:0045893: positive regulation of DNA-templated transcription), metabolic processes relating to nitrogen and phosphate compounds (GO:0044271: cellular nitrogen

compound biosynthetic process, GO:0051172: negative regulation of nitrogen compound metabolic process, GO:0006796: phosphate-containing compound metabolic process), circulatory system development (GO:0072359), animal organ development (GO:0048513), and cellular response to DNA damage (GO:0006974) among others (Additional file 1: Fig. S21; Additional file 2: Table S16). Underrepresented terms were predominantly associated with sensory perception (GO:0004984: olfactory receptor activity, GO:0004930: G protein-coupled receptor activity, GO:0050911: detection of chemical stimulus involved in sensory perception of smell) and immunity (GO:0019730: antimicrobial humoral response, GO:0019814: immunoglobulin complex, GO:0003823: antigen binding; GO:0002250: adaptive immune response) (Additional file 1: Fig. S21; Additional file 2: Table S16). Semantic clustering of over and underrepresented GO terms using REVIGO [35] produced a reduced set of broad, related functional categories (Fig. 6A-F; Additional file 2: Table S17).

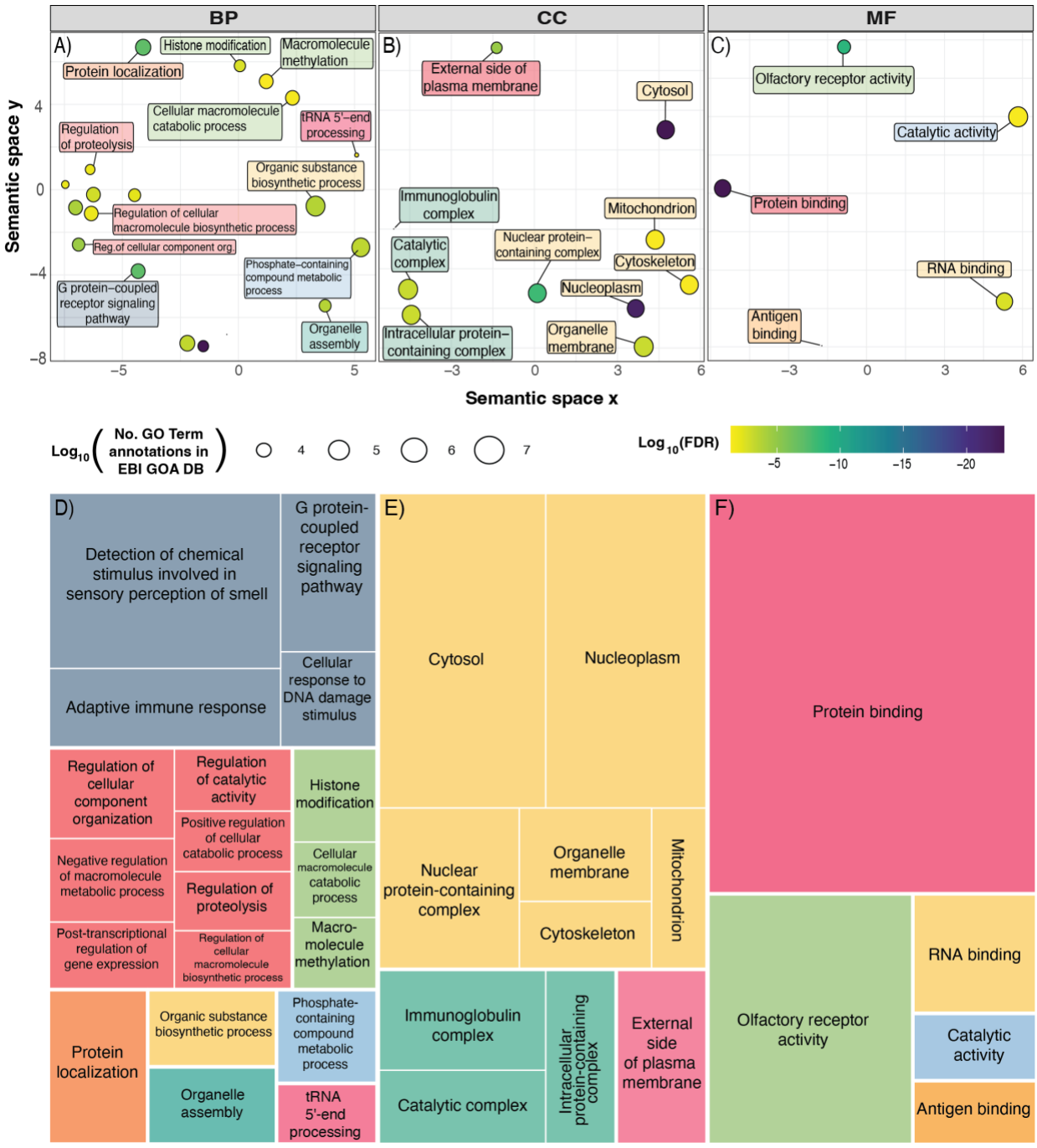


Fig. 6: *REVIGO* multidimensional scaling (MDS) and TreeMap plots. A-C) MDS plots for the Gene ontologies BP, CC and MF, respectively. The x- and y-axes represent arbitrary values for the semantic space. GO terms that are semantically similar cluster together. The colour of the circles represents the Log_{10}(FDR) value, while the size of the

circles Log_{10} value of the number of annotations for the GO Term in the selected species in the EBI GOA database. Point labels have been coloured to match the TreeMap figures. D-F) TreeMap figures generated by REVIGO. Semantically similar terms are clustered into broad categories, with the top-left term being the representative term for the group. Colours are ontology specific, and do not match across ontologies.

Discussion

Comparative and phylo-genomics

Non-avian reptiles have emerged as an important group for studying karyotype evolution (reviewed in [36]). Our study contributes to this expanding literature by examining syntenic changes over a much more recent timeframe than most previous analyses of the group. The sequenced *Hydrophis* span less than 10 million years of evolution [37] and show extensive synteny across their macro- and micro-chromosome sequences. Most variation in karyotype is attributable to uncertainty in the assembly and identification of micro-chromosomes due to their unique sequence content, close proximity within the nucleus and propensity to interact with one-another [38]. Chromosome counts within the assembled *Hydrophis* genomes are consistent with ancestrally reconstructed karyotype configurations [39] and earlier cytological work [40]. While snakes appear to have settled on a relatively stable ancestral chromosome configuration (2n=36), the clade encompassing *Hydrophis* and another fully marine sea snake genus, *Aipysurus*, are both modelled as having a reduced karyotype (2n=32). This deviation of karyotype in sea snakes is consistent with the reported increased rate of chromosome change within elapid snakes [41].

Structural variations, most commonly inversions, were observed along all *Hydrophis* chromosome sequences, with considerable variation along the total length of the Z-chromosome. Structural variants have been hypothesised to establish reproductive isolation and genomic diversity between incipient species by disrupting recombination, reducing introgression and potentially assisting in the development of genomic barriers during speciation [42,43]. The intra-chromosomal variations identified here will be important targets for future adaptation and speciation genomic studies of this system.

The proportion of repetitive sequence content within squamate genomes varies greatly [44], with previous work on repetitive elements in snakes showing that their abundance in the genome is variable and that transposable elements likely contribute to increases in genome size [44–46]. Based on the six species of *Hydrophis* presented between this study and Li et al. [6], both genome size and repeat content appear relatively consistent at this phylogenetic scale. Galbraith et al. [46] showed that the proportion of repetitive elements in genome assemblies is strongly influenced by assembly quality, with short-read assemblies providing underestimations of genome size and repeat content due to collapsing recent, highly similar repetitive elements. This can be seen in *Hydrophis;* the initial draft genome of *H. curtus* had an assembly size of 1.62Gbp and repeat content of ~34% of genome length [7], before long-read technologies resulted in an assembly of 1.96Gbp with 55.6% of the genome being reported as repetitive [6]. As more high-quality genomes become available, it will be important to distinguish the effects of unequal assembly quality from genuine variation in genome size and repeat content across squamates.

Our phylogenomic analyses include only six of ~50 *Hydrophis* species, but these bridge the unresolved backbone of the genus and our results provide some new insights. Species tree and network analyses resolved two reciprocally monophyletic lineage pairs that have been recovered in previous analyses of mitochondrial and nuclear loci (two deep lineages within *H. curtus*, and a close affinity between *H. ornatus* and *H. major* with respect to the other sampled species), plus a sister relationship between morphologically similar species *H. cyanocinctus* and *H. elegans* that was not present in previous molecular trees [37]. The two approaches yielded discordant relationships with low concordance among genealogies. This topological conflict is despite maximal bootstrap and posterior support values for internal nodes, perhaps reflecting the tendency for support values to be inflated with increasing sampling of the genome [47]. The preferred network (versus tree) model recovered four reticulations that were mostly concentrated early in the *Hydrophis* radiation and among geographically overlapping lineages. It is likely that gene flow during speciation, post-speciation hybridisation, and incomplete lineage sorting all contribute to these complex relationships, especially given the short intervals between speciation and large and overlapping populations within *Hydrophis* [48,49]. However, it also remains to be determined whether the deepest speciation events in the clade conform to a bifurcating tree. *Hydrophis* presents a rich system for future studies to explore the challenges of distinguishing between weak phylogenetic resolution and near-simultaneous speciation.

Gene selection during the land-sea transition of elapids

Our strict screens for gene selection in comparison to terrestrial snake genomes identified marine-associated genes in clusters of semantically similar terms related to

marine phenotypes including hypoxia tolerance, sensory perception, immunity, osmoregulation and morphological traits. Below we explore our marine-associated genes in more detail, specifically in the context of marine adaptation in sea snakes and other secondarily marine vertebrates. This fills a conspicuous taxonomic gap in genomic studies of marine adaptation and has broad relevance for understanding how divergent and convergent changes to a shared genetic toolkit underpin ecological transitions in vertebrates.

Adaptations to hypoxia

The transition to a hypoxic and energetically demanding marine environment must have exerted significant pressures on the respiratory, cardiovascular and metabolic traits of sea snakes. In line with this expectation, GO terms such as *circulatory system development* (GO:0072359), *animal organ development* (GO:0048513), and *cellular response to DNA damage stimulus* (GO:0006974) were identified as being over-represented in the set of PSGs.

Circulatory system development (GO:0072359) contains genes related not just to cardiovascular and respiratory systems but also the circulation of nutrients and hormones. Several PSGs in this over-represented set, however, have specificities to cardiovascular and pulmonary function; for example *HAND2* and *SGCG* belong to the *dilated cardiomyopathy* KEGG pathway [50,51] and epinephrine-binding receptor *ADRA2B* is a key gene in studies of broncho- and vasodilation in response to exercise in humans [52]. Selection on these genes might contribute to various aspects of the bimodal respiration of sea snakes. The respiratory tissues of the trachea and right lung

extend almost the length of the sea snake body and the perfusion of these surfaces is enhanced by ‘anticipatory’ breathing tachycardia during surfacing [53]. Cutaneous gas exchange with seawater provides approximately a third of the oxygen requirements of sea snakes, and a significant portion of their CO_2 and N_2 loss [16,17,19,54]. The skin is supplied by a dense network of capillaries, and a redistribution of blood flow (bypassing the lung) ensures the delivery of blood with favourable partial pressure gradients for gas exchange [9,18]. Further facilitating gas exchange is modification of the permeability barrier of the skin [55,56], specifically the inner layer composed of lipids and α-keratin [57,58]. Our PSG set includes the α-keratin encoding *KRT24* gene, which has been lost independently in birds, crocodiles and mammals [59,60] but may play a role in cutaneous respiration in snakes.

Respiratory adaptations of the blood are also prominent in studies of gene selection in secondarily marine taxa. Sea snakes, however, are reported to have haemoglobin concentrations, haemoglobin-oxygen affinities and haematocrit levels that are generally within the ranges reported for terrestrial snakes [19]. Nevertheless, PSGs linked to iron transport and homeostasis were included in several over-represented sets, including *nuclear protein-containing complex* (GO:0140513) and *organic cyclic compound binding* (GO:0097159). Heme-binding protein *HEBP1* and heme-exporter *FLVCR2* are involved in the regulation and transport of heme - an essential component of haemoglobin [61,62]. Positive selection on these and other genes involved in heme biosynthesis and blood cell development suggests molecular adaptation involving oxygen storage and transport capabilities in sea snakes.

A notable sea snake PSG is Lactate dehydrogenase A (*LDHA*). This is a key

enzyme in the energy metabolism of hypoxia tolerant species [63,64]. It mediates pyruvate-lactate interconversion in blood, allowing the metabolism of lactate accumulated during periods of low oxygen availability. Selection on this gene in several fish and cetaceans is thought to increase energy production under hypoxic conditions by increasing affinity for pyruvate (which allows increased lactate production) [65,66]. Relevant here is that *Hydrophis* also show positive selection on *SLC16A14* - a transporter of pyruvate and lactate across cells [67].

Six additional metabolism-associated PSGs (*NDUFA13*, *NDUFAF1*, *NDUFB1*, *NDUFV2*, *NDUFS7* and *NDUFAF3*) are nuclear-encoded components of the mitochondrial oxidative phosphorylation (OXPHOS) pathway responsible for generating ATP. These and other OXPHOS pathway genes have been reported to be under selection in several, independent lineages of marine diving mammals (e.g. Tian et al. [66]).

Despite adaptations to meet oxygen demands, the fluctuating levels of hypoxia and reoxygenation experienced by diving animals causes a build-up of reactive oxygen species that damage DNA, proteins and lipids. There are 75 PSGs annotated to the enriched set *cellular response to DNA damage stimulus* (GO:0006974). Many of these genes are linked to hypoxia-induced oxidative stress and DNA damage repair, among other functions. Included in this PSG set are *DMAP1*, *SMARCAL1* and *USP7*, which are among twelve genes reported to be under positive selection for hypoxia tolerance in the high altitude adapted Tibetan hot-spring snake [68].

Salt and water balance

Sea snakes eliminate excess salts using specialised sublingual glands. Extrarenal salt glands have also evolved in birds, turtles and crocodiles [69] and use membrane transport proteins, notably the sodium-potassium pump (Na+/K+-ATPase), to transport ions across the epithelial cells lining the secretory tubules of the gland. Our PSG set includes *ATP1B4*, which encodes a subunit of the sodium/potassium-transporting ATPase [70]. However, this molecular pathway is involved in non-osmoregulatory functions, and altered selection on *ATP1B4* has been linked to traits such as toxin resistance in toads [71] and muscle development in tigers [72]. Expression studies are therefore needed to identify whether *ATP1B4* is upregulated in sea snake salt glands.

A PSG closely linked to osmoregulation is the renal chloride channel gene *CLCN5*, which encodes a member of voltage-gated chloride channel family that, in humans, is predominantly expressed in the kidney and plays a critical role in renal chloride reabsorption [73]. Selection on this and other renal genes indicates adaptation of kidney function in sea snakes despite observations of low concentrations and rates of excretion of salt in sea snake urine [74].

Neural and sensory adaptations

Related to sensory perception are under-represented terms *detection of chemical stimulus* (GO:0009593) and *G protein-coupled receptor activity* (GO:0004930); and one over-represented term - *anatomical structure development* (GO:0048856) (Fig. 6A-B; Additional file 1: Fig. S21; Additional file 2: Table S16). Many sensory PSGs relate to aspects of visual function. Among these are the aquaporin-encoding *MIP* gene, which

has a specific and critical function in the transparency and refractive properties of the lens [75]. In aquatic vision, only the lens is responsible for focusing light onto the retina because the cornea has negligible refractive power in water [76]. Other vision PSGs include Retinitis Pigmentosa 2 (*RP2*), and Retinal Degeneration 3 (*RD3*), which function in the development of retinal photoreceptors [77–79]; *FSCN2*, involved in photoreceptor disk morphogenesis [80,81]; *RDH10*, which encodes an enzyme that is involved in the production of retinoic acid in the retina [82]; and the eye development gene *TMX3*, which is also a candidate PSG in cichlid fish [83].

Sea snake PSG *BARHL1* is primarily expressed in the vertebrate cerebellum, and is a key component of the proprioceptive pathway, which is the sense that provides information concerning movements, orientation and position of the body within the environment [84]. The proprioceptive capacity of sea snakes is virtually unknown but is likely to have been subject to selection during their transition to the three-dimensional marine environment. Three deafness-related PSGs might also perform roles in balance and spatial orientation: *PDZD7* and *GRXCR2* are involved in morphogenesis of the sensory hair cells (‘stereocilia’) in the inner ear [85,86], and *TECTA* encodes a key protein of the tectorial membrane in which the stereocilia are embedded [87,88]. These results provide the first candidate genes to complement morphological analyses of cerebellum [89] and inner ear [90] evolution during aquatic transitions in snakes.

Finally, several PSGs are associated with neural function and behaviour. These include *DCDC2*, which is linked to perception and memory [91,92]; and *TIMELESS*, which regulates circadian rhythms [93,94].

Morphological marine innovations

All sea snakes show major modifications of head, body and tail shape associated with marine locomotion and prey capture (e.g. Voris and Jayne [95]). However, *Hydrophis* show accelerated rates of head and body shape change that are linked to trophic specialisation, high sympatric diversity, and elevated speciation rates compared to other sea snakes and terrestrial elapids [96]. These rapid morphological shifts in *Hydrophis* are underpinned by developmental mechanisms involving the presence and positioning of Hox boundaries and heterochronic changes in segmentation of the vertebrae [96,97]. Our PSG set includes several genes that have key roles in cranial and axial patterning during embryonic development and are annotated to multiple over-represented GO terms. Three PSGs are bone morphogenetic proteins (*BMPER*, *BMPR1A* and *BMPR2*), which belong to a pathway of prolific candidate genes for morphological differentiation in numerous vertebrae systems (e.g. Darwin's finches [98]; phyllostomid bats [99]). *BMPR1A*, for example, mediates craniofacial development, including tooth and palate formation [100–102]. Several PSGs belong to the Wnt signalling pathway (e.g. *AXIN2*, *PITX2*, *TCFL5*, *WNT4*, *WNT10A*, *WNT10B*, *WNT11*, *WNT2B*) and HOX family (*HOXA11 HOXD11*), both of which play critical roles in skeletal patterning [103–108]. Transcriptional studies will be needed to determine the specific roles of these candidate genes in the evolutionary development of sea snakes.

Conclusion

We have generated four high quality genomes for the prolific *Hydrophis* sea snake radiation. Three of these are at chromosome-scale and represent the best reference

genomes that have so far been provided for the medically important Elapidae based on assembly metrics relating to contiguity, base-accuracy, sequence completeness and gene completeness. Our strict screens for gene selection in comparison to terrestrial snakes identified marine-associated genes in sea snakes that are linked to functional categories including hypoxia adaptation, sensory perception, immune response and morphological development. This fills a conspicuous taxonomic gap in genomic studies of marine adaptation and has broad relevance for understanding how divergent and convergent changes to a shared genetic toolkit underpin ecological transitions in vertebrates. Other results include phylogenomic tree and network analyses, which highlight the possibility of near-simultaneous speciation at the root of *Hydrophis*, and synteny maps showing intra-chromosomal variations that will be important targets for future adaptation and speciation genomic studies of this system.

Methods

Public genomes and sequence datasets

Existing genome assemblies and annotation files were downloaded for the following organisms: *Anolis carolinensis* [109], *Crotalus tigris* [110], *Naja naja* [111,112], *Notechis scutatus [113]*, *Pantherophis guttatus* [114], *Protobothrops mucrosquamatus [115]*, *Pseudonaja textilis [116]*, *Python bivittatus [117]* and *Thamnophis elegans* [118]. The genome assemblies of *Hydrophis curtus* (East) and *Hydrophis cyanocinctus* were also downloaded [6,119,120], along with their corresponding RNA-sequencing data [121,122], which were filtered for contamination against the Kraken 2 (v2.1.2) standard database (dated 2021-05-17) [123] and quality filtered using Fastp (v0.23.2) [124].

Sample collection

H. major and *H. elegans* were collected and sampled during fieldwork in Western Australia, using procedures approved by The University of Adelaide's Animal Ethics Committee (approval number S-2015-119/34903), under a fauna taking license (regulation 25, number FO25000393) granted by the Department of Biodiversity, Conservation and Attractions of Western Australia. *H. ornatus* and *H. curtus* (West) were sourced by collaborators from commercial fisheries operating in coastal waters of the Emirate of Fujairah, United Arab Emirates.

Library construction and Sequencing

PacBio HiFi sequence data was generated from high-molecular weight DNA from *Hydrophis major*. DNA was extracted from blood using the Monarch HMW DNA Extraction Kit for Tissue (New England BioLabs Inc. - #T3060S/L) as per the manufacturer's instructions at Australian Genome Research Facility (AGRF) Adelaide. Two sequence libraries were constructed using the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, Menlo Park, CA, USA), which were each sequenced on a 8M SMRT cell on a PacBio Sequel II system. HiFi circular consensus sequences (CCS) were generated from the subreads using the SMRT link software (v10.1.0.119588). Both the library construction and sequencing were performed at AGRF-University of Queensland PacBio facility (AGRF-UQ PacBio). The HiFi reads were then filtered for adapter contamination using the program HifiAdapterFilt (v2.0.0) [125] (Additional file 2: Table S1). Hi-C libraries were prepared from kidney tissue using the Arima-HIC 2.0 protocol at the Australian Cancer Research Foundation Biomolecular Resource Facility

(ACRF BRF). Libraries underwent size selection and were quality checked for concentration and size using Bioanalyzer and Qbit before being sequenced on an Illumina NovaSeq 6000 machine (2x150bp paired-end) (Additional file 2: Table S1). RNA from brain, skin, liver and vitellogenic follicles was extracted using RNeasy mini-kits (Qiagen), with sequence libraries prepared following the Illumina Stranded mRNA Prep, Ligation protocol (Illumina Inc., San Diego, USA). The libraries were then sequenced on an Illumina NovaSeq 6000 machine on S1 flowcells (NovaSeq 1.5 chemistry kits; 2 x 100bp paired-end). RNA-sequence data was then filtered for contaminants against Kraken 2 (v2.1.2) standard database (dated 2021-05-17) [123], with low quality reads being filtered using Fastp (v0.23.2) [124] (Additional file 2: Table S1).

To generate Nanopore whole genome data for *H. elegans, H. ornatus* and *H. curtus* (West), DNA was extracted from 25-50 µl of blood using the Circulomics Nanobind CBB Big DNA Kit, following the “Nanobind UHMW DNA extraction - Nucleated Blood Protocol”. Library preparation was performed on this DNA using the Ultra-Long DNA sequencing kit (SQK-ULK001) before the libraries were sequenced across 2 x promethION (FLO-PRO002) flow cells each for 72 hours. DNase washes (EXP-WSH003) were performed at 24 and 48 hours to help unblock pores and increase overall output. Basecalling was performed with Guppy (v4.0.11) for *H. ornatus* and *H. curtus* (West) and Guppy (v6.1.5) for *H. elegans* (Additional file 2: Table S1). For *H. ornatus* and *H. curtus* (West), 70 µl of blood was sent to ACRF Biomolecular Resource Facility, for generation of Hi-C libraries. Samples were prepared using the Arima Hi-C

preparation and run on NovaSeq S1 300 cycles (2x150bp Paired End), before trimming with Trim Galore (v0.6.6) [126] (Additional file 2: Table S1).

Short-read libraries were generated for *H. elegans, H. ornatus* and *H. curtus* (West). For *H. elegans*, DNA was extracted from tail tissue following the Gentra Puregene Tissue Kit protocol. DNA was sent to the South Australian Genomic Centre (SAGC), where libraries were generated according to Illumina DNA Prep (M) guidelines (Part No. 10000000254 v10) and Illumina to MGI Library Conversion (MGIEasy Universal Library Conversion Kit, Part No. MGI000004155), and were sequenced on a MGI DNBSEQ-G400 (2x150bp paired-end). The MGI sequence data was then filtered for contaminants and low quality against Kraken 2 and Fastp, respectively. For *H. ornatus* and *H. curtus* (West), DNA was extracted from blood using the Circulomics CBB Big Nanobind kit, following the “HMW DNA Extraction – Nucleated Blood" protocol. DNA was sent to the Ramaciotti Centre for Genomics, where libraries were prepared using the Illumina TruSeq DNA PCR-Free workflow, and paired end sequencing was performed on a NovaSeq 6000 with a SP 2x150 bp flow cell. Reads were then trimmed using Trim Galore (v0.6.6) (Additional file 2: Table S1).

Genome assembly

Genome size estimation

Genome size estimation was performed for each of the four sea snakes prior to assembly. High-accuracy sequence datasets for each snake (PacBio Hifi for *H. major*, Illumina for *H. ornatus* and *H. curtus* (West) and MGI for *H. elegans*) were passed to KMC (v3.2.1) to count 31-mers before being dumped to file using kmc_tools [127].

GenomeScope2 (v2.0) was then used to estimate genome sizes from the K-mer histograms [128].

Hydrophis major

The PacBio HiFi long-reads were assembled into a primary and dual assembly (pseudo-haplotypes) using hifiasm (v0.16.1) [129,130]. Hi-C reads were supplied to hifiasm to assist with phasing but were not incorporated into the initial assemblies. The Hi-C reads were then mapped to the primary and dual assemblies using an adapted version of the Arima Hi-C mapping pipeline. The cleaned Hi-C alignments were passed to pin_hic (v3.0.0) to perform iterative scaffolding and mis-join correction [131], followed by manual curation in Juicebox Assembly Tools (JBAT) to anchor scaffolds into chromosomes [132,133]. Following manual curation, TGS-GapCloser (v1.0.3) was used to perform gap-filling [134], utilising the HiFi long-reads, resulting in the final assemblies.

Hydrophis ornatus and Hydrophis curtus (West)

A total of 60x Nanopore long reads and 70x illumina short-reads and 50x Hi-C reads were generated for both *H. ornatus* and *H. curtus* (West). For assembly, initial contigs were generated from the ultra long ONT reads using the Flye (v2.8.3) assembler [135], including 2 polishing iterations. The resulting contigs were then polished using Hypo (v1.0.3) [136], utilising both the Nanopore long-reads and Illumina short-reads. The program Purge Haplotigs was then used to remove heterozygous, syntenic contigs from each of the primary assemblies to reduce redundancy before scaffolding [137]. The Hi-C data was processed with Juicer (v1.6) [133], then used as input for the 3d-DNA

pipeline (v180419) [138]. The resulting assembly was then manually reviewed and edited in JBAT to form the final chromosome sequences for each snake.

Hydrophis elegans

Raw Nanopore reads were assembled into contigs using Flye (v2.9-b1768) [135]. Assembled contigs were then polished using the Nanopore data and the program Medaka (v1.8.0) [139] before a final two rounds of polishing using the high-accuracy paired-end sequence data via Nextpolish (v1.4.1) [140].

Genome assessment

Multiple completeness measures were used to assess the quality of the four newly assembled genomes. General assembly metrics were generated using QUAST (v5.0.2) [141]. Reference free assembly evaluation was performed with Merqury (v1.3) [142], comparing the k-mer profile of each assembly to its respective sequence dataset, generating k-mer spectra plots, k-mer recovery rate tables and Phred quality consensus estimates (QV) for each genome. Finally, the level of gene-completeness in each assembly was assessed using BUSCO (v5.2.2) using the Tetrapoda (odb10) database of 5,310 single copy orthologs [143,144].

Repeat Annotation

The *de novo* repeat annotation pipeline Extensive De Novo TE Annotator (EDTA; v2.0.1) was first used to model and annotate repetitive elements in each of the four snake genomes [145]. *EDTA* was run in ‘divide and conquer’ mode, first identifying LTR, TIR and helitron elements, before running the remaining annotation steps. To reduce

the misclassification of gene sequences as repetitive elements, coding sequences from *Notechis scutatus* were provided as gene evidence from a somewhat evolutionarily close species. A combined library of EDTA modelled repeats and curated RepBase repeat sequences (v2018-10-26) were then passed to RepeatMasker to perform homology-based repeat annotation [146,147]. Kimura divergence repeat landscapes were generated from the RepeatMasker output using the accessory script calcDivergenceFromAlign.pl and a custom R-script for visualisation. LTR insertion times (*T*) were estimated from the EDTA output using the equation $T = d/2\mu$, where *d* is the sequence divergence between LTR pairs and μ is the mutation rate. The mutation rate 4.71×10^{-9} mutations per site/year was used for these calculations [6].

The quality of the annotated repeats were assessed using two approaches: by evaluating the loss of complete, single-copy BUSCOs from each genome after hard-masking repetitive elements, and via the LTR Assembly Index (LAI) [25], which is a formal measure of LTR-completeness within a genome assembly. The program LAI, which is packaged with the program LTR_retriever [148], was run on the output generated by EDTA.

Gene annotation

The pipeline Funannotate (v1.8.11) was used to predict protein coding genes in *H. major*, *H. cyanocinctus* and *H. curtus* (East) using transcriptomic, homology and *de novo* methods [149]. These snakes were *de novo* annotated as they all had transcriptomic data from the same sample that was assembled, and the Li et al. [6] snakes did not have public gene annotations at the time of writing. Prior to running the Funannotate pipeline, three external sources of supporting gene evidence were

generated. First, Liftoff (v1.6.3) [150] was used to lift the RefSeq gene annotations from *Notechis scutatus*, *Psudonaja textilis*, *Naja naja*, *Protobothrops mucrosquamatus*, *Thamnophis elegans* and *Anolis carolinensis* to each of the three snakes. Next, protein sequences from the six RefSeq annotations above, the reviewed SwissProt database [151] and the lifted-over annotations from each snake were pooled and passed to MMseqs2 easy-cluster (v14.7e284) [152] to create a non-redundant set of representative proteins. This representative protein set, along with the reviewed Swissprot database, were then used to generate homology-based gene predictions for each snake using MetaEuk easy-predict (v6.a5d39d9) [153].

Funannotate train was then used to assemble genome-guided transcripts for each snake using Trinity (v2.8.5) [154], which were converted into transcript-derived gene models by PASA (v2.4.1) [155]. Gene prediction was then performed via the Funannotate predict module, which performs *de novo* and homology-based gene prediction, before incorporating the transcriptomic, *de novo*, homology and external sources of evidence into a non-redundant gene-set using EVidenceModeler (v1.1.1) [156]. The external sources of gene and protein evidence were incorporated into the pipeline at this stage. Next, Funannotate update was used to refine the resulting gene models via two rounds of *PASA* annotation and compare, before filtering the gene-set using expression information estimated by Kallisto (v0.46.1) [157]. The predicted gene models were screened against the InterProScan5 database (v5.57-90.0) *[158]* and EggNOG database (v5.0) using EggNOGG-mapper (v2.1.9) [159,160] to obtain functional annotations, which were then compiled by Funannotate annotate into non-redundant functional annotations. As transcriptomic data was not available for the *H.*

ornatus, H. curtus (West) and *H. elegans* samples, we elected to lift gene annotations from *H. major* to each of the three snakes using Liftoff [150]. Finally, BUSCO [143,144] was used to assess the overall completeness of the predicted genes against Tetrapoda (odb10), while length distribution plots were used to compare the predicted gene models to the RefSeq annotated snakes.

Species tree estimation

A combination of phylogenomic tree and network analyses were performed to account for the possibility that the early rapid radiation within *Hydrophis* does not conform to a bifurcating tree.

First, a two-step approach was used to build a species tree while accounting for discordances with and among gene trees. Maximum likelihood gene trees were inferred for 9,277 *Hydrophis*-specific single-copy orthologs, all of which had at least one parsimony-informative site, using IQ-TREE (v2.2.0.3) [161]. For each ortholog, IQ-TREE was allowed to automatically select the best-fitting model and performed 1000 ultrafast bootstraps. These individual trees were then used to infer a single species tree with ASTRAL-III (v5.7.8) [162]. To explore discordance among gene trees, we calculated gene-concordance factors (gCF) using IQ-TREE, which represent the proportion of gene trees that contain each node of the species tree.

Species networks were inferred from a reduced set of IQ-TREE gene trees using PhyloNet (v3.8.2) [163] with the command InferNetworks_ML. Network searches were run with 5 iterations, allowing up to 4 reticulation events, with branch lengths and inheritance probabilities optimised for each proposed network. To reduce run times, and due to the limited amount of sequence variation per ortholog, the IQ-TREE gene trees

were filtered using PhyKIT (v1.11.15) [164] to find the set of 2,568 trees that included all six taxa and at least five parsimony informative sites.

Finally, the full *Hydrophis* assemblies were used as input to SANS serif (v2.3_3A) [165], which calculates a set of splits by comparing shared k-mer distributions between unaligned reference genomes. SANS serif was run using the default k-mer value (k=31), the geometric mean weight function and using the *weakly* filtering criterion to greedily filter the list of splits.

Genome synteny and structural rearrangements

Broad-scale genomic synteny was investigated between the five-chromosome scale *Hydrophis* assemblies (*H. ornatus, H. major, H. curtus* (West)*, H. curtus* (East) and *H. cyanocinctus)* and the chromosome scale outgroup *Thamnophis elegans* using MCscan [166]. Protein sequences from each snake were extracted using AGAT (v0.9.2) [167] and were pairwise aligned using LAST [168] using the JCVI python module [169]. Initial alignments between snakes were used to identify chromosomes assembled in the reverse complement, which were corrected in one of the snakes by reverse complementing the sequence using SAMtools faidx (v1.16.1) [170]**.** Gene annotations were lifted over to the newly reversed sequences using Liftoff, before re-running the MCscan pipeline to generate the final, oriented synteny plots.

Structural variation between the *Hydrophis* snakes was assessed using the program Syri (v1.6.3) [171]. The curated genomes used in the MCscan alignments were further edited to have the same karyotype and chromosome identifiers (a requirement of Syri). *Hydrophis ornatus* was used as the anchoring reference, with the karyotypes and chromosome identifiers of the other *Hydrophis* snakes edited to match. Pairwise

alignments were then performed between the snakes using Minimap2 (v2.24-r1122) [172], with the resulting alignment files passed to Syri to identify structural variation. Plotsr (v1.1.0) was then used to visualise the structural rearrangements in a single plot [173].

Single copy ortholog detection

The snakes *Hydrophis major, H. elegans, H. ornatus, H. cyanocinctus, H. curtus* (East)*, H. curtus* (West)*, Crotalus tigris, Notechis scutatus, Pantherophis guttatus, Protobothrops mucrosquamatus, Pseudonaja textilis, Python bivittatus and Thamnophis elegans* were used for single-copy ortholog identification. Snakes not part of this study were selected due to having RefSeq gene annotations and a minimum *BUSCO* score of >85%. For each snake, AGAT (v0.9.2) was used to extract each gene's longest isoform in peptide and nucleotide format. OrthoFinder (v2.5.2) was then used to find orthogroups, using the protein sequences and estimated species tree (see above) [174]. Single-copy orthologs were then aligned using Mafft (v7.505) [175] and codon-translated using Pal2Nal (v14) [176], before dynamic trimming using ClipKIT (v1.3.0) [177]. Orthogroup annotations were generated from a range of data sources. Gene symbols were parsed from RefSeq and Funannotate gene annotation files, along with BLASTP (v2.12.0) [178] results after screening protein sequences against SwissProt [179].

Selection testing

Signatures of positive selection in *Hydrophis*

Two separate techniques were used to identify *Hydrophis* specific signals of positive selection using the single-copy orthologs: PAML drop-out experiments and BUSTED-PH.

PAML is a maximum-likelihood method that infers positive selection using the non-synonymous/synonymous rate ratio (*dN/dS* or ω*)*, where ω > 1 is indicative of positive selection, ω < 1 signifies purifying selection and ω ≈ 1 is neutral [30,180]. PAML Branch-Site tests for positive selection were run for each single-copy ortholog, with the *Hydrophis* snakes (and all subtending branches) marked as foreground. To account for signatures of pervasive selection (i.e. tree-wide selection), we performed drop-out tests as recommended by Kowalczyk et al. [28], whereby foreground branches were removed from the species tree and Site models were run for each ortholog on the background branches. Likelihood ratio tests (LRT) were then used to compare the null and alternate models within the Branch-Site and Site tests, respectively, with a Benjamini-Hochberg correction being applied to account for multiple testing (FDR ≤ 0.01) [181]. Genes were only considered as under positive selection within *Hydrophis* when the Site model failed to reach significance (FDR > 0.01), but the Branch-Site did (FDR < 0.01).

The second approach involved running the program BUSTED-PH for each single-copy ortholog. BUSTED (Branch-site Unrestricted Statistical Test for Episodic Diversification) is a model which provides a gene-wide test for positive selection, asking if a gene has experienced positive selection at least one site along at least one branch [31]. The workflow BUSTED-PH (BUSTED-Phenotype) builds on this model by testing if

a specific phenotype/trait is associated with positive selection by performing a series of selection tests that are similar to the drop-out tests above [29]. BUSTED-PH not only uses a different selection testing framework, but it also provides context relating to the selective regimes between the *Test* (*Hydrophis*) and *Background* (*Terrestrial*) species. BUSTED-PH was run for each single-copy ortholog using the same tree partitioning scheme as the PAML drop-out analyses, with the same FDR correction being applied (FDR ≤ 0.01). The final set of *Hydrophis* specific positively selected genes was obtained by taking only genes that reached significance in both methods (FDR ≤ 0.01) and showed no sign of pervasive selection across the tree.

Testing the strength of natural selection

Understanding the selective regime of genes is equally as important as simply identifying selection candidates. RELAX is a hypothesis testing framework for detecting selective strength in a codon-based phylogenetic framework [32]. It is a relative measure, testing if the strength of selection in a *Test* partition is different to that of the *Reference* group. It first fits a null model where the *Test* and *Reference* branches share the same selective regime, before fitting the alternate model which incorporates a selection intensity parameter - *k* - a free parameter that is used to adjust the *Test* ω-rate-classes while being fixed to $k = 1$ in the *Reference* set. The null and alternate models are then compared using an LRT statistic, with a significant LRT statistic indicating differing selective regimes between *Test* and *Reference* branches. The strength of selection is then interpreted from the selection intensity parameter *k*, where $k > 1$ represents an intensification of positive selection and $k < 1$ indicates a relaxation of positive selection. RELAX was run on each single-copy ortholog using a species tree

where *Hydrophis* snakes were marked as the *Test* and all remaining branches marked as *Reference*. FDR corrections were applied to LRT *p*-values using a filtering criterion of FDR ≤ 0.01.

Gene ontology over-representation and semantic similarity analysis

Gene ontology over-representation analysis was used to identify biologically relevant GO terms within the *Hydrophis* positively selected gene set. The program PANTHER *(GO database released 10 May 2023)* was used to perform the over-representation tests [33,34], to identify functional classes that were significantly over- or under-represented in the input list of genes. The *Hydrophis* PSGs were passed to PANTHER to perform the over-representation analysis, specifying *Homo sapiens* as the reference organism, using a Fisher's exact test with FDR multiple test correction (*FDR* < 0.05) to perform the over-representation test. Significantly enriched GO terms were returned as a hierarchical table, with the most specific GO Terms being reported first, followed by more general parental terms occurring as nested entries. The program REVIGO (GO database released 10 May 2023; UniProt-to-GO mapping database released 15 March 2023) was then used to reduce the over-represented GO terms into an informative set of non-redundant, representative biological terms [35]. Over-represented GO terms and their *p*-values were passed to REVIGO, setting the semantic threshold to 'small', filtering obsolete GO terms and setting UniProt as the database to compare to.

Abbreviations

QV: Consensus quality value

LAI: LTR Assembly Index

LTR: Long terminal repeat

MYA: Million years ago

SV: Structural variant

PSG: Positively selected gene

GO: Gene Ontology

BP: Biological process

CC: Cellular component

MF: Molecular function

MDS: Multidimensional scaling

SRA: Sequence read archive

Supplementary information

Additional file 1:

Fig. S1. Cumulative length of assembled sequences. **Fig. S2** Assembly Nx plots. **Fig. S3** *H. major* Hi-C contact map. **Fig. S4** *H. curtus* (West) and *H.* cyanocinctus Hi-C contact maps. **Fig. S5** Genome size estimation. **Fig. S6** *H. major* k-mer spectra and multiplicity. **Fig. S7** K-mer spectra for *H. elegans, H. curtus* (West) and *H. cyanocinctus.* **Fig. S8** Assembly BUSCO completeness. **Fig. S9** Length distributions of gene features. **Fig. S10** Protein BUSCO completeness. **Fig. S11** Hard-mask BUSCO completeness. **Fig. S12** *H.* major sliding window LAI scores. **Fig. S13** *H. ornatus* sliding window LAI scores. **Fig. S14** *H. curtus* (West) sliding window LAI scores. **Fig. S15** *H. elegans* sliding window LAI scores. **Fig. S16** PhyloNet networks for varying reticulation values.

Fig. S17 Chromosomal synteny before manual curation. **Fig. S18** Structural variation between *Hydrophis* snakes. **Fig. S19** Overlap between PAML drop-out and BUSTED-PH methods. **Fig. S20** Summary of PAML branch-site (alternate model) results. **Fig. S21** Fold enrichment reported by PANTHER.

Additional file 2:

Table S1 Sequence data summary. **Table S2** Estimated genome size statistics. **Table S3** Genome assembly summary metrics. **Table S4** K-mer completeness and QV statistics. **Table S5** Gene annotation summary statistics. **Table S6** RepeatMasker summary tables. **Table S7** LTR Assembly Index results for each of the newly assembled genomes. **Table S8** PhyloNet output for each reticulation value. **Table S9** A summary of structural variants identified by Syri between sea snakes. **Table S10** OrthoFinder summary statistics. **Table S11** Orthogroup summary statistics. **Table S12** PAML drop-out results for each single-copy ortholog. **Table S13** BUSTED-PH results for each single-copy ortholog. **Table S14** Selection testing results for marine-specific positively selected genes. **Table S15** RELAX results for each single-copy ortholog. **Table S16** Over-representation results generated by PANTHER. **Table S17** REVIGO clustering of the 46 most specifically enriched GO terms.

Declarations

Ethics approval and consent to participate

Sampling procedures were approved by The University of Adelaide's Animal Ethics Committee (approval number S-2015-119 / 34903), under a fauna taking license

(regulation 25, number FO25000393) granted by the Department of Biodiversity, Conservation and Attractions of Western Australia. *H. ornatus* and *H. curtus* (West) were sourced by collaborators from commercial fisheries operating in coastal waters of the Emirate of Fujairah, United Arab Emirates.

Consent for publication

Not applicable

Availability of data and materials

All data generated or analysed during this study are included in this published article, its supplementary information files, and publicly accessible repositories. Genome assemblies for *H. major* (JAUCBL000000000) and *H. elegans* (JAUPSR000000000) have been uploaded to NCBI GenBank under the BioProject accession PRJNA984433 [182]. Raw sequencing reads generated for *H. elegans* have additionally been uploaded to the NCBI Sequence Read Archive (SRA) under the same BioProject. Raw sequencing data for *H. major* is available through the Bioplatforms Data Portal under the following dataset identifiers: PacBio HiFi (102.100.100/351827 and 102.100.100/351778), Hi-C (102.100.100/351780) and RNA (102.100.100/351777). Genome assemblies and sequencing data (Nanopore, Illumina and Hi-C) for *H. ornatus* (JAWKAR000000000) and *H. curtus* (West) (JAWKAS000000000) have been uploaded to NCBI GenBank and SRA under BioProject PRJNA780942 [183]. Scripts and pipelines used to conduct the analyses are available through GitHub [184], with additional data files deposited to figshare [185].

Competing interests

The authors declare that they have no competing interests.

Funding

Australian Research Council Discovery project (DP180101688; KLS)

Australian Research Council Future Fellowship (FT130101965; KLS)

Medical Research Future Fund (MRF1173594; IWD)

Author contributions

Study design: KLS, AJL. Genome assembly and curation: AJL, JMH. Genome annotation, synteny, selection and overrepresentation analyses: AJL. Phylogenetic analyses: KLS, AJL. Manuscript preparation: AJL, KLS. Funding: KLS, IWD. All authors read and approved the final manuscript.

Acknowledgements

We would like to acknowledge the contribution of the Australian Amphibian and Reptile Genomics consortium in the generation of data used in this publication. The Australian amphibian and reptile genomics initiative is supported by funding from Bioplatforms Australia through the Australian Government's National Collaborative Research Infrastructure Strategy (NCRIS), the Australian National University, the University of Canberra, the Australian Museum, Museums Victoria and the South Australian Museum. Funding for research and sequencing was also provided by Hydrophis UK Ltd as part of the company's research and development into sea snake cutaneous

respiration. The research was also supported by an Australian Research Council Future Fellowship (FT130101965) and Discovery Project grant (DP180101688) to KLS. We are grateful to the Department of Biodiversity, Conservation and Attractions of Western Australia for permission to collect sea snake DNA samples. Finally, we thank Vinay Udyawer, Jenna Crowe-Riddell, James Nankivell and Balazs Buzas for their help with sample collection.

References

1. McGowen MR, Grossman LI, Wildman DE. Dolphin genome provides evidence for adaptive evolution of nervous system genes and a molecular rate slowdown. Proc R Soc B. 2012;279(1743):3643–51.

2. Sun YB, Zhou WP, Liu HQ, Irwin DM, Shen YY, Zhang YP. Genome-Wide Scans for Candidate Genes Involved in the Aquatic Adaptation of Dolphins. Genome Biology and Evolution. 2013;5(1):130–9.

3. Foote AD, Liu Y, Thomas GWC, Vinař T, Alföldi J, Deng J, et al. Convergent evolution of the genomes of marine mammals. Nat Genet. 2015;47(3):272–5.

4. Chikina M, Robinson JD, Clark NL. Hundreds of Genes Experienced Convergent Shifts in Selective Pressure in Marine Mammals. Mol Biol Evol. 2016;33(9):2182–92.

5. Gayk ZG, Le Duc D, Horn J, Lindsay AR. Genomic insights into natural selection in the common loon (Gavia immer): evidence for aquatic adaptation. BMC Evol Biol. 2018;18:64.

6. Li A, Wang J, Sun K, Wang S, Zhao X, Wang T, et al. Two Reference-Quality Sea Snake Genomes Reveal Their Divergent Evolution of Adaptive Traits and Venom Systems. Nowick K, editor. Molecular Biology and Evolution. 2021;38(11):4867-4883 .
7. Peng C, Ren JL, Deng C, Jiang D, Wang J, Qu J, et al. The genome of Shaw's sea snake (Hydrophis curtus) reveals secondary adaptation to its marine environment. Liu L, editor. Molecular Biology and Evolution. 2020;37(6):1744-1760.
8. Whiteley SL, Holleley CE, Wagner S, Blackburn J, Deveson IW, Marshall Graves JA, et al. Two transcriptionally distinct pathways drive female development in a reptile with both genetic and temperature dependent sex determination. Aboobaker AA, editor. PLoS Genet. 2021;17(4):e1009465.
9. Seymour RogerS. Physiological Adaptations to Aquatic Life. In: Biology of the Reptilia, Physiological Ecology. Academic Press; 1982. p. 1–51.
10. Rasmussen AR, Murphy JC, Ompi M, Gibbons JW, Uetz P. Marine Reptiles. Clifton J, editor. PLoS ONE. 2011;6(11):e27373.
11. Lee MSY, Sanders KL, King B, Palci A. Diversification rates and phenotypic evolution in venomous snakes (Elapidae). R Soc open sci. 2016;3(1):150277.
12. Sanders KL, Lee MSY, Leys R, Foster R, Keogh JS. Molecular phylogeny and divergence dates for Australasian elapids and sea snakes (hydrophiinae): evidence from seven genes for rapid evolutionary radiations. Journal of Evolutionary Biology. 2008;21(3):682–95.
13. Brischoux F, Shine R. Morphological adaptations to marine life in snakes. Journal of Morphology. 2011;272(5):566–72.

14. Sanders KL, Rasmussen AR, Elmberg J. Independent Innovation in the Evolution of Paddle-Shaped Tails in Viviparous Sea Snakes (Elapidae: Hydrophiinae). Integrative and Comparative Biology. 2012;52(2):311–20.

15. Heatwole H, Seymour R. Pulmonary and cutaneous oxygen uptake in sea snakes and a file snake. Comparative Biochemistry and Physiology Part A: Physiology. 1975;51(2):399–405.

16. Heatwole H, Seymour R. The Biology of Sea Snakes. University Park Press; 1975.

17. Graham JB. Aquatic respiration in the sea snake Pelamis Platurus. Respiration Physiology. 1974;21(1):1–7.

18. Palci A, Seymour RS, Van Nguyen C, Hutchinson MN, Lee MSY, Sanders KL. Novel vascular plexus in the head of a sea snake (Elapidae, Hydrophiinae) revealed by high-resolution computed tomography and histology. R Soc open sci. 2019;6(9):191099.

19. Seymour RS, Webster MED. Gas transport and blood acid-base balance in diving sea snakes. Journal of Experimental Zoology. 1975;191(2):169–81.

20. Simões BF, Gower DJ, Rasmussen AR, Sarker MAR, Fry GC, Casewell NR, et al. Spectral Diversification and Trans-Species Allelic Polymorphism during the Land-to-Sea Transition in Snakes. Current Biology. 2020;30(13):2608-2615.e4.

21. Crowe-Riddell JM, Williams R, Chapuis L, Sanders KL. Ultrastructural evidence of a mechanosensory function of scale organs (sensilla) in sea snakes (Hydrophiinae). R Soc open sci. 2019;6(4):182022.

22. Crowe-Riddell JM, Snelling EP, Watson AP, Suh AK, Partridge JC, Sanders KL. The evolution of scale sensilla in the transition from land to sea in elapid snakes. Open Biol. 2016;6(6):160054.

23. Crowe-Riddell JM, Simões BF, Partridge JC, Hunt DM, Delean S, Schwerdt JG, et al. Phototactic tails: Evolution and molecular basis of a novel sensory trait in sea snakes. Molecular Ecology. 2019;28(8):2013–28.

24. Ukuwela KDB, Silva A de, Mumpuni, Fry BG, Sanders KL. Multilocus phylogeography of the sea snake Hydrophis curtus reveals historical vicariance and cryptic lineage diversity. Zoologica Scripta. 2014;43(5):472–84.

25. Ou S, Chen J, Jiang N. Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Research. 2018;46(21):e126

26. Sanders KL, Mumpuni, Lee MSY. Uncoupling ecological innovation and speciation in sea snakes (Elapidae, Hydrophiinae, Hydrophiini). Journal of Evolutionary Biology. 2010;23(12):2685–93.

27. Zaher H, Murphy RW, Arredondo JC, Graboski R, Machado-Filho PR, Mahlow K, et al. Large-scale molecular phylogeny, morphology, divergence-time estimation, and the fossil record of advanced caenophidian snakes (Squamata: Serpentes). Joger U, editor. PLoS ONE. 2019;14(5):e0216148.

28. Kowalczyk A, Chikina M, Clark NL. A cautionary tale on proper use of branch-site models to detect convergent positive selection. bioRxiv. 2021. http://biorxiv.org/lookup/doi/10.1101/2021.10.26.465984.

29. iGEM/UCSD evolutionary biology and bioinformatics group. hyphy-analyses. https://github.com/veg/hyphy-analyses. Accessed 22 Mar 2023.

30. Álvarez-Carretero S, Kapli P, Yang Z. Beginner's Guide on the Use of PAML to Detect Positive Selection. Crandall K, editor. Molecular Biology and Evolution. 2023;40(4):msad041.

31. Murrell B, Weaver S, Smith MD, Wertheim JO, Murrell S, Aylward A, et al. Gene-Wide Identification of Episodic Selection. Molecular Biology and Evolution. 2015;32(5):1365–71.
32. Wertheim JO, Murrell B, Smith MD, Kosakovsky Pond SL, Scheffler K. RELAX: Detecting Relaxed Selection in a Phylogenetic Framework. Molecular Biology and Evolution. 2015;32(3):820–32.
33. Mi H, Muruganujan A, Huang X, Ebert D, Mills C, Guo X, et al. Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). Nat Protoc. 2019;14(3):703–21.
34. Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou L, Mi H. PANTHER : Making genome-scale phylogenetics accessible to all. Protein Science. 2022;31(1):8–22.
35. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. Gibas C, editor. PLoS ONE. 2011;6(7):e21800.
36. Card DC, Jennings WB, Edwards SV. Genome Evolution and the Future of Phylogenomics of Non-Avian Reptiles. Animals. 2023;13(3):471.
37. Sanders KL, Lee MSY, Mumpuni, Bertozzi T, Rasmussen AR. Multilocus phylogeny and recent rapid radiation of the viviparous sea snakes (Elapidae: Hydrophiinae). Molecular Phylogenetics and Evolution. 2013;66(3):575–91.
38. Waters PD, Patel HR, Ruiz-Herrera A, Álvarez-González L, Lister NC, Simakov O, et al. Microchromosomes are building blocks of bird, reptile, and mammal chromosomes. Proc Natl Acad Sci USA. 2021;118(45):e2112494118.

39. Oguiura N, Ferrarezzi H, Batistic RF. Cytogenetics and Molecular Data in Snakes: A Phylogenetic Approach. Cytogenet Genome Res. 2009;127(2–4):128–42.

40. Singh L. Evolution of karyotypes in snakes. Chromosoma. 1972;38(2):185–236.

41. Olmo E. Rate of Chromosome changes and Speciation in Reptiles. Genetica. 2005;125(2–3):185–203.

42. Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, Hohenlohe PA, et al. Genomics and the origin of species. Nat Rev Genet. 2014;15(3):176–92.

43. Zhang L, Reifová R, Halenková Z, Gompert Z. How Important Are Structural Variants for Speciation? Genes. 2021;12(7):1084.

44. Pasquesi GIM, Adams RH, Card DC, Schield DR, Corbin AB, Perry BW, et al. Squamate reptiles challenge paradigms of genomic repeat element evolution set by birds and mammals. Nat Commun. 2018;9(1):2774.

45. Ahmad SF, Singchat W, Panthum T, Srikulnath K. Impact of Repetitive DNA Elements on Snake Genome Biology and Evolution. Cells. 2021;10(7):1707.

46. Galbraith JD, Ludington AJ, Sanders KL, Amos TG, Thomson VA, Enosi Tuipulotu D, et al. Horizontal Transposon Transfer and Its Implications for the Ancestral Ecology of Hydrophiine Snakes. Genes. 2022;13(2):217.

47. Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, et al. Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. Penny D, editor. PLoS Biol. 2011;9(3):e1000602.

48. Sanders KL, Rasmussen AR, Mumpuni, Elmberg J, Silva A de, Guinea ML, et al. Recent rapid speciation and ecomorph divergence in Indo-Australian sea snakes. Molecular Ecology. 2013;22(10):2742–59.

49. Ludington AJ, Sanders KL. Demographic analyses of marine and terrestrial snakes (Elapidae) using whole genome sequences. Molecular Ecology. 2021;30(2):545–54.

50. Rutschow D, Bauer R, Göhringer C, Bekeredjian R, Schinkel S, Straub V, et al. S151A δ-sarcoglycan mutation causes a mild phenotype of cardiomyopathy in mice. Eur J Hum Genet. 2014;22(1):119–25.

51. Schoenebeck JJ, Keegan BR, Yelon D. Vessel and Blood Specification Override Cardiac Potential in Anterior Mesoderm. Developmental Cell. 2007;13(2):254–67.

52. Snyder EM, Johnson BD, Joyner MJ. Genetics of A2-Adrenergic Receptors and the Cardiopulmonary Response to Exercise. Exerc Sport Sci Rev. 2008;36(2):98-105.

53. Heatwole H, Seymour RS, Webster MED. Heart rates of sea snakes diving in the sea. Comparative Biochemistry and Physiology Part A: Physiology. 1979;62(2):453–5.

54. Heatwole H, Seymour RS. Cutaneous oxygen uptake in three groups of aquatic snakes. Aust J Zool. 1978;26(3):481.

55. Edqvist PHD, Fagerberg L, Hallström BM, Danielsson A, Edlund K, Uhlén M, et al. Expression of Human Skin-Specific Genes Defined by Transcriptomics and Antibody-Based Profiling. J Histochem Cytochem. 2015;63(2):129–41.

56. Feder ME, Burggren WW. Cutaneous Gas Exchange in Vertebrates: Design, Patterns, Control and Implications. Biological Reviews. 1985;60(1):1–45.

57. Lillywhite HB, Menon GK. Structure and function of skin in the pelagic sea snake, Hydrophis platurus. Journal of Morphology. 2019;280(4):544–54.

58. Toni M, Alibardi L. Alpha- and beta-keratins of the snake epidermis. Zoology. 2007;110(1):41–7.

59. Ehrlich F, Laggner M, Langbein L, Burger P, Pollreisz A, Tschachler E, et al. Comparative genomics suggests loss of keratin K24 in three evolutionary lineages of mammals. Sci Rep. 2019;9(1):10924.
60. Greenwold MJ, Bao W, Jarvis ED, Hu H, Li C, Gilbert MTP, et al. Dynamic evolution of the alpha (α) and beta (β) keratins has accompanied integument diversification and the adaptation of birds into novel lifestyles. BMC Evol Biol. 2014;14(1):249.
61. Duffy SP, Shing J, Saraon P, Berger LC, Eiden MV, Wilde A, et al. The Fowler Syndrome-Associated Protein FLVCR2 Is an Importer of Heme. Molecular and Cellular Biology. 2010;30(22):5318–24.
62. Goodfellow BJ, Freire F, Carvalho AL, Aveiro SS, Charbonnier P, Moulis JM, et al. The SOUL family of heme-binding proteins: Structure and function 15 years later. Coordination Chemistry Reviews. 2021;448:214189.
63. Firth JD, Ebert BL, Ratcliffe PJ. Hypoxic Regulation of Lactate Dehydrogenase A. Journal of Biological Chemistry. 1995;270(36):21021–7.
64. Semenza GL, Jiang BH, Leung SW, Passantino R, Concordet JP, Maire P, et al. Hypoxia Response Elements in the Aldolase A, Enolase 1, and Lactate Dehydrogenase A Gene Promoters Contain Essential Binding Sites for Hypoxia-inducible Factor 1. Journal of Biological Chemistry. 1996;271(51):32529–37.
65. Davis RW. A review of the multi-level adaptations for maximizing aerobic dive duration in marine mammals: from biochemistry to behavior. J Comp Physiol B. 2014;184(1):23–53.
66. Tian R, Yin D, Liu Y, Seim I, Xu S, Yang G. Adaptive Evolution of Energy Metabolism-Related Genes in Hypoxia-Tolerant Mammals. Front Genet. 2017;8:205.

67. Felmlee MA, Jones RS, Rodriguez-Cruz V, Follman KE, Morris ME. Monocarboxylate Transporters (SLC16): Function, Regulation, and Role in Health and Disease. Daws LC, editor. Pharmacol Rev. 2020;72(2):466–85.
68. Li JT, Gao YD, Xie L, Deng C, Shi P, Guan ML, et al. Comparative genomic investigation of high-elevation adaptation in ectothermic snakes. Proc Natl Acad Sci USA. 2018;115(33):8406–11.
69. Babonis LS, Brischoux F. Perspectives on the Convergent Evolution of Tetrapod Salt Glands. Integrative and Comparative Biology. 2012;52(2):245–56.
70. Li Y, Yang J, Li S, Zhang J, Zheng J, Hou W, et al. N-myc Downstream-regulated Gene 2, a Novel Estrogen-targeted Gene, Is Involved in the Regulation of Na+/K+-ATPase. Journal of Biological Chemistry. 2011;286(37):32289–99.
71. Lu B, Jiang J, Wu H, Chen X, Song X, Liao W, et al. A large genome with chromosome-scale assembly sheds light on the evolutionary success of a true toad (Bufo gargarizans). Mol Ecol Resour. 2021;21(4):1256–73.
72. Shukla H, Suryamohan K, Khan A, Mohan K, Perumal RC, Mathew OK, et al. Near-chromosomal de novo assembly of Bengal tiger genome reveals genetic hallmarks of apex predation. GigaScience. 2022;12:giac112.
73. Devuyst O, Luciani A. Chloride transporters and receptor-mediated endocytosis in the renal proximal tubule: Renal Cl^- transporters and receptor-mediated endocytosis. J Physiol. 2015;593(18):4151–64.
74. Dunson W. Control mechanisms in reptiles. In: Mechanisms of osmoregulation in animals. R. Gilles (ed.). In New York: Wiley Interscience; 1979. p. 273–322.

75. Shiels A, Bassnett S, Varadaraj K, Mathias R, Al-Ghoul K, Kuszak J, et al. Optical dysfunction of the crystalline lens in aquaporin-0-deficient mice. Physiological Genomics. 2001;7(2):179–86.

76. Kröger RHH, Fritsches KA, Warrant EJ. Lens optical properties in the eyes of large marine predatory teleosts. J Comp Physiol A. 2009;195(2):175–82.

77. Dizhoor AM, Olshevskaya EV, Peshenko IV. Retinal degeneration-3 protein promotes photoreceptor survival by suppressing activation of guanylyl cyclase rather than accelerating GMP recycling. Journal of Biological Chemistry. 2021;296:100362.

78. Grayson C. Localization in the human retina of the X-linked retinitis pigmentosa protein RP2, its homologue cofactor C and the RP2 interacting protein Arl3. Human Molecular Genetics. 2002;11(24):3065–74.

79. Jane Evans R, Hardcastle AJ, Cheetham ME. Focus on Molecules: X-linked Retinitis Pigmentosa 2 protein, RP2. Experimental Eye Research. 2006;82(4):543–4.

80. Tanaka T, Tsujimura T, Takeda K, Sugihara A, Maekawa A, Terada N, et al. Targeted disruption of ATF4 discloses its essential role in the formation of eye lens fibres. Genes to Cells. 1998;3(12):801–10.

81. Yokokura S, Wada Y, Nakai S, Sato H, Yao R, Yamanaka H, et al. Targeted Disruption of FSCN2 Gene Induces Retinopathy in Mice. Invest Ophthalmol Vis Sci. 2005;46(8):2905.

82. Wu BX, Moiseyev G, Chen Y, Rohrer B, Crouch RK, Ma J xing. Identification of RDH10, an All- trans Retinol Dehydrogenase, in Retinal Müller Cells. Invest Ophthalmol Vis Sci. 2004;45(11):3857.

83. Hahn C, Genner MJ, Turner GF, Joyce DA. The genomic basis of cichlid fish adaptation within the deepwater “twilight zone” of Lake Malawi. Evolution Letters. 2017;1(4):184–98.

84. Bermingham NA, Hassan BA, Wang VY, Fernandez M, Banfi S, Bellen HJ, et al. Proprioceptor Pathway Development Is Dependent on MATH1. Neuron. 2001;30(2):411–22.

85. Avenarius MR, Jung JY, Askew C, Jones SM, Hunker KL, Azaiez H, et al. Grxcr2 is required for stereocilia morphogenesis in the cochlea. Sokolowski B, editor. PLoS ONE. 2018;13(8):e0201713.

86. Zhou X, Sun F, Xu S, Fan G, Zhu K, Liu X, et al. Baiji genomes reveal low genetic variability and new insights into secondary aquatic adaptations. Nat Commun. 2013;4(1):2708.

87. Booth KT, Azaiez H, Kahrizi K, Simpson AC, Tollefson WTA, Sloan CM, et al. PDZD7 and hearing loss: More than just a modifier. Am J Med Genet. 2015;167(12):2957–65.

88. Goodyear RJ, Richardson GP. Extracellular matrices associated with the apical surfaces of sensory epithelia in the inner ear: Molecular and structural diversity. J Neurobiol. 2002;53(2):212–27.

89. Macrì S, Savriama Y, Khan I, Di-Poï N. Comparative analysis of squamate brains unveils multi-level variation in cerebellar architecture associated with locomotor specialization. Nat Commun. 2019;10(1):5560.

90. Palci A, Hutchinson MN, Caldwell MW, Lee MSY. The morphology of the inner ear of squamate reptiles and its bearing on the origin of snakes. R Soc open sci. 2017;4(8):170685.

91. Marino C, Mascheretti S, Riva V, Cattaneo F, Rigoletto C, Rusconi M, et al. Pleiotropic Effects of DCDC2 and DYX1C1 Genes on Language and Mathematics Traits in Nuclear Families of Developmental Dyslexia. Behav Genet. 2011;41(1):67–76.

92. Truong DT, Che A, Rendall AR, Szalkowski CE, LoTurco JJ, Galaburda AM, et al. Mutation of Dcdc2 in mice leads to impairments in auditory processing and memory ability. Genes, Brain and Behavior. 2014;13(8):802–11.

93. Gekakis N, Saez L, Delahaye-Brown AM, Myers MP, Sehgal A, Young MW, et al. Isolation of timeless by PER Protein Interaction: Defective Interaction Between timeless Protein and Long-Period Mutant PER^{L}. Science. 1995;270(5237):811–5.

94. Sehgal A, Price JL, Man B, Young MW. Loss of Circadian Behavioral Rhythms and per RNA Oscillations in the Drosophila Mutant timeless. Science. 1994;263(5153):1603–6.

95. Voris HK, Jayne BC. The Costocutaneous Muscles in Some Sea Snakes (Reptilia, Serpentes). Journal of Herpetology. 1976;10(3):175–80.

96. Sherratt E, Nash-Hahn T, Nankivell JH, Rasmussen AR, Hampton PM, Sanders KL. Macroevolution in axial morphospace: innovations accompanying the transition to marine environments in elapid snakes. R Soc open sci. 2022;9(12):221087.

97. Sherratt E, Sanders KL. Patterns of intracolumnar size variation inform the heterochronic mechanisms underlying extreme body shape divergence in microcephalic sea snakes. Evolution & Development. 2020;22(3):283–90.

98. Kulkarni P, Mohanty A, Salgia R, Uversky VN. Intrinsically disordered BMP4 morphogen and the beak of the finch: Co-option of an ancient axial patterning system. International Journal of Biological Macromolecules. 2022;219:366–73.

99. Camacho J, Lin JD, McCormack M, Moon R, Smith SK, Rasweiler JJ, et al. BMP signaling underlies the craniofacial heterochrony in phyllostomid bats, a hyperdiverse mammal group. bioRxiv. 2021. http://biorxiv.org/lookup/doi/10.1101/2021.05.17.444516.

100. Andl T, Ahn K, Kairo A, Chu EY, Wine-Lee L, Reddy ST, et al. Epithelial Bmpr1a regulates differentiation and proliferation in postnatal hair follicles and is essential for tooth development. Development. 2004;131(10):2257–68.

101. Baek JA, Lan Y, Liu H, Maltby KM, Mishina Y, Jiang R. Bmpr1a signaling plays critical roles in palatal shelf growth and palatal bone formation. Developmental Biology. 2011;350(2):520–31.

102. Maruyama T, Stevens R, Boka A, DiRienzo L, Chang C, Yu HMI, et al. BMPR1A maintains skeletal stem cell properties in craniofacial development and craniosynostosis. Sci Transl Med. 2021;13(583):eabb4416.

103. Boulet AM, Capecchi MR. Multiple roles of Hoxa11 and Hoxd11 in the formation of the mammalian forelimb zeugopod. Development. 2004;131(2):299–309.

104. Huelsken J, Birchmeier W. New aspects of Wnt signaling pathways in higher vertebrates. Current Opinion in Genetics & Development. 2001;11(5):547–53.

105. Mallo M, Wellik DM, Deschamps J. Hox genes and regional patterning of the vertebrate body plan. Developmental Biology. 2010;344(1):7–15.

106. Pearson JC, Lemons D, McGinnis W. Modulating Hox gene functions during animal body patterning. Nat Rev Genet. 2005;6(12):893–904.

107. Veeman MT, Axelrod JD, Moon RT. A second canon. Functions and mechanisms of beta-catenin-independent Wnt signaling. Dev Cell. 2003;5(3):367–77.

108. Wellik DM. Hox patterning of the vertebrate axial skeleton. Dev Dyn. 2007;236(9):2454–63.

109. Di Palma F., Alfoldi J., Heiman D, Young S., Grabherr M., Johnson J., et al. An arboreal lizard native to southeastern United States and some Caribbean islands. GenBank https://www.ncbi.nlm.nih.gov/bioproject/PRJNA18787/ (2010)

110. Margres MJ. Crotalus tigris isolate:CLP2741 Genome sequencing and assembly. GenBank https://www.ncbi.nlm.nih.gov/bioproject/PRJNA558767 (2019)

111. Suryamohan K, Krishnankutty SP, Guillory J, Jevit M, Schroder MS, Wu M, et al. Naja naja Genome sequencing and assembly. GenBank https://www.ncbi.nlm.nih.gov/bioproject/PRJNA527614/ (2019)

112. Suryamohan K, Krishnankutty SP, Guillory J, Jevit M, Schröder MS, Wu M, et al. The Indian cobra reference genome and transcriptome enables comprehensive identification of venom toxins. Nat Genet. 2020;52(1):106–17.

113. Edwards RE. De novo whole genome sequencing of the mainland tiger snake, Notechis scutatus. GenBank https://www.ncbi.nlm.nih.gov/bioproject/PRJEB27871 (2018)

114. Ullate-Agote A, Burgelin I, Debry A, Langrez C, Montange F, Peraldi R, et al. Pantherophis guttatus Genome sequencing and assembly. GenBank https://www.ncbi.nlm.nih.gov/bioproject/PRJNA268069/ (2019)

115. Aird SD, Arora J, Barua A, Qiu L, Terada K, Mikheyev AS. Protobothrops mucrosquamatus and Protobothrops elegans genome. GenBank https://www.ncbi.nlm.nih.gov/bioproject/304719 (2015)

116. Edwards RE. De novo whole genome sequencing of the eastern brown snake, Pseudonaja textilis. GenBank https://www.ncbi.nlm.nih.gov/bioproject/483162 (2018)

117. Castoe TA, de Koning JA, Hall KT, Yokoyama KD, Gu W, Smith EN, et al. Python bivittatus Genome sequencing and assembly. GenBank https://www.ncbi.nlm.nih.gov/bioproject/61243 (2014)

118. Bronikowski A, Fedrigo O, Fungtammasan C, Rhie A, Mountcastle J, Haase B, et al. Thamnophis elegans (Western terrestrial garter snake) genome, rThaEle1, primary haplotype. GenBank https://www.ncbi.nlm.nih.gov/bioproject/561996 (2019)

119. Li A, Wang J, Sun K, Wang S, Zhao X, Wang T, et al. Hydrophis curtus Genome sequencing and assembly. GenBank https://www.ncbi.nlm.nih.gov/bioproject/PRJNA616080/ (2021)

120. Li A, Wang J, Sun K, Wang S, Zhao X, Wang T, et al. Hydrophis cyanocinctus Genome sequencing and assembly. GenBank https://www.ncbi.nlm.nih.gov/bioproject/PRJNA573877/ (2021)

121. Hydrophis cyanocinctus Transcriptome or Gene expression. SRA https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA608244 (2020)

122. Hydrophis curtus Transcriptome or Gene expression. SRA https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA622900 (2020)

123. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20(1):257.

124. Chen S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. iMeta. 2023;2(2):e107.

125. Sim SB, Corpuz RL, Simmonds TJ, Geib SM. HiFiAdapterFilt, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly. BMC Genomics. 2022;23(1):157.

126. Krueger F, James F, Ewels P, Afyounian E, Weinstein M, Schuster-Boeckler B, et al. FelixKrueger/TrimGalore: v0.6.10. Zenodo https://doi.org/10.5281/zenodo.7598955 (2023)

127. Kokot M, Długosz M, Deorowicz S. KMC 3: counting and manipulating k -mer statistics. Berger B, editor. Bioinformatics. 2017;33(17):2759–61.

128. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. 2020;11(1):1432.

129. Cheng H, Jarvis ED, Fedrigo O, Koepfli KP, Urban L, Gemmell NJ, et al. Haplotype-resolved assembly of diploid genomes without parental data. Nat Biotechnol. 2022;40(9):1332–5.

130. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18(2):170–5.

131. Guan D, McCarthy SA, Ning Z, Wang G, Wang Y, Durbin R. Efficient iterative Hi-C scaffolder based on N-best neighbors. BMC Bioinformatics. 2021;22(1):569.

132. Dudchenko O, Shamim MS, Batra SS, Durand NC, Musial NT, Mostofa R, et al. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian

genomes with chromosome-length scaffolds for under $1000. bioRxiv. 2018. http://biorxiv.org/lookup/doi/10.1101/254797.

133. Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, et al. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Systems. 2016;3(1):95–8.

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

135. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37(5):540–6.

136. Kundu R, Casey J, Sung WK. HyPo: Super Fast & Accurate Polisher for Long Read Genome Assemblies. bioRxiv. 2019. http://biorxiv.org/lookup/doi/10.1101/2019.12.19.882506.

137. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 2018;19(1):460.

138. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science. 2017;356(6333):92–5.

139. Nanoporetech. Medaka. https://github.com/nanoporetech/medaka. Accessed 26 Jul 2022

140. Hu J, Fan J, Sun Z, Liu S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. Berger B, editor. Bioinformatics. 2020;36(7):2253–5.

141. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072–5.

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

143. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Kelley J, editor. Molecular Biology and Evolution. 2021;38(10):4647–54.

144. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.

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

146. Smit A, Hubley R, Green P. RepeatMasker open-4.0. 2013. http://www.repeatmasker.org. Accessed Jun 1.

147. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA. 2015;6(1):11.

148. Ou S, Jiang N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. Plant Physiol. 2018;176(2):1410–22.

149. Palmer JM, Stajich JE. Funannotate. https://github.com/nextgenusfs/funannotate. Accessed 29 Jul 2022.

150. Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. Valencia A, editor. Bioinformatics. 2021;37(12):1639–43.

151. The UniProt Consortium, Bateman A, Martin MJ, Orchard S, Magrane M, Ahmad S, et al. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research. 2023;51(D1):D523–31.

152. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35(11):1026–8.

153. Levy Karin E, Mirdita M, Söding J. MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. Microbiome. 2020;8(1):48.

154. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644–52.

155. Haas BJ. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Research. 2003;31(19):5654–66.

156. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 2008;9(1):R7.

157. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34(5):525–7.

158. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9):1236–40.

159. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Molecular Biology and Evolution. 2021;8.

160. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research. 2019;47(D1):D309–14.

161. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution. 2015;32(1):268–74.

162. Zhang C, Rabiee M, Sayyari E, Mirarab S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics. 2018;19(S6):153.

163. Wen D, Yu Y, Zhu J, Nakhleh L. Inferring Phylogenetic Networks Using PhyloNet. Posada D, editor. Systematic Biology. 2018;67(4):735–40.

164. Steenwyk JL, Buida TJ, Labella AL, Li Y, Shen XX, Rokas A. PhyKIT: a broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data. Schwartz R, editor. Bioinformatics. 2021;37(16):2325–31.

165. Rempel A, Wittler R. SANS serif: alignment-free, whole-genome-based phylogenetic reconstruction. Schwartz R, editor. Bioinformatics. 2021;37(24):4868–70.

166. Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. Synteny and Collinearity in Plant Genomes. Science. 2008;320(5875):486–8.

167. Dainat J, Hereñú D, Davis E, Crouch K, LucileSol, Agostinho N, et al. NBISweden/AGAT: AGAT-v0.9.2. Zenodo. https://doi.org/10.5281/zenodo.6621429. (2022)

168. Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res. 2011;21(3):487–93.

169. Tang H, Krishnakumar V, Li J, Tiany, MichelMoser, Maria, et al. tanghaibao/jcvi: JCVI v1.3.4. Zenodo. https://doi.org/10.5281/zenodo.846919 (2022)

170. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10(2):giab008.

171. Goel M, Sun H, Jiao WB, Schneeberger K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol. 2019;20(1):277.

172. Li H. Minimap2: pairwise alignment for nucleotide sequences. Birol I, editor. Bioinformatics. 2018;34(18):3094–100.

173. Goel M, Schneeberger K. plotsr: visualizing structural similarities and rearrangements between multiple genomes. Robinson P, editor. Bioinformatics. 2022;38(10):2922–6.

174. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20(1):238.

175. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution. 2013;30(4):772–80.

176. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Research. 2006;34:W609–12.

177. Steenwyk JL, Buida TJ, Li Y, Shen XX, Rokas A. ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. Hejnol A, editor. PLoS Biol. 2020;18(12):e3001007.

178. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10(1):421.

179. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. UniProtKB/Swiss-Prot. In: Edwards D, editor. Plant Bioinformatics: Methods and Protocols. Totowa, NJ: Humana Press; 2007. p. 89–112.

180. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution. 2007;24(8):1586–91.

181. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological). 1995;57(1):289–300.

182. Hydrophis genome sequencing and assembly. GenBank https://www.ncbi.nlm.nih.gov/bioproject/PRJNA984433/ (2023)

183. Hydrophis assemblies. GenBank https://www.ncbi.nlm.nih.gov/bioproject/PRJNA780942/ (2021)

184. Ludington AJ, Hammond JillianM. Sea snake selection. Available from: https://github.com/a-lud/sea-snake-selection. Accessed 24 Jul 2023.

185. Ludington AJ, Hammond JM, Breen J, Deveson IW, Sanders KL. New chromosome-scale genomes provide insights into marine adaptations of sea snakes (Hydrophis: Elapidae). figshare https://doi.org/10.25909/23557194.v1 (2023)

Supplementary tables

Supplementary tables are presented in ‘Additional file 2.xlsx’ and are available at https://github.com/a-lud/Thesis

Supplementary figures

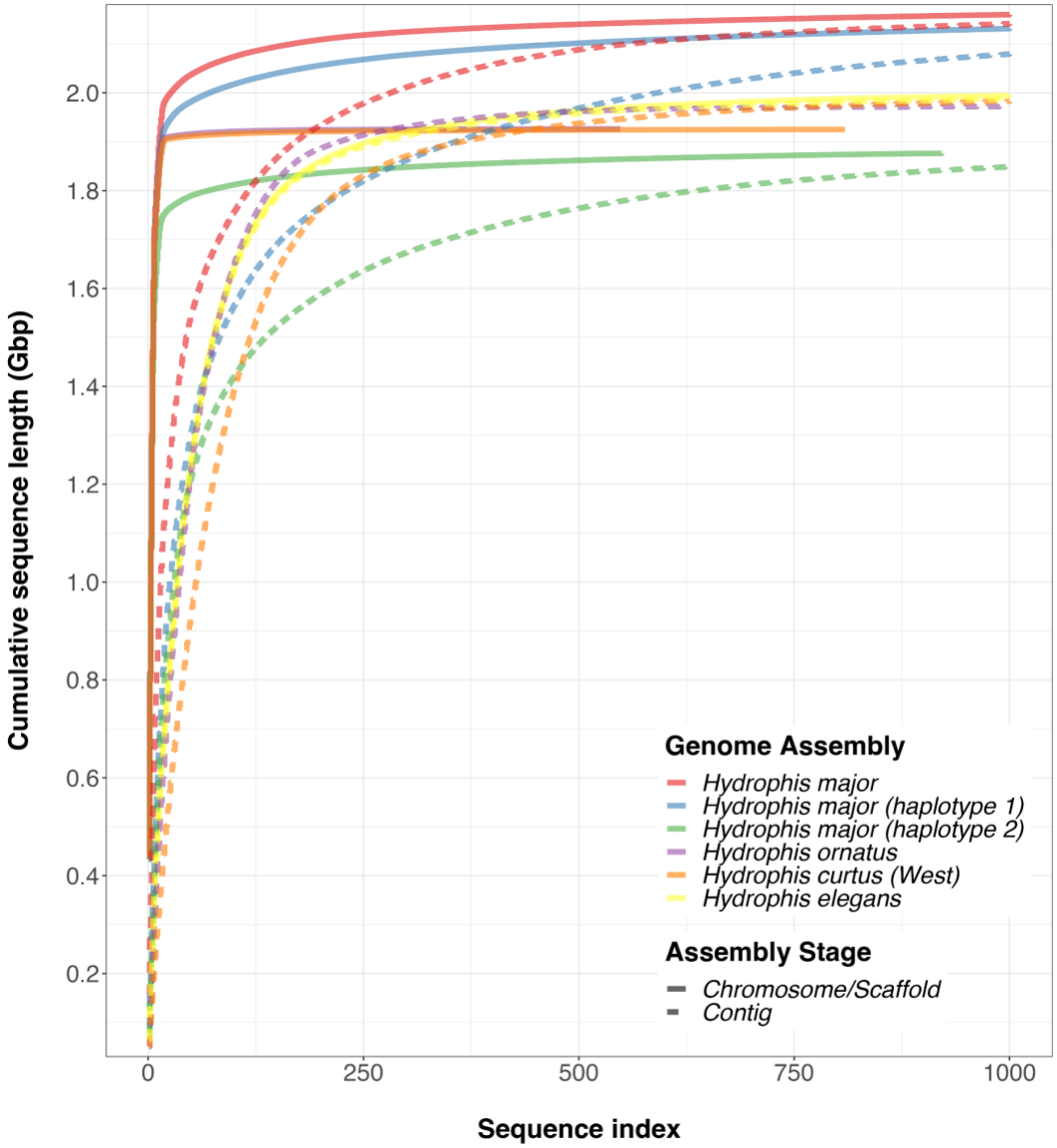


Fig. S1: Cumulative length of all assembled sequences (chromosomes, scaffolds and unplaced scaffolds). Assembled sequences are indexed from longest-to-shortest (x-axis), with the cumulative total length of the assembled sequences plotted along the y-axis. The final assemblies (chromosome/scaffold level) are represented by solid lines, with the initial contig assemblies represented by dashed lines.

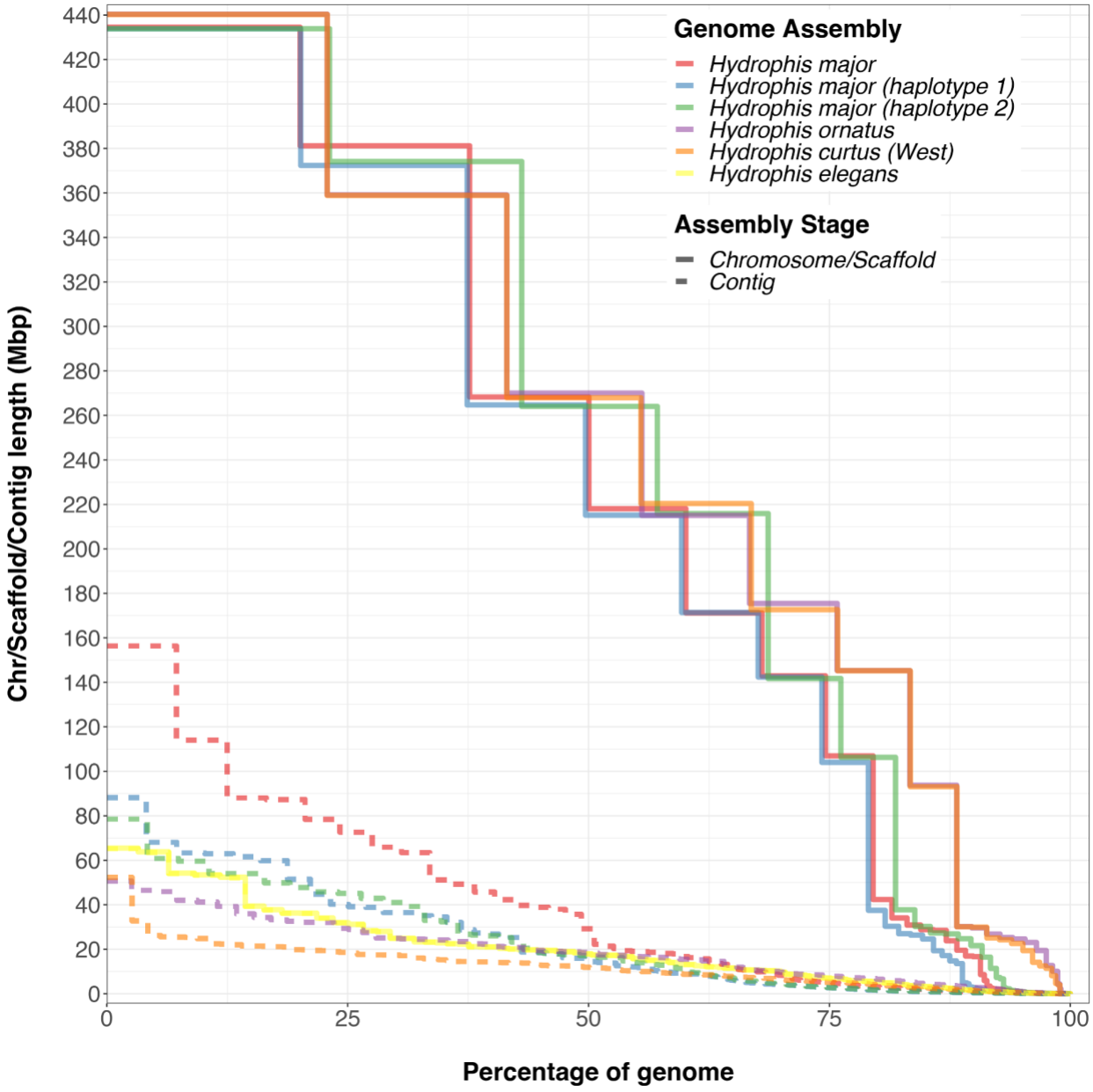


Fig. S2: Nx plot of assembled genomes representing overall contiguity. The figure displays the percentage of the genome (x-axis) that is composed of sequences of a minimum length or longer (y-axis). Initial contig assemblies are represented by the dotted lines, with the solid lines representing the final scaffold/chromosome scale assemblies.

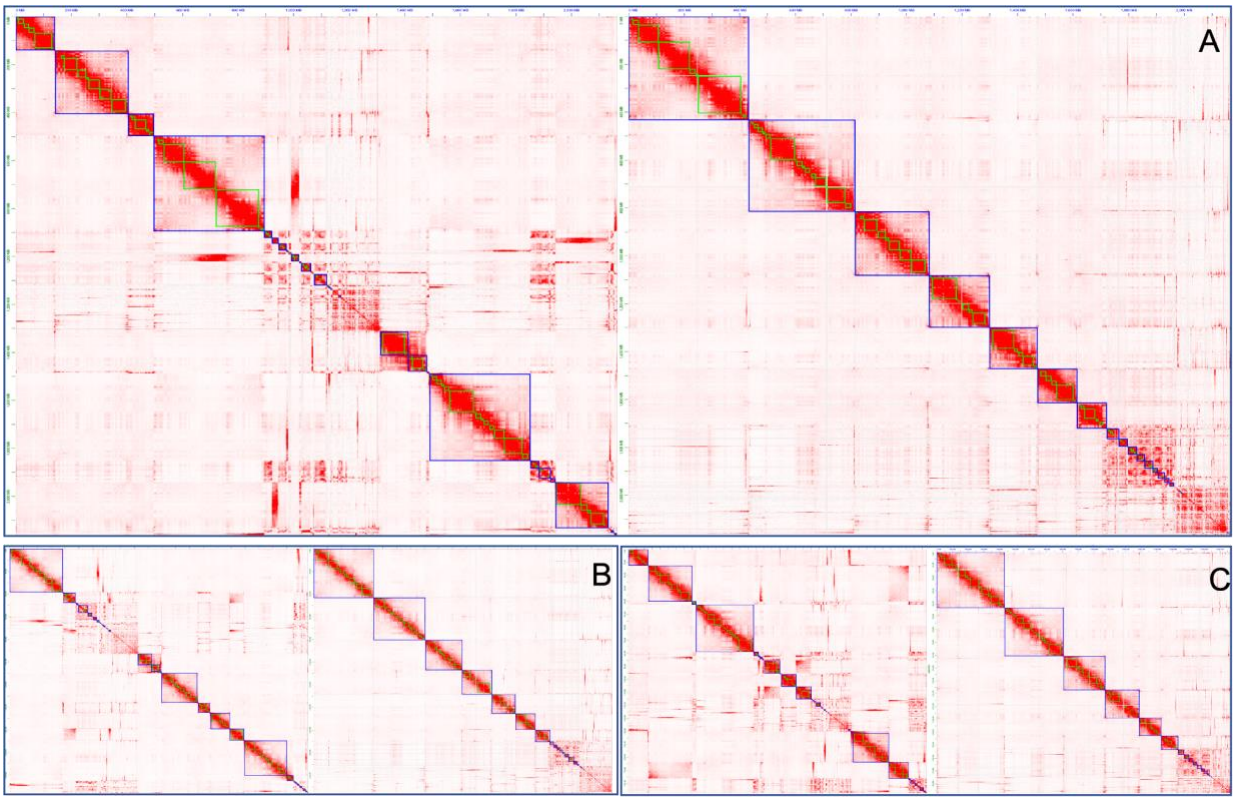


Fig. S3: Hi-C contact maps before and after manual curation for the *Hydrophis* major primary and haplotype assemblies. A) The pin_hic contact map for the primary contig assembly of *H. major*. The contact map on the left represents the raw contact before manual orientation via JBAT. The figure on the right represents the manually ordered contigs forming the chromosome scaffolds. Figures B) and C) represent the same process for haplotype-1 and haplotype-2 of *H. major*, respectively.

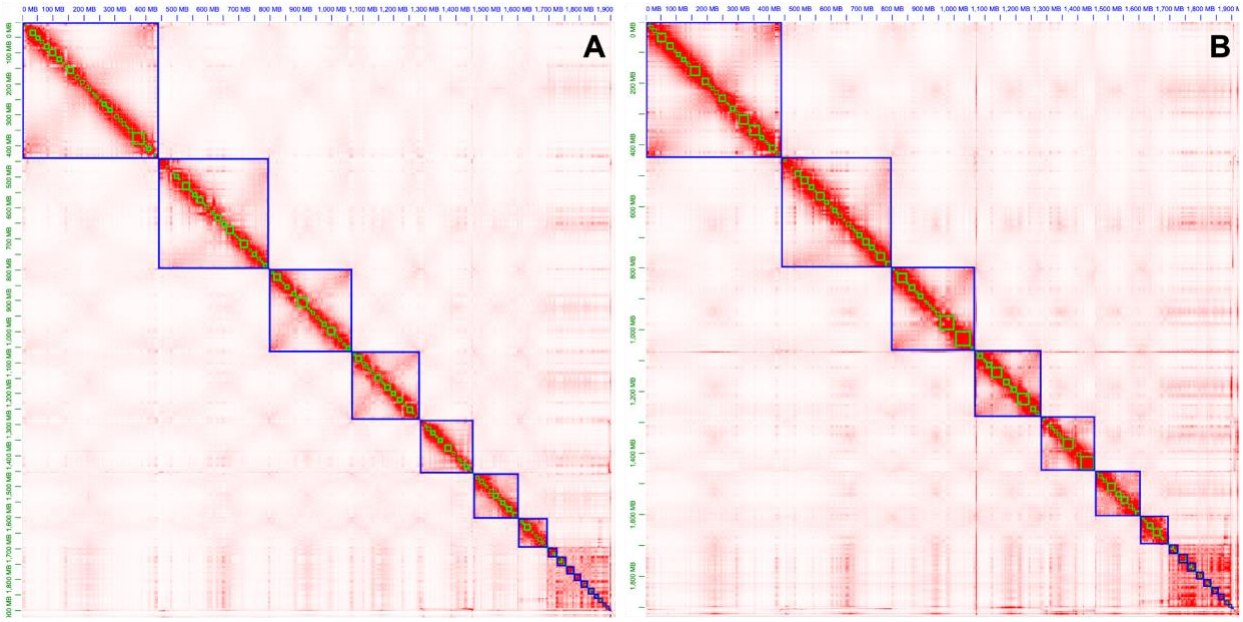


Fig. S4: Hi-C contact maps for (A) *Hydrophis curtus (West)* and (B) *Hydrophis ornatus*. Assembly files from 3d-DNA were edited in *JBAT* to form the final chromosome sequences for each snake.

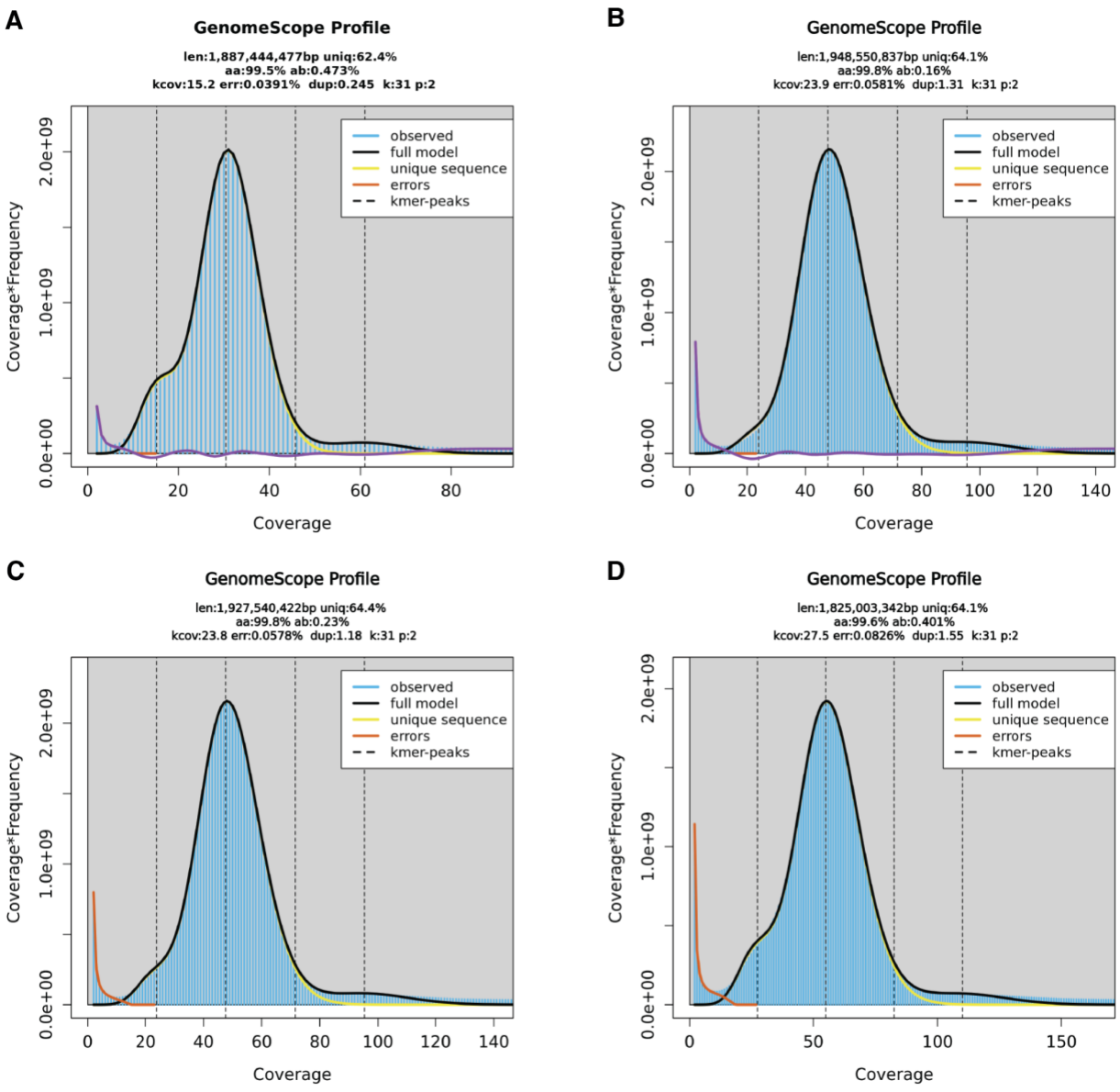


Fig. S5: Genome size estimation. *GenomeScope2* profiles for the four snakes assembled in this study. The main peak in each figure represents the homozygous portion of each genome, with the center of the peak corresponding to the approximate coverage of the homozygous portion of the genome. The shoulder to the left of the main peak corresponds to the heterozygous portion of the genome. A) Genome profile for *Hydrophis major* B) *H. ornatus* C) *H. curtus (West)* and D) *H. elegans.*

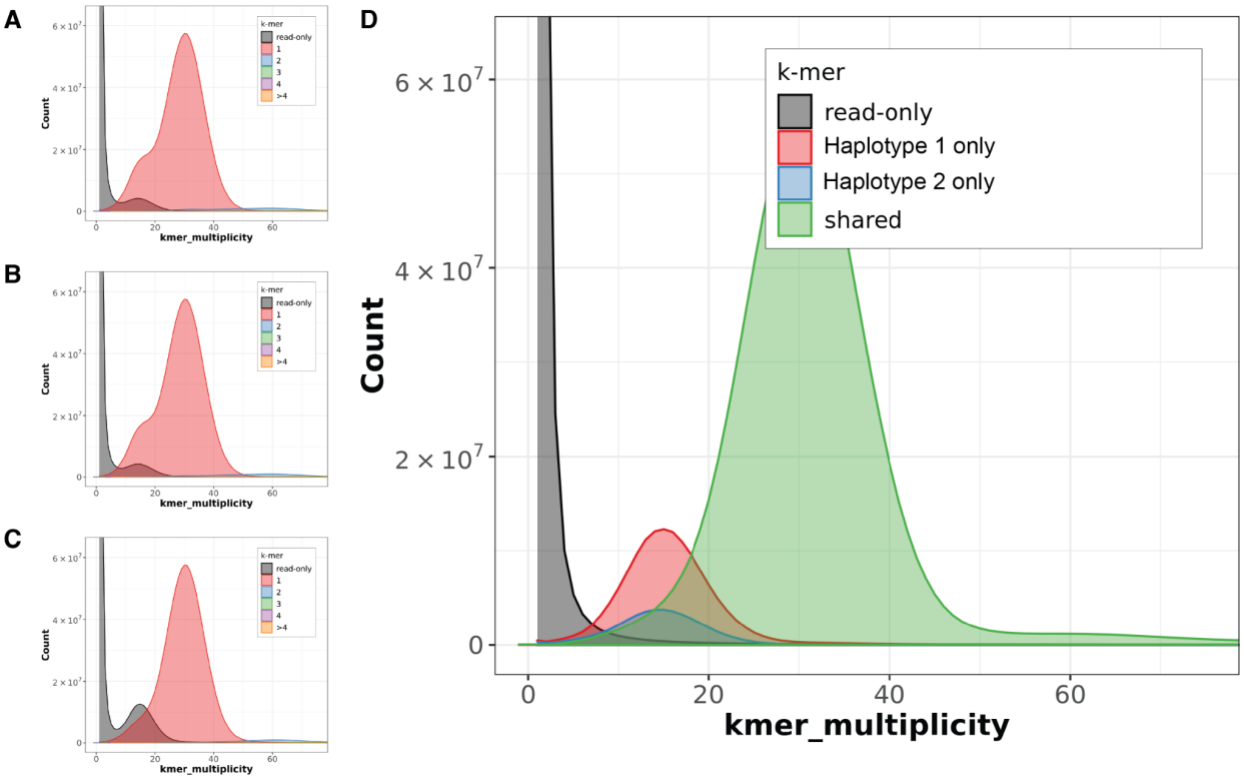


Fig. S6: Merqury k-mer spectra analysis for *H. major*. A-C) Per-assembly spectra plots showing the k-mer multiplicity in the primary assembly, haplotype-1 and haplotype-2, respectively. K-mers appearing in the genome once are in red, twice are in blue, three times in green and four or more times in purple and orange respectively. D) Assembly spectra plot shows the distinct k-mers shared between the assembled haplotypes and the reads (green), the distinct k-mers shared between the reads and each respective haplotype (red and blue), along with the distinct k-mers only found in the reads (black).

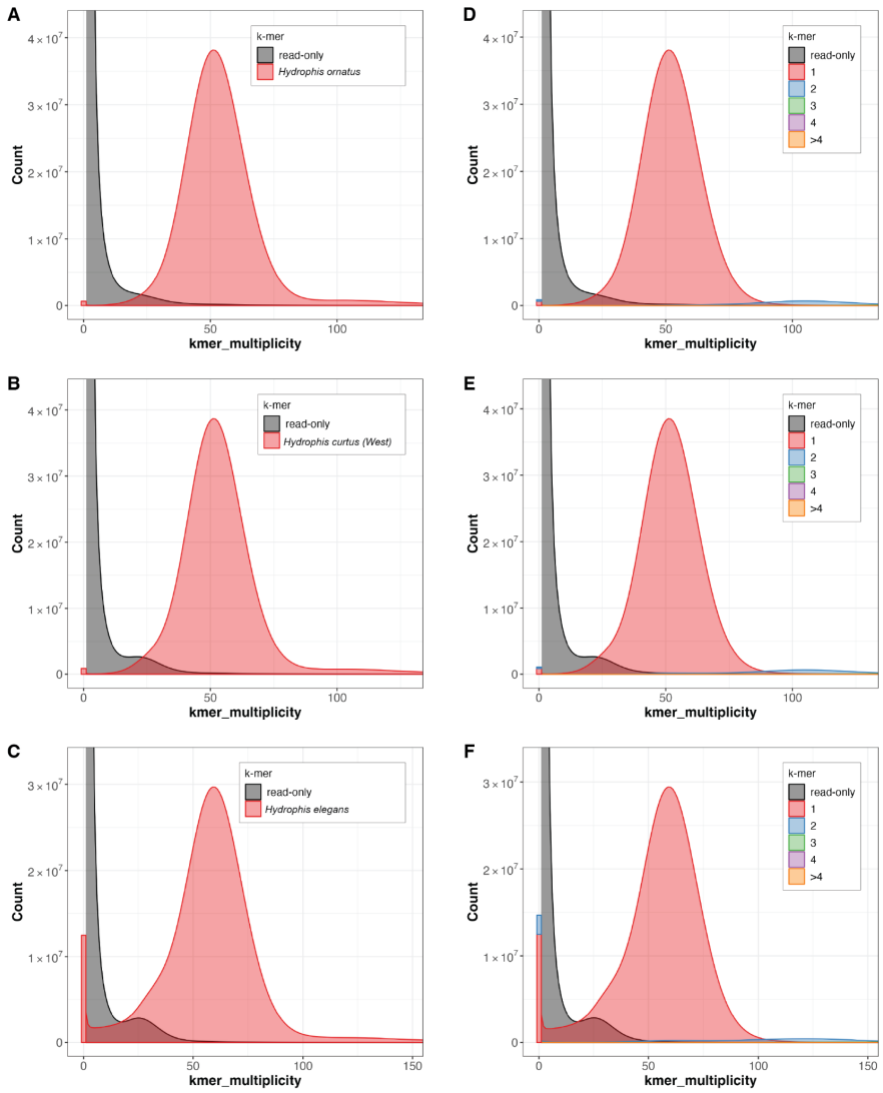


Fig. S7: Merqury k-mer spectra analysis for Nanopore assembled snakes. A-C) Assembly spectra plots showing the distinct k-mers shared between each assembly (*H. ornatus, H. curtus (West)* and *H. elegans* respectively*)* and their reads (red). K-mers only found in the read-set are shown in black. D-F) Spectra plots showing the k-mer multiplicity for the distinct k-mers in each respective genome. K-mers appearing once in each genome are coloured red, twice in blue, three times are green and four or more times in purple and orange respectively. K-mers that are not shared between the reads and assembly are shown in black.

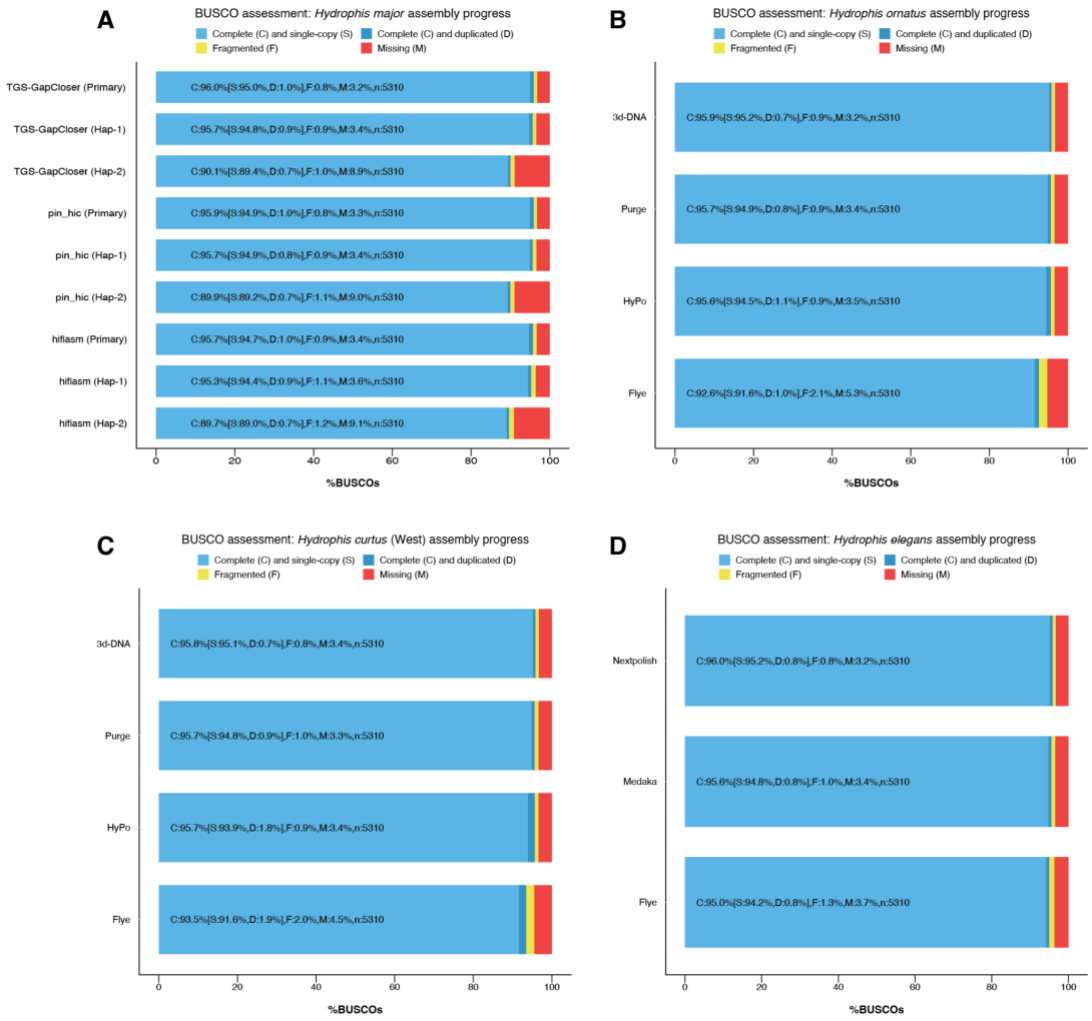


Fig. S8: BUSCO completeness during assembly. Assembly stage is presented along the y-axis, and percentage of BUSCO's is along the x-axis. A) BUSCO was run on the primary and dual assemblies of *Hydrophis major* at each stage of the assembly pipeline from the initial contigs (*hifiasm*) to the final gap-closed assembly (*TGS-GapCloser*). B-C) BUSCO scores for each assembly stage of the *H. ornatus* and *H. curtus (West)* assemblies. *Flye* represents the initial contig assemblies, with *3d-DNA* representing the final chromosomes. D) The assembly progression for *H. elegans*, in which Nextploish represents the final assembly, with Flye and Medaka representing the initial assembly and Nanopore-polished stages respectively.

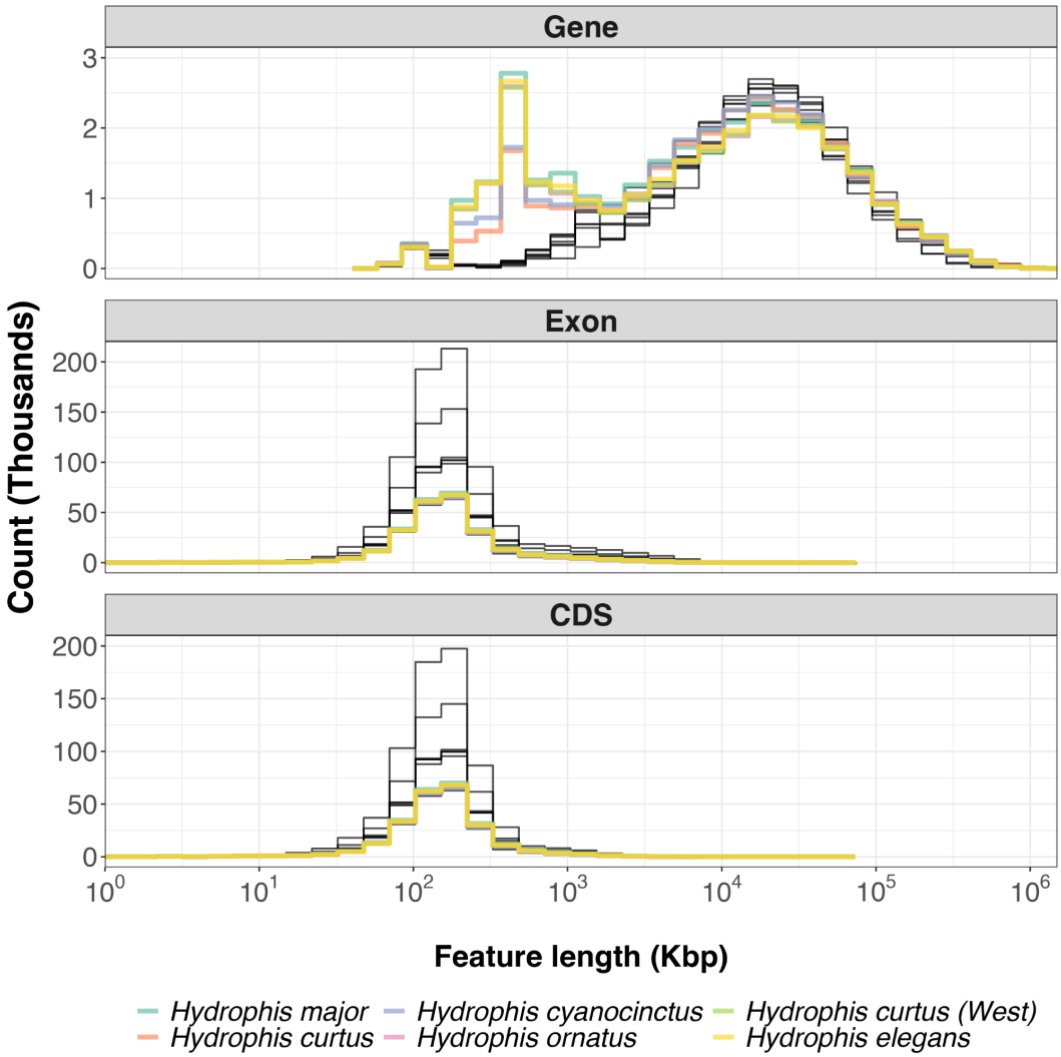


Fig. S9: Length distribution of gene features. Gene, exon and coding-sequence (CDS) lengths (x-axis) are plotted against their total count (y-axis) for each of the six newly assembled *Hydrophis* snakes (coloured lines). The black lines represent gene annotations for RefSeq annotated snakes including *Notechis scutatus*, *Pseudonaja textilis*, *Thamnophis elegans*, *Pantherophis guttatus*, *Protobothrops mucrosquamatus*, *Crotalus tigris* and *Python bivittatus*. Each of the features show similar length profiles, however there is an excess of shorter gene models only found in the *de novo* gene annotations that aren't present in the RefSeq annotated snakes.

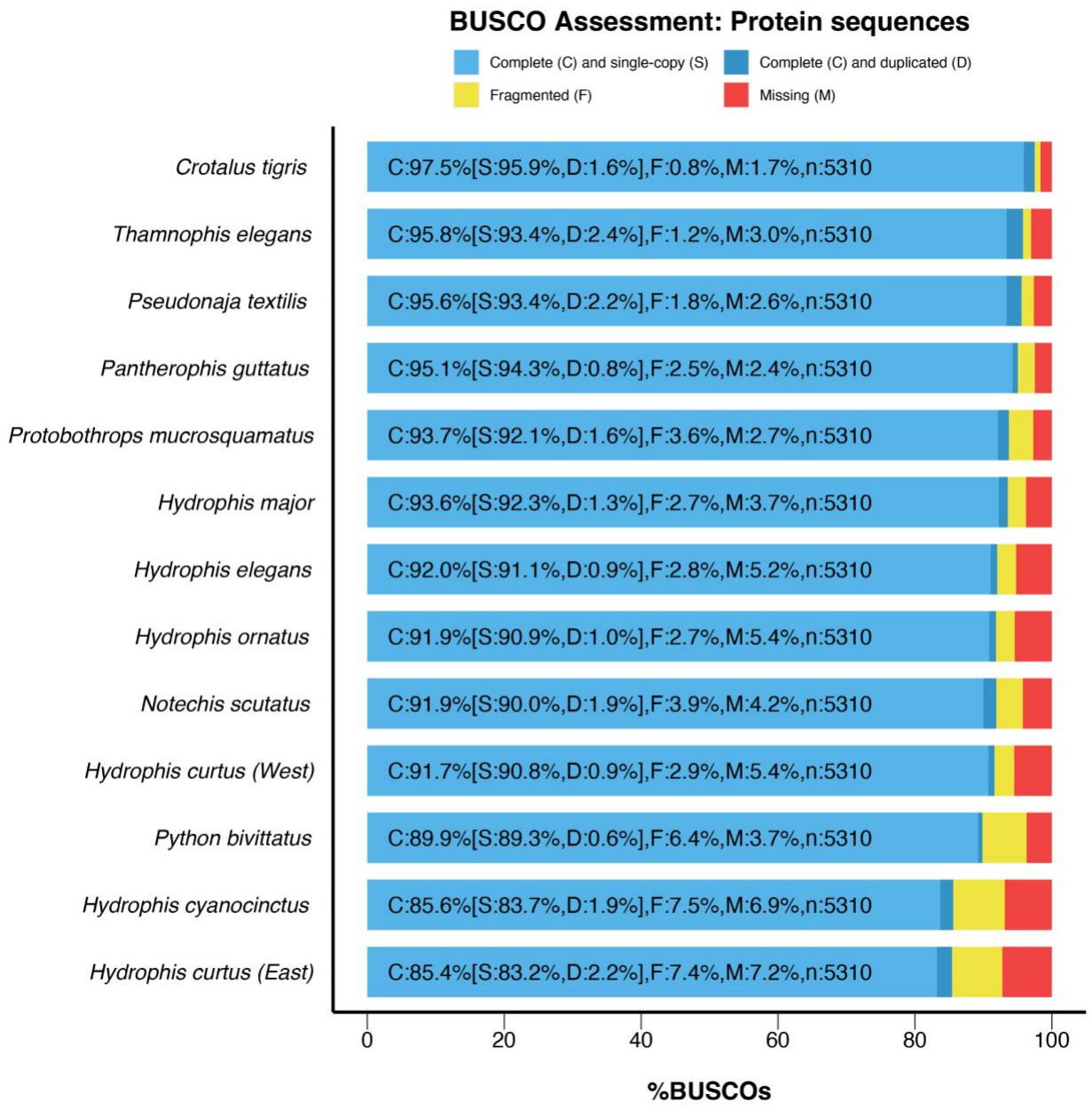


Fig. S10: BUSCO completeness of longest-isoform protein sequences. The longest isoform of each gene was extracted using AGAT, before running BUSCO on the resulting peptide files.

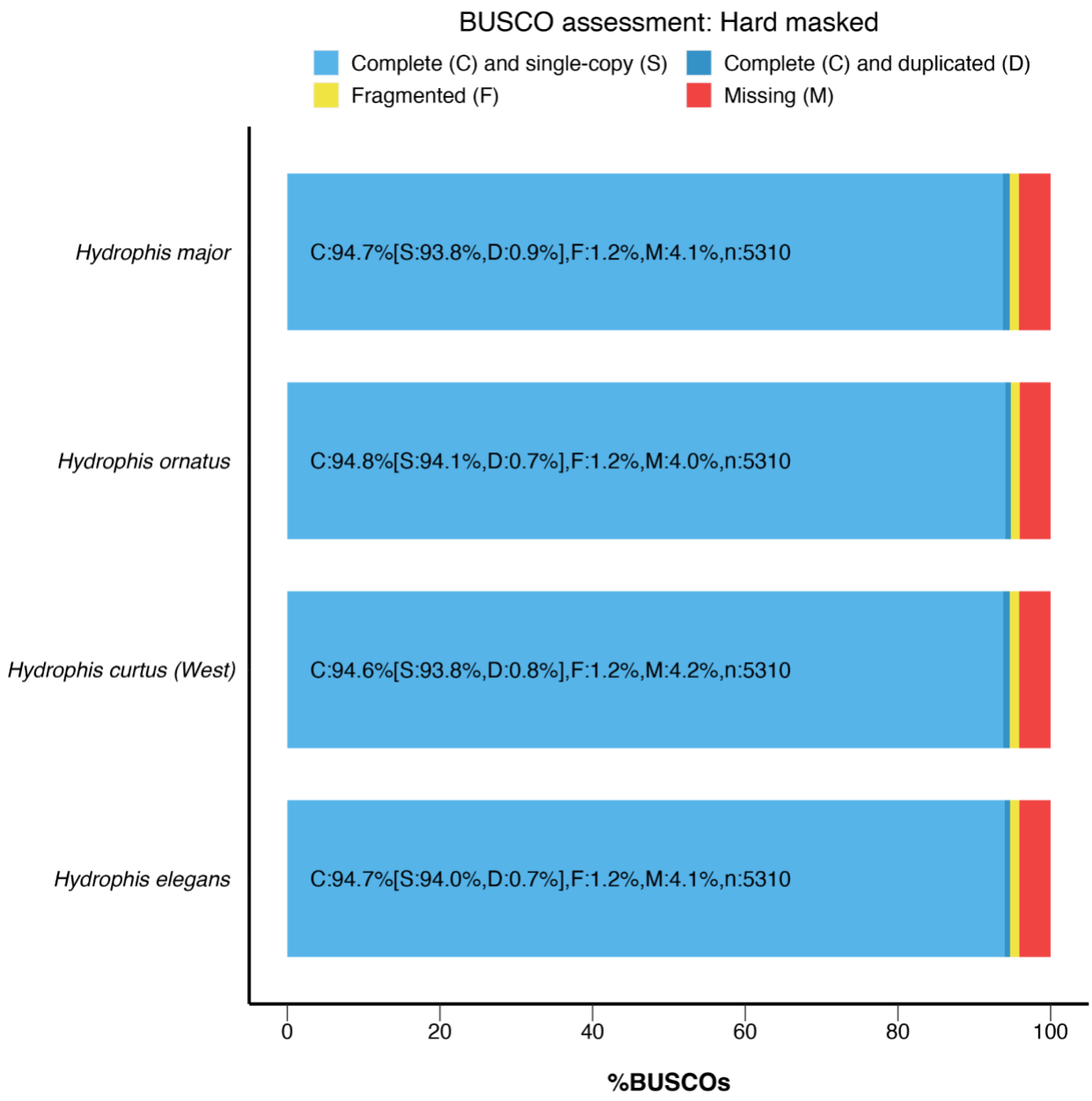


Fig. S11: *BUSCO* completeness post hard-masking. Each genome had annotated repeat elements hard-masked (replaced with N's), before running *BUSCO* using the tetrapoda_odb10 database. Hard-masking the genomes has little effect on the gene content, dropping the complete *BUSCO* measure by an average of 1.3% across all four snakes, indicating the annotated repeats constitute repetitive elements rather than non-repetitive genetic sequence.

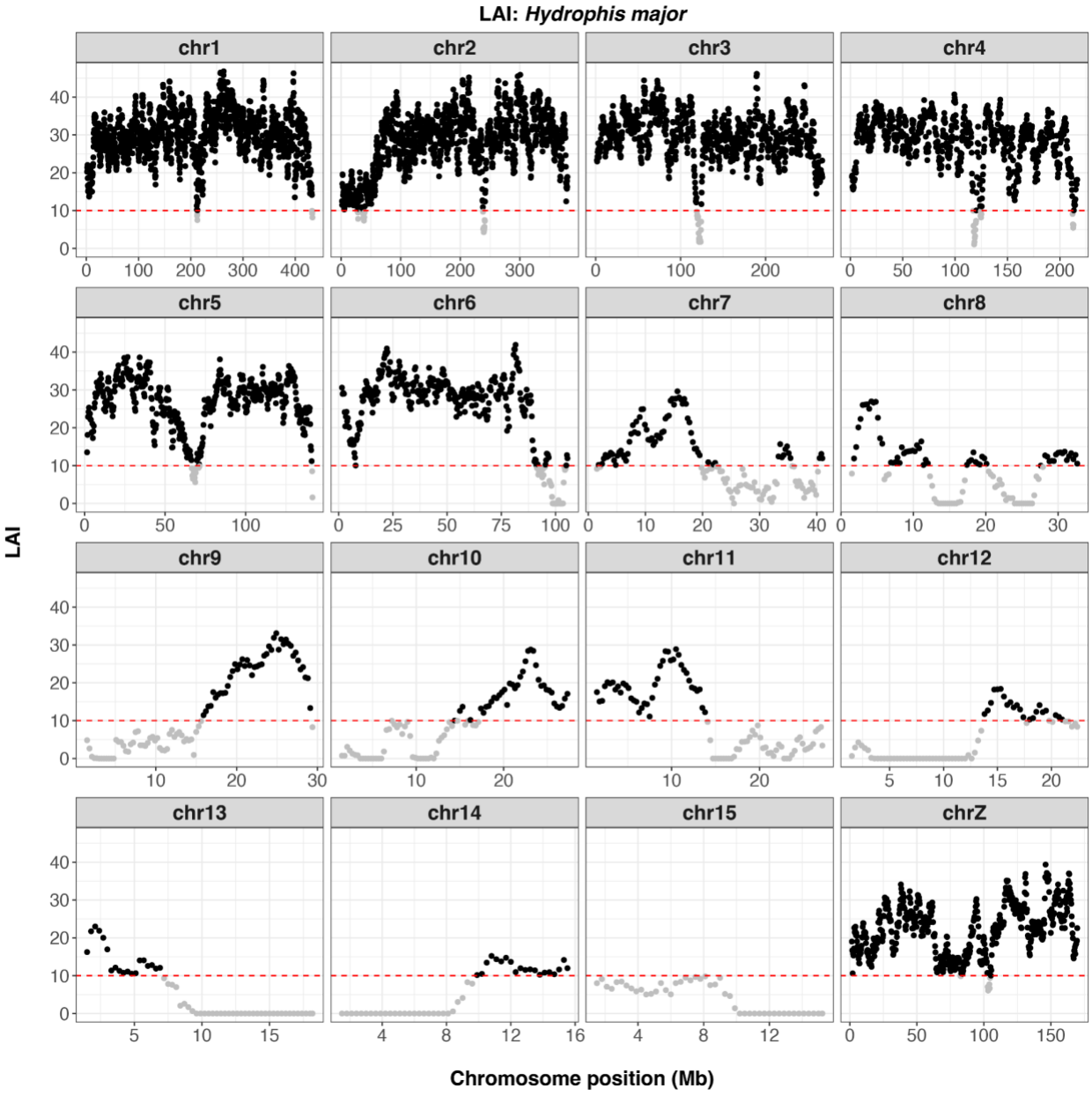


Fig. S12: LTR Assembly Index (LAI) scores for *Hydrophis major* chromosomal sequences. LAI scores were calculated in 3Mbp windows with a 300Kbp step-size. The red dashed line at y=10 indicates the 'reference quality' cut-off.

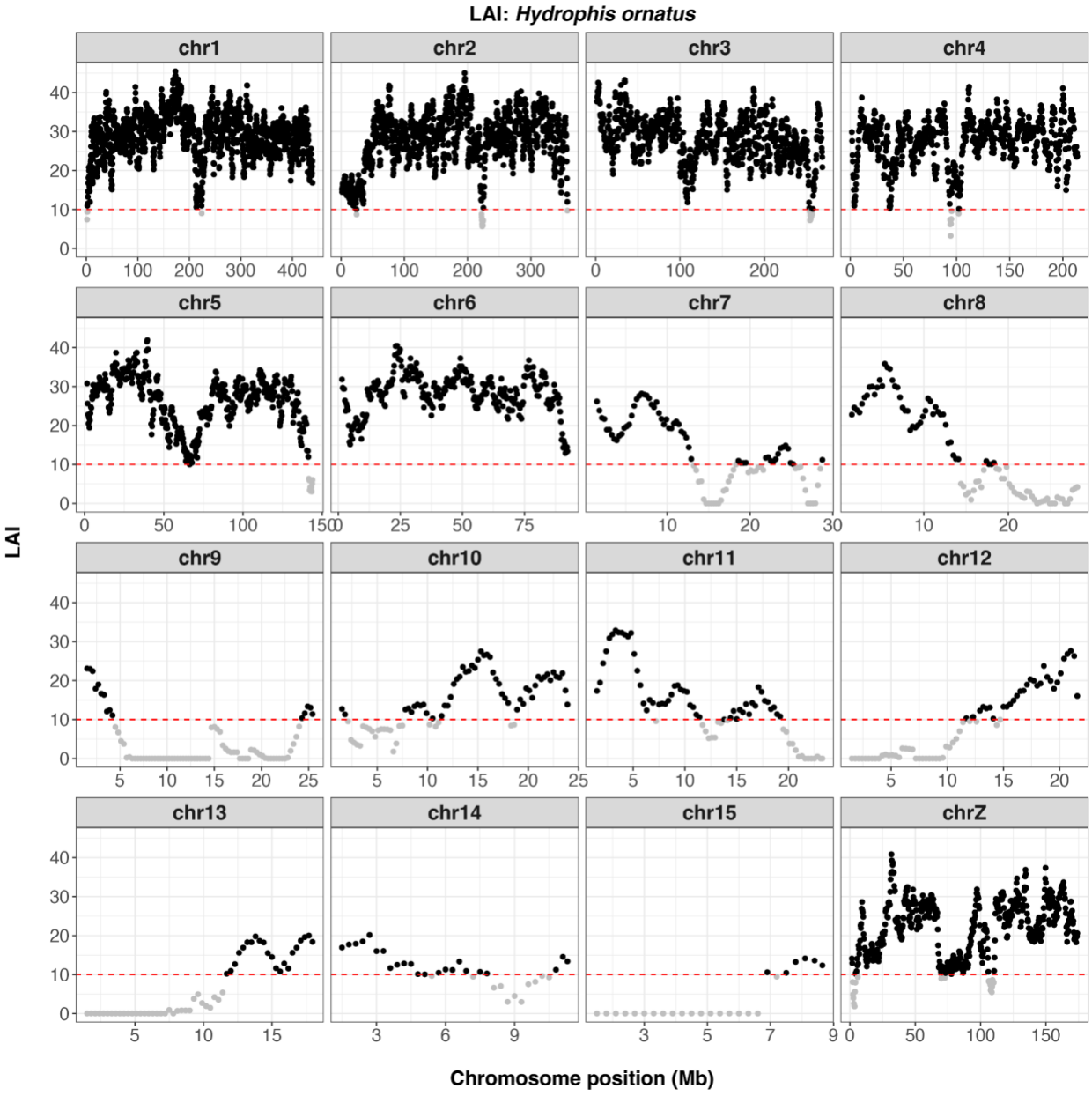


Fig. S13: LTR Assembly Index (LAI) scores for *Hydrophis ornatus* chromosomal sequences. LAI scores were calculated in 3Mbp windows with a 300Kbp step-size. The red dashed line at y=10 indicates the 'reference quality' cut-off.

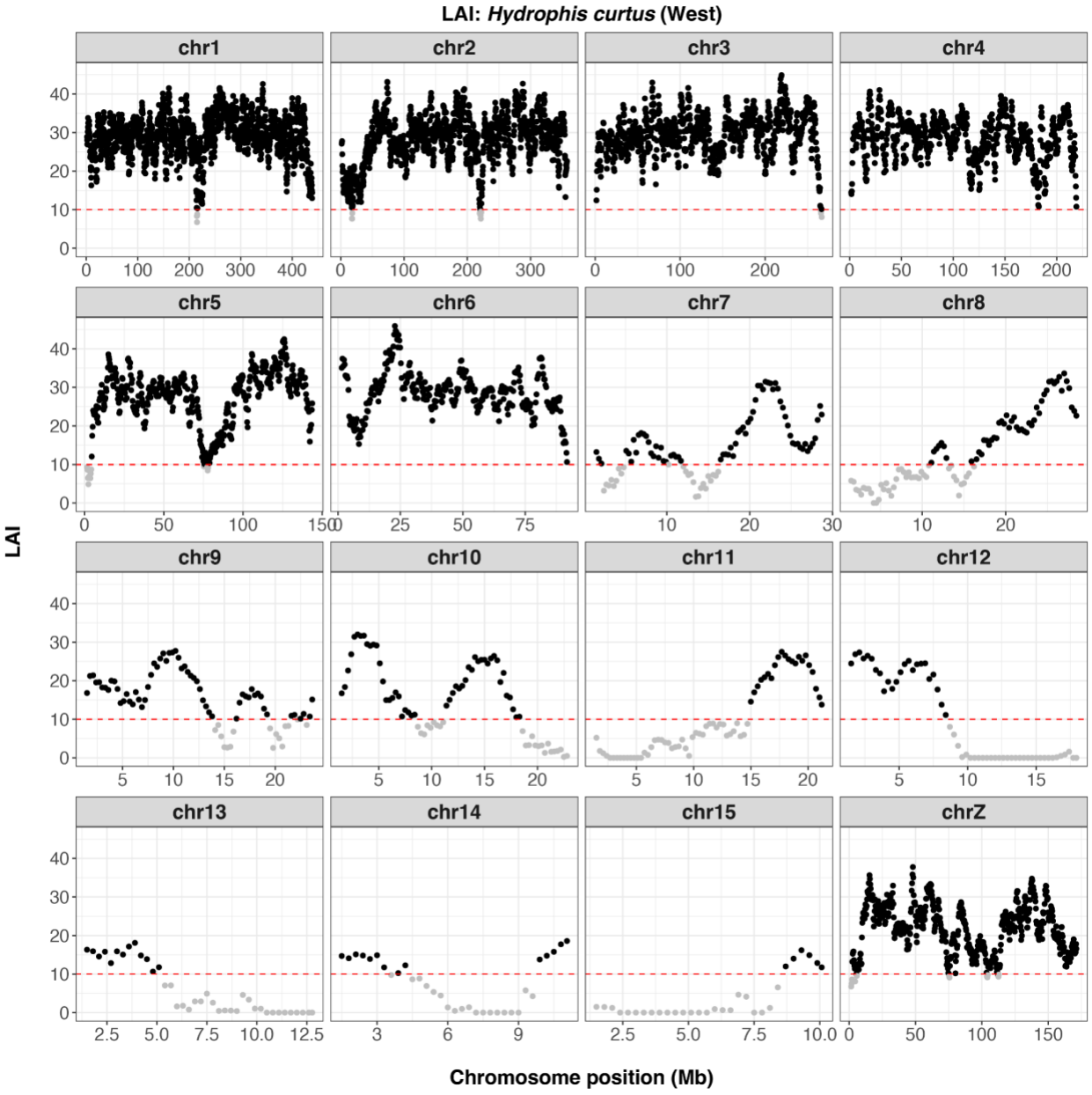


Fig. S14: LTR Assembly Index (LAI) scores for *Hydrophis curtus* chromosomal sequences. LAI scores were calculated in 3Mbp windows with a 300Kbp step-size. The red dashed line at y=10 indicates the 'reference quality' cut-off.

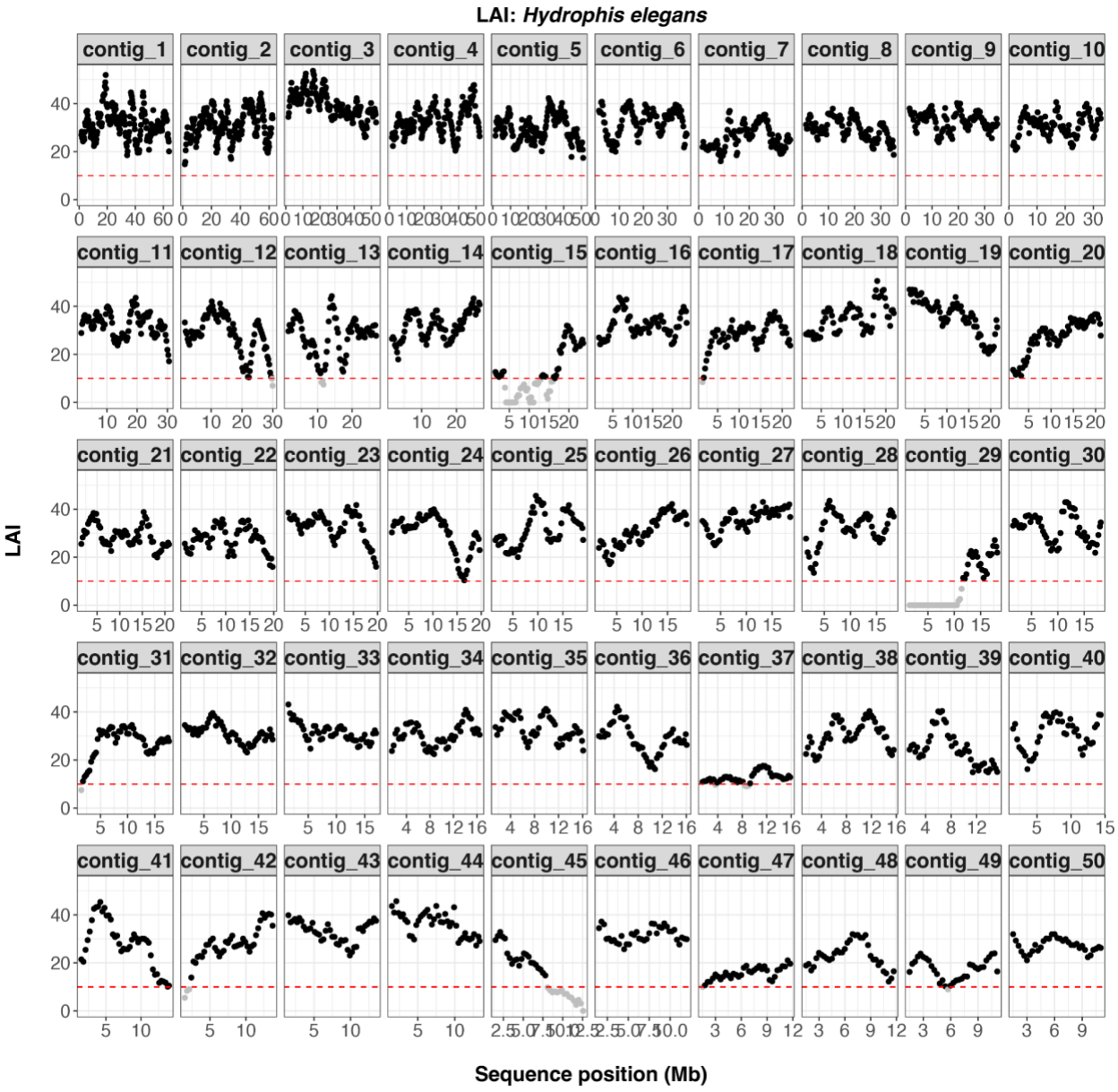


Fig. S15: LTR Assembly Index (LAI) scores for the fifty largest *Hydrophis elegans* contig sequences. LAI scores were calculated in 3Mbp windows with a 300Kbp step-size. The red dashed line at y=10 indicates the 'reference quality' cutoff.

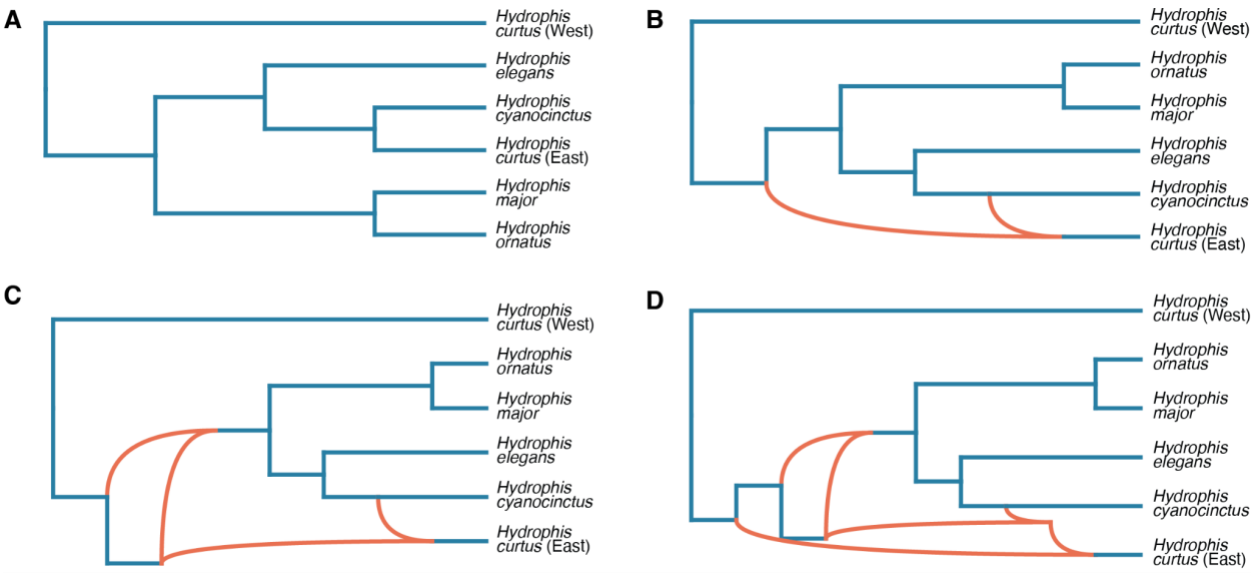


Fig. S16: *PhyloNet* networks for varying reticulation values. Multiple *PhyloNet* networks were produced by adjusting the number of reticulation events that were allowed to occur. Facets A-D represent zero through three reticulation events, respectively.

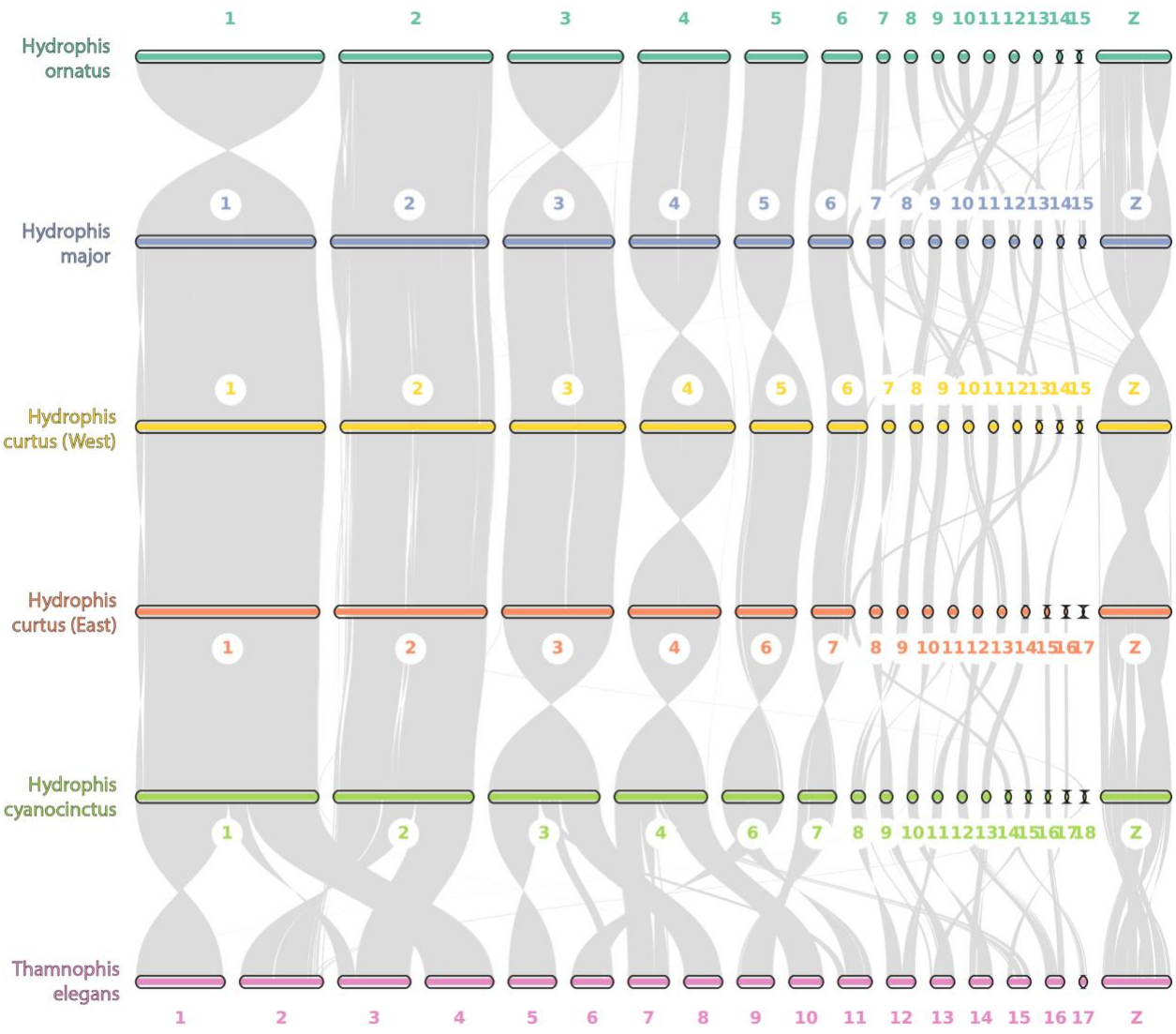


Fig. S17: MCscan synteny before reverse complementing select sequences and arranging chromosomes in order. Some examples of fully inverted chromosomes include chromosomes 1 and 3 between *H. ornatus* and *H. major*, chromosomes 3, 4, 6 and 7 between *H. curtus* (East) and *H. cyanocinctus* and chromosome 4 between *H. curtus* (West) and *H. curtus* (East).

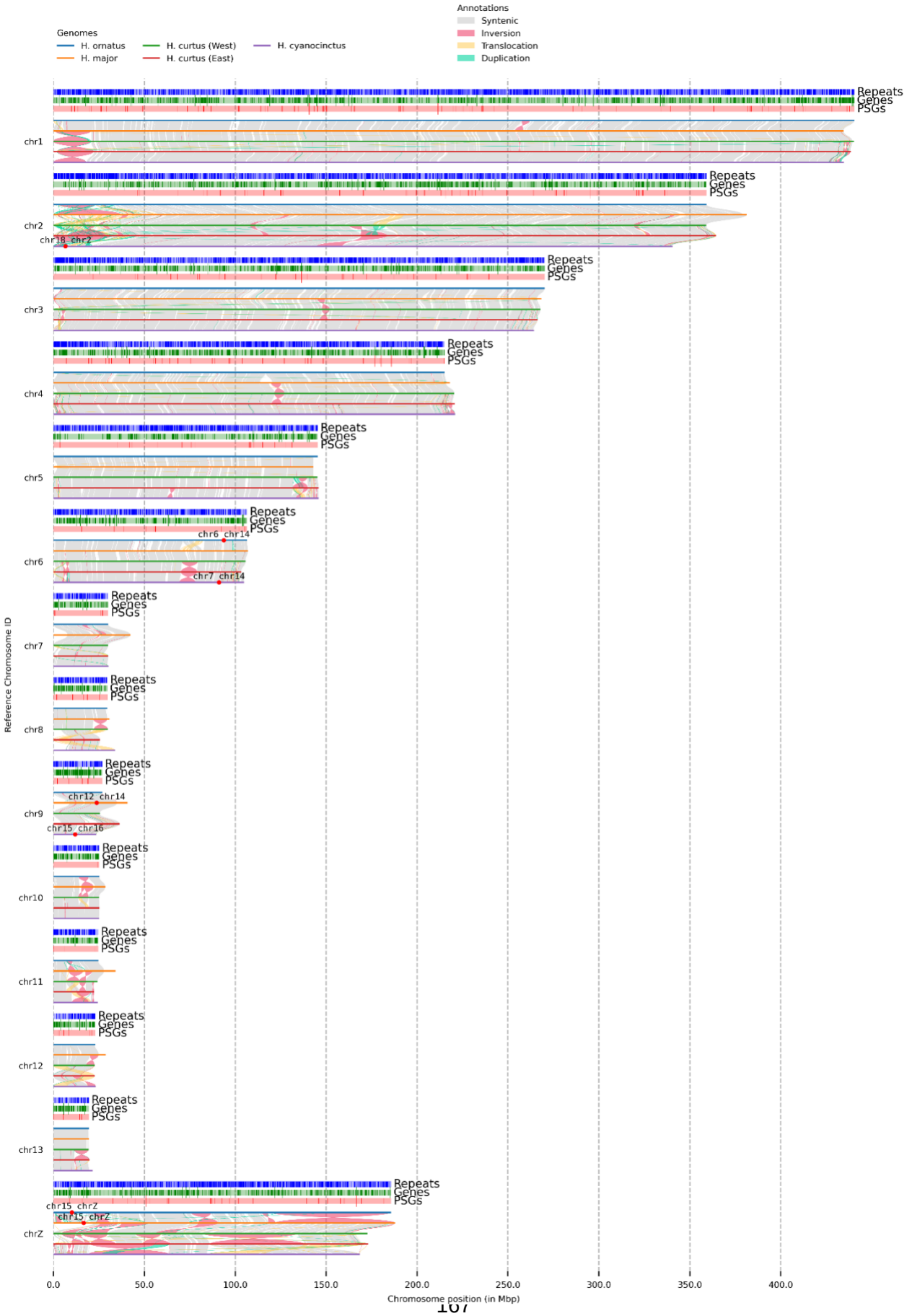

Genomes
H. ornatus
H. major
H. curtus (West)
H. curtus (East)
H. cyanocinctus
Annotations
Syntenic
Inversion
Translocation
Duplication
Repeats
Genes
PSGs
chr1
chr2
chr18 chr2
chr3
chr4
chr5
chr6 chr14
chr6
chr7 chr14
chr7
chr8
chr12 chr14
chr9
chr15 chr16
chr10
chr11
chr12
chr13
chr15 chrZ
chrZ
Reference Chromosome ID
0.0
50.0
100.0
150.0
200.0
250.0
300.0
350.0
400.0
Chromosome position (in Mbp)

Fig. S18: Chromosome synteny and structural variation between five *Hydrophis* snakes. Chromosome synteny and structural rearrangements were visualised using *Syri* and *Plotsr* from pairwise whole genome alignments between the *Hydrophis* snakes. Each snake in the figure is compared to the genome below it, with *Hydrophis ornatus* being used as the reference anchor, with all annotation tracks relative to it. Chromosome sequences were modified in each of the snakes so that their karyotypes were identical. Chromosomes that were concatenated together are labeled in the figure for each respective snake. Structural variation is coloured with inversions in red, translocations in yellow and duplications in green.

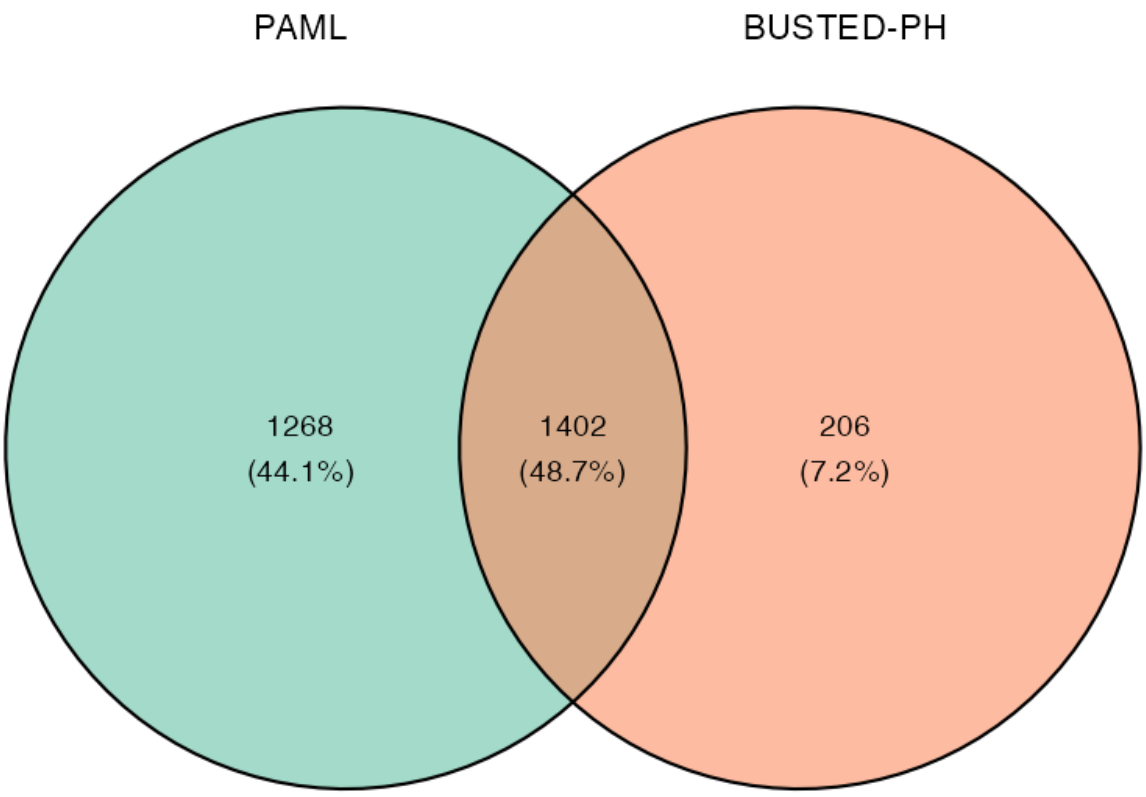


Fig. S19: Venn diagram of marine-specific PSGs identified by PAML (drop-out) and BUSTED-PH.

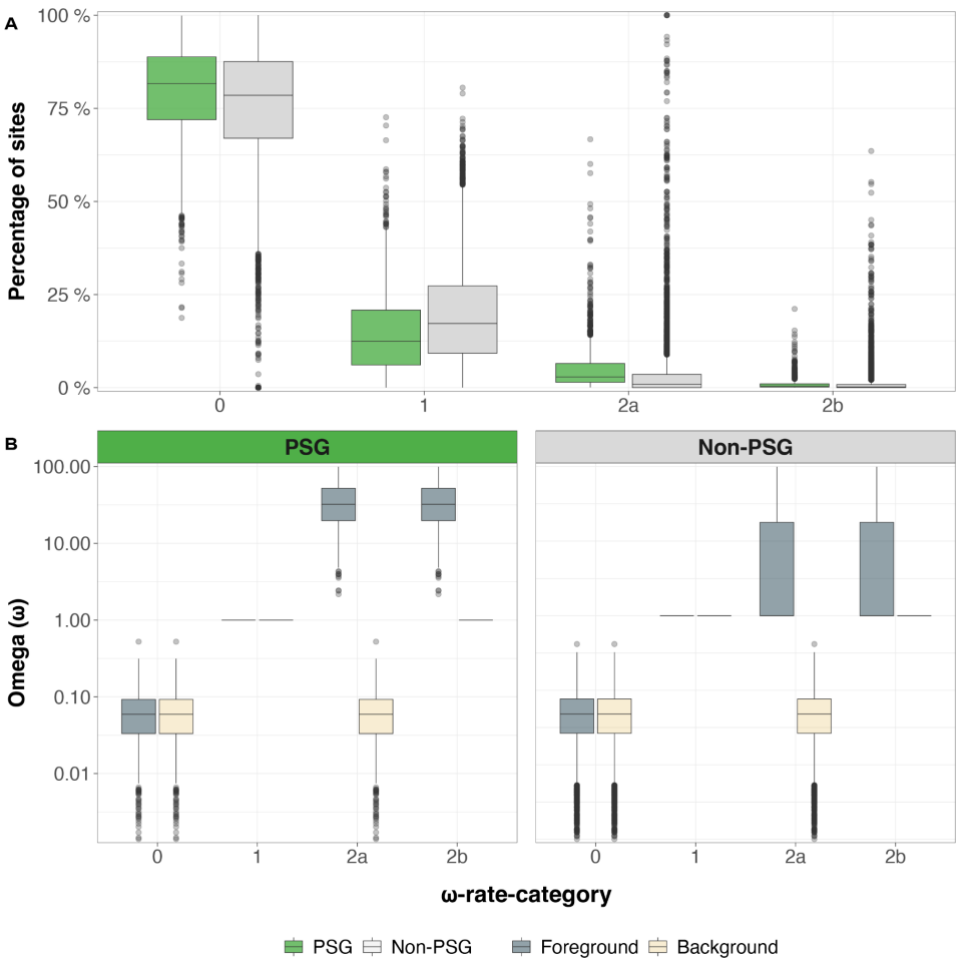


Fig. S20: PAML Branch-Site (alt.) summary. A) Percentage of sites in each multiple sequence alignment belonging to each *ω*-rate class. Green boxes represent *Hydrophis* positively selected genes with grey representing terrestrial snakes. Classes 0 and 1 represent instances where positive selection is prohibited ($\omega < 1$ and $\omega = 1$, respectively), while 2a and 2b represent situations where $\omega > 1$ is allowed in the foreground branches, but not in the background branches. B) Comparison of *ω* values between PSG and non-PSG genes for each rate-category. PSGs have considerably higher *ω* values in 2a and 2b relative to the background branches, while non-PSGs show very low *ω* values for the foreground snakes relative to the background snakes, which show higher *ω* values in 2a and 2b.

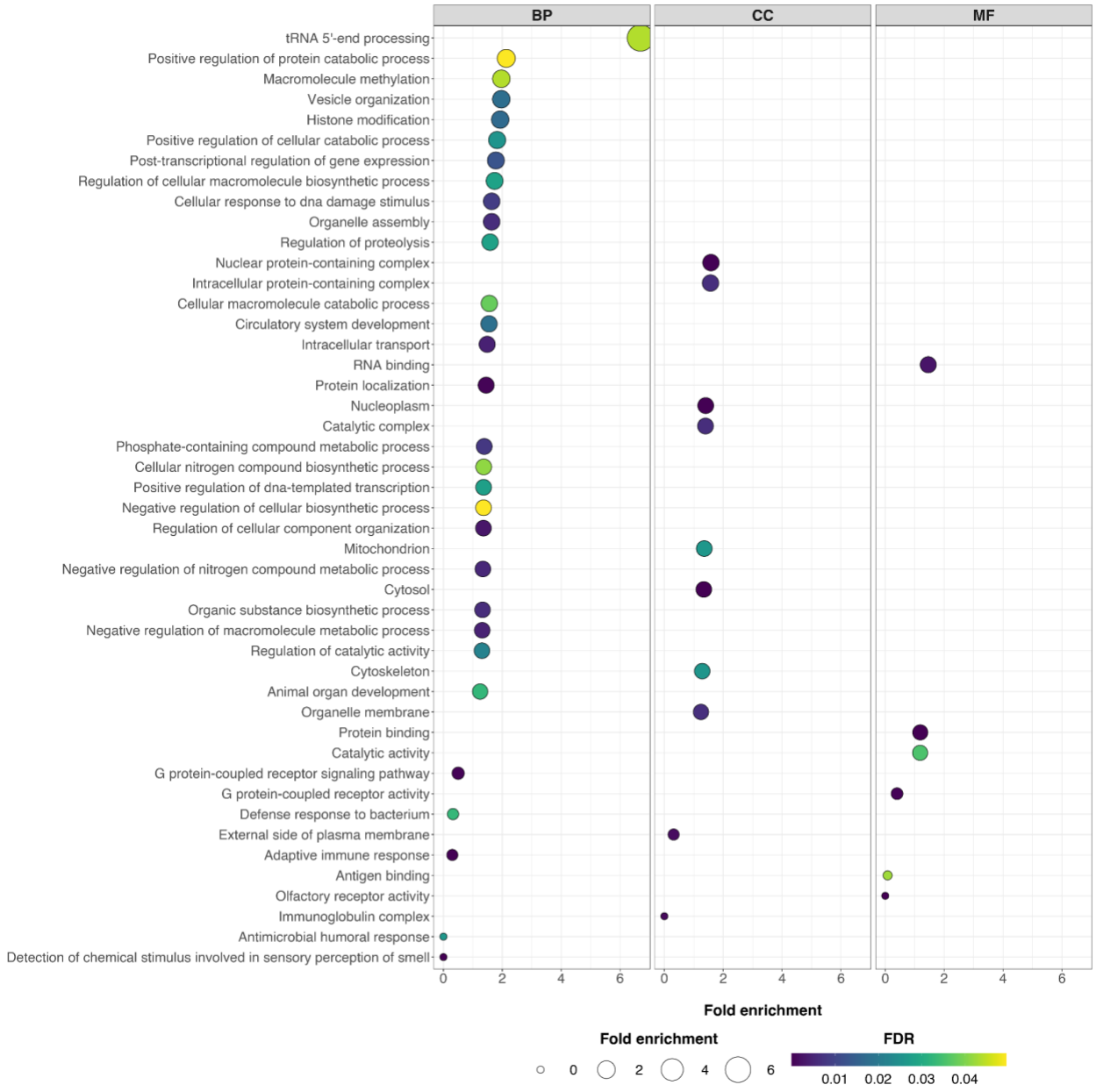


Fig. S21: *PANTHER* fold enrichment for the most specific over-represented terms (level 0). Fold enrichment represents the observed number of genes associated with a GO term divided by the expected number of genes for the term. Values greater than 1 indicate overrepresentation while values less than 1 represent underrepresentation. Facets represent the three ontologies with the size of the circles representing the fold enrichment, and the colour indicating the FDR value.

Chapter 4:

Genome scans reveal shared islands of differentiation and balancing selection on chemosensory genes in a rapid sea snake radiation (*Hydrophis*)

Ludington, A. J. & Sanders K. L (2023). Genome scans reveal shared islands of differentiation and balancing selection on chemosensory genes in a rapid sea snake radiation (*Hydrophis*). **Unpublished.**

Statement of Authorship

Title of Paper	Genome scans reveal shared islands of differentiation and balancing selection on chemosensory genes in a rapid sea snake radiation (Hydrophis)
Publication Status	☐ Published ☐ Accepted for Publication ☐ Submitted for Publication ☒ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Written in manuscript style for publication.

Principal Author

Name of Principal Author (Candidate)			
Contribution to the Paper	Conceived the research (with KLS), performed quantitative analyses, interpreted results, wrote the first version of the manuscript and edited subsequent versions.		
Overall percentage (%)	80		
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	28/08/2023

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	Kate Laura Sanders		
Contribution to the Paper	Conceived the research, interpreted results, wrote the first version of the manuscript and edited subsequent versions.		
Signature		Date	28/08/2023

Genome scans reveal shared islands of differentiation and balancing selection on chemosensory genes in a rapid sea snake radiation (*Hydrophis*)

Alastair J. Ludington[1,*] and Kate L. Sanders[1,7,*]

[1] School of Biological sciences, The University of Adelaide, Adelaide, SA 5005, Australia

[7]The South Australian Museum, Adelaide, Australia

[*] Corresponding authors: A.J.L: alastair.ludington@adelaide.edu.au; K.L.S: kate.sanders@adelaide.edu.au

Abstract

Hydrophis sea snakes comprise a recent and exceptionally diverse radiation of fully marine reptiles. We provide the first full sequence level patterns of diversity and differentiation for this group, focusing on three ecologically distinct and coexisting species that diverged near-simultaneously 3-8 million years ago (*Hydrophis major, H. elegans* and *H. curtus*). Analyses of introgression and genome-wide diversity and differentiation point to early admixture and shared evolutionary processes in shaping conserved genomic patterns among these species. Genome scans identified numerous islands of differentiation ("genomic islands") that are shared among the three species comparisons and reflect varied evolutionary scenarios. Most islands show elevated Fst, low π, and neutral *Dxy* against the genome average. These patterns best fit a scenario of long-term linked background and/or positive selection across a stable genomic

landscape. Low *Fst* islands, on the other hand, fit model expectations for balanced selection on ancestral polymorphisms (high *Dxy* and π, positive *Tajima's D*). Notable here is a 1.2Mb region that coincides with numerous chemosensory genes in the vomeronasal type-2 receptor (V2R) family. These genes are critical in mediating both reproductive and feeding behaviours (via tongue flicking) in sea snakes. We suggest that the maintenance of adaptive variation in chemosensory genes facilitates efficient adaptive responses, such as to new trophic opportunities and/or the reinforcement of reproductive barriers. This work on a promising new snake system adds to our growing understanding of the genomic basis of rapid adaptive radiation.

Keywords: Sea snakes; divergence; adaptation; speciation; chemosensory

Introduction

The advent of next generation sequencing has extended the scope of genomic studies of adaptation and speciation to include an array of systems such as flycatchers (Ellegren et al., 2012; Nadachowska-Brzyska et al., 2016), cichlid fish (Kautt et al., 2020; Meier et al., 2017, 2018), *Heliconius* butterflies (Kronforst et al., 2013; Martin et al., 2013, 2016), poplar trees (Ma et al., 2018), and sunflowers (Andrew & Rieseberg, 2013; Huang et al., 2020; Renaut et al., 2013). These systems have provided valuable opportunities to observe lineages at the early stages of divergence, often at multiple points along the speciation continuum. In many such cases, genome-wide scans have identified local regions of pronounced genetic differentiation that stand out against a background of relatively low differentiation between diverging lineages. These so-called "islands of differentiation" can reflect various divergence processes depending on the

selection regime, timing of gene flow, and underlying genomic architecture. Alternative models have been developed to identify and explain the formation of genomic islands, particularly using population genetic summary statistics of diversity and divergence.

The earliest models invoked differential gene flow in elevating differentiation at loci involved in reproductive isolation ("speciation genes") above a relatively homogeneous genetic background. Support for this model is found if genomic islands show elevated differentiation for both relative (*Fst*) and absolute (*Dxy*) measures of divergence, and low intraspecific diversity (π) (Han et al., 2017; Irwin et al., 2018). Only a few cases meet these criteria but in doing so have provided valuable insights into the loci underlying speciation and adaptation (Christmas et al., 2021; Kulmuni et al., 2020; Liu et al., 2020). Other models explain heterogeneous patterns of differentiation without reference to reproductive isolation. These propose that lineage sorting and selection build differentiation at some regions of the genome, while low differentiation is maintained at other regions due to the sharing of ancestral polymorphisms. Under a scenario of "selection in allopatry", high *Fst* and low π (but neutral *Dxy*) indicate that differentiation was driven by recent positive selection and/or background selection (i.e. selection against harmful mutations). The scenario of balancing selection, on the other hand, predicts high *Dxy* and π but low *Fst* - signatures of the sorting of ancestral balanced polymorphisms among descendant lineages (Han et al., 2017; Irwin et al., 2018).

All models of the formation of genomic islands emphasise the importance of genomic architecture and divergence timescale. The effects of linked positive and/or background selection are promoted in regions with low recombination rates, especially

those that also have high functional densities (i.e. harbour more targets for selection) (Seehausen et al., 2014; Wolf & Ellegren, 2017). This is seen particularly in systems that have retained stable patterns of recombination and gene density over successive speciation events; common genomic islands occur across multiple pairwise comparisons in these systems due to the long-term effects of linked selection in the same (low recombination) regions.

Here, we explore patterns of genomic differentiation in a powerful new system that is phylogenetically and ecologically distinct from previously studied taxa. *Hydrophis* sea snakes share a common ancestor dated at only ~3-8 million years old and exhibit the classic hallmarks of adaptive radiation (Lee et al., 2016; Sanders, Lee, et al., 2013; Sanders, Rasmussen, et al., 2013). The ~50 species in this fully marine clade have undergone three-fold faster speciation compared to all other hydrophiine snakes, including the contemporaneous and geographically overlapping *Aipysurus*-*Emydocephalus* sea snakes (Lee et al., 2016; Sanders et al., 2010). *Hydrophis* species also show pronounced ecomorphological diversity accompanied by high sympatric species richness throughout their Indo-Australian distribution (Voris & Voris, 1983). Recently generated genomes showed conserved chromosome-scale synteny, but widespread intra-chromosomal variations, and high repeat content in five *Hydrophis* species spanning the clade's phylogeny (Chapter 2, this thesis; Li et al., 2021; Peng et al., 2023). These findings prompted the current study to identify the evolutionary processes and genomic regions that have contributed to rapid differentiation in *Hydrophis*.

We re-sequenced single populations of three eco-morphologically distinct *Hydrophis* species that have overlapping distributions centred in northern Australia. *H. elegans* is a gracile, long bodied (up to 3m long) specialist on crevice and burrowing eels in depths up to 150m (C. Fry et al., 2001; Sherratt et al., 2018). *Hydrophis major* is of moderate length with a large head and strong jaws, and specialises on spiny toxic prey, especially plotosid catfish, in depths up to 50m (Shine et al., 2019). *Hydrophis curtus* is a short, thick-bodied generalist predator in shallow water (~5-15m) and, unlike the other species (Sherratt et al., 2018), is predominantly day-active (Simões et al., 2020). The clear eco-morphological boundaries among these co-occurring species suggest near-complete or complete reproductive isolation; however, reticulate relationships and high levels of gene tree discordance indicate extensive incomplete lineage sorting and potential gene flow during their recent speciation (Chapter 3, this thesis). Comparisons of these species therefore provide an excellent opportunity to examine shared and lineage-specific patterns of differentiation and selection in *Hydrophis*.

Methods

Sample collection and sequencing

Samples from each of the three *Hydrophis* species were collected and sampled during fieldwork in Western Australia, using procedures approved by The University of Adelaide's Animal Ethics Committee (approval number S-2015-119 / 34903), under a fauna taking license (regulation 25, number FO25000393) granted by the Department of Biodiversity, Conservation and Attractions of Western Australia. *Microcephalophis*

gracilis was sourced by collaborators from commercial fisheries operating in coastal waters of the Emirate of Fujairah, United Arab Emirates. DNA was extracted from liver and tail tissues following the Gentra Puregene Tissue Kit protocol. DNA was sent to the South Australian Genomic Center (SAGC), where libraries were generated according to Illumina DNA Prep (M) guidelines (Part No. 10000000254 v10) and Illumina to MGI Library Conversion (MGIEasy Universal Library Conversion Kit, Part No. MGI000004155), and were sequenced on a MGI DNBSEQ-G400 (2x150bp paired-end). A single *H. curtus* sample (R36639) and *M. gracilis* were sequenced separately at the Ramaciotti Centre for Genomics. Libraries were prepped using Illumina DNA PCR-Free prep kit before being sequenced on an Illumina NovaSeq 6000 S4 flow cell (2x150bp paired-end). All sequence data were filtered for contaminants and low-quality reads using BBduk v39.01 (BBMap – Bushnell B. – sourceforge.net/projects/bbmap), Kraken2 v2.1.2 (Wood et al., 2019) and Fastp v0.23.2 (Chen, 2023).

Alignment and variant calling

The whole genome sequence data for each snake were aligned to the *Hydrophis major* reference genome (JAUCBL000000000). The *H. major* reference genome was chosen as it has high base accuracy, sequence completeness and is chromosome-scale. The sequence aligner BWA-mem2 v2.2.1 (Vasimuddin et al., 2019) was used to align the sequence data, filtering for reads aligned in proper pairs with a minimum mapping quality of 20 (--require-flags 3 –exclude-flags 4 –min-MQ 20). Duplicate alignments were marked by Sambamba v0.8.2 (Tarasov et al., 2015) before depth statistics were calculated using MosDepth v0.3.3 (Pedersen & Quinlan, 2018). Alignments were then subset for only autosomal sequences.

Retaining as many sites as possible (variant and invariant) is crucial to accurately estimating many divergence statistics. We performed joint genotyping across all samples, generating an all-sites VCF, before applying filters to each species or species comparison separately to retain as many sites as possible. Joint-genotyping was performed across all samples using BCFtools mpileup v1.15.1 to estimate genotype likelihoods from the alignment files, requiring a minimum mapping and base quality of 20 (-q 20 -Q 20) before BCFtools call was used to call variant and invariant sites using the updated multiallelic caller (-m) (Danecek et al., 2021). We then applied a general filter to remove any sites within 10 bases of an INDEL (--SnpGap 10), before filtering out INDELs entirely (--exclude-types indels).

For each species, or species-comparison, we then applied the following filtering to obtain high quality all-sites VCFs, consisting of invariant sites and bi-allelic SNPs. High confidence invariant sites were first obtained by filtering for sites with no alternate alleles, zero missing data across samples, a minimum per-sample depth of 10 and a minimum mapping quality of 20 using VCFtools v0.1.6 (--recode –recode-INFO-all –max-maf 0 –max-missing 1 –minDP 10 –minQ 20) (Danecek et al., 2011). Bi-allelic variant sites were then obtained separately using the same missingness, depth and quality filters as above, but instead requiring a minimum and maximum of two alleles at each site (–min-alleles 2 –max-alleles 2). The invariant and variant VCFs were then merged using BCFtools concat (--allow-overlaps) to form the final output file.

Divergence statistics and correlation analyses

Relative (*Fst*) and absolute divergence (*Dxy*) measures were estimated between species pairs (*H. major - H. elegans, H. major - H. curtus* and *H. elegans - H. curtus)*,

while nucleotide diversity (π) was estimated within each species (*H. major, H. curtus* and *H. elegans)*. Divergence statistics *Fst* and *Dxy* were estimated across the genome in 50kbp non-overlapping windows using the program *Pixy* (Korunes & Samuk, 2021), grouping samples by their respective populations (--stats dxy fst –populations groups.tsv –window_size 50000). The program VCFtools was also used to calculate the average genome-wide weighted *Fst* for each species-comparison. Within each species, nucleotide diversity (π*)* and *Tajima's D* were estimated in 50kbp non-overlapping windows by Pixy and VCFtools, respectively. Genomic windows were required to have at least 5000 called sites (10% of the window size) to be included in the downstream analyses. *Fst* values were Z-transformed within species pairs to enable comparison of divergence landscapes between species pairs with different divergence times (Han et al., 2017). Diversity statistics were correlated within species-comparisons to explore the genome-wide signal of divergence, assessing the relationship between *Fst-Dxy*, average π*-Fst* and average π*-Dxy*. Additionally, genome-wide *Fst* and *Dxy* measures were correlated between each of the three species-comparisons to establish if shared or lineage-specific evolutionary processes were driving divergence across *Hydrophis*.

Detecting genomic outliers and assigning divergence patterns to genomic windows

Two separate approaches were used to identify and classify divergence across the genome. We first Z-transformed windowed *Fst* values in each species comparison to account for variable divergence times between species, allowing us to accurately compare across species pairs. Windows with *ZFst* scores of ≥ 3, which are windows

with *Fst* scores that are three standard deviations (SD) from the mean, formed our set of genomic outliers based solely on elevated relative divergence. We also identified windows where Z*Fst* ≤ -3, representing extremely low divergence regions within the genome. Adjacent outlier windows within species comparisons were reduced into a single representative region. We also explored shared outlier windows between species comparisons by assessing the level of overlap in outlier windows reported across each of the three species comparisons.

A second approach we used to classify divergence across the genome involved assigning divergence profiles to 50kb windows using information from both *Fst* and *Dxy*. The four divergence profiles included allopatric speciation, balancing selection, recurrent selection, and divergence-with-gene flow (Han et al., 2017; Irwin et al., 2018; Shang et al., 2023). Windowed *Fst* and *Dxy* estimates were filtered for values that fell within the top one percent of values. Windows where both *Fst* and *Dxy* were in the top one percent were classified as “divergence-with-gene flow”. Windows with *Fst* values in the top one percent, but *Dxy* between the top and bottom one percent were classified as “allopatric selection”. An extension of the “allopatric” profile were “recurrent selection” windows, which showed almost identical divergence profiles to “allopatric” windows but had *Dxy* values in the bottom one percent. Finally, windows with *Dxy* in the top one percent and *Fst* in the bottom one percent were classified as “*Balancing selection”*. Summary profiles were generated for within and between species comparisons in *R*, while a pairwise Wilcoxon rank sum test was used to assess if shared genomic regions across species comparisons were larger than unique classified regions within comparisons.

Genome-wide and sliding window introgression tests

We used a genome-wide test for introgression in the form of Patterson's D statistics (also called ABBA-BABA tests) to test for gene flow between the three *Hydrophis* species. *D* statistics use three ingroup populations and an outgroup population with the relationship (((P1, P2), P3), O), and tests if there is elevated shared variation between P3 and P2 compared to that shared by P3 and P1. The test represents the genotype states using "ABBA" and "BABA" patterns, where "A" is the ancestral allele and "B" is the derived allele. ABBA patterns represent sites where P2 and P3 share the derived allele and P1 shares the ancestral state with the outgroup. Alternatively, BABA patterns represent sites whereby P1 and P3 share the derived allele and P2 has the ancestral genotype. We used the program *Dsuite dtrios* v0.5r52 to calculate *D* statistics for all possible tree topologies, fixing *Microcephalophus gracilis* as the outgroup (Malinsky et al., 2021). From these tests, *Dmin* was calculated, representing the smallest possible *D statistic* from the possible tree topologies. When *Dmin* is statistically significant, it signifies that allele sharing between the three species is inconsistent with a simple bifurcating topology, even in the face of incomplete lineage sorting. *Dsuite* also calculated f_4 admixture statistics to quantify the level of admixture between species.

We then used the program *ABBABABBAwindows.py* to calculate f_d statistics in 50kbp sliding windows across the genome, requiring a minimum of 250 SNPs per window (https://github.com/simonhmartin/genomics_general). f_d is a modified f-statistic that is suitable for identifying introgressed regions using sliding windows, as it is more robust to the error introduced by using fewer SNPs per-window (Martin et al., 2015). The statistic uses the same tree topology and genotype patterns (ABBA and BABBA) to

test for introgression. The statistics is interpreted as the amount of shared variation between P3 and P2 that is not also shared by P1. We used f_d to estimate the level of shared genetic variation across the genome between all combinations of *H. major*, *H. elegans* and *H. curtus*, again fixing *M. gracilis* as the outgroup.

Manual curation and investigation of ZFst outliers and profiled genomic windows

Genomic outliers identified by *ZFst* scans, along with genomic windows classified as "Allopatric selection" were curated and investigated further. These regions were focussed on as they represent loci where significant differentiation may be occurring between pairs of species. Sequence divergence statistics (*ZFst*, *Dxy*, *π*, *Tajima's D*), along with sliding window estimates of introgression (f_d), were visualised within and around each candidate loci in *R*. Genes within divergent regions were identified and annotated if they lacked a gene annotation. Annotations were obtained by searching all *H. major* proteins against *UniProt* using *BLASTP* (chapter 2, this thesis) and *EggNOG* protein database v5.1 using *emapper* v2.1.9 (Cantalapiedra et al., 2021; Huerta-Cepas et al., 2019). Protein searches against *EggNOG* were only retained if 50% identity was achieved over at least 70% of the protein's length. UniProt accessions were mapped to gene symbols using the ID Mapping portal on the UniProt website (https://www.uniprot.org/id-mapping). Annotation sources were aggregated into a non-redundant set of gene annotations in *R*. Vomeronasal annotations for the genes in the balanced loci at the start of chromosome were refined by comparing every gene within the 1.2Mbp region against all vomeronasal sequences from UniProt-SwissProt and

TrEMBL. The best hit for each sequence was taken and used to annotate the sequence. Sequences in this region that were not vomeronasal genes were searched against the BLAST NR database to obtain curated gene annotations.

Results

Genome wide divergence statistics

Sixteen samples were sequenced from three *Hydrophis* species: five *H. major*, five *H. elegans* and six *H. curtus*. All samples were aligned to the *Hydrophis major* reference genome, with resulting alignment coverage ranging between 12-54x, while breadth-of-coverage ranged from 81 to 86% (Table 1). The number of called sites (biallelic variants and invariant sites) within each species ranged from 656,347,619 to 951,899,031, while the number of sites between species-pairs ranged from 577,984,682 to 876,005,580 (supplementary table S1). These data were then used to scan the whole genome using 50kbp non-overlapping windows to estimate the following divergence parameters: genomic divergence (π), Tajima's D, relative divergence (*Fst*) and absolute divergence (*Dxy*).

Table 1: Coverage statistics for the sixteen *Hydrophis* snakes and outgroup *Microcephalophus gracilis*.

Species	Sample ID	Aligned bases (Gb)	Average depth of coverage	Breadth of coverage (%)
H. curtus	Cur1VT	29.8	14.0	81.1
	Cur2VT	30.2	14.2	81.4
	Cur3VT	31.4	14.7	81.6
	Cur4VT	39.3	18.5	81.8
	Cur7VT	27.7	13.0	81.5
	R36639	36.8	17.3	82.9
H. elegans	KLS1121	116.8	54.8	83.2
	KLS1034	49.2	23.1	82.1
	KLS1024	37.3	17.6	81.4
	KLS1027	35.1	16.5	81.4
	KLS1025	34.4	16.2	81.5
H. major	KLS0629	35.9	16.9	85.1
	KLS0630	32.8	15.4	85.2
	KLS0631	31.5	14.8	83.9
	KLS1020	26.7	12.5	86.1
	KLS0647	25.4	12.0	84.3
M. gracilis	UAE_835	80.2	37.4	82.9

The level of nucleotide diversity was highest in *H. curtus* (π = 0.00254) and *H. elegans* (π = 0.00211), and lowest in *H. major* (π = 0.00163) (Figure 1A; Table 2). Genome-wide estimates of Tajima's D were predominantly negative for all three species (Figure 1A), indicating an excess of segregating sites relative to the total number of haplotypes within each species. Weighted *Fst* between species ranged from 0.54 to 0.66, while *Dxy* showed considerably less variability, ranging from 0.0050 to 0.0053 (Figure 1A; Table 2). Correlating windowed *Fst*, *Dxy* and π revealed a significant weak positive correlation between *Fst* and *Dxy* in each comparison (r = 0.01 to 0.05, p-value ≤ 0.04), a significant moderate negative correlation between average π and *Fst* (r = *-0.59* to -0.67, p-value <

0.001), and a significant positive correlation between average π and *Dxy* (*r = 0.65 to 0.73,* p-value < 0.001) (Figure 1B). Notably, genomic windows reporting high *Fst* typically also had low genetic divergence and low to moderate *Dxy* values (Figure 1C). Comparing *Fst* across the genome between each of the three species comparisons revealed a significant moderate correlation in all comparisons (*r = 0.61 to 0.66, p values < 0.001)* (Figure 1D). A significant positive correlation was also observed for *Dxy* (*r = 0.72 to 0.77, p-values < 0.001),* indicating that evolutionary processes shaping *Fst* and *Dxy* are likely common to each species.

Table 2: Genome-wide diversity statistics for within and between measures

Within population statistics		**Between population statistics**			
Species	Nucleotide diversity	Comparison	*Dxy*	*Fst* (weighted)	*Fst* (unweighted)
H. curtus	0.0025	HCU-HEL	0.0050	0.54	0.23
H. elegans	0.0021	HCU-HMA	0.0052	0.60	0.26
H. major	0.0016	HEL-HMA	0.0053	0.66	0.33

[1] Comparison column uses three letter species identifiers which correspond to: *H. curtus* (HCU), *H. elegans* (HEL) and *H. major* (HMA)

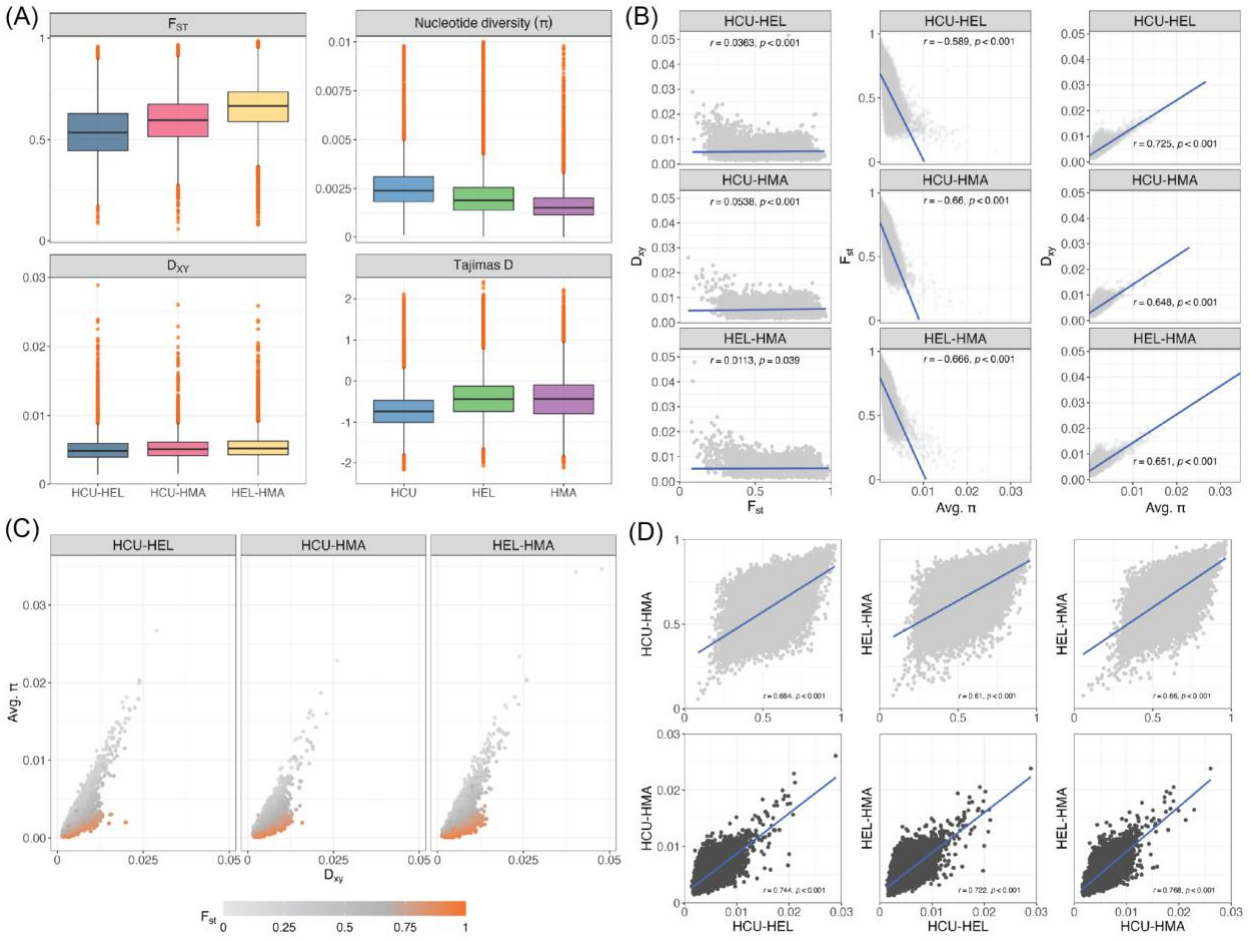


Figure 1: Nucleotide diversity statistics and correlation of genome-wide estimates of relative and absolute measures between species comparisons. (A) Boxplots of between and within species divergence. Between species statistics include *Fst* and *Dxy* in the left two facets for HCU-HEL (blue), HCU-HMA (pink) and HEL-HMA (yellow). Within species nucleotide diversity, along with *Tajima's D* are shown in the facets on the right for *H. curtus* (light blue), *H. elegans* (green) and *H. major* (purple). Outliers are coloured in orange, with the middle horizontal black bar representing the median value, with the top and bottom of the box representing the third and first quartiles, respectively. (B) genome-wide correlations between *Fst* and *Dxy* (left-most facets), average-*π* and *Fst* (middle column) and average-*π* and *Dxy* (right-most facets) for each species

comparison. Points represent 50kb windows, while the blue line is the line of best fit. (C) Pearson correlations between *Dxy* and π, coloured by *Fst* values. Each point represents a 50kb genomic window. (D) Pearson correlations between all combinations of species comparisons for *Fst* (top row, light grey) and *Dxy* (bottom row, dark grey). Each point represents a 50kbp window.

Identifying genomic outlier windows in species pairs

Few genomic islands were identified after filtering windows for *ZFst* ≥ 3. Following the merging of adjacent windows, eight 50kb high-divergence windows were identified between *H. curtus* - *H. elegans*, one 100kb islands was identified between *H. curtus* - *H. major* and no islands were identified between *H. elegans* - *H. major* (supplementary table S2). Only a single 100kbp genomic outlier region was shared across comparisons at position 214.55Mb to 216.65Mb of chromosome 1, being identified in the *H. curtus* - *H. elegans* and *H. curtus* - *H. major* comparisons (supplementary table S3). This same region in the *H. elegans* - *H. major* comparison was elevated but did not surpass the *ZFst* threshold (ZFst: 2.49 - 2.54). Alternatively, numerous low-divergence regions were identified (*ZFst* ≤ -3), with a total of 69 regions identified between *H. elegans* - *H. major*, 19 between *H. curtus* - *H. major* and four regions between *H. curtus* - *H. elegans*. Nine low diversity regions were shared between the *H. elegans* - *H. major* and *H. curtus* - *H. major* comparisons (supplementary table S3), with another four low divergence regions being shared by all three comparisons. Notably, of the 128 low-divergence windows identified across all species three comparisons, 94 (73.4%) resided within the first 70Mbp of chromosome two, with 71 belonging to the *H. elegans* - *H. major* comparison,

19 to the *H. curtus - H. major* comparison and four to the *H. curtus - H. elegans* comparison.

Absolute divergence and nucleotide diversity was then compared between outlier loci and the genomic background. In the scenario of speciation in the face of ongoing gene flow, we would expect to see elevated *Dxy* in genomic outliers, as these regions become genetically isolated as the rest of the genome is homogenised though gene flow (Han et al., 2017; Irwin et al., 2018). This was not observed, with genomic islands instead reporting low *Dxy* and π (Figure 2A), matching the general genome wide trend (Figure 1C). Genomic windows that showed exceptionally low *Fst* between species showed the opposite pattern, instead having elevated *Dxy* and π relative to the genomic background. This signature supports a “balancing selection” model, whereby ancestral polymorphisms are maintained between species pairs, resulting in the higher than neutral π and *Dxy* (Shang et al., 2023).

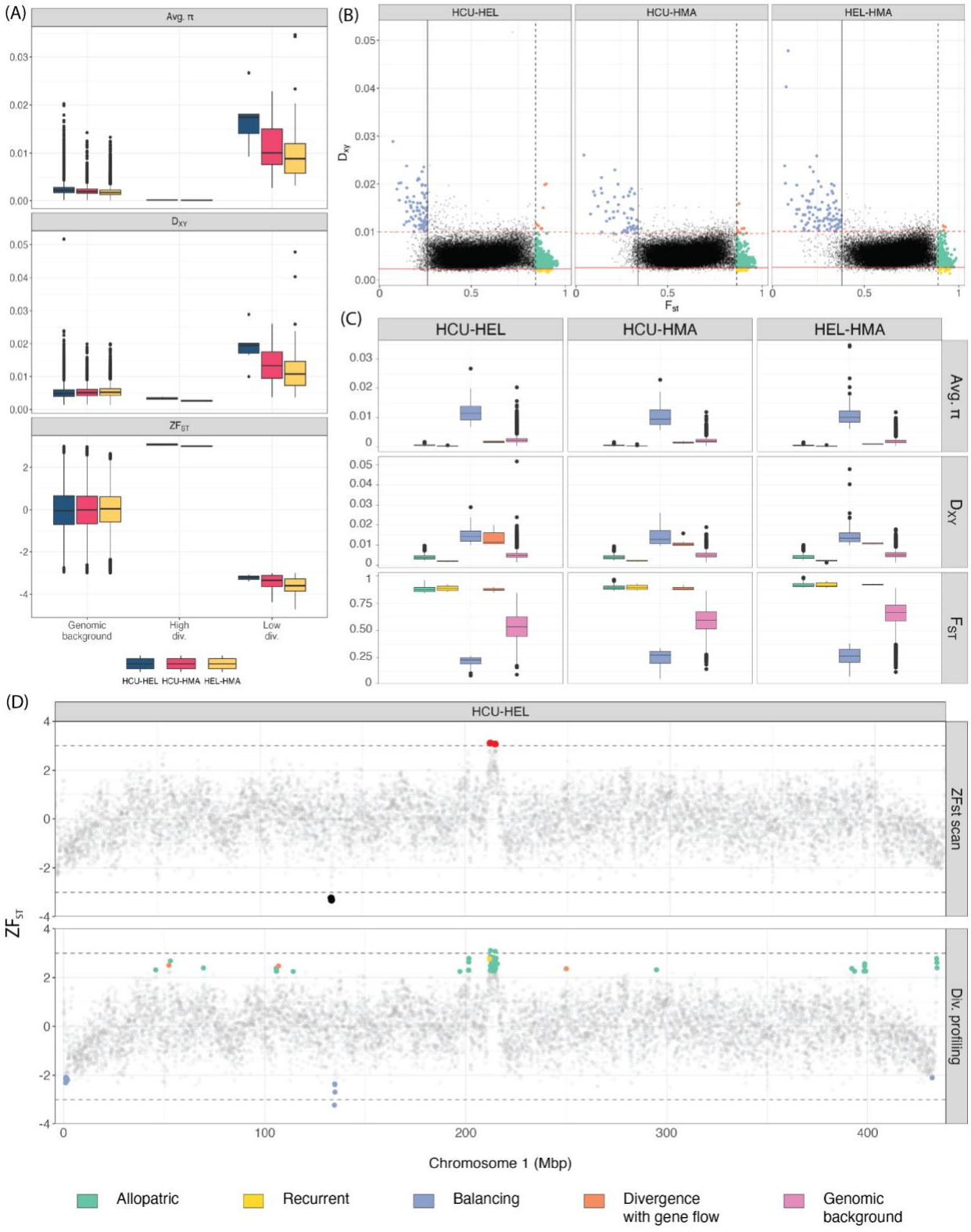


Figure 2: Classification of outlier loci and comparison of outliers/classified windows relative to the genomic background (A) Boxplots representing the difference in

divergence measures between the high-diversity (*ZFst* ≥ 3) and low-diversity (*ZFst* ≤ -3) outlier windows. Coloured boxes correspond to species comparisons: *H. curtus - H. elegans (HCU-HEL: blue*), *H. curtus - H. major (HCU-HMA; pink)* and *H. elegans - H. major* (*HEL-HMA; yellow*). (B) Scatter plot of *Fst* against *Dxy*. The dashed lines represent the top one percent of values, with the colours corresponding to the respective statistic (*Fst*; black, *Dxy; red)*. The solid lines represent the bottom one percent of values and follow the same colour scheme. Points in each of the four sections correspond to the type of divergence profile based on *Fst* and *Dxy*: Divergence-with-gene flow (orange), Allopatric selection (green), Recurrent selection (yellow) and balancing selection (light purple). (C) Boxplots showing the difference in divergence statistics between the four divergence profiles and the genomic background. Colours in figures (B), (C) and (D) are shared. (D) Genome-wide view of outlier regions classified by each of the two methods. This figure shows chromosome one, with the length on the x-axis and *ZFst* score on the y-axis. The top facet shows outliers identified using *ZFst* (*ZFst* ≥ 3; red, *ZFst* ≤ -3; dark green), with the bottom highlighting the classified windows.

Classifying divergence profiles across the genome

Genomic windows residing in the top and bottom one percent of *Fst* and *Dxy* values were assigned divergence profiles based on expectations described in Shang et al., (2023) (Figure 2B-C). Between 383 to 429 50kb windows were assigned a divergence profile in each of the species' comparisons (supplementary table S4). Allopatric selection was the most common classification, accounting for 298 to 310 of all classified windows, with an additional 20 to 27 windows classified as "Recurrent selection", a

regime that is a specific subset of allopatric selection. Balancing selection was the next most common classification (50 to 94 windows), with almost all windows localised to the start of chromosome 2 (supplementary figure S1). Divergence-with-gene flow loci were the least common classification, with only 3 to 8 windows being identified across all comparisons. Merging adjacent windows resulted in a reduced set of classified regions of increased size (supplementary table S5). This reduction process highlighted the tendency of windows classified as recurrent being situated amongst allopatric peaks, identifying 6-13 adjacent pairs of recurrent and allopatric windows. These results illustrate a varied divergence landscape across the genome (Figure 2D), indicating that single genome-wide divergence profiles do not capture the full extent of genetic differentiation (Figure 1B).

We next assessed how many classified windows were shared across the three species comparisons. Due to this approach identifying significantly more outlier windows than the *ZFst* approach, we opted to perform this analysis on the 50kb windows rather than the collapsed set to make the interpretation of overlapping windows easier. One hundred and seventeen allopatric windows were shared between at least two different comparisons, while 107 were shared by all three. An additional 32 windows were classified as allopatric or recurrent in more than one comparison. Twenty-one windows classified as “balancing selection” were shared by all comparisons, while windows classified as “divergence-with-gene flow” remained mostly unique to their respective comparison, with a single window overlapping across the comparison pairs (supplementary table S6).

We then merged adjacent classified windows across all comparisons and tested if the size of the shared regions were significantly larger than the regions that remained unique to each comparison. The general pattern was that shared regions were significantly larger after correcting for multiple testing, with this being especially true for overlapping regions amongst all three comparisons (pairwise Wilcoxon rank sum test, $p < 0.001$; supplementary table S7). Further, windows that were unique to specific comparisons often resided amongst windows of one classification type and were situated in genomic regions with similar divergence profiles across all comparisons (supplementary figure S2). In combination with the correlation analyses (Figure 1D), these results further support a heterogeneous divergence landscape that likely has shared ancestry within *Hydrophis*.

Historical gene flow and violation of tree-like species relationship

Hydrophis are believed to share a recent common ancestor within the last 5 million years. Previous work has established that species within *Hydrophis* likely experienced reticulate evolution, with new species occurring almost simultaneously (chapter 3). We formally tested for deviation from a strict bifurcating evolutionary history, along with signatures of introgression between *H. major, H. curtus* and *H. elegans*. Regardless of topology, *Dmin*, the minimum obtainable Patterson's D statistics after testing all possible topologies, remained positive and significant (*Dmin = 0.026, Z = 7.99, p-value < 0.001)*, further supporting that the evolutionary relationship between these snakes cannot be explained by a simple species tree (Figure 3A). Genome-wide estimates of admixture proportions were relatively low for all tests, ranging from 1.4 to 4.8%. Correlating windowed estimates of gene flow (f_d) with *Fst* resulted in a consistently weak negative

correlation in each species comparison (r = -0.19 to -0.30, p < 0.001), indicating that gene flow has likely had a role in homogenising genomic differentiation throughout the genome (Figure 3B).

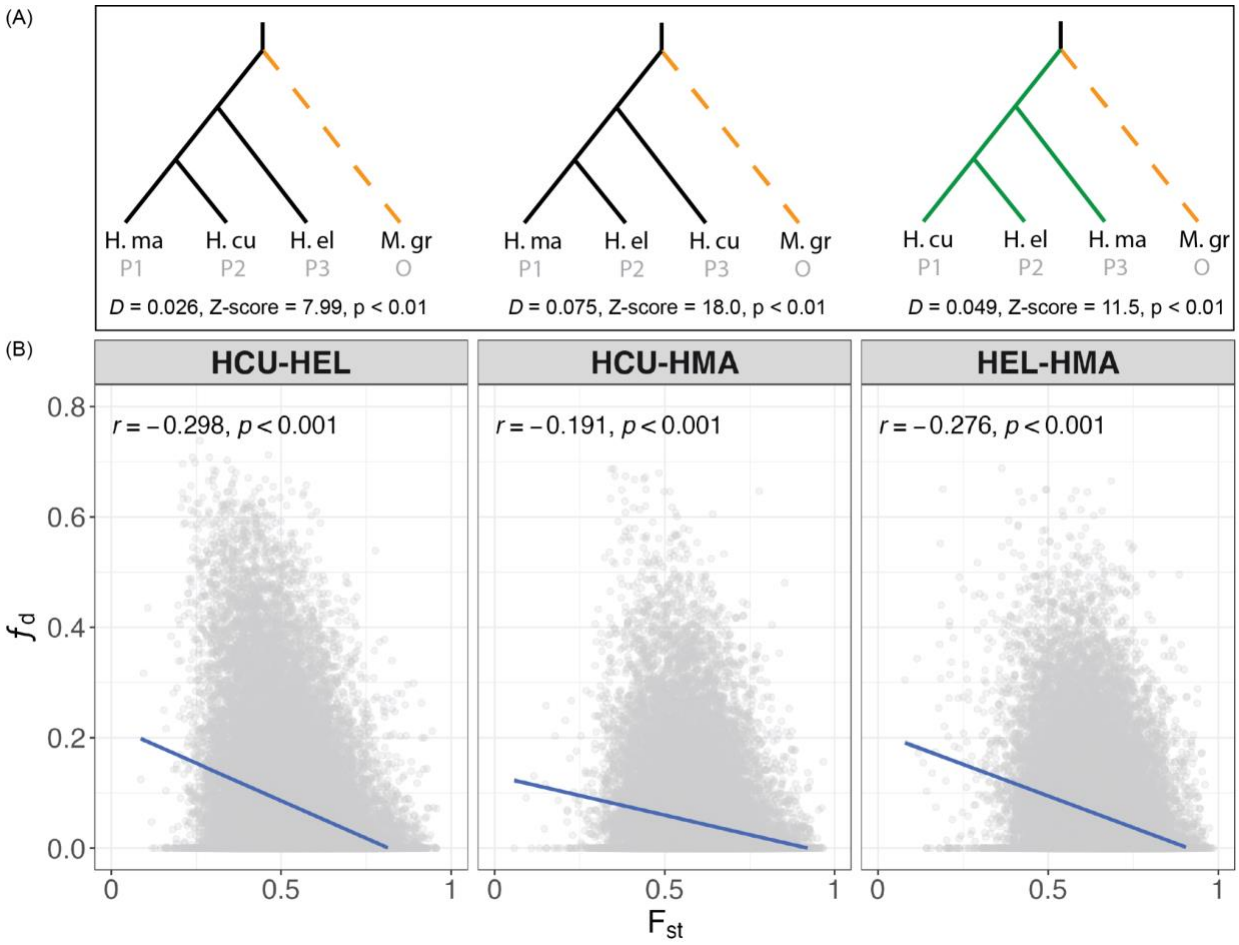


Figure 3: Three topologies used in *D* statistic calculations and correlation between windowed *Fst* against f_d. (A) Tree topologies and *D* statistic results for the three topologies tested to obtain *Dmin*. We note that rearranging P1 and P2 in each tree will simply reverse the sign of the *D*-statistic and have opted to use the arrangement output by *Dsuite*, which maintains positive *D* (i.e. excess of ABBA patterns). *Microcephalophus gracilis* was fixed as an outgroup in each test and is shown as the dashed orange branch in each tree. The tree highlighted in green represents the topology that produced

the most BBAA patterns. (B) Correlation plots between *Fst* and f_d for each species comparison. Windowed estimates of *Fst* are along the x-axis, with windowed f_d along the y-axis. Each point represents a 50kb window, with a line of best fit shown in blue.

Divergence profiles of curated outlier regions

Regions of high and low divergence were manually curated to identify key loci with functional relevance to speciation and/or adaptation. We note that while the start of chromosome two shows an excessive signal of low divergence/balancing selection, we opted to focus our curation efforts instead on regions outside of this region. The start of chromosome two is known to house extensive structural variation between *Hydrophis* snakes (Chapter 3, this thesis), adding a significant degree of difficulty in trying to interpret possible evolutionary scenarios. In total, we identified eight curated high-divergence regions, three of which were identified by both approaches, along with five low-divergence regions, of which three were identified by both the ZFst scans and divergence profiling. The size of high-divergence regions ranged from ~800Kbp to 12Mbp in size, and were significantly larger than low-divergence regions, whose sizes ranged from 0.4 to 1.1Mbp (Wilcoxon rank sum test, $p < 0.01$).

In all curated regions, the three species comparisons record almost identical patterns of sequence divergence, typically only differing slightly in the magnitude of a peak or trough. Regions of high-divergence conformed to “allopatric” models of speciation, with *Dxy* showing little deviation from the genomic background and π dipping over the span of the divergent windows (supplementary figures S3-S5). However, measures of *Tajima’s D* across high-divergence regions conflicted with this classification, often not deviating from the genome-wide pattern by substantial amounts

(Figure 4A), and in many cases, often increasing in at least one of the species rather than decreasing, as would be expected after a selective sweep. Conserved genomic regions all fit a profile matching “balancing selection”, recording increased π and *Dxy,* along with prominent increases in *Tajima’s D* which is indicative of “balancing” selection (Figure 4A; supplementary figures S3-5). We observed slightly elevated f_d at three high-divergence loci on chromosomes two between positions ~238 - 242Mbp and two at positions ~121.8-124.8Mb and ~138.9 - 144.9Mbp, providing potential target for future investigations. Notably, we also identified three instances of nested “balanced” loci within divergent peaks at positions ~399 - 400Mb on chromosome one, ~145 -146Mbp on chromosome three and ~68 - 69Mbp on chromosome five (supplementary figures S5).

We then explored the number of genes residing within the curated high- and low-divergence peaks. The number of genes within high-divergence peaks ranged from 11 to 116, while the smaller low-divergence regions contained anywhere from 4 to 34 genes (supplementary tables S8). There was a strong positive correlation in high-divergence regions between gene count and region size (r = 0.82, p-value ≤ 0.01), while conserved genomic loci reported an insignificant moderate positive correlation (r = 0.40, p-value = 0.44) (supplementary figure S6). Through manual inspection of the annotated genes within high- and low-divergent regions, we identified one balanced region of particular interest on chromosome one (0.8-2Mbp) that contained 33 vomeronasal type-2 receptors (V2R) genes (Figure 4B-C), which are important in feeding and mating behaviours.

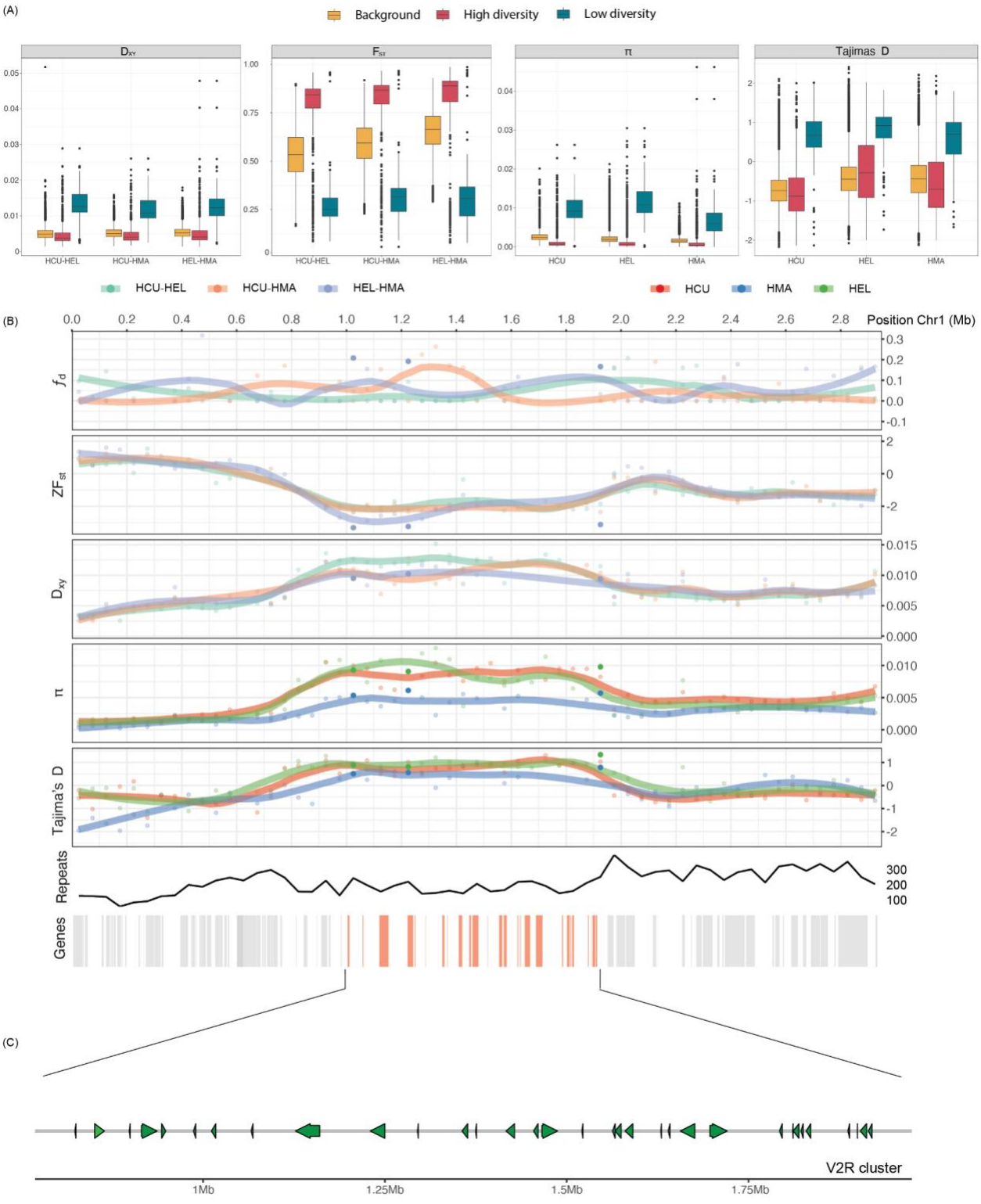


Figure 4: Divergence profile for all outlier windows (*ZFst* and profiling approach) detailing the patterns of divergence in each class relative to the background. (A)

Boxplots of diversity statistics between (*Fst/Dxy*) and within (*π/Tajima's D*) populations for high and low divergence regions, along with the genomic background. Yellow boxes represent the genomic background, which correspond to windows that were not identified either through the *ZFst* scan or divergence profiling analyses. Red boxes correspond to windows in high-divergence regions, while blue boxes correspond to windows in low-divergence regions. (B) Divergence landscape for a 1.2Mb "balanced" region containing a cassette of Vomeronasal genes. Genomic coordinates are presented along the x-axis (top) with the seven facets corresponding to: f_d, *ZFst*, *Dxy*, *π,* Tajima's D, repeat sequence density in 50kb bins and gene positions. Larger highlighted points (blue and green) represent *ZFst* identified windows. (C) Zoomed in view of the genes residing within the 'balanced" loci. Multiple vomeronasal genes are situated within the region on both the forward and reverse strand.

Discussion

We analysed full sequence level patterns of genome diversity and differentiation in *Hydrophis* sea snakes to provide new insights into the processes and genomic regions that underlie their anomalously fast rates of speciation and ecological adaptation. Introgression analyses recovered an excess of allele sharing among our three sampled species, a significant negative relationship between windowed estimates of introgression and relative differentiation, and low to moderate levels of shared variation among species. These results suggest significant interspecific hybridisation during the onset of the *Hydrophis* radiation (which coincides with the initial divergence of the sampled species) followed by increasing reproductive isolation. Such a scenario is not

unexpected given molecular clock estimates of ~3-8 million years for the origin of *Hydrophis*, and the clear morphological and molecular boundaries among the three species in sympatry. The evidence of early admixture is also consistent with recent phylogenomic analyses that rejected a bifurcating tree model for these and three other *Hydrophis* species (Chapter 3, this thesis), indicating that network approaches will be needed to understand interspecific relationships in this system.

Our interspecific comparisons of diversity and differentiation all contain a shared species. However, several consistent patterns in the three combinations of species pairs suggest conserved genome-wide processes in *Hydrophis major, H. elegans* and *H. curtus*. Correlations of windowed estimates of *Fst* and (separately) *Dxy* are strongly positive, and each comparison shows similar distributions of *Fst* and *Dxy*. Genome-wide *Dxy* and *Fst* were significantly correlated in each species comparison, but all showed a weak positive relationship. This pattern is closest to the expectations for models of allopatric selection, where species-specific background and/or positive selection drives *Fst* differentiation and *Dxy* reflects ancestral selection and diversity. However, the low strength of this correlation suggests that the signal has eroded with increasing divergence, presumably via the accumulated effects of lineage-specific processes. Nucleotide diversity was significantly positively correlated with *Dxy*, and significantly negatively correlated with *Fst*, in all comparisons. Very similar relationships between diversity versus absolute and relative differentiation have been reported in other systems (Shang et al., 2023), and follow expectations where the long-term impact of linked selection has accumulated across a stable genomic landscape. Finally, within-species estimates of nucleotide diversity and Tajima's D overlapped across species.

Together, these patterns suggest that the three species, despite their relatively deep divergence, have shared similar patterns of genomic diversity from ancestral to relatively recent timeframes.

Common genomic landscapes have been shaped by allopatric selection

Genomic islands of differentiation were also highly conserved among the species and showed heterogeneous patterns that conformed to the four main evolutionary scenarios. The majority (72-78%) of classified genomic windows in all comparisons fit a scenario of 'allopatric selection', where selection acts on each species independently causing low within-species diversity that drives high *Fst* and leaves *Dxy* unchanged. Windows of 'recurrent selection' were detected at the centre of many allopatric peaks. Recurrent selection reflects a similar process to allopatric selection but predicts lower *Dxy* due to the influence of linked selection in accelerating lineage sorting. Ancestral and species-specific selective sweeps have been shown to contribute to *Fst* peaks by reducing intraspecific variation at linked neutral sites. In all our *Fst* peaks, *Tajima's D* showed variable but mostly neutral patterns that do not conform with expectations of negative values (due to an excess of rare variants) after a recent selective sweep (Montejo-Kovacevich et al., 2022). However, as linkage disequilibrium decays with increasing time since speciation, it becomes difficult to detect sweeps using tests of the SFS such as *Tajima's D* (Stephan, 2019). The *Fst* peaks in sea snakes may have once experienced such selective pressures, but have since undergone a breakdown of linkage, and are now diverging under neutral processes.

High differentiation islands can be shaped by positive and/or background selection, i.e. negative selection against harmful variants (Cruickshank & Hahn, 2014; Ravinet et al., 2017). Background selection is often implicated where differentiation islands are conserved across speciation events due to the retention of stable patterns of recombination and gene density (Wolf & Ellegren, 2017). We were unable to measure recombination rates in this study, yet our high differentiation windows were substantial in size, which is a common characteristic of long term linked selection occurring in low recombination regions (Noor & Bennett, 2009). This observation, neutral *Dxy*, and the absence of species-specific differentiation islands, together suggest a dominant role for background selection in generating the high differentiation islands identified here.

Balancing selection maintains ancestral variation in chemoreception genes

Genomic windows showing exceptionally low *Fst* fit a scenario of balancing selection on ancestral polymorphisms, with π, *Dxy* and *Tajima's D* values that are all elevated above the genomic background. A large majority of these windows are located at the start of chromosome 2 (supplementary figure S1) and coincide with high densities of repeat elements and structural variations that might have suppressed recombination, extending this ~70Mb hotspot of balancing selection (Llaurens et al., 2017). However, due to the complexity of this region, we focused on curating the remaining balanced windows.

Curation of a 1.2Mb window at the start of chromosome one revealed a cluster of vomeronasal type-2 receptor (V2R) genes that underpin the highly tuned chemosensory systems of advanced snakes (Figure 4B-C). Odorants are transferred from the forked

tongue to a paired vomeronasal organ comprising an epithelium with receptor proteins encoded by V2R genes (Schwenk, 1994). Evidence of balancing selection on these genes is a significant finding because chemoreception via tongue flicking is critical in mediating the feeding and mating behaviours of sea snakes. Trophic specialisation is the primary axis of ecological differentiation in *Hydrophis,* and these species rely on chemoreception to find their (mostly) concealed prey in crevices and burrows (Kutsuma et al., 2018). Long-range mate searching is unlikely to involve chemoreception in marine environments (Shine, 2005). However, sea snakes recognise mates via chemoreception of pheromonal skin secretions during close courtship. Accordingly, very large numbers of intact V2R genes have been reported in some *Hydrophis* (including *H. curtus*) genomes (Kishida et al., 2019; Li et al., 2021).

Given these observations, V2Rs are highly plausible candidates as targets of balancing selection in sea snakes. Adaptive variation in this gene family is likely to be maintained in *Hydrophis* due to ontogenetic and environmental heterogeneity, and perhaps also frequency-dependent predator-prey interactions. In turn, retained polymorphisms must allow efficient adaptive responses, such as to new trophic opportunities or reinforcement of reproductive barriers. *Hydrophis* exhibit rapid evolutionary shifts in diet specialisation, such as between families of burrowing and crevice sheltering eels (Ophichthidae versus Congridae), and among species of toxic catfishes (Plotosidae) (Sherratt et al., 2018; Voris & Voris, 1983). Each sensory neuron expresses one or very few V2R genes (Wynn et al., 2012), and the repertoire of vomeronasal genes is expected to vary in proportion to the variety of odorants that can be detected (Young et al., 2010). Hence, it is likely that chemoreception in sea snakes

requires the preservation of adaptive combinations of V2Rs (Malnic et al., 1999). Understanding the recombination landscape of this region is therefore a high priority. It will also be important to understand the relationships among V2Rs that mediate mate versus prey recognition, particularly whether these co-occur in this window and are linked. Physical linkage of loci involved in ecological divergence and mate recognition appears to have accelerated speciation in some taxa (Boughman, 2002; Hench et al., 2019; Seehausen et al., 2008), but has not yet been reported in reptiles.

Author contributions

AJL and KLS conceived the research. AJL performed quantitative analyses. Both AJL and KLS interpreted results and drafted the manuscript.

Acknowledgements

We thank Vinay Udyawer, Jenna Crowe-Riddell, James Nankivell and Balazs Buzas for their help with sample collection. We would also like to thank Chris Ward for their valuable discussions relating to this work.

Data availability

Code and result files used in these analyses are available on GitHub at https://github.com/a-lud/hydrophis-genome-scans/tree/main

References

Andrew, R. L., & Rieseberg, L. H. (2013). Divergence Is Focused on Few Genomic

Regions Early in Speciation: Incipient Speciation of Sunflower Ecotypes. *Evolution*, *67*(9), 2468–2482. https://doi.org/10.1111/evo.12106

Boughman, J. W. (2002). How sensory drive can promote speciation. *Trends in Ecology & Evolution*, *17*(12), 571–577. https://doi.org/10.1016/S0169-5347(02)02595-8

C. Fry, G., A. Milton, D., & J. Wassenberg, T. (2001). The reproductive biology and diet of sea snake bycatch of prawn trawling in northern Australia: Characteristics important for assessing the impacts on populations. *Pacific Conservation Biology*, *7*(1), 55–73.

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., & Bork, P. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, 8.

Chen, S. (2023). Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *IMeta*, *2*(2), e107. https://doi.org/10.1002/imt2.107

Christmas, M. J., Jones, J. C., Olsson, A., Wallerman, O., Bunikis, I., Kierczak, M., Peona, V., Whitley, K. M., Larva, T., Suh, A., Miller-Struttmann, N. E., Geib, J. C., & Webster, M. T. (2021). Genetic Barriers to Historical Gene Flow between Cryptic Species of Alpine Bumblebees Revealed by Comparative Population Genomics. *Molecular Biology and Evolution*, *38*(8), 3126–3143. https://doi.org/10.1093/molbev/msab086

Cruickshank, T. E., & Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, *23*(13). https://doi.org/10.1111/mec.12796

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A.,

Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Ellegren, H., Smeds, L., Burri, R., Olason, P. I., Backström, N., Kawakami, T., Künstner, A., Mäkinen, H., Nadachowska-Brzyska, K., Qvarnström, A., Uebbing, S., & Wolf, J. B. W. (2012). The genomic landscape of species divergence in Ficedula flycatchers. *Nature*, *491*(7426), 756–760. https://doi.org/10.1038/nature11584

Han, F., Lamichhaney, S., Grant, B. R., Grant, P. R., Andersson, L., & Webster, M. T. (2017). Gene flow, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence among Darwin's finches. *Genome Research*, *27*(6), 1004–1015. https://doi.org/10.1101/gr.212522.116

Hench, K., Vargas, M., Höppner, M. P., McMillan, W. O., & Puebla, O. (2019). Inter-chromosomal coupling between vision and pigmentation genes during genomic divergence. *Nature Ecology & Evolution*, *3*(4), 657–667. https://doi.org/10.1038/s41559-019-0814-5

Huang, K., Andrew, R. L., Owens, G. L., Ostevik, K. L., & Rieseberg, L. H. (2020). Multiple chromosomal inversions contribute to adaptive divergence of a dune sunflower ecotype. *Molecular Ecology*, *29*(14), 2535–2549.

https://doi.org/10.1111/mec.15428

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., & Bork, P. (2019). eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, *47*(D1), D309–D314. https://doi.org/10.1093/nar/gky1085

Irwin, D. E., Milá, B., Toews, D. P. L., Brelsford, A., Kenyon, H. L., Porter, A. N., Grossen, C., Delmore, K. E., Alcaide, M., & Irwin, J. H. (2018). A comparison of genomic islands of differentiation across three young avian species pairs. *Molecular Ecology*, *27*(23), 4839–4855. https://doi.org/10.1111/mec.14858

Kautt, A. F., Kratochwil, C. F., Nater, A., Machado-Schiaffino, G., Olave, M., Henning, F., Torres-Dowdall, J., Härer, A., Hulsey, C. D., Franchini, P., Pippel, M., Myers, E. W., & Meyer, A. (2020). Contrasting signatures of genomic divergence during sympatric speciation. *Nature*. https://doi.org/10.1038/s41586-020-2845-0

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. R. Soc. B.*, *286*(1910), 20191828. https://doi.org/10.1098/rspb.2019.1828

Korunes, K. L., & Samuk, K. (2021). pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular Ecology Resources*, *21*(4), 1359–1368. https://doi.org/10.1111/1755-0998.13326

Kronforst, M. R., Hansen, M. E. B., Crawford, N. G., Gallant, J. R., Zhang, W.,

Kulathinal, R. J., Kapan, D. D., & Mullen, S. P. (2013). Hybridization Reveals the Evolving Genomic Architecture of Speciation. *Cell Reports*, *5*(3), 666–677. https://doi.org/10.1016/j.celrep.2013.09.042

Kulmuni, J., Nouhaud, P., Pluckrose, L., Satokangas, I., Dhaygude, K., & Butlin, R. K. (2020). Instability of natural selection at candidate barrier loci underlying speciation in wood ants. *Molecular Ecology*, *29*(20), 3988–3999. https://doi.org/10.1111/mec.15606

Kutsuma, R., Sasai, T., & Kishida, T. (2018). How Snakes Find Prey Underwater: Sea Snakes Use Visual and Chemical Cues for Foraging. *Zoological Science*, *35*(6), 483. https://doi.org/10.2108/zs180059

Lee, M. S. Y., Sanders, K. L., King, B., & Palci, A. (2016). Diversification rates and phenotypic evolution in venomous snakes (Elapidae). *R. Soc. Open Sci.*, *3*(1), 150277. https://doi.org/10.1098/rsos.150277

Li, A., Wang, J., Sun, K., Wang, S., Zhao, X., Wang, T., Xiong, L., Xu, W., Qiu, L., Shang, Y., Liu, R., Wang, S., & Lu, Y. (2021). Two Reference-Quality Sea Snake Genomes Reveal Their Divergent Evolution of Adaptive Traits and Venom Systems. *Molecular Biology and Evolution*, msab212. https://doi.org/10.1093/molbev/msab212

Liu, X., Glémin, S., & Karrenberg, S. (2020). Evolution of putative barrier loci at an intermediate stage of speciation with gene flow in campions (*Silene*). *Molecular Ecology*, *29*(18), 3511–3525. https://doi.org/10.1111/mec.15571

Llaurens, V., Whibley, A., & Joron, M. (2017). Genetic architecture and balancing selection: The life and death of differentiated variants. *Molecular Ecology*, *26*(9),

2430–2448. https://doi.org/10.1111/mec.14051

Ma, T., Wang, K., Hu, Q., Xi, Z., Wan, D., Wang, Q., Feng, J., Jiang, D., Ahani, H., Abbott, R. J., Lascoux, M., Nevo, E., & Liu, J. (2018). Ancient polymorphisms and divergence hitchhiking contribute to genomic islands of divergence within a poplar species complex. *Proc Natl Acad Sci USA*, *115*(2), E236–E243. https://doi.org/10.1073/pnas.1713288114

Malinsky, M., Matschiner, M., & Svardal, H. (2021). Dsuite—Fast D-statistics and related admixture evidence from VCF files. *Molecular Ecology Resources*, *21*(2), 584–595. https://doi.org/10.1111/1755-0998.13265

Malnic, B., Hirono, J., Sato, T., & Buck, L. B. (1999). Combinatorial Receptor Codes for Odors. *Genome Research*, *9*, 713–723. https://doi.org/10.1016/S0092-8674(00)80581-4

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., Blaxter, M., Manica, A., Mallet, J., & Jiggins, C. D. (2013). Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res.*, *23*(11), 1817–1828. https://doi.org/10.1101/gr.159426.113

Martin, S. H., Davey, J. W., & Jiggins, C. D. (2015). Evaluating the Use of ABBA–BABA Statistics to Locate Introgressed Loci. *Molecular Biology and Evolution*, *32*(1), 244–257. https://doi.org/10.1093/molbev/msu269

Martin, S. H., Möst, M., Palmer, W. J., Salazar, C., McMillan, W. O., Jiggins, F. M., & Jiggins, C. D. (2016). Natural Selection and Genetic Diversity in the Butterfly *Heliconius* melpomene. *Genetics*, *203*(1), 525–541. https://doi.org/10.1534/genetics.115.183285

Meier, J. I., Marques, D. A., Wagner, C. E., Excoffier, L., & Seehausen, O. (2018). Genomics of Parallel Ecological Speciation in Lake Victoria Cichlids. *Molecular Biology and Evolution*, *35*(6), 1489–1506. https://doi.org/10.1093/molbev/msy051

Meier, J. I., Sousa, V. C., Marques, D. A., Selz, O. M., Wagner, C. E., Excoffier, L., & Seehausen, O. (2017). Demographic modelling with whole-genome data reveals parallel origin of similar Pundamilia cichlid species after hybridization. *Molecular Ecology*, *26*(1), 123–141. https://doi.org/10.1111/mec.13838

Montejo-Kovacevich, G., Meier, J. I., Bacquet, C. N., Warren, I. A., Chan, Y. F., Kucka, M., Salazar, C., Rueda-M, N., Montgomery, S. H., McMillan, W. O., Kozak, K. M., Nadeau, N. J., Martin, S. H., & Jiggins, C. D. (2022). Repeated genetic adaptation to altitude in two tropical butterflies. *Nature Communications*, *13*(1), 4676. https://doi.org/10.1038/s41467-022-32316-x

Nadachowska-Brzyska, K., Burri, R., Smeds, L., & Ellegren, H. (2016). PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white *Ficedula* flycatchers. *Mol Ecol*, *25*(5), 1058–1072. https://doi.org/10.1111/mec.13540

Noor, M. A. F., & Bennett, S. M. (2009). Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity*, *103*(6), 439–444. https://doi.org/10.1038/hdy.2009.151

Pedersen, B. S., & Quinlan, A. R. (2018). Mosdepth: Quick coverage calculation for genomes and exomes. *Bioinformatics*, *34*(5), 867–868. https://doi.org/10.1093/bioinformatics/btx699

Peng, C., Wu, D.-D., Ren, J.-L., Peng, Z.-L., Ma, Z., Wu, W., Lv, Y., Wang, Z., Deng, C.,

Jiang, K., Parkinson, C. L., Qi, Y., Zhang, Z.-Y., & Li, J.-T. (2023). Large-scale snake genome analyses provide insights into vertebrate development. *Cell*, *186*(14), 2959-2976.e22. https://doi.org/10.1016/j.cell.2023.05.030

Ravinet, M., Faria, R., Butlin, R. K., Galindo, J., Bierne, N., Rafajlović, M., Noor, M. a. F., Mehlig, B., & Westram, A. M. (2017). Interpreting the genomic landscape of speciation: A road map for finding barriers to gene flow. *Journal of Evolutionary Biology*, *30*(8), 1450–1477. https://doi.org/10.1111/jeb.13047

Renaut, S., Grassa, C. J., Yeaman, S., Moyers, B. T., Lai, Z., Kane, N. C., Bowers, J. E., Burke, J. M., & Rieseberg, L. H. (2013). Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nature Communications*, *4*(1), 1827. https://doi.org/10.1038/ncomms2833

Sanders, K. L., Lee, M. S. Y., Mumpuni, Bertozzi, T., & Rasmussen, A. R. (2013). Multilocus phylogeny and recent rapid radiation of the viviparous sea snakes (Elapidae: Hydrophiinae). *Molecular Phylogenetics and Evolution*, *66*(3), 575–591. https://doi.org/10.1016/j.ympev.2012.09.021

Sanders, K. L., Mumpuni, & Lee, M. S. Y. (2010). Uncoupling ecological innovation and speciation in sea snakes (Elapidae, Hydrophiinae, Hydrophiini). *Journal of Evolutionary Biology*, *23*(12), 2685–2693. https://doi.org/10.1111/j.1420-9101.2010.02131.x

Sanders, K. L., Rasmussen, A. R., Mumpuni, Elmberg, J., Silva, A. de, Guinea, M. L., & Lee, M. S. Y. (2013). Recent rapid speciation and ecomorph divergence in Indo-Australian sea snakes. *Molecular Ecology*, *22*(10), 2742–2759. https://doi.org/10.1111/mec.12291

Schwenk, K. (1994). Why Snakes Have Forked Tongues. *Science*, *263*(5153), 1573–1577. https://doi.org/10.1126/science.263.5153.1573

Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. A., Peichel, C. L., Saetre, G.-P., Bank, C., Brännström, Å., Brelsford, A., Clarkson, C. S., Eroukhmanoff, F., Feder, J. L., Fischer, M. C., Foote, A. D., Franchini, P., Jiggins, C. D., Jones, F. C., … Widmer, A. (2014). Genomics and the origin of species. *Nat Rev Genet*, *15*(3), 176–192. https://doi.org/10.1038/nrg3644

Seehausen, O., Terai, Y., Magalhaes, I. S., Carleton, K. L., Mrosso, H. D. J., Miyagi, R., Van Der Sluijs, I., Schneider, M. V., Maan, M. E., Tachida, H., Imai, H., & Okada, N. (2008). Speciation through sensory drive in cichlid fish. *Nature*, *455*(7213), 620–626. https://doi.org/10.1038/nature07285

Shang, H., Field, D. L., Paun, O., Rendón-Anaya, M., Hess, J., Vogl, C., Liu, J., Ingvarsson, P. K., Lexer, C., & Leroy, T. (2023). Drivers of genomic landscapes of differentiation across a Populus divergence gradient. *Molecular Ecology*, *32*(15), 4348–4361. https://doi.org/10.1111/mec.17034

Sherratt, E., Rasmussen, A. R., & Sanders, K. L. (2018). Trophic specialization drives morphological evolution in sea snakes. *Royal Society Open Science*, *5*(3), 172141. https://doi.org/10.1098/rsos.172141

Shine, R. (2005). All at sea: Aquatic life modifies mate-recognition modalities in sea snakes (Emydocephalus annulatus, Hydrophiidae). *Behavioral Ecology and Sociobiology*, *57*(6), 591–598. https://doi.org/10.1007/s00265-004-0897-z

Shine, R., Shine, T., & Goiran, C. (2019). Morphology, reproduction and diet of the

greater sea snake, Hydrophis major (Elapidae, Hydrophiinae). *Coral Reefs*, *38*(5), 1057–1064. https://doi.org/10.1007/s00338-019-01833-5

Simões, B. F., Gower, D. J., Rasmussen, A. R., Sarker, M. A. R., Fry, G. C., Casewell, N. R., Harrison, R. A., Hart, N. S., Partridge, J. C., Hunt, D. M., Chang, B. S., Pisani, D., & Sanders, K. L. (2020). Spectral Diversification and Trans-Species Allelic Polymorphism during the Land-to-Sea Transition in Snakes. *Current Biology*, *30*(13), 2608-2615.e4. https://doi.org/10.1016/j.cub.2020.04.061

Stephan, W. (2019). Selective Sweeps. *Genetics*, *211*(1), 5–13. https://doi.org/10.1534/genetics.118.301319

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., & Prins, P. (2015). Sambamba: Fast processing of NGS alignment formats. *Bioinformatics*, *31*(12), 2032–2034. https://doi.org/10.1093/bioinformatics/btv098

Vasimuddin, Md., Misra, S., Li, H., & Aluru, S. (2019). Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 314–324. https://doi.org/10.1109/IPDPS.2019.00041

Voris, H. K., & Voris, H. H. (1983). Feeding Strategies in Marine Snakes: An Analysis of Evolutionary, Morphological, Behavioral and Ecological Relationships. *American Zoologist*, *23*(2), 411–425. https://doi.org/10.1093/icb/23.2.411

Wolf, J. B. W., & Ellegren, H. (2017). Making sense of genomic islands of differentiation in light of speciation. *Nat Rev Genet*, *18*(2), 87–100. https://doi.org/10.1038/nrg.2016.133

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with

Kraken 2. *Genome Biology*, *20*(1), 257. https://doi.org/10.1186/s13059-019-1891-0

Wynn, E. H., Sánchez-Andrade, G., Carss, K. J., & Logan, D. W. (2012). Genomic variation in the vomeronasal receptor gene repertoires of inbred mice. *BMC Genomics*, *13*(1), 415. https://doi.org/10.1186/1471-2164-13-415

Young, J. M., Massa, H. F., Hsu, L., & Trask, B. J. (2010). Extreme variability among mammalian V1R gene families. *Genome Research*, *20*(1), 10–18. https://doi.org/10.1101/gr.098913.109

Supplementary figures

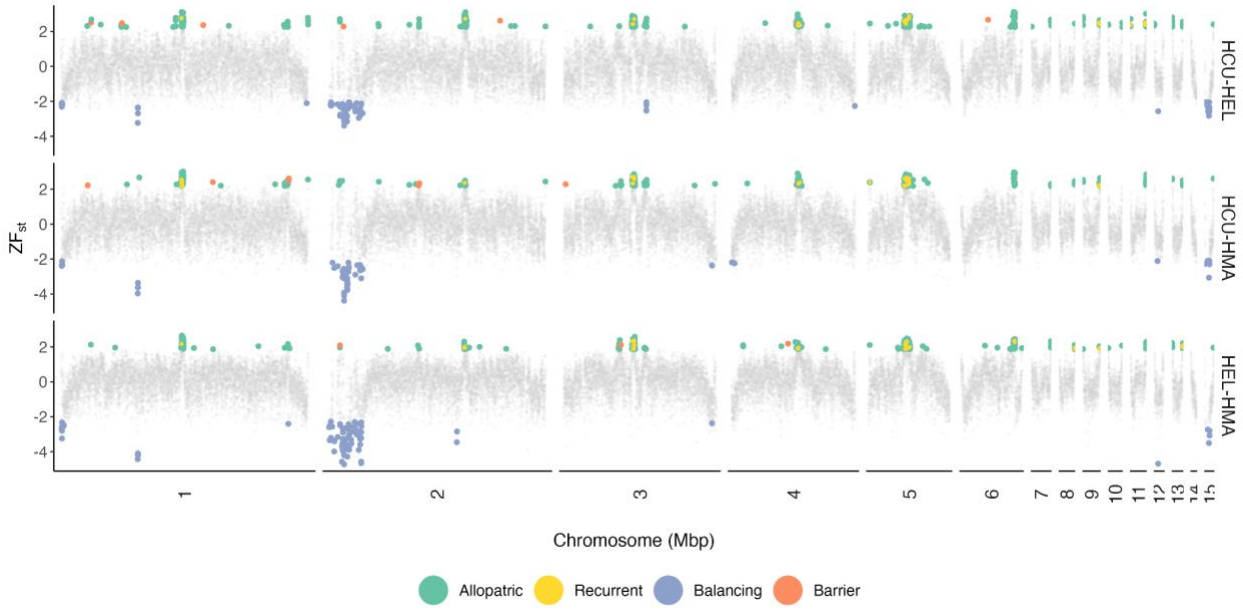


Figure S1: Genome-wide view of classified windows. Chromosomes are represented along the x-axis and *ZFst* is along the y-axis. Each row represents a different species comparison. Green points correspond to windows classified as “allopatric” selection, yellow points represent “recurrent” selection and are typically at the centre of allopatric peaks. Purple points correspond to windows classified as “balancing” selection and are predominantly localised to the start of chromosome two. Finally, orange points correspond to “barrier” loci, or “divergence-with-gene flow” loci.

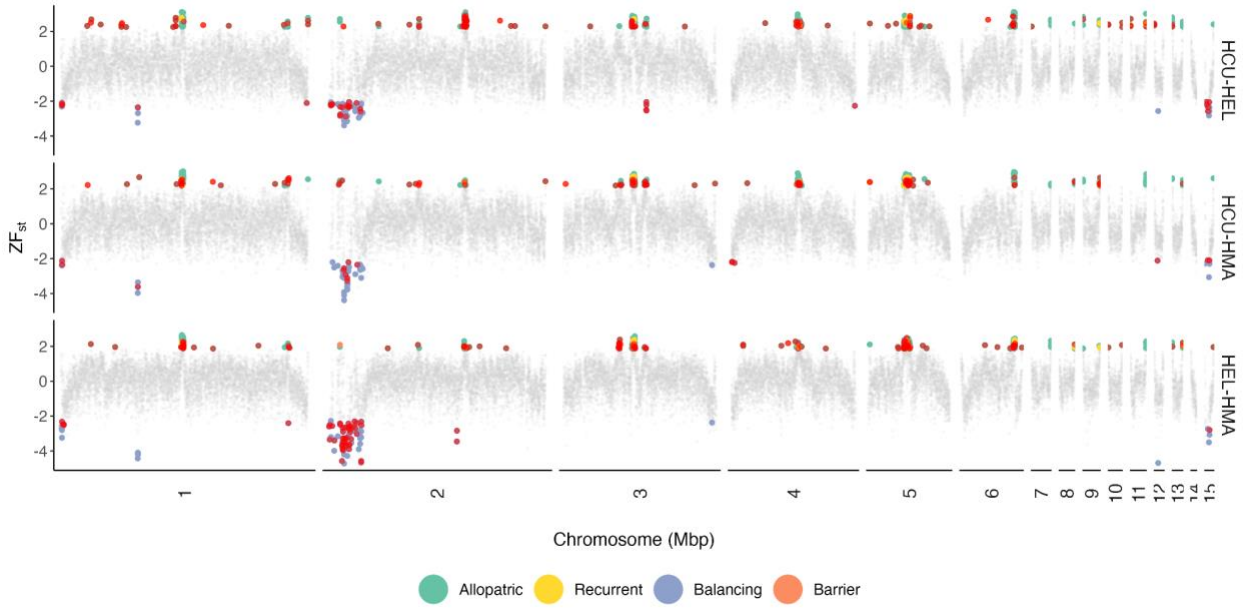


Figure S2: Genome wide plot highlighting genomic windows that were not adjacent to any other classified window. Chromosomes are represented along the x-axis and *ZFst* is along the y-axis. Each row represents a different species comparison. Red points represent “unique” windows; however, it is evident that nearly all these windows reside in classified regions.

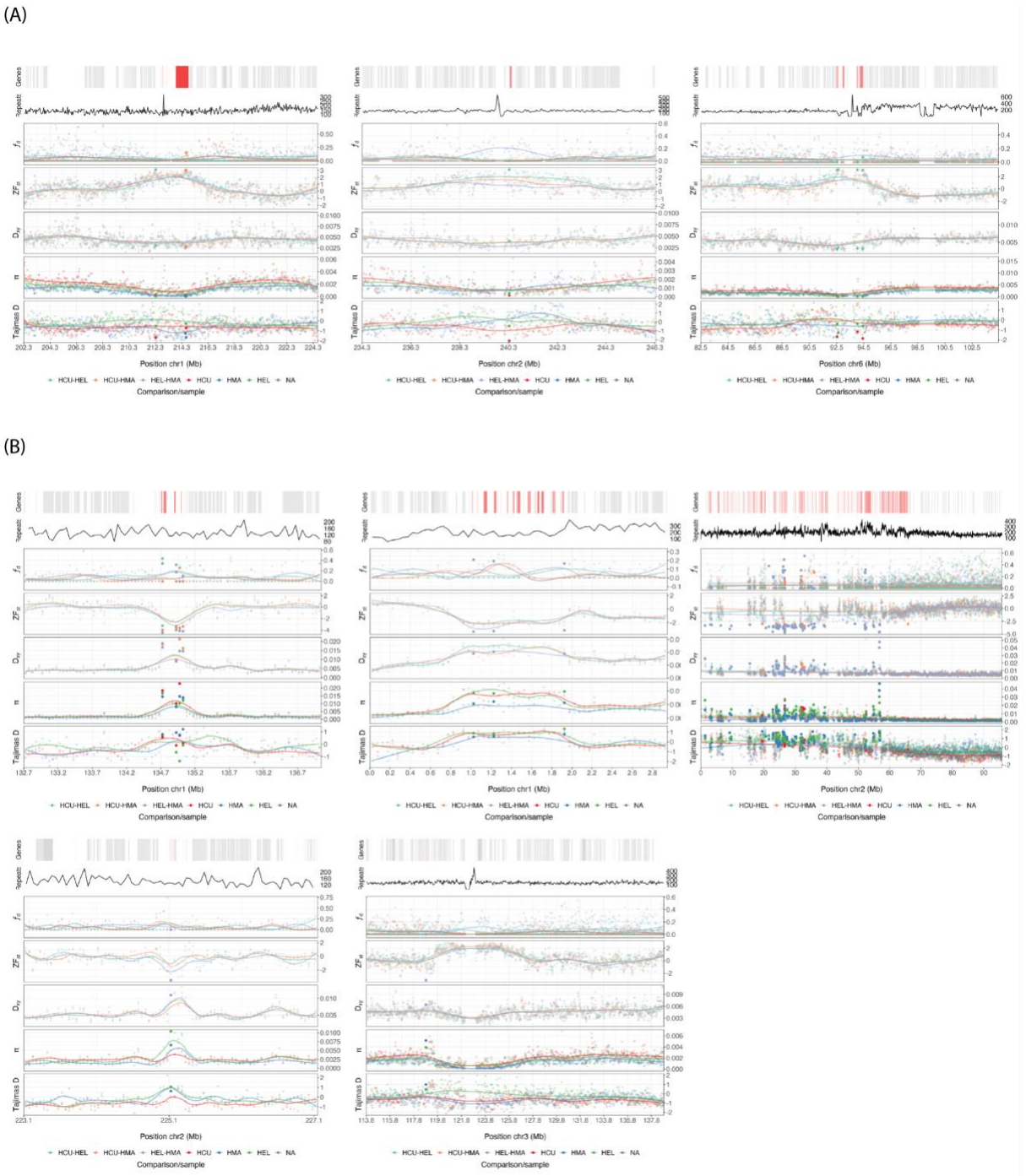


Figure S3: Curated *ZFst* regions for both high (A) and low (B) divergence regions. Some of the regions identified above overlap with those reported in supplementary figures S4 and S5. In both high- and low-divergence figures, the facets correspond to the following (from top to bottom): Gene track, repeat density track, f_d, *ZFst*, *Dxy*, π and Tajima's D.

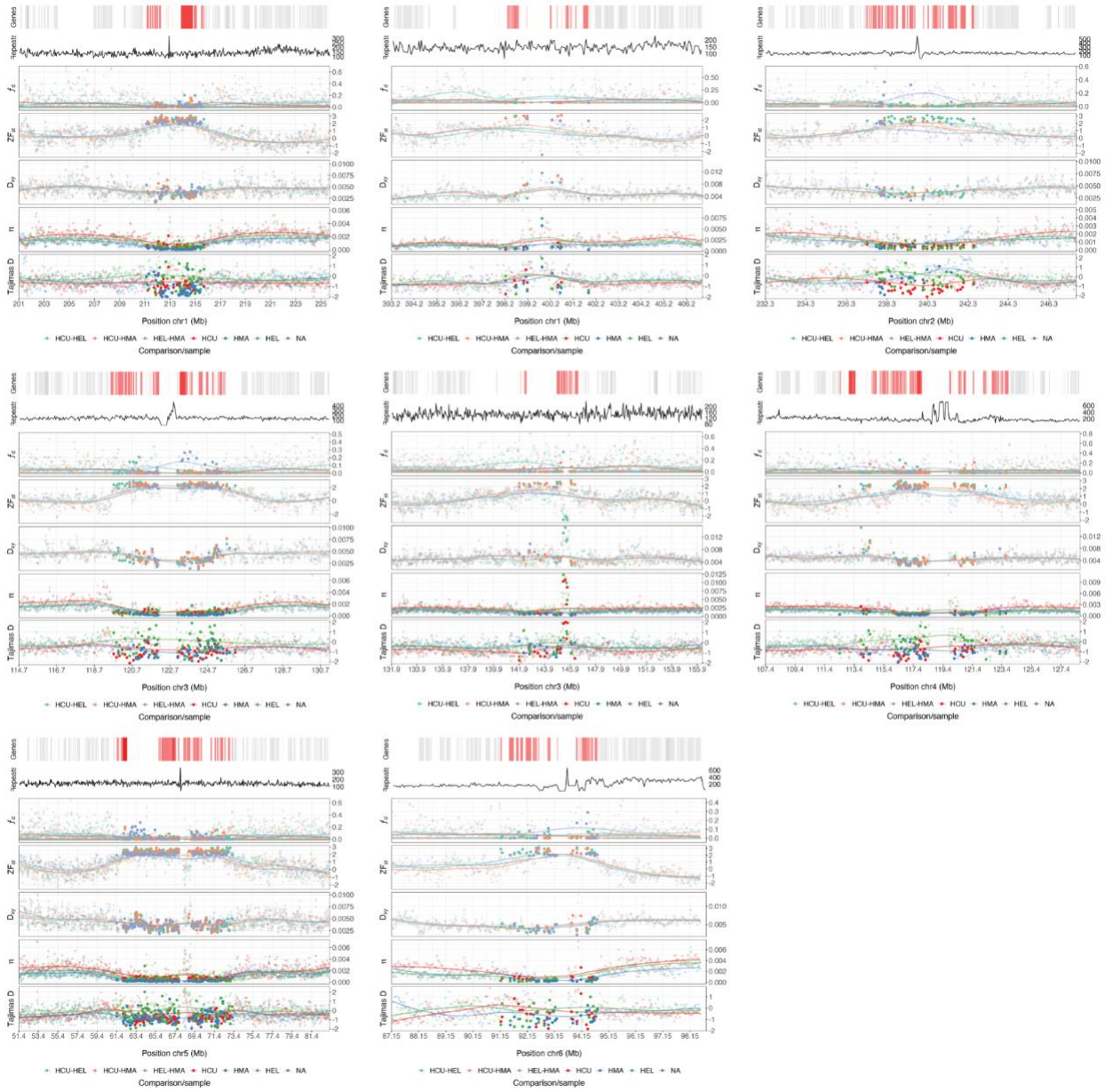


Figure S4: Curated high-divergence (allopatric) loci identified by the profiling approach. Each facet corresponds to a different genomic region. Notably, there are considerably more windows within each of the regions that have been identified using this approach (saturated points) compared to the *ZFst* approach in supplementary figure S3. The facets correspond to the following (from top to bottom): Gene track, repeat density track, f_d, *ZFst*, *Dxy*, π and Tajima's D.

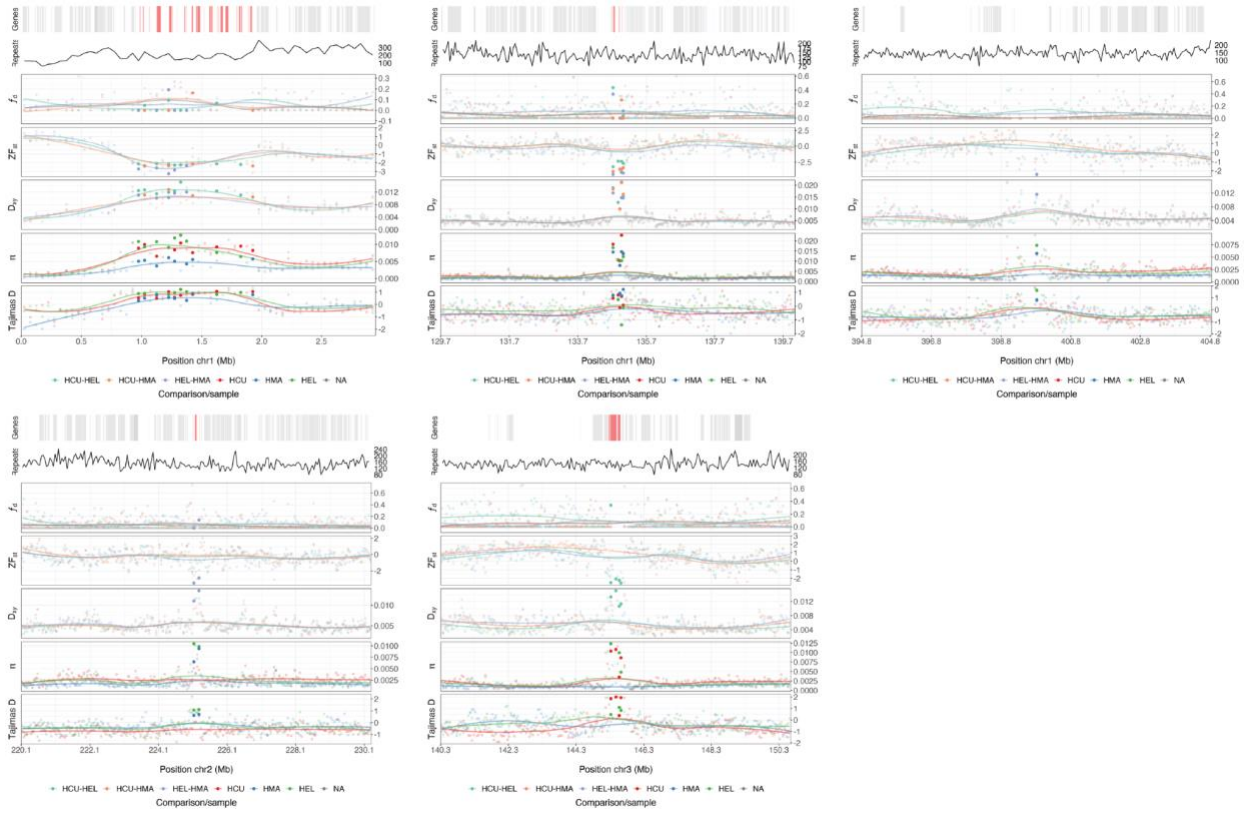


Figure S5: Curated low-divergence regions identified by the profiling approach. Each facet corresponds to a different genomic region. Notably, there are considerably more windows within each of the regions that have been identified using this approach (saturated points) compared to the *ZFst* approach in supplementary figure S3. The facets correspond to the following (from top to bottom): Gene track, repeat density track, f_d, *ZFst*, *Dxy*, π and Tajima's D.

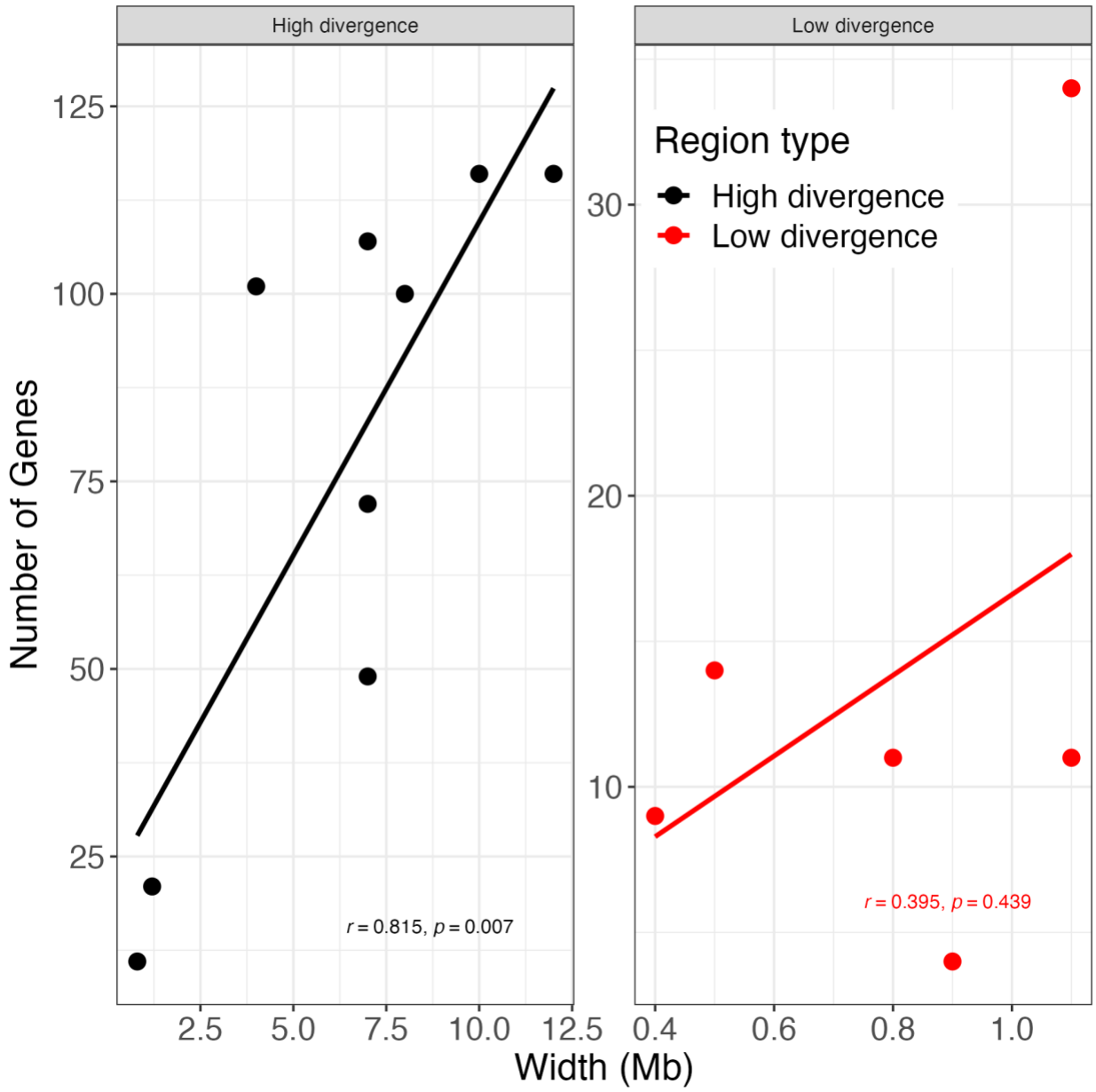


Figure S6: Pearson correlation between the size of the genomic regions and gene number. High divergence regions have a significant positive correlation between region size and gene number. Low divergence regions see a moderate positive correlation, but it is insignificant.

Supplementary tables

Table S1: Number of sites called (variant and invariant).

Within populations	Genome size (autosomes)	Combined	Invariant	Variant
HCU	1,793,465,688	951,899,031	943,369,442	8,529,589
HMA	1,793,465,688	747,014,765	743,295,430	3,719,335
HEL	1,793,465,688	656,347,619	652,140,093	4,207,526
Between populations				
HCU-HEL	1,793,465,688	876,005,580	862,227,509	13,778,071
HEL-HMA	1,793,465,688	673,611,934	664,833,250	8,778,684
HCU-HMA	1,793,465,688	577,984,682	569,230,940	8,753,742

Table S2: Number of outlier regions after collapsing adjacent 50kb outlier windows within comparisons.

Comparison	Window Type	50kb	100kb	150kb	200kb	250kb
HCU-HEL	*ZFst* ≥ 3	8	0	0	0	0
HCU-HEL	*ZFst* ≤ -3	3	0	1	0	0
HCU-HMA	*ZFst* ≤ -3	15	2	1	1	0
HCU-HMA	*ZFst* ≥ 3	0	1	0	0	0
HEL-HMA	*ZFst* ≤ -3	53	11	2	0	3

Table S3: Number of shared outlier regions after collapsing overlapping or adjacent windows between comparisons.

Comparisons	Window type	50kb	100kb	150kb	250kb	200kb
HCU-HEL	ZFst >= 3	7	0	0	0	0
HCU-HMA	ZFst <= -3	5	1	0	0	0
HEL-HMA	ZFst <= -3	44	9	1	1	0
HCU-HEL|HCU-HMA	ZFst >= 3	0	1	0	0	0
HCU-HMA|HEL-HMA	ZFst <= -3	6	1	0	1	1
HCU-HEL|HCU-HMA|HEL-HMA	ZFst <= -3	1	0	2	1	0

Table S4: Number of 50kb windows assigned to a profile in each species comparison.

Comparisons	Allopatric	Recurrent	Balancing	Div. with gene flow	Total
HCU-HEL	307	20	75	8	410
HCU-HMA	298	27	50	8	383
HEL-HMA	310	22	94	3	429

Table S5: Number of regions after collapsing adjacent windows within comparisons of species.

Comparisons	Allopatric	Allopatric|Recurrent	Allopatric|Div. with gene flow	Recurrent	Balancing	Barrier	Total
HCU-HEL	206	6	0	11	56	8	287
HCU-HMA	165	13	0	9	38	8	233
HEL-HMA	158	12	1	9	65	2	247

Table S6: Number of 50kb profiled windows shared between the species comparisons.

Comparisons	Allopatric	Recurrent	Balancing	Div. with gene flow	Allopatric|Recurrent	Allopatric|Div. with gene flow	Total
HCU-HEL	107	1	32	7	0	0	147
HCU-HMA	82	4	16	7	0	0	109
HEL-HMA	117	1	52	1	0	0	171
HCU-HEL HCU-HMA	48	2	7	1	2	0	60
HCU-HEL HEL-HMA	25	0	15	0	3	1	44
HCU-HMA HEL-HMA	44	1	6	0	3	1	55
HCU-HEL HCU-HMA HEL-HMA	107	7	21	0	24	0	159

Table S7: Wilcoxon Rank Sum Test p-value statistics

Comparison	HCU-HEL	HCU-HMA	HEL-HMA	HCU-HEL HCU-HMA	HCU-HEL HEL-HMA	HCU-HMA HEL-HMA
HCU-HMA	1					
HEL-HMA	1	1				
HCU-HEL						
HCU-HMA	0.054	0.048	0.076			
HCU-HEL						
HEL-HMA	0.055	0.048	0.081	1		
HCU-HMA						
HEL-HMA	0.0038	0.0039	0.0064	1	1	
HCU-HEL						
HCU-HMA						
HEL-HMA	< 0.001	< 0.001	< 0.001	< 0.001	0.0091	0.0081

Table S8: Number of genes in curated regions

Classification	Chromosome	Start	End	Width	No. genes
High. div	chr1	400,200,000	401,000,000	800,000	11
High. div	chr1	398,000,000	399,200,000	1,200,000	21
High. div	chr6	91,150,000	95,150,000	4,000,000	100
High. div	chr2	236,300,000	243,300,000	7,000,000	107
High. div	chr3	119,700,000	126,700,000	7,000,000	72
High. div	chr3	139,900,000	146,900,000	7,000,000	49
High. div	chr1	210,300,000	218,300,000	8,000,000	100
High. div	chr4	113,400,000	123,400,000	10,000,000	115
High. div	chr5	61,400,000	73,400,000	12,000,000	116
Low div.	chr2	225,000,000	225,400,000	400,000	9
Low div.	chr1	399,200,000	40,010,0000	900,000	4
Low div.	chr1	800,000	2,000,000	1,200,000	38

Chapter 5:
General discussion

5.1 General discussion

My thesis comprises three data chapters in the format of full-length research papers. Each chapter generated substantial new datasets and involved the development of analysis pipelines, contributing important resources for future research on this and other study systems. However, in addition to their individual impacts, these chapters collectively provide a new understanding of sea snake evolution that links multiple complex biological processes in the same study taxon.

Coalescent models of historical patterns of effective population size (N_e) using three marine and three terrestrial genomes showed that sea snakes have large effective population sizes, in line with expectations for adaptive radiations of species with extensive geographic ranges (Chapter 2). However, counter to the expectation that marine species should show higher overall N_e and less pronounced fluctuations in N_e compared to terrestrial species (Carr et al., 2003; Grosberg et al., 2012; Lourie & Vincent, 2004; Mayr, 1954), demographic patterns were found to be highly variable among species and did not clearly correspond to major ecological divisions. Future studies could usefully reconstruct ancestral changes in N_e during the initial land-sea transition of sea snakes and at the early stages of the explosive *Hydrophis* radiation. These results would contribute to a general understanding of the shifts in N_e that accompany major colonisation events and upticks in species diversification (MacArthur et al., 1972). Genomic estimates of N_e can also provide valuable information for assessments of conservation status. A priority here should be to apply the analysis pipelines developed in Chapter 2 to the critically endangered sea snakes *Aipysurus*

apraefrontalis and *A. foliosquama*, which have undergone dramatic population declines and local extinctions in the Timor Sea.

The large effective population sizes of sea snakes were consistent with subsequent discoveries of extensive incomplete lineage sorting and introgression (Chapters 3 and 4), both of which are more likely when species descend from large ancestral populations. In Chapter 3, phylogenomic analyses of six *Hydrophi*s supported a reticulate versus bifurcating tree model, only finding between 21.6 to 29.7% concordance amongst gene trees and the species tree. These results were mirrored by introgression tests between three ecologically distinct *Hydrophis* species, which identified evidence of early admixture between the three *Hydrophis* (Chapter 4). Future work could build on these results by quantifying the extent and direction of gene flow, and the role of adaptive introgression in more recently diverged pairs of *Hydrophis* species.

Chapter 3 additionally provided the first detailed view of gene selection in the recent transition from terrestrial (e.g. Australian tiger and brown snakes) to fully marine (*Hydrophis*) snakes. Positively selected genes span functional categories including hypoxia adaptation, sensory perception, immune response, and morphological development. These results fill a conspicuous taxonomic gap in genomic studies of marine adaptation. However, many questions remain to be addressed, particularly how the genetic signatures reported here are linked to the marine phenotypes of sea snakes.

While identifying candidate genes under divergent selection pressures is important for understanding the genetic basis of adaptive radiations, analyses of sequence level patterns of diversity and differentiation, and genomic architecture,

provide complementary insights into the evolution of species and adaptive traits (Campbell et al., 2018; Seehausen et al., 2014; Wolf & Ellegren, 2017). Comparisons of whole genome sequences between six *Hydrophis* snakes identified a relatively consistent karyotype (2n = 32) and significant levels of chromosomal synteny (chapter 3). While the karyotype of *Hydrophis* varied significantly to a terrestrial outgroup (*Thamnophis elegans)*, chromosomal sequences still appeared to share significant synteny across their lengths. Consequently, establishing when and how this karyotypic variation originated between terrestrial and marine lineages will require denser sampling of fully marine, semi-aquatic and terrestrial snakes, while exploring this architectural variation between *Hydrophis* snakes may elucidate if genomic architecture played a role in their rapid radiation and adaptation.

In addition to exploring macro-variation at the chromosome level, we also investigated divergence landscapes between three *Hydrophis* species in chapter 4. Through these scans, we showed that divergence landscapes in *Hydrophis* were likely shaped by shared evolutionary processes, with numerous shared high and low-divergence peaks being identified across all three species. Notably, we found evidence of balanced selection acting on chemosensory receptor genes, which are functionally important in their feeding and mating behaviours (Kutsuma et al., 2018; Shine, 2005). Like above, the generation of whole genome resequencing for a diverse spread of *Hydrophis* snakes will be crucial to exploring the phylogenetic depth of these balanced chemosensory receptor genes. Further, such diverse sampling of *Hydrophis* will also significantly aid in identifying common and lineage specific islands of divergence across

their speciation continuum, which are likely to shed light on the evolutionary dynamics of this rapid radiation.

References

Campbell, C. R., Poelstra, J. W., & Yoder, A. D. (2018). What is Speciation Genomics? The roles of ecology, gene flow, and genomic architecture in the formation of species. *Biological Journal of the Linnean Society*, *124*(4), 561–583. https://doi.org/10.1093/biolinnean/bly063

Carr, M. H., Neigel, J. E., Estes, J. A., Andelman, S., Warner, R. R., & Largier, J. L. (2003). Comparing Marine and Terrestrial Ecosystems: Implications for the Design of Coastal Marine Reserves. *Ecological Applications*, *13*(sp1), 90–107. https://doi.org/10.1890/1051-0761(2003)013[0090:CMATEI]2.0.CO;2

Grosberg, R. K., Vermeij, G. J., & Wainwright, P. C. (2012). Biodiversity in water and on land. *Current Biology*, *22*(21), R900–R903. https://doi.org/10.1016/j.cub.2012.09.050

Kutsuma, R., Sasai, T., & Kishida, T. (2018). How Snakes Find Prey Underwater: Sea Snakes Use Visual and Chemical Cues for Foraging. *Zoological Science*, *35*(6), 483. https://doi.org/10.2108/zs180059

Lourie, S. A., & Vincent, A. C. J. (2004). Using Biogeography to Help Set Priorities in Marine Conservation. *Conservation Biology*, *18*(4), 1004–1020. https://doi.org/10.1111/j.1523-1739.2004.00137.x

MacArthur, R. H., Diamond, J. M., & Karr, J. R. (1972). Density Compensation in Island Faunas. *Ecology*, *53*(2), 330–342. https://doi.org/10.2307/1934090

Mayr, E. (1954). Geographic Speciation in Tropical Echinoids. *International Journal of Organic Evolution*, *8*(1), 19.

Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. A., Peichel, C. L., Saetre, G.-P., Bank, C., Brännström, Å., Brelsford, A., Clarkson, C. S., Eroukhmanoff, F., Feder, J. L., Fischer, M. C., Foote, A. D., Franchini, P., Jiggins, C. D., Jones, F. C., … Widmer, A. (2014). Genomics and the origin of species. *Nat Rev Genet*, *15*(3), 176–192. https://doi.org/10.1038/nrg3644

Shine, R. (2005). All at sea: Aquatic life modifies mate-recognition modalities in sea snakes (Emydocephalus annulatus, Hydrophiidae). *Behavioral Ecology and Sociobiology*, *57*(6), 591–598. https://doi.org/10.1007/s00265-004-0897-z

Wolf, J. B. W., & Ellegren, H. (2017). Making sense of genomic islands of differentiation in light of speciation. *Nat Rev Genet*, *18*(2), 87–100. https://doi.org/10.1038/nrg.2016.133

Appendix A:
Contributing author publications

Co-authored publications related to the work presented in this thesis.

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

Galbraith, J. D., Ludington, A. J., Suh, A., Sanders, K. L., & Adelson, D. L. (2020). New Environment, New Invaders—Repeated Horizontal Transfer of LINEs to Sea Snakes. *Genome Biology and Evolution*, *12*(12), 2370–2383. https://doi.org/10.1093/gbe/evaa208

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

Galbraith, J. D., Ludington, A. J., Sanders, K. L., Suh, A., & Adelson, D. L. (2021). Horizontal transfer and subsequent explosive expansion of a DNA transposon in sea kraits (*Laticauda*). *Biology Letters*, *17*(9), 20210342. https://doi.org/10.1098/rsbl.2021.0342

Horizontal Transposon Transfer and Its Implications for the Ancestral Ecology of Hydrophiine Snakes

Galbraith JD, Ludington A. J., Sanders K. L., Amos T. G., Thomson V. A., Enosi Tuipulotu D, Dunstan N, Edwards R. J., Suh A, Adelson D. L. Horizontal Transposon Transfer and Its Implications for the Ancestral Ecology of Hydrophiine Snakes. Genes. 2022; 13(2):217. https://doi.org/10.3390/genes13020217

Functional Duplication of the Short-Wavelength-Sensitive Opsin in Sea Snakes: Evidence for Reexpanded Color Sensitivity Following Ancestral Regression

Rossetto, I. H., Sanders, K. L., Simões, B. F., Nguyen Van Cao, & Ludington, A. J. (2023). Functional Duplication of the Short-Wavelength-Sensitive Opsin in Sea Snakes: Evidence for Reexpanded Color Sensitivity Following Ancestral Regression. 15(7). https://doi.org/10.1093/gbe/evad107