Leibniz Institut für Zoo und Wildtierforschung

Abteilung Evolutionsgenetik

Genome structure analysis and patterns of transposable elements evolution in the slow-evolving Testudines clade.

Publikationsbasierte

DISSERTATION

zur Erlangung des akademischen Grades
“doctor rerum naturalium” (Dr. rer. nat.)
in der Wissenschaftsdisziplin “Evolutionsgenetik”

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Universität Potsdam

von

Tomas Carrasco-Valenzuela

Potsdam, February 2023

First supervisor Prof. Dr. Jörns Fickel
Second supervisor Dr. Camila Mazzoni

First reviewer Prof. Dr. Jörns Fickel
Second reviewer Prof. Dr. Ralph Tiedemann
Third reviewer Dr. Aaron Vogan

Published online on the
Publication Server of the University of Potsdam:
https://doi.org/10.25932/publishup- 60657
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4- 606577

This dissertation is based in the following manuscripts:

1.- Bentley, B. P., **Carrasco-Valenzuela, T.**, Ramos, E. K., Pawar, H., Souza Arantes, L., Alexander, A., ... & Komoroske, L. M. (2023). Divergent sensory and immune gene evolution in sea turtles with contrasting demographic and life histories. *Proceedings of the National Academy of Sciences*, 120(7), e2201076120.

2.- **Carrasco-Valenzuela, T.**, Marins, L., Ramos, E., Suh, A., Mazzoni, C. (2023). Recent expansion of Penelope-like retrotransposons in the leatherback turtle Dermochelys coriacea. Submitted to Mobile DNA. Under review in *Mobile DNA*

3.- **Carrasco-Valenzuela**,**T**., Ramos,E., Mazzoni C. (2023). Testudine-wide Transposable element exploration: A history of slow evolution and conserved genomes. In preparation (formatted for submission)

Table of Contents

Acknowledgments

I would like to express my gratitude to my supervisors, Prof. Dr. Jörns Fickel and Dr. Camila Mazzoni, who guided me throughout this project. Together with the people from BeGenDiv, and IZW that help me with their knowledge and company. To my understanding family who have always been supportive of my challenges. To my friends who have offered their company and engaging conversations during the long winter months. And lastly, to my partner in life and main editor, Elisa, without whom this journey would have been much more arduous.

Summary

Transposable elements (TEs) are *loci* that can replicate and multiply within the genome of their host. Within the host, TEs through transposition are responsible for variation on genomic architecture and gene regulation across all vertebrates. Genome assemblies have increased in numbers in recent years. However, to explore in deep the variations within different genomes, such as SNPs (single nucleotide polymorphism), INDELs (Insertion-deletion), satellites and transposable elements, we need high-quality genomes. Studies of molecular markers in the past 10 years have limitations to correlate with biological differences because molecular markers rely on the accuracy of the genomic resources. This has generated that a substantial part of the studies of TE in recent years have been on high quality genomic resources such as Drosophila, zebrafinch and maize. As testudine have a slow mutation rate lower only to crocodilians, with more than 300 species, adapted to different environments all across the globe, the testudine clade can help us to study variation. Here we propose Testudines as a clade to study variation and the abundance of TE on different species that diverged a long time ago. We investigated the genomic diversity of sea turtles, identifying key genomic regions associated to gene family duplication, specific expansion of particular TE families for Dermochelyidae and that are important for phenotypic differentiation, the impact of environmental changes on their populations, and the dynamics of TEs within different lineages. In chapter 1, we identify that despite high levels of genome synteny within sea turtles, we identified that regions of reduced collinearity and microchromosomes showed higher concentrations of multicopy gene families, as well as genetic distances between species, indicating their potential importance as sources of variation underlying phenotypic differentiation. We found that differences in the ecological niches occupied by leatherback and green turtles have led to contrasting evolutionary paths for their olfactory receptor genes. We identified in leatherback turtles a long-term low population size. Nonetheless, we identify no correlation between the regions of reduced collinearity with abundance of TEs or an accumulation of a particular TE group. In chapter 2, we identified that sea turtle genomes contain a significant proportion of TEs, with differences in TE abundance between species, and the discovery of a recent expansion of Penelope-like elements (PLEs) in the highly conserved sea turtle genome provides new insights into the dynamics of TEs within Testudines. In chapter 3, we compared the proportion of TE across the Testudine clade, and we identified that the proportion of transposable elements within the clade is stable, regardless of the quality of the assemblies. However, we identified that the proportion of TEs orders has correlation with genome quality depending of their expanded abundancy. For retrotransposon, a highly abundant

element for this clade, we identify no correlation. However, for DNA elements a rarer element on this clade, correlate with the quality of the assemblies.

Here we confirm that high-quality genomes are fundamental for the study of transposable element evolution and the conservation within the clade. The detection and abundance of specific orders of TEs are influenced by the quality of the genomes. We identified that a reduction in the population size on *D. coriacea* had left signals of long-term low population sizes on their genomes. On the same note we identified an expansion of TE on *D. coriacea*, not present in any other member of the available genomes of Testudines, strongly suggesting that it is a response of deregulation of TE on their genomes as consequences of the low population sizes.

Here we have identified important genomic regions and gene families for phenotypic differentiation and highlighted the impact of environmental changes on the populations of sea turtles. We stated that accurate classification and analysis of TE families are important and require high-quality genome assemblies. Using TE analysis we manage to identify differences in highly syntenic species. These findings have significant implications for conservation and provide a foundation for further research into genome evolution and gene function in turtles and other vertebrates. Overall, this study contributes to our understanding of evolutionary change and adaptation mechanisms.

Zusammenfassung

Transponierbare Elemente (TEs) sind Loci, die sich im Genom ihres Wirts replizieren und vermehren können. Innerhalb des Wirts sind TEs durch Transposition für die Variation der genomischen Architektur und der Genregulation bei allen Wirbeltieren verantwortlich. In den letzten Jahren hat die Zahl der Genomassemblies zugenommen. Um jedoch die Variationen innerhalb verschiedener Genome, wie SNPs, INDELs, Satelliten und transponierbare Elemente, eingehend zu untersuchen, benötigen wir qualitativ hochwertige Genome. Studien über molekulare Marker in den letzten 10 Jahren haben nur begrenzt mit biologischen Unterschieden korreliert, da molekulare Marker von der Genauigkeit der genomischen Ressourcen abhängen. Dies hat dazu geführt, dass ein großer Teil der TE-Studien der letzten Jahre an qualitativ hochwertigen genomischen Ressourcen wie *Drosophila*, Zebrafinken und Mais durchgeführt wurde. Da die Testudinen eine langsame Mutationsrate haben, die nur bei Krokodilen niedriger ist, aber mehr als 300 Arten umfassen, die an verschiedene Umgebungen auf der ganzen Welt angepasst sind, kann uns diese Gruppe bei der Untersuchung der Variation helfen. Hier schlagen wir Testudinen als Klade vor, um die Variation und die Häufigkeit von TE bei verschiedenen Arten zu untersuchen, die sich vor langer Zeit auseinanderentwickelt haben. Wir untersuchten die genomische Vielfalt der Meeresschildkröten und identifizierten genomische Schlüsselregionen, die mit der Duplikation von Genfamilien, der spezifischen Ausbreitung bestimmter TE-Familien bei den Dermochelyidae verbunden und für die phänotypische Differenzierung wichtig sind, sowie die Auswirkungen von Umweltveränderungen auf ihre Populationen und die Dynamik transponierbarer Elemente (TEs) innerhalb verschiedener Linien.

In Kapitel 1 stellen wir fest, dass trotz des hohen Maßes an Genomsyntenie innerhalb der Meeresschildkröten Regionen mit geringerer Kollinearität und Mikrochromosomen eine höhere Konzentration von Genfamilien mit mehreren Kopien sowie genetische Abstände zwischen den Arten aufweisen, was auf ihre potenzielle Bedeutung als Variationsquellen für die phänotypische Differenzierung hinweist. Wir fanden heraus, dass die Unterschiede in den ökologischen Nischen, die Lederschildkröten und Suppenschildkröten besetzen, zu gegensätzlichen evolutionären Pfaden für ihre Geruchsrezeptorgene geführt haben. Bei Lederschildkröten haben wir Anzeichen für langfristig niedrige Populationsgrößen festgestellt. Dennoch konnten wir keine Korrelation zwischen den Regionen mit reduzierter Kollinearität und der Häufigkeit von TEs oder einer Akkumulation einer bestimmten TE-Gruppe feststellen. In Kapitel 2 haben wir festgestellt, dass die Genome von Meeresschildkröten einen beträchtlichen Anteil an TEs enthalten, mit Unterschieden in der TE-Häufigkeit zwischen den

Arten, und die Entdeckung einer kürzlichen Ausbreitung von Penelope-ähnlichen Elementen (PLEs) im hochkonservierten Genom von Meeresschildkröten bietet neue Einblicke in die Dynamik von TEs innerhalb der Testudinen. In Kapitel 3 haben wir den Anteil der TE innerhalb der Testudinenklade verglichen und festgestellt, dass der Anteil der transponierbaren Elemente innerhalb der Klade stabil ist, unabhängig von der Qualität der Assemblies. Allerdings haben wir festgestellt, dass der Anteil der TEs Bestellungen hat Korrelation mit Genom Qualität in Abhängigkeit von ihrer erweiterten Häufigkeit, wie auf Retrotransposon, ein sehr häufig Element für diese Klade, wir identifizieren keine Korrelation, aber, DNA-Elemente ein seltener Element auf dieser Klade, korrelieren mit der Qualität der Baugruppen.

Hier bestätigen wir, dass qualitativ hochwertige Genome für die Untersuchung der Entwicklung transponierbarer Elemente und der Erhaltung innerhalb der Gruppe von grundlegender Bedeutung sind. Der Nachweis und die Häufigkeit bestimmter Ordnungen von TEs werden durch die Qualität der Genome beeinflusst. Wir haben festgestellt, dass eine Verringerung der Populationsgröße bei *D. coriacea* Signale für langfristig niedrige Populationsgrößen in ihren Genomen hinterlassen hat. Gleichzeitig haben wir bei *D. coriacea* eine Ausdehnung der TE festgestellt, die in keinem anderen Mitglied der verfügbaren Genome der Testudinen vorkommt, was stark darauf hindeutet, dass es sich um eine Reaktion auf die Deregulierung der TE auf ihren Genomen als Folge der geringen Populationsgrößen handelt.

Hier haben wir wichtige genomische Regionen und Genfamilien für die phänotypische Differenzierung identifiziert und die Auswirkungen von Umweltveränderungen auf die Populationen von Meeresschildkröten aufgezeigt. Wir haben festgestellt, dass eine genaue Klassifizierung und Analyse von TE-Familien wichtig ist und qualitativ hochwertige Genomassemblies erfordert. Mit Hilfe der TE-Analyse gelingt es uns, Unterschiede in hochsynthetischen Arten zu identifizieren. Diese Ergebnisse sind von großer Bedeutung für den Artenschutz und bilden eine Grundlage für die weitere Erforschung der Genomevolution und der Genfunktionen bei Schildkröten und anderen Wirbeltieren. Insgesamt trägt diese Studie zu unserem Verständnis des evolutionären Wandels und der Anpassungsmechanismen bei.

General introduction

Genomes are a resourceful type of data for the study of evolutionary biology, and the analysis and comparison of genomes from related species has proven to be an effective method for studying molecular evolution (Ekblom and Wolf 2014; Koepfli et al. 2015). However, the effectiveness of linking molecular diversity to evolutionary processes is dependent on the quality of the genomic data used. Complete genomes provide access to the molecular evolution of different types of genetic markers whose evolutionary changes could shape and maintain genetic variation in organisms (Shahid and Slotkin 2020). Nonetheless, genomes generated through short-read sequencing technologies alone have limitations in comprehending the evolutionary patterns found in repetitive regions, sub-telomeric regions of chromosomes, and in grasping chromosomal structure and synteny (Damas et al. 2017; Rhie et al. 2021). Therefore, improving the contiguity of genome assemblies is a critical aspect of genome research, providing greater completeness of genes and genomic elements and enabling a more in-depth examination of the evolution of countless species.

Structural variations in the genomes provide different information on species evolution that could not be recovered only from conserved regions of the genome. Therefore, investigating the modifications such as gene duplications, chromosomal rearrangements and transposable elements also contribute to understand the role of repetitive regions of the genome in species adaptation (Mérot et al. 2020).

In genomics, transposable elements (TEs) refer to *loci* that can replicate and multiply within the genome of their host (Boissinot et al. 2019). These elements are incredibly diverse and can be grouped into orders, superfamilies, families, and subfamilies based on their sequence, length, structure, and distribution (Wicker et al. 2007). Also, TEs can be divided into two main classes based on their mechanism of transposition and subsequently subdivided into superfamily, family, and subfamily according to the mechanism of chromosomal integration. Based on their transposition mechanism, two categories of TEs were described: Class I and Class II. Class I elements are retrotransposons that use an RNA intermediate to create a cDNA copy which is integrated into the genome through a "copy-and-paste" mechanism, as described by Boeke (1985) and reviewed by Bourque (2018). On the other hand, Class II elements, which are also known as DNA transposons, move through a “cut-and-paste” mechanism or a “peel-and-paste” replicative mechanism involving a circular DNA intermediate (Grabundzija et al. 2016; Greenblatt and Alexander Brink 1963; Rubin, Kidwell, and Bingham 1982).

The proportion of TEs in eukaryotic genomes can vary widely, with estimates ranging from 30-60% of reptile and mammal genomes (Canapa et al. 2015). Furthermore, the presence

of TEs is a major contributor to variations in haploid genome size (Margaret G. Kidwell 2002; Elliott and Gregory 2015). Differences in the abundance of TEs across genomes can also contribute to other genome features, such as differences in base composition in distinct regions or ectopic recombination (Symonová and Suh 2019; Robberecht et al. 2013). However, TEs and their host are in a constant battle, in which both suppression of TE expression and increased mutations in TEs may be employed to combat TE invasions (Skipper et al. 2013). Over time, as TE families become more evolutionarily ancient, they may acquire mutations that render them inactive. This happens due to mutations or fragmentation that occur during or after insertion or due to an active role of the host through different mechanisms (Bruno, Mahgoub, and Macfarlan 2019; Jacobs et al. 2014), and the extent of this inactivation can be measured using the Kimura 2-parameter distance to consensus (K-value) (Kimura 1980).

Furthermore, TEs show non-random patterns in their integration into host genomes. For instance, there is evidence that recent TE insertions in *A. thaliana* in regions enriched with genes related to environmental response (Baduel et al. 2021), while *Mutator* elements in *Drosophila* target open chromatin regions near recombination spots (S. Liu et al. 2009). *P* elements in *Drosophila* have also been found to associate with replication origins (Spradling, Bellen, and Hoskins 2011). Also *Penelope-like* elements have been described as associated with telomeric regions of the chromosomes helping to extend the telomeres (Gladyshev and Arkhipova 2007). This selective integration is not limited to regulatory regions, as *Ty3-Gypsy* LTR retroelements can bind specific methylation on histone H3 to only transpose to heterochromatin, a phenomenon seen in fungi to vertebrates (Malik and Eickbush 1999). Another example of integration into gene-poor regions is seen with the Ty5 LTR retrotransposon, with approximately 90% of its insertions in *S. cerevisiae* found within silent mating type loci or near silent heterochromatin at telomeres (Zou and Voytas 1997; Zou et al. 1996; Zou, Wright, and Voytas 1995).

Despite the rapid generation of high-quality genomes, the majority of reptile genomic resources have been applied to avian species, leaving non-avian reptiles severely underrepresented (Kelley et al. 2016; Card, Jennings, and Edwards 2023). The Testudine clade is seen as a good subject for the examination of TE dynamics (Sotero-Caio et al. 2017). Despite this, the progress in generating high-quality genomes for this group is limited and there is restricted information on TE composition, only available for a few turtle species such as the western painted turtle (Shaffer et al. 2013), the Chinese softshell turtle (Wang et al. 2013), the Asian yellow pond turtle (X. Liu et al. 2022), the Common Snapping Turtle (Das et al. 2020), and sea turtles (Wang et al. 2013). Hence, investigating TE evolution in the turtle clade is essential to understand how TE ratios may have impacted turtle evolution and diversity, as well

as provide a deeper understanding of the evolution of TEs in Testudines by including information from this understudied group. Given that turtle genomes have longer generation times and slower mutation rates compared to mammals and most reptiles (Janes et al. 2010), this clade provides an unique opportunity to examine mobilome diversification. A comparison of TE genome compositions in turtles can provide answers to questions about the relationship between TEs and functional genomic regions, ultimately contributing to a better understanding of TE evolution.

One specific group of Testudines that have a fascinating evolutionary history and no high-quality reference genome is the lineage of sea turtles. The sea turtle group is one of the most widely distributed vertebrates on the planet and has recolonized the seas over 100 million years ago (Hirayama 1998; Shaffer et al. 2017; Pike 2013). Of the seven species of sea turtles that exist today, leatherback turtles (*Dermochelys coriacea*) are the only living species from the Dermochelidae family, which diverged from other sea turtles (Cheloniidae) over 60 million years ago (Thomson, Spinks, and Shaffer 2021). Leatherbacks have unique characteristics that set them apart from other sea turtles, including a soft shell and the ability to feed in cool and productive pelagic habitats (Frair, Ackman, and Mrosovsky 1972; Davenport 1997). In contrast, green turtles (*Chelonia mydas*) are a species of hard-shell (Cheloniidae family) and are found in warmer water and nearshore habitats (Bentley et al. 2023).

As mentioned before, turtles in general exhibit slow rates of nucleotide substitution compared to other vertebrates (Green et al. 2014; Avise et al. 1992). In particular, sea turtles from the superfamily Chelonioidea exhibit low levels of genetic divergence in various genome-wide studies (Komoroske, Miller, and O'Rourke 2019; Vilaça et al. 2021; Zbinden et al. 2007; van der Zee et al. 2022; Driller, Vilaca, and Arantes 2020). However, the underlying genomic differences between these two sea turtle groups are not well understood.

For species like green and leatherback turtles to succeed in diverse environments, they must have the ability to regulate the expression of different genes. This occurs through random changes, allowing the best-adapted individuals to survive. The coordination of various genomic elements, such as promoters, enhancers, silencers, and insulators, which are non-coding sequences that control gene expression, plays a role in this process (Ali, Han, and Liang 2021; Conley, Piriyapongsa, and Jordan 2008). Several studies have demonstrated that TEs, which play a role in regulating gene expression, can contribute to changes in gene expression by altering their transcription machinery (Franchini et al. 2011; Samuelson et al. 1990; Brini, Lee, and Kinet 1993; Hambor et al. 1993). Moreover, one of the most variable genomic features among vertebrates is the number and diversity of TEs (Sotero-Caio et al. 2017; Tollis and Boissinot 2012). TEs are known to be a significant source of genetic variation in living

organisms (M. G. Kidwell and Lisch 2001) and can be a valuable source of data for comparing genomes of closely related species or species with slow evolution (Green et al. 2014), such as green and leatherback turtles.

Due to their high levels of conservation, sea turtles are ideal models for studying the evolution of TEs since speciation, which has been of interest for over 30 years (Endoh and Okada 1986). Despite the early discovery of the role of Short Interspersed Elements (SINEs) in hijacking the retropositional machinery of LINEs by acquiring 3' sequence fragments from LINEs on turtles (Kajikawa, Ohshima, and Okada 1997), there is limited understanding of the dynamics of TEs in the sea turtles clade. Although a draft-level genome of the green turtle was sequenced a decade ago (Wang et al. 2013), only recently has there been a focus on producing reference genomes for this group, offering quality data to enhance our understanding of their evolutionary history using this type of genetic markers.

Therefore, the creation of nearly-complete, high-quality, chromosome-level genomes for sea turtles presents a valuable opportunity to fully characterise the sea turtle genome and understand its evolution through comparative genomics of transposable element regions.

Aims of this study

Complete gapless reference genomes are a valuable resource in genetics, enabling the identification of genomic variations among closely related species. In recent years, numerous projects have successfully produced complete genome assemblies in various organisms. However, there is a lack of genomic resources for non-avian reptiles. Therefore, the first objective of this thesis was to generate high-quality genome assemblies for the understudied sea turtles.

In recent years, a comprehensive analysis of transposable elements has led to a greater understanding of their significance in adaptation, gene regulation, copy number variation, and other regulatory modifications resulting from their transposition. Based on this understanding, we aim to investigate the genomic divergence that this group of elements can generate in long-time diverged species with a slow mutation rate.

1 - Sequence, assemble and describe high-quality genomes of sea turtles. Subsequently, compare different features of these genomes in order to identify potential regions that may have contributed to phenotypical differences between two species that have important morphological, ecological and behavioural differences and a deep divergence time.

2 - To conduct a thorough comparison of transposable elements between the two newly assembled high-quality sea turtles genomes and identify potential regions for genomic divergence between highly syntenic sea turtles.

3 - To conduct a comprehensive analysis of transposable elements across the entire clade of Testudines, with a particular emphasis on identifying variations within this slowly evolving group. Furthermore, we aim to examine the association between transposable elements and genomic attributes such as genes and exons.

Chapter 1

Divergent sensory and immune gene evolution in sea turtles with contrasting demographic and life histories

PNAS RESEARCH ARTICLE | GENETICS

Divergent sensory and immune gene evolution in sea turtles with contrasting demographic and life histories

Blair P. Bentley[a,1], Tomás Carrasco-Valenzuela[b,c], Elisa K. S. Ramos[b,c,d], Harvinder Pawar[e], Larissa Souza Arantes[b,c], Alana Alexander[f], Shreya M. Banerjee[a], Patrick Masterson[g], Martin Kuhlwilm[e,h], Martin Pippel[i,j], Jacquelyn Mountcastle[k], Bettina Haase[k], Marcela Uliano-Silva[b,c], Giulio Formenti[k,l], Kerstin Howe[m], William Chow[m], Alan Tracey[m], Ying Sims[m], Sarah Pelan[m], Jonathan Wood[m], Kelsey Yetsko[n], Justin R. Perrault[o], Kelly Stewart[p,q], Scott R. Benson[p,r], Yaniv Levy[s], Erica V. Todd[t], H. Bradley Shaffer[u,v], Peter Scott[u,w], Brian T. Henen[x], Robert W. Murphy[y], David W. Mohr[z], Alan F. Scott[z], David J. Duffy[n,aa], Neil J. Gemmell[f], Alexander Suh[bb,cc], Sylke Winkler[j,k], Françoise Thibaud-Nissen[g], Mariana F. Nery[d], Tomas Marques-Bonet[e,dd,ee,ff], Agostinho Antunes[gg,hh], Yaron Tikochinski[i], Peter H. Dutton[p], Olivier Fedrigo[k], Eugene W. Myers[i,j,ii], Erich D. Jarvis[k,l,kk], Camila J. Mazzoni[b,c,1,2], and Lisa M. Komoroske[a,1,2]

Edited by Kerstin Lindblad-Toh, Broad Institute, Cambridge, MA; received January 21, 2022; accepted November 17, 2022

Sea turtles represent an ancient lineage of marine vertebrates that evolved from terrestrial ancestors over 100 Mya. The genomic basis of the unique physiological and ecological traits enabling these species to thrive in diverse marine habitats remains largely unknown. Additionally, many populations have drastically declined due to anthropogenic activities over the past two centuries, and their recovery is a high global conservation priority. We generated and analyzed high-quality reference genomes for the leatherback (Dermochelys coriacea) and green (Chelonia mydas) turtles, representing the two extant sea turtle families. These genomes are highly syntenic and homologous, but localized regions of noncollinearity were associated with higher copy numbers of immune, zinc-finger, and olfactory receptor (OR) genes in green turtles, with ORs related to waterborne odorants greatly expanded in green turtles. Our findings suggest that divergent evolution of these key gene families may underlie immunological and sensory adaptations assisting navigation, occupancy of neritic versus pelagic environments, and diet specialization. Reduced collinearity was especially prevalent in microchromosomes, with greater gene content, heterozygosity, and genetic distances between species, supporting their critical role in vertebrate evolutionary adaptation. Finally, diversity and demographic histories starkly contrasted between species, indicating that leatherback turtles have had a low yet stable effective population size, exhibit extremely low diversity compared with other reptiles, and harbor a higher genetic load compared with green turtles, reinforcing concern over their persistence under future climate scenarios. These genomes provide invaluable resources for advancing our understanding of evolution and conservation best practices in an imperiled vertebrate lineage.

marine turtle | gene evolution | conservation genomics | genetic diversity | demographic history

Sea turtles recolonized marine environments over 100 Mya (1, 2) and are now one of the most widely distributed vertebrate groups on the planet (3). Leatherback turtles (Dermochelys coriacea) represent the only remaining species of the family Dermochelyidae, which diverged from the Cheloniidae (hard-shelled sea turtles) about 60 Mya (4). Unique morphological (Fig. 1A) and physiological traits allow leatherback turtles to exploit cool, highly productive pelagic habitats (5, 6), while green turtles (Chelonia mydas) and other hard-shelled species largely inhabit warmer nearshore habitats following an early pelagic life stage. Most previous research in this group has focused on organismal and ecological adaptations (7), but the genomic basis of traits that differentiate or unite these species is not well understood.

Anthropogenic pressures have caused substantial population declines in sea turtles, with contemporary populations representing mere fractions of their historical abundances (8, 9). Although sea turtles spend most of their life in the ocean, they also exhibit long-distance migrations to natal rookeries for terrestrial reproduction (7, 10, 11). Consequently, they are threatened by human activities in both terrestrial and marine environments, including direct harvest of meat and eggs (12), fisheries bycatch (13), coastal development (14, 15), pollution (16), disease (17), and climate change (18, 19), which is exacerbated by their temperature-dependent mechanism of sex determination (TSD) altering population dynamics (20, 21). The IUCN lists most sea turtle species as vulnerable or endangered, and while decades of conservation efforts have fueled positive trends for some populations (22), others continue to decline (23). In particular, leatherback turtles have undergone extensive declines (>95% in some populations) over the last century (24–27), including the extirpation of the Malaysian nesting population (28). Leatherback turtle

Significance

Sea turtle populations have undergone recent global declines. We analyzed de novo assembled genomes for both extant sea turtle families through the Vertebrate Genomes Project to inform their conservation and evolutionary biology. These highly conserved genomes were differentiated by localized gene-rich regions of divergence, particularly within microchromosomes, suggesting that these genomic elements play key functional roles in the evolution of sea turtles and possibly other vertebrates. We further demonstrate that dissimilar evolutionary histories impact standing genomic diversity and genetic load, and are critical to consider when using these metrics to assess adaptive potential and extinction risk. Our results also demonstrate how reference genome quality impacts inferences of comparative and conservation genomics analyses that need to be considered in their application.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: bbentley@umass.edu, mazzoni@izw-berlin.de, or lkomoroske@umass.edu.

[2]C.J.M. and L.M.K. contributed equally to this work.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2201076120/-/DCSupplemental.

Published February 7, 2023.

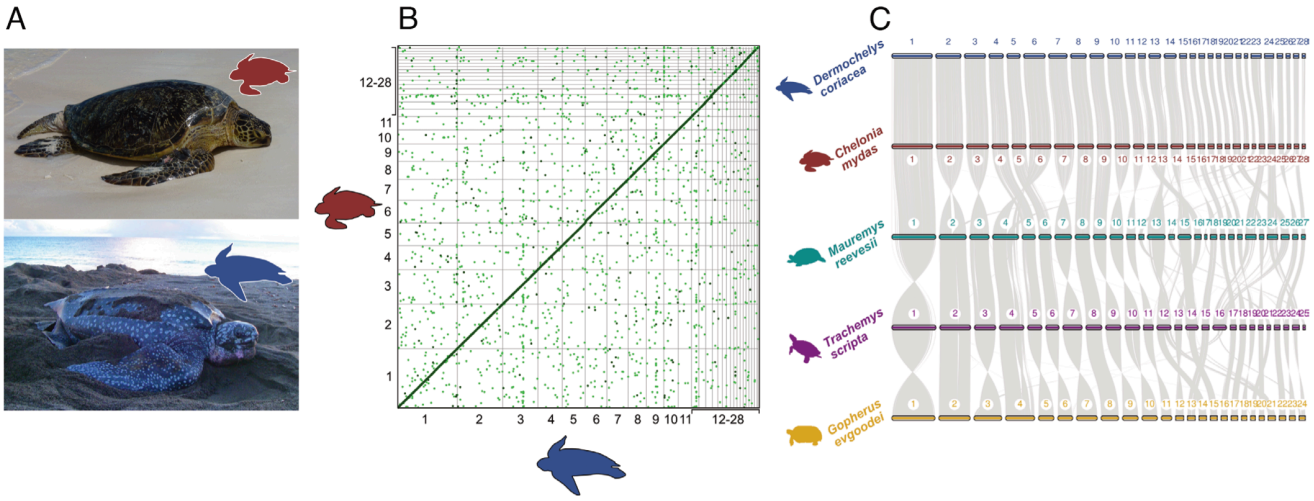


Fig. 1. (A) Green turtle (C. mydas); photo credit: NOAA NMFS PIFSC under USFWS Permit #TE-72088A-3, and leatherback turtle (D. coriacea); photo credit: Ricardo Tapilatu. (B) Dot plot showing regions with an identity greater than 0.5 across the entire genomes of green (red) and leatherback (blue) turtles. (C) Gene synteny and collinearity among leatherback turtle (blue), green turtle (red), Chinese pond turtle (Mauremys reevesii; green), pond slider turtle (Trachemys scripta; purple) and Goode's thornscrub tortoise (Gopherus evgoodei; yellow). Each bar represents chromosomes with respective numbers, and gray lines represent homolog gene connections.

recovery is impeded by relatively low hatching success compared with other sea turtle species (29). In contrast, many green turtle populations have recently increased following conservation actions (22), but their continued recovery remains threatened by anthropogenic activities and high incidence of the neoplastic disease fibropapillomatosis (FP), a viral-mediated tumor disease that disproportionately impacts this species (30).

Genomic data have been instrumental in advancing understanding of species' evolutionary histories and ecological adaptations (31–33), and providing critical information for conservation management (34–37). However, this research has been hampered in taxa where genomic resources remain limited. In particular, the lack of high-quality reference genomes, which are essential for accurate comparative evolutionary analyses (38, 39) and robust estimates of a range of metrics to inform conservation biology such as inbreeding, hybridization, disease susceptibility, genetic load, and adaptation (36, 40, 41), impede this work in threatened species. A draft genome for the green turtle was assembled almost a decade ago (42), and provided important insights into turtle evolution. However, errors, gaps, misassemblies, and fragmentation in draft genomes can lead to spurious inferences, potentially masking signals of interest (38, 43) and impeding effective management strategies (41). Well-annotated, chromosomal-level reference genomes can resolve these issues, improving our understanding of the genomic underpinnings of ecological and evolutionary adaptations (39, 44). For example, high-quality genomes with accurate annotations have enabled examination of gene changes associated with recolonization of the marine environment by terrestrial vertebrates, including the loss of olfactory receptor (OR) gene families (32, 45). Comparative genomic analyses have also demonstrated adaptive diversity in genes underlying reptilian immunity (46), and high-quality genomes have provided key insights into mammalian disease susceptibility (33, 47, 48). Equivalent investigations are critical for sea turtles, with diseases such as FP adversely impacting populations across the globe (30), information on immune genes is needed for devising effective conservation strategies (49).

We assembled chromosome-level reference genomes for leatherback and green turtles as part of the Vertebrate Genomes Project (VGP), and leveraged these resources to address questions centered around evolutionary history and conservation. Specifically, we provide insights into the genomic underpinnings of phenotypic traits that separate and unite these two species by examining genome synteny and regions of divergence. Given the contrasting recent population trends of these two species, we additionally used whole genome resequencing data of individuals representative of global populations to compare key conservation-relevant metrics, including patterns of diversity and deleterious variants, and reconstructed demographic histories to inform assessments of future vulnerability. These genomes represent two of the most contiguous reptilian genomes assembled to date, and our results provide a foundation for further hypothesis-driven investigations into the evolutionary adaptation and conservation of this imperiled vertebrate lineage.

Results

Genome Quality. Reference genomes for the leatherback and green turtles were generated using four genomic technologies following the VGP pipeline v1.6 (39), with minor modifications (see Methods). A total of 100% of the leatherback and 99.8% of the green turtle assembled sequences were placeable within chromosomes. The assembled genomes were near full-length (~2.1 GB), with annotations of all 28 known chromosomes for both species, composed of 11 macrochromosomes (>50 Mb) and 17 microchromosomes (<50 Mb) (SI Appendix, Table S1 and Fig. S1). These genomes are among the highest quality genomes assembled for nonavian reptiles to date in terms of both contiguity and completeness (Dataset S1), with the leatherback turtle assembly representing the first reptile genome where all scaffolds were assigned to chromosomes. Scaffold N50s were high for both genomes (SI Appendix, Table S1). We annotated 18,775 protein-coding genes in the leatherback and 19,752 in the green turtle genomes (see below for analysis of these gene differences). For the leatherback and green turtles, 96.9% and 97.5% of these genes were supported at >95% of their length from experimental evidence and/or high-quality protein models from related species (see Methods). The numbers of protein-coding genes are within the range of other reptiles (Dataset S1) and include 97.7% and 98.2% complete BUSCO copies for leatherback and green turtles based

on Sauropsida models (50), which are similar to or higher than all other assembled reptilian genomes to date (SI Appendix, Fig. S2).

Genome Architecture. Despite diverging over 60 Mya (4), leatherback and green turtles show extremely high genome synteny and collinearity (Fig. 1 B and C and SI Appendix, Figs. S6 and S7), with Progressive CACTUS revealing 95% sequence identity across the length of the genomes (SI Appendix, Table S3). After multiple rounds of manual curation to correct artifacts of misassemblies, few large structural rearrangements between the two species remained, including inversions of up to 7 Mb on chromosomes 12, 13, 24, and 28 (SI Appendix, Fig. S6). The high collinearity between species included near-complete end-to-end contiguous synteny for nine of 28 chromosomes (SI Appendix, Fig. S6). The remaining 19 chromosomes exhibited at least one small region of reduced collinearity (RRC) between the species, with RRCs representing a total of ~83.4 Mb (~3.9%) and ~110.5 Mb (~5.2%) of the leatherback and green turtle genome lengths, respectively. Eight chromosomes exhibited small RRCs (0.1 to 3 Mb), and 11 contained RRCs that were between 3 and 18 Mb in length (Fig. 2 A–D and Dataset S3). Analyses of coding regions revealed a similar pattern of strong collinearity between the two species (Fig. 1C and SI Appendix, Fig. S6), particularly within the macrochromosomes, which contain more than 80% of the total length of the genomes. The two genomes also displayed similar percentages of repetitive elements (REs), which were almost exclusively transposable elements (TEs) and unclassified repeats (SI Appendix, Fig. S8). The landscape of TE superfamily composition over evolutionary time was also similar between the two species, with the exception of REs with low Kimura values (<5%), which appeared at a higher frequency in the leatherback turtle genome (see SI Appendix, section I for full analyses).

Gene Families and Gene Functional Analysis. Gene function analysis of localized RRCs revealed that most contained genes with higher copy numbers in the green turtle compared with the leatherback (Fig. 2 A–D and Dataset S3). Nineteen chromosomes had RRCs with higher gene copy numbers in the green turtle, and of these, ten contained genes associated with immune system, olfactory reception, and/or zinc-finger protein-coding genes. Many of the same gene families were also detected as high-diversity exonic regions via separate, independent analyses (SI Appendix, section I), reinforcing their importance in the divergent evolution of these species. In addition to localized RRCs, higher gene copy numbers in the green turtle occurred in many gene orthologous groups (orthogroups) across the entire genome, and generally in variable multicopy genes (Fig. 2 F and G). Copy number variation accounted for most of the nearly one thousand more genes annotated in the green turtle genome relative to the leatherback (Fig. 2 F and G and SI Appendix, Table S1). We detected no evidence of collapsed multicopy genes in the leatherback turtle assembly across multiple analyses (see Methods and SI Appendix, Table S4), supporting this as a biological signal rather than technical artifact of the assemblies.

Olfactory receptors (ORs) represented the largest orthogroups in both genomes, and differences in copy numbers were connected to many of the identified RRCs. All OR class I genes were clustered at the beginning of chromosome 1, and the green turtle had higher copy numbers in this region (Fig. 2 A–D). This area also contained a cluster of OR class I genes in at least three additional testudinid species (SI Appendix, Fig. S10), and is the only divergent region across the very large chromosome 1 in the turtles analyzed. In contrast, OR class II genes were spread across several chromosomes in both sea turtle species, with higher copy numbers again in the green turtle found within RRCs (Fig. 2 B–D). The instability and rapid evolution of OR gene numbers in turtles is further illustrated in the expansion–contraction analysis of orthogroups (Fig. 2E and Dataset S6 A–D), which showed that OR class I genes underwent a modest contraction in the ancestral sea turtle lineage, followed by an expansion in the green turtle but a further contraction in the leatherback turtle. Similar trends were detected for OR class II genes, but with a greater magnitude of contraction in the ancestral sea turtle lineage followed by a further contraction for the leatherback turtle and only a small expansion for the green turtle (Fig. 2E).

Another important RRC (RRC14) encompassed the major histocompatibility complex (MHC), which plays a critical role in vertebrate immunity and is particularly relevant to sea turtle conservation due to the threat of FP and other diseases (32). In addition to the MHC region, this RRC includes several copies of OR class II genes, zinc-finger protein-coding genes and other genes involved with immunity, such as butyrophilin subfamily members and killer cell lectin-like receptors (Fig. 2D and Dataset S3). Invariably, the green turtle carried higher numbers of all the multicopy genes present in RRC14. RRCs on other chromosomes similarly showed increased levels of zinc-finger protein genes in the green turtle, including the RRCs labeled 6A, 11A, 14A, and 28 (Dataset S3). In particular, zinc-finger protein genes were highly prevalent on chromosomes 14 and 28 in both sea turtles, representing more than 50% of all the protein domains present on these chromosomes (SI Appendix, Fig. S11). Finally, all but three genes with known roles in TSD in reptiles (Dataset S7) were located as single-copy genes within both sea turtle genomes, with homologous copies located in the same region of the chromosomes in both species (see SI Appendix, section I for full analyses).

Macro and Microchromosomes. Microchromosomes contained significantly higher proportions of genes than macrochromosomes (Fig. 3 A and B; green turtle: $F_{(2,\ 25)} = 16.46$, $P < 0.01$; leatherback turtle: $F_{(2,\ 25)} = 16.35$, $P < 0.01$), and gene content was strongly positively correlated with GC content (SI Appendix, Fig. S13; green turtle $R^2 = 0.81$, $P < 0.01$; leatherback turtle $R^2 = 0.87$, $P < 0.01$). These patterns were particularly apparent in small (<20 Mb) microchromosomes, where GC content reached 50%, compared with the 43 to 44% genome-wide averages. Within chromosome groups, larger proportions of multicopy genes were generally associated with higher total gene counts (green turtle: R^2 0.84, $P < 0.01$; leatherback turtle: $R^2 = 0.92$, $P < 0.01$), and chromosomes with the highest multicopy genes numbers had increased proportions of RRCs (Fig. 3 A and B; green turtle: $R^2 = 0.69$, $P < 0.01$; leatherback turtle: $R^2 = 0.81$, $P < 0.01$).

Mean genetic distances for single-copy regions between the two sea turtles were also higher in small microchromosomes (0.053) compared with both intermediate (>20 Mb) microchromosomes (0.047), and macrochromosomes (0.045) (Fig. 3C; $F_{(2,\ 25)} = 21.98$, $P < 0.01$). However, examination of intermediate microchromosome and macrochromosome RRCs revealed elevated genetic distances in these regions that approached the values observed in small microchromosomes (SI Appendix, Table S5). Genetic distances were also significantly positively correlated with heterozygosity (green turtle: $R^2 = 0.97$, $P < 0.01$; leatherback turtle $R^2 = 0.97$, $P < 0.01$), which was significantly higher in small microchromosomes for both species (Fig. 3D; green turtle: $F_{(2,\ 25)} = 15.72$, $P < 0.01$; leatherback turtle: $F_{(2,\ 25)} = 5.09$, $P < 0.05$).

Genome Diversity. Genome-wide nucleotide diversity was almost a magnitude of order lower in leatherback compared with green turtles (mean repeat masked $\pi = 2.86 \times 10^{-4}$ and 2.46×10^{-3}, respectively; $t_{(5.52)} = 36.9$, $P < 0.001$; Fig. 4A and SI Appendix, Figs. S15–S17 and Table S7). Despite having largely similar gene

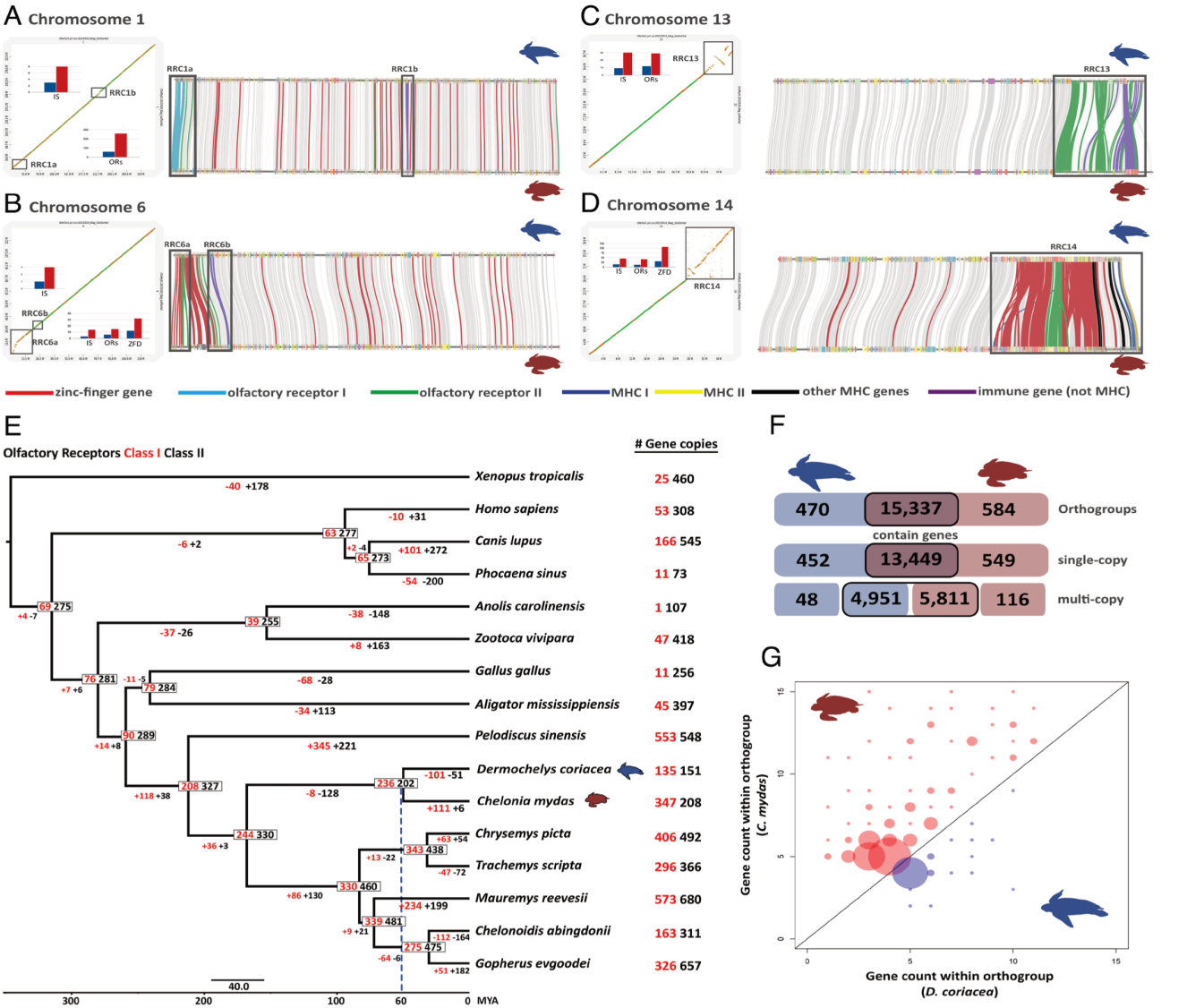


Fig. 2. (A–D) Dotplots (identity values; dark green =1 to 0.75, green =0.75 to 0.5, orange =0.5 to 0.25 and yellow =0.25 to 0) showing four of the regions with reduced collinearity (RRC) identified within chromosomes and associated with higher copy numbers of immune system (IS), ORs, or zinc finger domain genes in the green turtle relative to leatherback turtle (see also SI Appendix, Fig. S6 and Tables S3–S5 and Dataset S3 for full details of all RRCs). RRC positions are marked with gray squares on the dot plots (Left; with leatherback turtle on the X-axes and green turtle on the Y-axes) and gene collinearity maps (Right) for each chromosome highlighting the connections among specific gene families in different colors. (E) Gene family evolution of ORs class I (red) and class II (black) for amniote phylogeny. Gene numbers are presented on the nodes and gain/loss along each branch are presented below branches. Small scale bar represents substitutions/site, and big scale bar represents divergence times (MA). The blue dashed line shows the estimated divergence between the two sea turtle families. (F) Number of unique and shared orthogroups and single- and multicopy genes between the two sea turtles (coding genes including genes with rearrangement). The boxes outlined in black denote shared orthogroups, with the higher multicopy in the green turtle due to greater gene copies within orthogroups. (G) Comparison of gene counts between both species per multigenic orthogroup, depicting only those orthogroups where both species have different numbers of genes and a minimum number of five genes for one of the species. Bubbles above the diagonal represent higher counts for the green turtle and below for the leatherback turtle. The size of the bubbles represents the number of orthogroups with the same gene count combination.

content identified in the annotation, this strong pattern was also observed in coding regions (Fig. 4A; $t_{(5.52)} = 37.7$, $P < 0.001$), such that leatherback turtles possess much less standing functional variation, possibly impacting their adaptive capacity to future novel environmental conditions. The strikingly low genomic diversity of leatherback turtles is also less than almost all other reptile species examined (SI Appendix, Fig. S19; but see ref. 51), including Chelonoidis abingdonii, where low diversity has been considered a contributing factor to their extinction (52). Contrastingly, genomic diversity of the green turtle fell in the midrange for reptiles, as well as for mammals examined using similar methods (53, 54). Finally, within both species, heterozygosity was lower in coding regions (mean $\pi = 2.77 \times 10^{-4}$ and 2.18×1^{-3} for leatherback and green turtles; Fig. 4A) relative to noncoding regions (mean $\pi = 3.18 \times 10^{-4}$ and 2.64×10^{-3}; leatherbacks: $[t_{(4)} = -8.9, P < 0.01]$ and greens: $[t_{(5)} = -30.9, P < 0.01]$), as expected from selection pressures driving higher sequence conservation in these functional genomic regions.

Runs of Homozygosity (ROH). In addition to lower genome-wide heterozygosity, leatherbacks had a greater total number of ROHs (>50 kb) than green turtles (mean $N_{ROH} = 4{,}510$ and

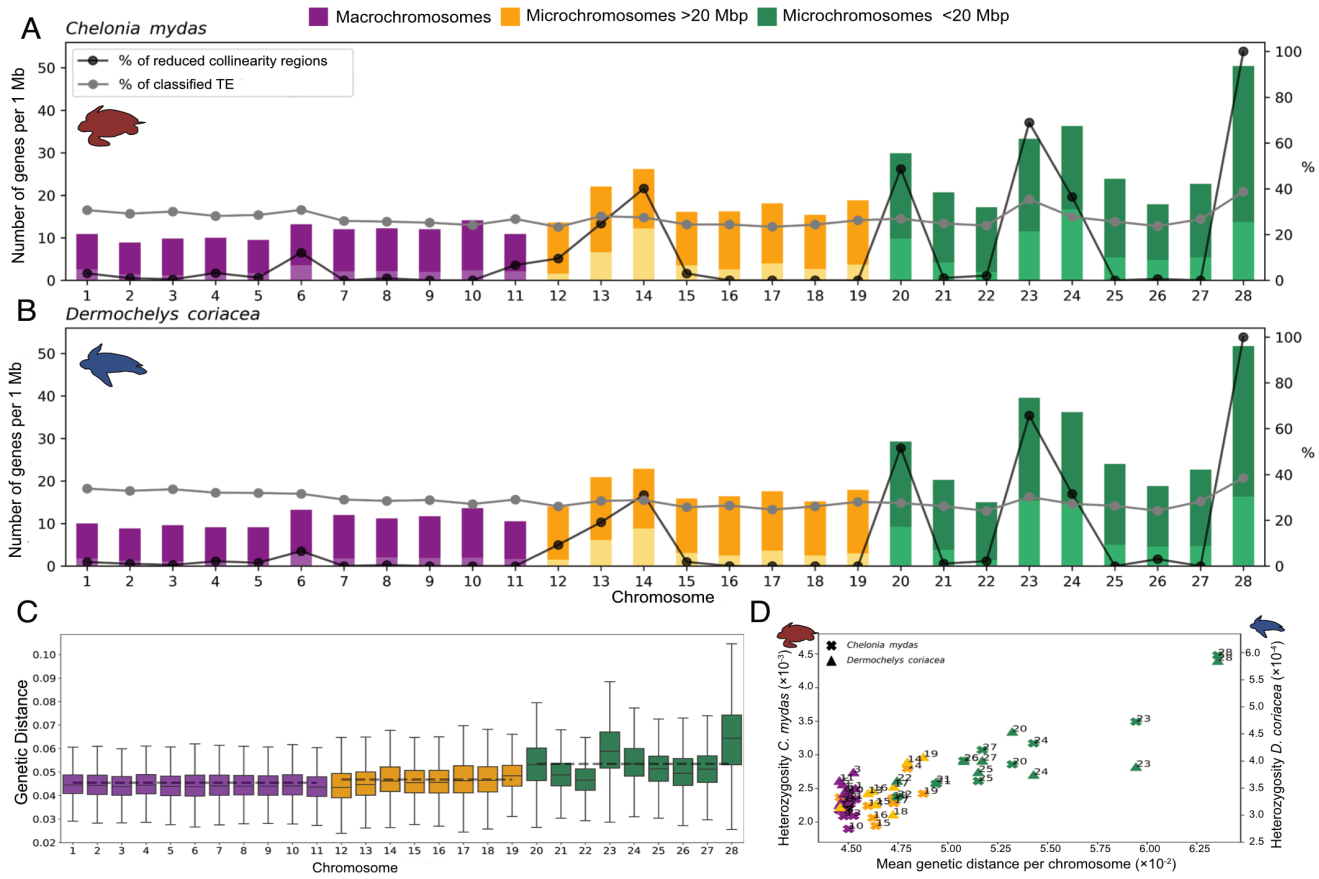


Fig. 3. Number of genes, genetic distance between species, and heterozygosity within species in macrochromosomes, small (<20Mb), and intermediate (>20Mb) microchromosomes. (A) Relation between the number of genes, percentage of RRCs, and classified TE per chromosome for the green and (B) leatherback turtles. Dark colors indicate the total number of genes and light colors indicate the number of multicopy genes. (C) Average genetic distance between green and leatherback turtles per chromosome. (D) Relation between genetic distance and heterozygosity per chromosome for each species.

829, respectively), as well as a greater total aggregate length of the genome in ROH (range = 26.1 to 45.5% in leatherback turtles; 1.8 to 17.7% in green turtles). The mean length of ROHs was also significantly higher in leatherback (L_{ROH} = 183.9 kb) compared with green turtles (L_{ROH} = 154.9 kb) ($t_{(7429.4)} = -8.85$, $P < 0.01$). Length distribution breakdown showed that leatherbacks have a higher aggregate length of all categories of ROHs relative to the green turtles (Fig. 4B and SI Appendix, Fig. S22). Short ROHs (50 to 500 kb) had the highest total aggregate length in leatherbacks, with a mean aggregate length of 597 Mb (Fig. 4B), suggesting long-term low population sizes in the leatherback turtle.

Within species, overall ROH distributions were generally similar between samples representative of different populations for leatherback turtles, although individuals from the Northwest Atlantic and East Pacific populations displayed slightly higher total aggregate lengths of ROHs than those from the West Pacific population, primarily due to greater aggregate lengths of medium and long ROHs (Fig. 4B). Among green turtles, the aggregate length of ROHs in all categories were generally small and similar across individuals, with the clear exception of the genome reference sample that originated from the Mediterranean population. This individual displayed higher numbers and lengths of long ROHs (>1 Mb) compared with all other green turtles (n = 50 compared with <5, and aggregate length = 74 Mb compared with <4 Mb), suggesting higher levels of recent inbreeding relative to the other green turtle populations represented in our dataset. Comparative analyses mapping this individual to the two previous green turtle assemblies failed to detect these long ROHs (SI Appendix, Fig. S23), demonstrating the importance of highly contiguous reference genomes for detecting biologically important patterns using this conservation-relevant metric.

Genetic Load. Coding region variants with predicted high (e.g., stop-codon gain or loss) or moderate impacts were significantly more common in leatherback compared with green turtles (Fig. 4C; high-impact variants: $t_{(4.18)} = -65.7$, $P < 0.001$; moderate impact variants: $t_{(4.51)} = -29.5$, $P < 0.001$). Conversely, low-impact and modifier (i.e., variants predicted to cause negligible impacts) variants were significantly more common in green turtles (Fig. 4C; low-impact variants: $t_{(5.88)} = 4.0$, $P < 0.01$; modifier variants: $t_{(5.33)} = 31.8$, $P < 0.001$). The missense-to-silent mutation ratio was also higher in leatherbacks than green turtles ($t_{(7.19)} = -72.3$, $P < 0.001$; mean = 0.99 and 0.70), further suggesting that genetic load is higher in the leatherback turtles. Within species, there was limited variation between individuals for all variant categories (Fig. 4C).

Demographic History. Pairwise Sequential Markovian Coalescence (PSMC) analyses indicated different historical effective population sizes (N_e) between the two sea turtle species (Fig. 4D). N_e for all leatherback turtle populations represented in our dataset have been relatively small and sustained over time, ranging in size from approximately 2,000 to 21,000 over the last 10 My, up until the

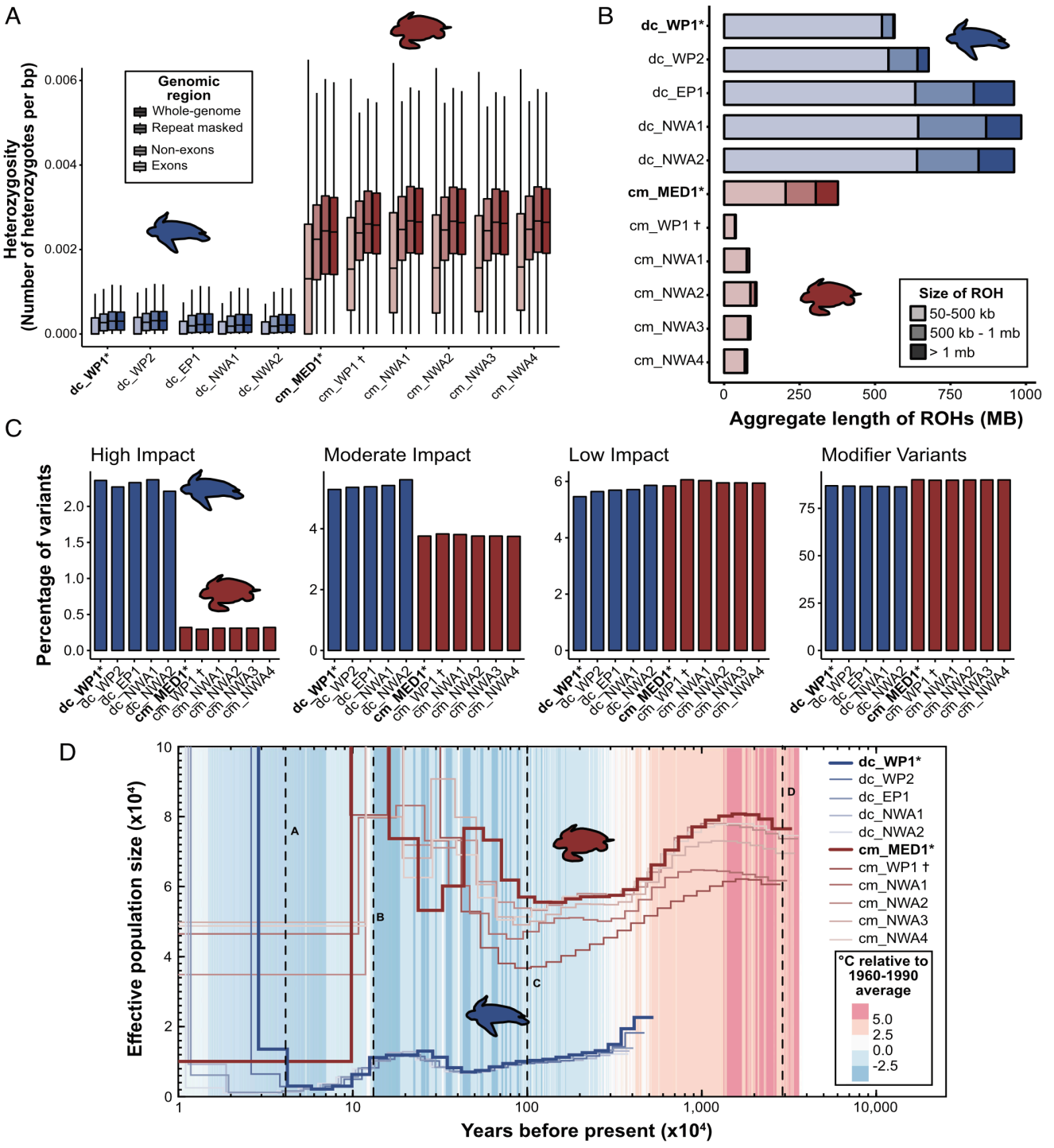


Fig. 4. Data are presented for the leatherback (blue) and green (red) turtle genomes, including reference individuals for both species (*), and the individual used to generate the draft genome (†; Wang et al. 2013). (A) estimates of heterozygosity for the entire genome, repeat-masked genome, exon and nonexon regions, with outliers removed. (B) accumulated lengths of runs of homozygosity (ROH). (C) predicted impacts of variants from within coding regions. (D) Pairwise sequential Markovian coalescent plot (PSMC) of demographic history of both species overlayed with temperature. Letters indicating portions of the PSMC curves (A–D) are geological events referred to in the main text and SI Appendix, section I.

Last Glacial Maximum (LGM) and at the lower end of this range for most of the last 5 My. This pattern is consistent between all individuals examined, with similar timings and magnitudes of N_e fluctuations until recent history (Fig. 4D). In contrast, green turtles have experienced wider variation and a higher overall N_e in general, fluctuating between approximately 50,000 and 125,000, until the late Pleistocene, with estimates varying by population (Figs. 4D and SI Appendix, Fig. S24). While N_e for leatherback turtles is relatively low, it modestly increased prior to the Eemian warm period (Fig. 4D [B]), followed by a subsequent

decreased during this period until the LGM (Fig. 4D [A]) when all populations exhibit sharp spikes in N_e possibly due to interocean gene flow following warming after the LGM. In contrast, green turtles generally displayed three distinct peaks in N_e (Fig. 4D), associated with ocean connectivity changes following the closure of the Tethys Sea [D], during the Pleistocene period [C], and prior to the Eemian warming period (Fig. 4D [B]). While the patterns of N_e are broadly similar within green turtles, the timing and magnitude of these fluctuations varied between populations (SI Appendix, Fig. S24).

Discussion

Divergence in Localized RRCs and Microchromosomes amidst High Global Genome Synteny. The ancestral lineage leading to leatherback and green turtles diverged over 60 Mya (4), giving rise to species that are adapted to dissimilar habitats, diets, and modes of life. Despite high overall levels of genome synteny between the species, RRCs and small microchromosomes were associated with higher concentrations of multicopy gene families, as well as heightened nucleotide diversity and genetic distances between species, suggesting that these genomic elements may be important sources of variation underlying phenotypic differentiation. Higher heterozygosity despite richer gene content in the microchromosomes suggests that these regions accumulate variation and may have a high adaptation value. Though our results here do not demonstrate direct causality, we have identified candidate regions and gene families that can be targeted in further studies quantifying evidence for positive selection and their roles in sea turtle adaptation and speciation.

The high global stability of macro- and microchromosomes between sea turtle families aligns with recent work showing similar patterns across reptiles, including birds, emphasizing the importance of microchromosomes in vertebrate evolution (55). Higher evolutionary rates in microchromosomes have been documented in intraspecific (56) and interspecific (57) avian studies, so it is possible that the characteristics of microchromosomes and RRCs we observed are not unique to sea turtles, but rather, are prevalent among vertebrates. The mechanisms driving these patterns are not well-understood, but could be related to higher recombination rates in micro- compared with macrochromosomes (58) that result in higher nucleotide diversity and lower haplotype sharing. Once generated, balancing selection may play a role in maintaining variation in these gene-dense regions, but more work is needed across taxa to determine the broad support for these hypotheses. The prevalence of localized genomic differentiation and underlying mechanisms among other closely or more distantly related vertebrate groups has yet to be widely evaluated due to a lack of equivalent quality genomic resources, but this is rapidly changing. Our detailed analyses of RRCs, microchromosomes, and their associated genes were only possible due to the high-quality of the assembled genomes because these analyses can be sensitive to genome fragmentation and misassemblies (39). For example, the RRCs and many microchromosomes could not be detected using the draft green turtle genome due to fragmentation and sequence gaps (SI Appendix, Figs. S3 and S4). As chromosomal-level genomes across all vertebrate lineages become available, our work provides a roadmap for identifying genomic regions harboring contrasting expansion/contractions of gene families and diversity levels. For taxa with highly conserved genomes like sea turtles, analyses of RRCs and microchromosomes are likely important to understand their divergent evolutionary histories and the phenotypic connections of the genes within them.

Contrasting Sensory and Immune gene Evolution between Sea Turtle Families. Sea turtles have complex sensory systems and can detect both volatile and water-soluble odorants, which are imperative for migration, reproduction, and identification of prey, conspecifics, and predators (59–63). However, leatherback and green turtles occupy dissimilar ecological niches, depending on different sensory cues. While leatherback turtles almost exclusively inhabit the pelagic environment posthatching, performing large horizontal and vertical migrations to seek out patches of gelatinous prey (64), green turtles recruit to neritic coastal and estuarine habitats as juveniles, and can have highly variable diets (65, 66). Sea turtle nasal cavity morphology also differs between species, with leatherback turtle cavities relatively short, wide, and more voluminous than chelonids (67–69), suggesting reduced requirements for olfactory reception. OR genes encode proteins used to detect olfactory cues, with the number of genes correlated with the number of detectable odorants (70), and linked to the chemical complexity of the inhabited environment (71). The two major groups of ORs in amniotic vertebrates are separated by their affinities with hydrophilic molecules (class I) or hydrophobic molecules (class II) (72). Class I OR genes may be particularly important in aquatic adaptation (32), and expansions of class I ORs in testudines, including green turtles, have been previously reported. However, the accuracy of these estimates for complex gene families using short-read assemblies has been uncertain because they may be prone to misassembly (32, 42, 73). We detected an additional 93 class I OR genes in our green turtle genome compared with those reported in the draft green turtle genome (42), suggesting they can be erroneously collapsed in short-read assemblies. Our reconstruction of both classes of OR gene evolution throughout the sea turtle lineage revealed that after ancestral contractions, gene copy evolution diverged in opposite directions between the two sea turtles. The greater loss of class II compared with class I OR genes in the ancestral sea turtle lineage likely reflects relaxed selection for detection of airborne odorants, as has been observed in other lineages that recolonized marine environments (74). However, as sea turtles continue to use terrestrial habitats for reproduction, they may need to retain some of these capabilities, which could explain why the observed contraction was weaker than those in exclusively marine species (e.g., the vaquita Phocaenasinus; Fig. 2E).

The strong class I OR expansion in the green turtle may be related to its distribution in complex neritic habitats and variable diet, requiring detection of a high diversity of waterborne odorants, while the continued loss of ORs in the leatherback turtle could be a consequence of its more specialized diet and the lower chemosensory-complexity of pelagic habitats. Although leatherback turtles can detect chemical cues from their prey, sensory experiments have indicated that visual cues are more important for food recognition in this species (75, 76). Additionally, while the precise mechanisms underpinning philopatry in sea turtles remain unclear, green turtles are thought to use olfactory cues to reach specific natal nesting beaches following long-distance navigation guided by magnetoreception (61, 63). In contrast, leatherback turtles exhibit more 'straying' from natal rookeries than other species, and such relaxed philopatry may be related to reduced reliance on olfactory cues to hone in on specific beaches.

Diversity within the highly complex MHC region is a key component in the vertebrate immune response to pathogens, with greater gene copy numbers and heterozygosity linked to lower disease susceptibility (77). While both sea turtle species contained most of the core MHC-related genes, the green turtle had more copies of genes involved in adaptive and innate immunity. Pathogen prevalence and persistence is often greater in neritic habitats than

open ocean habitats (78), so green turtles may be exposed to higher pathogen loads and diversity than leatherback turtles (79). However, reptilian immune systems are understudied compared with other vertebrates, with few studies of MHC genes conducted in turtles (80). Thus, it is not yet understood how immune gene diversity translates into disease susceptibility or ecological adaptation in sea turtles, which is critical for their conservation as FP continues to threaten population recoveries around the globe (30). Although this viral-mediated tumor disease occurs in all sea turtle species, prevalence and recovery greatly vary between and within species, making it plausible that harboring certain genes, copy numbers, or specific alleles may play important roles in disease dynamics. Despite decades of research on this disease (30) only one study on the immunogenomic factors governing FP susceptibility or resilience has been conducted (81), in part due to difficulty in accurately quantifying hypervariable and complex MHC loci with short-read sequencing technologies (82). Our reference genomes now enable studies to accurately interrogate these complex gene families to advance our fundamental understanding of immune gene evolution in testudines.

Differential Genomic Diversity and Demographic Histories. Genomic diversity is a critical metric for evaluating extinction risk and adaptive potential to environmental perturbation (83–85), with heterozygosity positively correlated with individual fitness (see reviews by refs. 86 and 87). Understanding the causes and consequences of genomic diversity is imperative for leatherback turtles in particular, where contemporary populations have sharply declined due to human activities (25). The exceptionally low genomic diversity observed in leatherback turtles broadly aligns with previous estimates (88, 89), but our PSMC and ROH results indicate that this is likely a consequence of long-term low effective population sizes and historical bottleneck events, rather than losses during recent population declines. This is consistent with mitochondrial analyses suggesting that contemporary populations radiated from a small number of matriarchal lineages within a single refugium following the Pleistocene (89). In contrast, higher heterozygosity, limited ROHs, and larger, more variable historical N_e in green turtles likely reflects radiation from many refugia and frequent admixing of populations (90).

Regardless of the causes of current genomic diversity levels, the amount of standing variation can have important implications for species' future persistence (91), especially given the adaptive capacity likely required to keep pace with rapid anthropogenic global change. Although genome-wide diversity estimation does not require high-quality reference genomes, they enable deeper examination of diversity patterns relevant to conservation. The use of our reference genomes demonstrated that diversity is very low within coding regions of leatherback turtle genomes, indicating limited standing functional variation that may have implications for their adaptive potential to novel conditions. Additionally, leatherback turtles exhibited a higher genetic load compared with green turtles, and this signal was consistent across all samples, regardless of population. Leatherback turtles have substantially lower hatching success compared with other sea turtle species (29), potentially related to the heightened genetic load and low heterozygosity (92, 93), and may combine with other factors to slow population recoveries despite conservation efforts. However, other species with low diversity have rebounded following population declines and/or appear to have purged deleterious alleles through long-term low population sizes (94–96), thereby limiting the impacts on viability (54, 94, 97). Although our results of greater genetic load despite long-term low N_e suggest this is not the scenario for leatherback turtles, further in-depth research on these topics enabled by the presented reference genomes will clarify these relationships for leatherback and other sea turtle species to guide conservation recommendations.

Although patterns of diversity, genetic load, and demographic histories were generally consistent within species, ROH analyses revealed a striking exception of the green turtle reference individual from the Mediterranean. This isolated population has undergone severe decline over the last century due to human exploitation (98), and our results indicate that consequent inbreeding is likely occurring, which may impact recovery. The specific individual was from the Israel green turtle rookery, estimated to have only 10 to 20 nesting females in the last decade (99, 100), but it is unclear whether Israel is demographically isolated from other rookeries in the region (100, 101). Further research is needed to understand whether inbreeding is a concern only for this nesting aggregation, or the Mediterranean population more broadly. These findings also highlight the utility of ROH metrics even in animals with longer generation times, and the importance of using highly contiguous genomes for accurate ROH assessment to inform conservation.

While it is widely documented that environmental changes can strongly impact species' abundances and distributions (102–104), following an initial decrease associated with declining temperatures, N_e of leatherback turtles remained relatively constant throughout substantial temperature fluctuations in the Pleistocene. As ectotherms, reptiles are sensitive to climatic thermal fluctuations; however, leatherback turtles exhibit unique physiological adaptations that produce regional endothermy and facilitate exploitation of cold-water habitats (6), potentially leading them to being less susceptible to periods of cooler temperatures. The long-term lower N_e of leatherback turtles may be associated with this species' greater mass and trophic position (105). In contrast, wide fluctuations for green turtles appear correlated with climatic events, beginning with the closure of the Tethys Sea, which altered ocean connectivity and represented a period of increasing temperatures that may have opened more suitable habitat. As temperatures subsequently decreased, N_e also decreased; however, temperature fluctuations during the Pleistocene were associated with additional increases in N_e. While warmer temperatures presumably allowed for larger population sizes of green turtles, large spikes in N_e around the Eemian warming, particularly for the Mediterranean individual, are likely associated with admixing of previously isolated populations due to warm-water corridors allowing movement between populations and ocean basins (106). While our overall estimates and trends for both species were broadly concordant with previous studies (89, 107, 108), a recent study using multiple sequentially Markovian coalescent (MSMC2) analyses found steep declines in N_e for green turtles >100,000 y before present (108), which was not detected in our PSMC analyses. Since this decline was also not detected in a prior study using PSMC on the draft green turtle genome (107), and demographic inferences are generally robust to genome quality (109, 110), this is likely a consequence of the different methods, with MSMC analyses inferring larger N_e for more ancient time scales (109).

Enabling Future Research and Conservation Applications. In addition to the insights reported here, the reference genomes for both extant sea turtle families provide invaluable resources to enable a wide breadth of previously unattainable fundamental and applied research. Combined with other forthcoming genomes (39), comparative genomics analyses can reveal the genomic basis for long-standing traits of interest such as adaptation to saltwater, diving capacity, and long-distance natal homing. Studies leveraging these reference genomes alongside

whole–genome sequencing of archival samples can assess how genomic erosion, inbreeding, and mutational load are linked to population size, trajectories, and conservation measures in global sea turtle populations. For instance, the fact that leatherback turtles have persisted with low diversity and N_e for extended periods offers hope for their recovery, but given that some populations have now been reduced to only a few hundred individuals (111), research quantifying purging of deleterious alleles, inbreeding depression, and adaptive capacity within populations is urgently needed (112). We emphasize that high–quality reference genomes are not required for all research goals, and combined with other recent studies (109, 110, 113), our findings provide clear guidance on when they may, or may not, be necessary to generate accurate results to inform conservation. For example, genome–wide diversity estimates are typically robust to assembly quality, but detection of long ROHs can be strongly affected. As ROH metrics are increasingly used to guide species management plans (114–116), it is important for researchers to understand how genome quality may impact their analyses and inferences. Additionally, many conservation applications that may not explicitly require whole–genome data can also directly benefit from the utility of these reference genomes, including the development of amplicon panels and molecular assays to investigate TSD mechanisms and adaptive capacity under climate change, and assessing linkages between immune genes and disease risk. Finally, with global distributions and long–distance migratory connectivity, sea turtle conservation requires international collaboration that has been previously hampered by difficulty comparing datasets between laboratories. Existing anonymous markers (e.g., microsatellites and restriction–site based SNP markers) can now be anchored to these genomes, and new ones can be optimized for conservation–focused questions and shared across the global research community, facilitating large–scale syntheses and equitable capacity building for genomics research. While ongoing anthropogenic impacts continue to threaten the viability of sea turtles to persist, combined with the critical work of reducing major threats such as fisheries bycatch and habitat loss, these genomes will enable research that make critical contributions to recovering imperiled populations.

Methods

Reference Sample Collections, Genome Assembly, and Annotation. Ultra–high molecular weight DNA was isolated from blood collections, and biopsies of internal organs for RNA were collected opportunistically from recently deceased or euthanized animals. Raw data were deposited into the VGP Genome Ark and NCBI Short–Read Archive (SRA; see Data Accessibility Statement). We assembled both genomes using four genomic technologies following the VGP pipeline v1.6 (39) with minor modifications. Short– and long–read transcriptome data (RNA–Seq and Iso–Seq) were generated from tissues known for their high transcript diversity in each species to enable accurate, species–specific annotations. These data, plus homology–based mapping from other species, were used to annotate the genomes using the standardized NCBI pipeline (117). We performed annotation as previously described (39, 118), using the same RNA–Seq, Iso–Seq, and protein input evidence for the prediction of genes in both species. Full details for all methods are provided in SI Appendix, section I.

Genome Quality Analysis. We used the pipeline assembly–stats from https://github.com/sanger–pathogens/assembly–stats to estimate scaffold N50, size distributions, and assembly size. BUSCO analysis (115) and QV value estimations (116) were conducted to assess the overall completeness, duplication, and relative quality of the assemblies. We used D–GENIES (118) with default parameters to conduct dot plot mapping of the entire genomes and each individual chromosome to evaluate the synteny between leatherback and green turtle genomes, and Haibao Tang JCVI utility libraries following the MCScan pipeline (119) to verify their contiguity. Incongruences in gene synteny blocks were manually investigated using Artemis Comparison Tool (120), identifying possible regions that could be caused by artifacts during assembly, and correcting these. The final curated assemblies were analyzed using the genome evaluation pipeline (https://git.imp.fu–berlin.de/cmazzoni/GEP) to obtain all final QC plots and summary statistics.

Identification and Analysis of RRCs. Using dotplots, 20 Mb windows were visually screened to identify regions of reduced collinearity (RRCs; SI Appendix, Fig. S5). Several genomic features (e.g., GC content, repeat elements) were compared between RRCs and equisized regions directly up– and down–stream to determine whether these were influencing collinearity (Dataset S5). Interproscan (119) was used to identify the functions of genes found within RRCs, and overall GO–term proportions for each chromosome were estimated using PANTHER (120); SI Appendix, Fig. S25). The two sea turtle genomes were aligned using Progressive Cactus (121, 122) to examine whether RRCs presented patterns of sequence divergence and/or gene duplication between the species.

Gene Families and Gene Functional Analysis. To estimate the timing of OR gene family evolution in sea turtles, we used computational analysis of gene family evolution (CAFE v5; (123). CAFE uses phylogenomics and gene family sizes to identify expansions and contractions. We used a dataset containing 8 species of turtle, 4 nonturtle reptiles, 3 mammals, and 1 amphibian using OrthoFinder (124, 125). OR orthogroups were grouped based on subfamily (class I and class II; see ref. 73), and an ultrametric phylogeny was generated by gathering 1:1 orthologs. We then aligned OrthoFinder amino acid sequences for each orthogroup and generated a phylogenetic tree. See SI Appendix, section I for searches of other specific genes.

Genetic Diversity and Demographic History. The halSnps pipeline (126) was used to estimate genetic distance between species by computing interspecific single variants based on alignments obtained with Progressive Cactus (121, 122). Genetic distances were calculated for 10,000–bp windows across the genome, where each window included only single alignments in the Cactus output. Differences in genetic distance, gene content, GC content, and heterozygosity between macro–, intermediate micro–, and small microchromosomes were tested using one–way ANOVAs for each species. Regression analyses were used to test for correlations between these measures across chromosomes.

For genome diversity, ROH, demographic history, and genetic load analyses, we included whole–genome resequencing data for additional individuals representing multiple global populations in each species (SI Appendix, Table S6 and section I). We calculated genome–wide heterozygosity using a method adapted from (127) using 100–kb nonoverlapping windows. Heterozygosity was calculated for the entire genome, repeat–masked genome, exons, and nonexons. Statistical comparisons between species were made using t tests. We subsequently applied the heterozygosity pipeline to generate genome–wide heterozygosity for additional reptilian species sourced from NCBI SRA, where species–specific reference genomes were available (SI Appendix, section I).

ROHs were identified by generating a SNP–list using the analysis of next generation sequencing data [ANGSD; (128)] pipeline. ANGSD was parameterized to output files configured for use as input for the PLINK ROH analysis (129). ROHs were then further characterized into size classes approximately based on (130).

Estimates of deleterious allele accumulation were conducted using snpEff (131). We estimated the impacts of variants (SNPs and INDELs) from coding regions using the species–specific genome annotations generated for both species. gVCFs were generated for each individual followed by joint–genotyping using GATK (132), allowing the reference individuals to include homozygous alleles found in other individuals. Combined VCFs were separated for each individual and filtered using based on depth of coverage (⅓×–2× mean coverage). snpEff predicts variant impacts and bins them into 'high–', 'moderate', or 'low–' impact categories, and outputs a list of genes that have predicted variant effects. We ran the snpEff analysis on all individuals for both species, and compared the percentages of each variant type between species using t tests.

PSMC (133) analyses of demographic history were employed for all individuals for both species. We used SAMtools (134) and BCFtools (135) to call genotypes with base and mapping quality filters of >Q30, before filtering for insert size (50 to 5,000 bp) and allele balance (AB), and retaining only biallelic sites with an AB of <0.25 and >0.75. We then ran PSMC analysis using the first 10 scaffolds (84% of total genome length). We scaled our outputs using a generation time of 30 y (SI Appendix, section I), and a mutation rate of 1.2×10^{-8} (107).

Data Accessibility Statement. Genome assemblies have been deposited on NCBI GenBank. The NCBI GenBank accession numbers for the leatherback turtle assembly (rDerCor1) are GCF_009764565.3 and GCA_009762595.2 for the annotated primary and original alternate haplotypes in BioProject PRJNA561993, and for the green turtle assembly (rCheMyd1) are GCF_015237465.2 and GCA_015220195.2 for primary and alternate haplotypes respectively in BioProject PRJNA561941. The raw data used for assemblies are available on the Vertebrate Genome Ark (https://vgp.github.io/genomeark/). The leatherback turtle data generated for the purpose of assembly annotation was deposited in the SRA under accession numbers SRX8787564–SRX8787566 (RNA-Seq) and SRX6360706–SRX6360708 (ISO-Seq). Green turtle data generated for annotation were deposited in SRA under accessions SRX10863130–SRX10863133 (RNA-Seq) and as SRX11164043–SRX11164046 (ISO-Seq). The NovaSeq 6000 DNA-Seq data for the green turtle resequencing, including raw reads, are deposited in NCBI (https://www.ncbi.nlm.nih.gov/) under BioProject ID: PRJNA449022. All scripts used for downstream analyses following genome assembly and annotation have been deposited on GitHub under repository https://github.com/bpbentley/sea_turtle_genomes.

Data, Materials, and Software Availability. All genomic data and scripts data have been deposited in VGP GenomeArk (136, 137) Github (138).

ACKNOWLEDGMENTS. We thank the St. Croix Sea Turtle Program, USFWS, NOAA–SWFSC California in–water leatherback research team, and the New England Aquarium for assistance with leatherback sample collection, and the Israel National Sea Turtle Rescue Centre, NOAA PIFSC–MTBAP, and Thierry Work (USGS) for assistance with green turtle sample collection. We thank Estefany Argueta and Jamie Adkins for assistance with literature searches and library preparations, and Phillip Morin, Andrew Foote, Anna Brüniche–Olsen, Annabel Beichman, Morgan McCarthy, David Adelson, and Yuanyuan Cheng for invaluable discussions and comments on the manuscript. Green turtle sequencing was performed by the Long Read Team of the DRESDEN–concept Genome Center, DFG NGS Competence Center, part of the Center for Molecular and Cellular Bioengineering (CMCB), Technische Universität Dresden and MPI–CBG. For green turtle resequenced samples, we thank Jessica Farrell, Whitney Crowder, Brooke Burkhalter, Nancy Condron, and the veterinary and rehabilitation staff and volunteers of the University of Florida's (UF) Sea Turtle Hospital at Whitney Laboratories, staff of the South Carolina Aquarium, UF's Interdisciplinary Center for Biotechnology Research Core for sequencing services, and to the Florida FWC and South Carolina DNR for assistance with permitting. We thank Erin LaCasella for help pro–curing leatherback samples for resequencing from the NOAA–SWFSC Marine Mammal and Sea Turtle Research Collection. This work was completed in part with resources provided by the University of Massachusetts' Green High Performance Computing Cluster (GHPCC). Funding was provided by the University of Massachusetts Amherst, NSF–IOS (grant #1904439 to L.M.K.), NOAA–Fisheries, National Research Council postdoctoral fellowship program to L.M.K., VGP, Rockefeller University, to E.D.J., HHMI to E.D.J., the Sanger Institute, Max–Planck–Gesellschaft, and grant contributions from Tom Gilbert, Paul Flicek, R.W.M., Karen A. Bjorndal, Alan B. Bolten, Ed Braun, N.J.G., T.M.–B., and A.F.S. We acknowledge CONICYT–DAAD forscholarship support to T.C.–V., the São Paulo Research Foundation to E.K.S.R.–FAPESP (grant #2020/10372–6). BeGenDiv is partially funded by the German Federal Ministry of Education and Research (BMbF, Förderkennzeichen 033W034A). The work of F.T.–N. and P.M. was supported by the Intramural Research Program of the National Library of Medicine, NIH. The work of M.P. was partially funded through the Federal Ministry of Education and Research (grant 01IS18026C). H.P. was supported by a Formació de Personal Investigador fellowship from Generalitat de Catalunya (FI_B100131). M.K. was supported by "la Caixa" Foundation (ID 100010434; code LCF/BQ/PR19/11700002), the Vienna Science and Technology Fund (WWTF), and the City of Vienna (VRG20–001). Funding for green turtle resequencing was provided by a Welsh Government Sêr Cymru II and the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska–Curie grant agreement No. 663830–BU115 and the Sea Turtle Conservancy, Florida Sea Turtle Grants Program (17–033R).

Author affiliations: [a]Department of Environmental Conservation, University of Massachusetts, Amherst, MA 01003; [b]Evolutionary Genetics Department, Leibniz Institute for Zoo and Wildlife Research, Berlin 10315, Germany; [c]Berlin Center for Genomics in Biodiversity Research, Berlin 14195, Germany; [d]Department of Genetics, Evolution, Microbiology and Immunology, State University of Campinas, Campinas 13083–970, Brazil; [e]Institut de Biologia Evolutiva, (CSIC–Universitat Pompeu Fabra), PRBB, Barcelona, Catalonia 08003, Spain; [f]Department of Anatomy, School of Biomedical Sciences, University of Otago, Dunedin 9016, New Zealand; [g]National Center for Biotechnology Information, National Library of Medicine, NIH, Bethesda, MD 20894; [h]Department of Evolutionary Anthropology, University of Vienna, Vienna 1030, Austria; [i]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden 01307, Germany; [j]Center for Systems Biology, Dresden 01307, Germany; [k]Vertebrate Genome Lab, The Rockefeller University, New York, NY 10065; [l]Laboratory of Neurogenetics of Language, The Rockefeller University, New York, NY 10065; [m]Tree of Life, Wellcome Sanger Institute, Cambridge CB10 1SA, UK; [n]The Whitney Laboratory for Marine Bioscience and Sea Turtle Hospital, University of Florida, St. Augustine, FL 32080; [o]Loggerhead Marinelife Center, Juno Beach, FL 33408; [p]Marine Mammal and Turtle Division, Southwest Fisheries Science Center, National Oceanic and Atmospheric Administration, La Jolla, CA 92037; [q]The Ocean Foundation, Washington, DC 20036; [r]Moss Landing Marine Laboratories, Moss Landing, CA 95039; [s]Department of Marine Biology, Leon H. Charney School of Marine Sciences, University of Haifa, Haifa 3498838, Israel; [t]School of Life and Environmental Sciences, Deakin University, Queenscliff, VIC 3225, Australia; [u]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095; [v]La Kretz Center for California Conservation Science, Institute of the Environment and Sustainability, University of California, Los Angeles, CA 90095; [w]Department of Life, Earth, and Environmental Sciences, West Texas A&M University, Canyon, TX 79016; [x]Environmental Affairs, Marine Air Ground Task Force and Training Command, Marine Corps Air Ground Combat Center, Twentynine Palms, CA 92278; [y]Centre for Biodiversity Royal Ontario Museum, Toronto, ON M5S 2C6, Canada; [z]Genetic Resources Core Facility, School of Medicine McKusick–Nathans Department of Genetic Medicine Johns Hopkins University, Baltimore, MD 21287; [aa]Department of Biology University of Florida, Gainesville, FL 32611; [bb]School of Biological Sciences, University of East Anglia, Norwich NR4 7TU, UK; [cc]Department of Organismal Biology, Evolutionary Biology Centre, Science for Life Laboratory, Uppsala University, Uppsala 75105, Sweden; [dd]Barcelona Institute of Science and Technology (BIST), CNAG–CRG, Centre for Genomic Regulation, Barcelona 08028, Spain; [ee]Institucio Catalana de Recerca i Estudis Avançats, Barcelona, Catalonia 08010, Spain; [ff]Institut Català de Paleontologia Miquel Crusafont Universitat Autònoma de Barcelona, Barcelona 08193, Spain; [gg]Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Porto 4450–208, Portugal; [hh]Department of Biology, Faculty of Sciences, University of Porto, Porto 4169–007, Portugal; [ii]Faculty of Marine Sciences, Ruppin Academic Center, Michmoret 4025000, Israel; [jj]Faculty of Computer Science, Technical University Dresden, Dresden 01069, Germany; and [kk]HHMI, Chevy Chase, MD 20815

Author contributions: B.P.B., E.W.M., E.D.J., C.J.M., and L.M.K. designed research; B.P.B., T.C.–V., E.K.S.R., H.P., L.S.A., A. Alexander, S.M.B., P.M., M.K., M.P., J.M., B.H., M.U.–S., G.F., K.H., W.C., A.T., Y.S., S.P., J.W., K.Y., J.R.P., K.S., S.R.B., Y.L., H.B.S., P.S., B.T.H., R.W.M., D.W.M., A.F.S., D.J.D., N.J.G., S.W., F.T.–N., M.F.N., T.M.–B., A. Antunes, Y.T., P.H.D., O.F., C.J.M., and L.M.K. performed research; B.P.B., T.C.–V., E.K.S.R., H.P., L.S.A., A. Alexander, P.M., M.K., M.P., M.U.–S., G.F., K.H., W.C., A.T., Y.S., S.P., J.W., A.F.S., A.S., F.T.–N., T.M.–B., A. Antunes, O.F., C.J.M., and L.M.K. analyzed data; S.M.B. and Y.T. collected samples; K.Y. provided whole genome resequencing data for additional individuals that were added to the analyses.; Y.L. provided samples; D.J.D. provided whole genome resequencing data for additional individuals that were added to the analyses; and B.P.B., T.C.–V., E.K.S.R., H.P., L.S.A., A. Alexander, M.K., G.F., K.H., J.R.P., E.V.T., H.B.S., P.S., R.W.M., N.J.G., A.F.S., S.W., F.T.–N., T.M.–B., A. Antunes, P.H.D., E.W.M., E.D.J., C.J.M., and L.M.K. wrote the paper.

1. R. Hirayama, Oldest known sea turtle. Nature 392, 705–708 (1998).
2. H. B. Shaffer, E. McCartney–Melstad, T.J. Near, G. G. Mount, P. Q. Spinks, Phylogenomic analyses of 539 highly informative loci dates a fully resolved time tree for the major clades of living turtles (Testudines). Mol. Phylogenet. Evol. 115, 7–15 (2017).
3. D.A. Pike, Climate influences the global distribution of sea turtle nesting: Sea turtle nesting distributions. Glob. Ecol. Biogeogr. 22, 555–566 (2013).
4. R. C. Thomson, P. Q. Spinks, H. B. Shaffer, Aglobal phylogeny of turtles reveals a burst of climate–associated diversification on continental margins. Proc. Natl. Acad. Sci. U.S.A. 118, e2012215118 (2021).
5. J. Davenport, Temperature and the life–history strategies of sea turtles. J. Therm. Biol. 22, 479–488 (1997).
6. W. Frair, R. G. Ackman, N. Mrosovsky, Body temperature of Dermochelys coriacea: Warm turtle from cold water. Science 177, 791–793 (1972).
7. P. C. H. Pritchard, "Evolution, phylogeny, and current status" in The Biology of Sea Turtles, P. L. Lutz, J. A. Musick, Eds. (CRC Press, 1996), pp. 1–28.
8. L. McClenachan, J. B. C. Jackson, M. J. H. Newman, Conservation implications of historic sea turtle nesting beach loss. Front. Ecol. Environ. 4, 290–296 (2006).
9. J. B. Jackson et al., Historical overfishing and the recent collapse of coastal ecosystems. Science 293, 629–637 (2001).
10. M. A. Grassman, D.W. Owens, J. P. McVey, R. M. M., Olfactory–based orientation in artificially imprinted sea turtles. Science 224, 83–84 (1984).
11. K. J. Lohmann, C. M. F. Lohmann, There and back again: Natal homing by magnetic navigation in sea turtles and salmon. J. Exp. Biol. 222, jeb184077 (2019).
12. P. S. Tomillo, V. S. Saba, R. Piedra, F. V. Paladino, J. R. Spotila, Effects of illegal harvest of eggs on the population decline of leatherback turtles in Las Baulas Marine National Park, Costa Rica. Conserv. Biol. 22, 1216–1224 (2008).

13. S. Fossette et al., Pan–atlantic analysis of the overlap of a highly migratory species, the leatherback turtle, with pelagic longline fisheries. Proc. Biol. Sci. 281, 20133065 (2014).
14. Y. Kaska et al., Natural and anthropogenic factors affecting the nest–site selection of Loggerhead Turtles, Caretta caretta, on Dalaman–Sarigerme beach in South–west Turkey. (Reptilia: Cheloniidae). Zool. Middle East 50, 47–58 (2010).
15. B. Von Holle et al., Effects of future sea level rise on coastal habitat. J. Wildl. Manage. 83, 694–704 (2019).
16. N. Mrosovsky, G. D. Ryan, M. C. James, Leatherback turtles: The menace of plastic. Mar. Pollut. Bull. 58, 287–289 (2009).
17. M. Chaloupka, G. H. Balazs, T. M. Work, Rise and fall over 26 years of a marine epizootic in Hawaiian green sea turtles. J. Wildl. Dis. 45, 1138–1142 (2009).
18. L. A. Hawkes, A. C. Broderick, M. H. Godfrey, B. J. Godley, Climate change and marine turtles. Endanger. Species Res. 7, 137–154 (2009).
19. B. P. Wallace et al., Global conservation priorities for marine turtles. PLoS One 6, e24510 (2011).
20. C. L. Yntema, N. Mrosovsky, Incubation temperature and sex ratio in hatchling loggerhead turtles: a preliminary report. Mar. Turtle Newsl. 11, 9–10 (1979).
21. M. P. Jensen et al., Environmental warming and feminization of one of the largest sea turtle populations in the world. Curr. Biol. 28, 154–159.e4 (2018).
22. A. D. Mazaris, G. Schofield, C. Gkazinou, V. Almpanidou, G. C. Hays, Global sea turtle conservation successes. Sci Adv 3, e1600730 (2017).
23. IUCN, The IUCN red list of threatened species (2021). (April 16, 2021).
24. L. S. Martínez et al., Conservation and biology of the leatherback turtle in the Mexican Pacific. Chelonian Conserv. Biol. 6, 70–78 (2007).
25. P. Santidrián Tomillo et al., Reassessment of the leatherback turtle (Dermochelys coriacea) nesting population at Parque Nacional Marino Las Baulas, Costa Rica: Effects of conservation efforts. Chelonian Conserv. Biol. 6, 54–62 (2007).
26. J. R. Spotila, R. D. Reina, A. C. Steyermark, P. T. Plotkin, F. V. Paladino, Pacific leatherback turtles face extinction. Nature 405, 529–530 (2000).
27. Laúd OPO Network, Enhanced, coordinated conservation efforts required to avoid extinction of critically endangered Eastern Pacific leatherback turtles. Sci. Rep. 10, 4772 (2020).
28. E. Chan, H. Liew, Decline of the leatherback population in Terengganu, Malaysia, 1956–1995. Chelonian Conserv. Biol. 2, 196–203 (1996).
29. K. L. Eckert, B. P. Wallace, J. G. Frazier, S. A. Eckert, P. C. H. Pritchard, Synopsis of the Biological Data on the Leatherback Sea Turtle, Dermochelys coriacea (U.S. Department of Interior, Fish and Wildlife Service, 2012).
30. K. Jones, E. Ariel, G. Burgess, M. Read, A review of fibropapillomatosis in Green turtles (Chelonia mydas). Vet. J. 212, 48–57 (2016).
31. G. Zhang et al., Comparative genomics reveals insights into avian genome evolution and adaptation. Science 346, 1311–1320 (2014).
32. I. Khan et al., Olfactory receptor subgenomes linked with broad ecological adaptations in Sauropsida. Mol. Biol. Evol. 32, 2832–2843 (2015).
33. D. Jebb et al., Six reference–quality genomes reveal evolution of bat adaptations. Nature 583, 578–584 (2020).
34. B. J. McMahon, E. C. Teeling, J. Höglund, How and why should we implement genomics into conservation? Evol. Appl. 7, 999–1007 (2014).
35. M. A. Supple, B. Shapiro, Conservation of biodiversity in the genomics era. Genome Biol. 19, 131 (2018).
36. P. Brandies, E. Peel, C. J. Hogg, K. Belov, The value of reference genomes in the conservation of threatened species. Genes 10, 864 (2019).
37. P. A. Hohenlohe, W. C. Funk, O. P. Rajora, Population genomics for wildlife conservation and management. Mol. Ecol. 30, 62–82 (2020), 10.1111/mec.15720.
38. X. Zhang, J. Goodsell, R. B. Norgren Jr., Limitations of the rhesus macaque draft genome assembly and annotation. BMC Genomics 13, 206 (2012).
39. A. Rhie et al., Towards complete and error–free genome assemblies of all vertebrate species. Nature 592, 737–746 (2021).
40. A. P. Fuentes–Pardo, D. E. Ruzzante, Whole–genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. Mol. Ecol. 26, 5369–5406 (2017).
41. G. Formenti et al., The era of reference genomes in conservation genomics. Trends Ecol. Evol. 37, 197–202 (2022).
42. Z. Wang et al., The draft genomes of soft–shell turtle and green sea turtle yield insights into the development and evolution of the turtle–specific body plan. Nat. Genet. 45, 701–706 (2013).
43. A. Whibley, J. L. Kelley, S. R. Narum, The changing face of genome assemblies: Guidance on achieving high–quality reference genomes. Mol. Ecol. Resour. 21, 641–652 (2021).
44. V. Peona et al., Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird–of–paradise. Mol. Ecol. Resour. 21, 263–286 (2021).
45. Y. Yuan et al., Comparative genomics provides insights into the aquatic adaptations of mammals. Proc. Natl. Acad. Sci. U.S.A. 118, e2106080118 (2021).
46. N. J. Gemmell et al., The tuatara genome reveals ancient features of amniote evolution. Nature 584, 403–409 (2020).
47. R. N. Johnson et al., Adaptation and conservation insights from the koala genome. Nat. Genet. 50, 1102–1111 (2018).
48. J.–N. Hubert, T. Zerjal, F. Hospital, Cancer–and behavior–related genes are targeted by selection in the Tasmanian devil (Sarcophilus harrisii). PLoS One 13, e0201838 (2018).
49. L. M. Zimmerman, The reptilian perspective on vertebrate immunity: 10 years of progress. J. Exp. Biol. 223, jeb214171 (2020).
50. M. Seppey, M. Manni, E. M. Zdobnov, BUSCO: Assessing genome assembly and annotation completeness. Methods Mol. Biol. 1962, 227–245 (2019).
51. Q.–H. Wan et al., Genome analysis and signature discovery for diving and sensory properties of the endangered Chinese alligator. Cell Res. 23, 1091–1105 (2013).
52. V. Quesada et al., Giant tortoise genomes provide insights into longevity and age–related disease. Nat Ecol Evol 3, 87–95 (2019).
53. P. A. Morin et al., Reference genome and demographic history of the most endangered marine mammal, the vaquita. Mol. Ecol. Resour. 21, 1008–1020 (2021).
54. J. A. Robinson et al., Genomic flatlining in the endangered island fox. Curr. Biol. 26, 1183–1189 (2016).
55. P. D. Waters et al., Microchromosomes are building blocks of bird, reptile, and mammal chromosomes. Proc. Natl. Acad. Sci. U.S.A. 118, e2112494118 (2021).
56. H.–J. Megens et al., Comparison of linkage disequilibrium and haplotype diversity on macro–and microchromosomes in chicken. BMC Genet. 10, 86 (2009).
57. E. Axelsson, M. T. Webster, N. G. C. Smith, D. W. Burt, H. Ellegren, Comparison of the chicken and turkey genomes reveals a higher rate of nucleotide divergence on microchromosomes than macrochromosomes. Genome Res. 15, 120–125 (2005).
58. A. V. Rodionov, Micro vs. macro: Structural–functional organization of avian micro–and macrochromosomes. Genetika 32, 597–608 (1996).
59. C. S. Endres, K. J. Lohmann, Detection of coastal mud odors by loggerhead sea turtles: A possible mechanism for sensing nearby land. Mar. Biol. 160, 2951–2956 (2013).
60. C. S. Endres, N. F. Putman, K. J. Lohmann, Perception of airborne odors by loggerhead sea turtles. J. Exp. Biol. 212, 3823–3827 (2009).
61. M. Manton, A. Karr, D. W. Ehrenfeld, Chemoreception in the migratory sea turtle, Chelonia mydas. Biol. Bull. 143, 184–195 (1972).
62. C. Kitayama et al., Behavioral effects of scents from male mature Rathke glands on juvenile green sea turtles (Chelonia mydas). J. Vet. Med. Sci. 82, 1312–1315 (2020).
63. C. S. Endres et al., Multi–modal homing in sea turtles: Modeling dual use of geomagnetic and chemical cues in island–finding. Front. Behav. Neurosci. 10, 19 (2016).
64. K. L. Dodge, J. M. Logan, M. E. Lutcavage, Foraging ecology of leatherback sea turtles in the Western North Atlantic determined through multi–tissue stable isotope analyses. Mar. Biol. 158, 2813–2824 (2011).
65. K. E. Arthur, M. C. Boyle, C. J. Limpus, Ontogenetic changes in diet and habitat use in green sea turtle (Chelonia mydas) life history. Mar. Ecol. Prog. Ser. 362, 303–311 (2008).
66. J. A. Seminoff et al., Large–scale patterns of green turtle trophic ecology in the eastern Pacific Ocean. Ecosphere 12, e03479 (2021).
67. C. Kitayama et al., Morphological features of the nasal cavities of hawksbill, olive ridley, and black sea turtles: Comparative studies with green, loggerhead and leatherback sea turtles. PLoS One 16, e0250873 (2021).
68. D. Kondoh, C. Kitayama, Y. K. Kawai, The nasal cavity in sea turtles: Adaptation to olfaction and seawater flow. Cell Tissue Res. 383, 347–352 (2021).
69. Y. Yamaguchi et al., Computed tomographical analysis of internal structures within the nasal cavities of green, loggerhead and leatherback sea turtles. Anat. Rec. 304, 584–590 (2021).
70. Y. Niimura, M. Nei, Evolutionary dynamics of olfactory and other chemosensory receptor genes in vertebrates. J. Hum. Genet. 51, 505–517 (2006).
71. L. R. Yohe, M. Fabbri, M. Hanson, B.–A. S. Bhullar, Olfactory receptor gene evolution is unusually rapid across Tetrapoda and outpaces chemosensory phenotypic change. Curr. Zool. 66, 505–514 (2020).
72. H. Saito, Q. Chi, H. Zhuang, H. Matsunami, J. D. Mainland, Odor coding by a mammalian receptor repertoire. Sci. Signal. 2, ra9 (2009).
73. M. W. Vandewege et al., Contrasting patterns of evolutionary diversification in the olfactory repertoires of reptile and bird genomes. Genome Biol. Evol. 8, 470–480 (2016).
74. A. Liu et al., Convergent degeneration of olfactory receptor gene repertoires in marine mammals. BMC Genomics 20, 977 (2019).
75. M. A. Constantino, M. Salmon, Role of chemical and visual cues in food recognition by leatherback posthatchlings (Dermochelys coriacea L.). Zoology 106, 173–181 (2003).
76. N. Warraich, J. Wyneken, N. Blume, Feeding behavior and visual field differences in loggerhead and leatherback sea turtles may explain differences in longline fisheries interactions. Endanger. Species Res. 41, 67–77 (2020).
77. H. V. Siddle, J. Marzec, Y. Cheng, M. Jones, K. Belov, MHC gene copy number variation in Tasmanian devils: Implications for the spread of a contagious cancer. Proc. Biol. Sci. 277, 2001–2006 (2010).
78. L. E. Escobar et al., A global map of suitability for coastal Vibrio cholerae under current and future climate conditions. Acta Trop. 149, 202–211 (2015).
79. L. Zhang et al., Massive expansion and functional divergence of innate immune genes in a protostome. Sci. Rep. 5, 8693 (2015).
80. J. P. Elbers, S. S. Taylor, Others, Major histocompatibility complex polymorphism in reptile conservation. Herpetol. Conserv. Biol. 11, 1–12 (2016).
81. K. R. Martin, K. L. Mansfield, A. E. Savage, Adaptive evolution of major histocompatibility complex class I immune genes and disease associations in coastal juvenile sea turtles. R Soc. Open Sci. 9, 211190 (2022).
82. X. Vekemans et al., Whole–genome sequencing and genome regions of special interest: Lessons from major histocompatibility complex, sex determination, and plant self–incompatibility. Mol. Ecol. 30, 6072–6086 (2021).
83. C. Moritz, Strategies to protect biological diversity and the evolutionary processes that sustain it. Syst. Biol. 51, 238–254 (2002).
84. P. Fernandez–Fournier, J. M. M. Lewthwaite, A. Ø. Mooers, Do we need to identify adaptive genetic variation when prioritizing populations for conservation? Conserv. Genet. 22, 205–216 (2021).
85. Y. J. Borrell et al., Heterozygosity–fitness correlations in the gilthead sea bream Sparus aurata using microsatellite loci from unknown and gene–rich genomic locations. J. Fish Biol. 79, 1111–1129 (2011).
86. J. A. DeWoody, A. M. Harder, S. Mathur, J. R. Willoughby, The long–standing significance of genetic diversity in conservation. Mol. Ecol. 30, 4147–4154 (2021).
87. Y. Willi et al., Conservation genetics as a management tool: The five best–supported paradigms to assist the management of threatened species. Proc. Natl. Acad. Sci. U.S.A. 119, e2105076119 (2022).
88. L. M. Komoroske et al., Aversatile rapture (RAD–Capture) platform for genotyping marine turtles. Mol. Ecol. Resour. 19, 497–511 (2019).
89. P. H. Dutton, B. W. Bowen, D. W. Owens, A. Barragan, S. K. Davis, Global phylogeography of the leatherback turtle (Dermochelys coriacea). J. Zool. 248, 397–409 (1999).
90. M. P. Jensen et al., The evolutionary history and global phylogeography of the green turtle (Chelonia mydas). J. Biogeogr. 46, 860–870 (2019).
91. M. Kardos et al., The crucial role of genome–wide genetic variation in conservation. Proc. Natl. Acad. Sci. U.S.A. 118, e2104642118 (2021).
92. P. Dobrynin et al., Genomic legacy of the African cheetah, Acinonyx jubatus. Genome Biol. 16, 277 (2015).
93. A. L. K. Mattila et al., High genetic load in an old isolated butterfly population. Proc. Natl. Acad. Sci. U.S.A. 109, E2496–E2505 (2012).
94. J. A. Robinson, C. Brown, B. Y. Kim, K. E. Lohmueller, R. K. Wayne, Purging of strongly deleterious mutations explains long–term persistence and absence of inbreeding depression in island foxes. Curr. Biol. 28, 3487–3494.e4 (2018).

95. N. Dussex et al., Population genomics of the critically endangered kākāpō. Cell Genomics 1, 100002 (2021).
96. C. C. Kyriazis, R. K. Wayne, K. E. Lohmueller, Strongly deleterious mutations are a primary determinant of extinction risk due to inbreeding depression. Evol. Lett. 5, 33–47 (2021).
97. Y. D. Dewoody, J. A. Dewoody, On the estimation of genome-wide heterozygosity using molecular markers. J. Hered. 96, 85–88 (2005).
98. P. Casale et al., Mediterranean sea turtles: Current knowledge and priorities for conservation and research. Endanger. Species Res. 36, 229–267 (2018).
99. Y. Tikochinski et al., Mitochondrial DNA STR analysis as a tool for studying the green sea turtle (Chelonia mydas) populations: The Mediterranean Sea case study. Mar. Genomics 6, 17–24 (2012).
100. Y. Tikochinski et al., Mitochondrial DNA short tandem repeats unveil hidden population structuring and migration routes of an endangered marine turtle. Aquat. Conserv. 28, 788–797 (2018), 10.1002/aqc.2908.
101. S. Karaman et al., Population genetic diversity of green turtles, Chelonia mydas, in the Mediterranean revisited. Mar. Biol. 169, 77 (2022).
102. G. M. Hewitt, Genetic consequences of climatic oscillations in the Quaternary. Philos. Trans. R. Soc. Lond. B Biol. Sci. 359, 183–95 (2004), discussion 195.
103. F. V. Seersholm et al., Rapid range shifts and megafaunal extinctions associated with late Pleistocene climate change. Nat. Commun. 11, 2770 (2020).
104. I.-C. Chen, J. K. Hill, R. Ohlemüller, D. B. Roy, C. D. Thomas, Rapid range shifts of species associated with high levels of climate warming. Science 333, 1024–1026 (2011).
105. A. Brüniche-Olsen, K. F. Kellner, J. L. Belant, J. A. DeWoody, Life-history traits and habitat availability shape genomic diversity in birds: Implications for conservation. Proc. Biol. Sci. 288, 20211441 (2021).
106. J. P. van der Zee et al., The population genomic structure of green turtles (Chelonia mydas) suggests a warm-water corridor for tropical marine fauna between the Atlantic and Indian oceans during the last interglacial. Heredity 127, 510–521 (2021).
107. R. R. Fitak, S. Johnsen, Green sea turtle (Chelonia mydas) population history indicates important demographic changes near the mid-Pleistocene transition. Mar. Biol. 165, 110 (2018).
108. S. T. Vilaça et al., Divergence and hybridization in sea turtles: Inferences from genome data show evidence of ancient gene flow between species. Mol. Ecol. 30, 6178-6192 (2021), 10.1111/mec.16113.
109. A. H. Patton et al., Contemporary demographic reconstruction methods are robust to genome assembly quality: A case study in tasmanian devils. Mol. Biol. Evol. 36, 2906–2921 (2019).
110. K. Nadachowska-Brzyska, R. Burri, L. Smeds, H. Ellegren, PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white Ficedula flycatchers. Mol. Ecol. 25, 1058–1072 (2016).
111. United States National Marine Fisheries Service, U.S. Fish and Wildlife Service, "Endangered species act status review of the leatherback turtle (Dermochelys coriacea)" (United States National Marine Fisheries Service, 2020).
112. A. Khan et al., Genomic evidence for inbreeding depression and purging of deleterious genetic variation in Indian tigers. Proc. Natl. Acad. Sci. U.S.A. 118, e2023018118 (2021).
113. A. Prasad, E. D. Lorenzen, M. V. Westbury, Evaluating the role of reference-genome phylogenetic distance on evolutionary inference. Mol. Ecol. Resour. 22, 45–55 (2022).
114. A. Brüniche-Olsen, K. F. Kellner, C. J. Anderson, J. A. DeWoody, Runs of homozygosity have utility in mammalian conservation and evolutionary studies. Conserv. Genet. 19, 1295–1307 (2018).
115. A. D. Foote et al., Runs of homozygosity in killer whale genomes provide a global record of demographic histories. Mol. Ecol. 30, 6162–6177 (2021), 10.1111/mec.16137.
116. J. A. Robinson et al., The critically endangered vaquita is not doomed to extinction by inbreeding depression. Science 376, 635–639 (2022).
117. N. A. O'Leary et al., Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44, D733–D745 (2016).
118. K. D. Pruitt et al., RefSeq: An update on mammalian reference sequences. Nucleic Acids Res. 42, D756–D763 (2014).
119. M. Blum et al., The InterPro protein families and domains database: 20 years on. Nucleic Acids Res. 49, D344–D354 (2021).
120. H. Mi et al., PANTHER version 16: A revised family classification, tree-based classification tool, enhancer regions and extensive API. Nucleic Acids Res. 49, D394–D403 (2021).
121. B. Paten et al., Cactus graphs for genome comparisons. J. Comput. Biol. 18, 469–481 (2011).
122. J. Armstrong et al., Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature 587, 246–251 (2020).
123. F. K. Mendes, D. Vanderpool, B. Fulton, M. W. Hahn, CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics, 10.1093/bioinformatics/btaa1022 (2020).
124. D. M. Emms, S. Kelly, OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16, 157 (2015).
125. D. M. Emms, S. Kelly, OrthoFinder: Phylogenetic orthology inference for comparative genomics. Genome Biol. 20, 238 (2019).
126. G. Hickey, B. Paten, D. Earl, D. Zerbino, D. Haussler, HAL: A hierarchical format for storing and analyzing multiple genome alignments. Bioinformatics 29, 1341–1342 (2013).
127. J. A. Robinson et al., Genomic signatures of extensive inbreeding in Isle Royale wolves, a population on the threshold of extinction. Sci Adv 5, eaau0757 (2019).
128. T. S. Korneliussen, A. Albrechtsen, R. Nielsen, ANGSD: Analysis of next generation sequencing data. BMC Bioinformatics 15, 356 (2014).
129. S. Purcell et al., PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575 (2007).
130. F. C. Ceballos, P. K. Joshi, D. W. Clark, M. Ramsay, J. F. Wilson, Runs of homozygosity: Windows into population history and trait architecture. Nat. Rev. Genet. 19, 220–234 (2018).
131. P. Cingolani et al., A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly 6, 80–92 (2012).
132. A. McKenna et al., The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303 (2010).
133. H. Li, R. Durbin, Inference of human population history from individual whole-genome sequences. Nature 475, 493–496 (2011).
134. H. Li et al., The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078-2079 (2009).
135. H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27, 2987–2993 (2011).
136. Vertebrate Genomes Project, Reference genomes and associated read data for the green turtle (Chelonia mydas). GenomeArk. https://genomeark.github.io/vgp-curated-assembly/Chelonia_mydas/. Deposited 28 May 2021.
137. Vertebrate Genomes Project, Reference genomes and associated read data for the leatherback turtle (Dermochelys coriacea). GenomeArk. https://genomeark.github.io/vgp-curated-assembly/Dermochelys_coriacea/. Deposited 28 May 2021.
138. B. Bentley et al., Scripts associated with genome analyses of green and leatherback turtles. Github. https://github.com/bpbentley/sea_turtle_genomes. Deposited 24 June 2022.

Supplementary material

All scripts associated with these analyses have been deposited under GitHub repository https://github.com/bpbentley/sea_turtle_genomes.

Sample collection & data generation

The conservation status of leatherback and green turtles precludes the sacrifice of individuals to obtain tissue samples, so blood was collected using minimally invasive techniques for isolation of ultra-high molecular weight DNA from a male leatherback turtle off the coast of Monterey, California (NMFS ESA10a1A permit #21260 and USFWS Recovery Permit #TE-72088A-3) and a captive male green turtle in Israel National Sea Turtle Rescue Centre (INPA Permit worker 02457/2021 given to YL). Blood samples were flash frozen following collection and stored at -80°C until processing. Frozen subsamples of whole blood were placed in 1ml of 95-100% ethanol and processed using a modified version of the Bionano blood DNA isolation protocol optimized for frozen whole nucleated blood stored in ethanol (https://bionanogenomics.com/wp-content/uploads/2017/03/30215-Bionano-Prep-Frozen-Blood-Protocol.pdf). DNA quality was assessed using pulse field gel electrophoresis (PFGE) (Pippin Pulse, SAGE Science, Beverly, MA) or the Femto Pulse instrument (Agilent Technologies, Santa Clara, CA). DNA was then further prepared for the different library types (PacBio, 10X Chromium and Bionano optical map imaging) as described in Rhie et al. (2021). Hi-C of the green turtle was performed on flash-frozen blood following the Arima Hi-C protocol (Arima Hi-C user guide for Animal tissues, v01, Material Part Number: A510008).

Tissue samples of internal organs for RNA were collected opportunistically from recently deceased or euthanized animals in the US Virgin Islands, New England Aquarium, and the National Marine Fisheries Service Pacific Island Fisheries Science Center (NMFS permit #15685), flash frozen and stored at -80°C until processing. Total RNA was extracted placing 20-30mg of frozen tissue on dry ice and cut into 2mm pieces before being disrupted and homogenized with the Qiagen TissueRuptor II (Cat No./ID: 9002755), followed by extraction using Qiagen kits (leatherback turtle: gonad, lung and brain tissues using QIAGEN RNeasy kit, Cat. No. 74104; green turtle: brain, gonads, thymus, and spleen using QIAGEN RNeasy Protect kit, Cat. No. 74124). The quality and quantity of RNA were measured with a Qubit 3 Fluorometer (Qubit RNA BR Assay Kit, Cat no. Q33216; ThermoFisher Scientific, Waltham, MA) and a Fragment Analyzer (Agilent Technologies); RINs were within 7.5-9.5. Libraries were then prepared for short-read Illumina sequencing (RNA-Seq) and long-read PacBio sequencing (Iso-Seq). For RNA-Seq, aliquots of total RNA from each tissue and both species were sent to Psomagen (Rockville, MD) for library preparation (TruSeq stranded mRNA kits, Illumina) and sequencing. For the leatherback turtle, PacBio Iso-seq libraries were prepared according to the 'Procedure & Checklist - Iso-SeqTM Template Preparation for Sequel® Systems' (PN 101-070-200 version 05) without Blue Pippin size selection. Briefly, cDNA was reversely transcribed using the SMRTer PCR cDNA synthesis kit from 1 µg total RNA and amplified in a large-scale PCR. Two fractions of amplified cDNA were isolated using either 1x AMPure beads or 0,4x AMPure beads. Both fractions were pooled equimolar and went into the Pacbio SMRTbell template preparation v1.0 protocol following the manufacturer's instruction. For the green turtle, PacBio Iso-seq libraries were prepared according to the 'Procedure & Checklist – Iso-Seq™ Express Template Preparation for Sequel® and Sequel II Systems' (PN 101-763-800 Version 01). Briefly, cDNA was reverse transcribed using the NEBNext® Single Cell/Low Input cDNA Synthesis & Amplification Module (New England BioLabs, cat. no. E6421S) and Iso-Seq Express Oligo Kit (PacBio PN 10 1-737-500) from 300ng total RNA. Forward and reverse barcoded primers were used during cDNA amplification. PacBio Iso-seq

libraries were sequenced on one PacBio 8M SMRT Cell (PN: 101-389-001) on the Sequel II instrument with Sequencing Kit 2.0 (PN: 101-820-200) and Binding Kit 2.1 (PN: 101-843-000) and 24 hours movie with 2 hours pre-extension. Resulting raw data was deposited into the NCBI Short-Read Archive (SRA) for genome annotation (see Data Accessibility Statement).

Genome assembly & curation

Both genomes were assembled following the VGP pipeline v1.6 (Rhie et al. 2021) with a few modifications. Initially, all genomic data from each species were screened for low quality and contamination with Mash (Ondov et al. 2016) as described by Rhie et al. (2021). A preliminary analysis was performed using the 10X Illumina data (with 24bp-barcodes trimmed-off) and GenomeScope 2.0 (Vurture et al. 2017) to estimate the haploid genome length, repeat content, and heterozygosity and k-mer size of 21bp (Fig. S1). The predicted genome length was used to help select the amount of PacBio reads covering 50× of the genome. The selected PacBio reads were first corrected and subsequently assembled into partially phased contigs using FALCON and FALCON-unzip (Chin et al. 2016). The primary assembly was further purged of false haplotype duplications using purge_dups (Guan et al. 2020) and all removed regions were assumed to represent haplotype retention and added to the alternative assembly (Fig. S1). Scaffolding of the primary assembly was performed in three major steps. First, the 10XG linked reads were aligned to the primary contigs, and two scaffolding rounds were performed using scaff10x v2.2 (https://github.com/wtsi-hpag/Scaff10X). Subsequently, Bionano cmaps were generated using the Bionano Pipeline in non-haplotype assembly mode and used to further scaffold the assembly with Bionano Solve v3.2.1. We used the DLE-1 one enzyme non-nicking approach, and scaffold gaps were sized according to the software estimate. Finally, Hi-C reads were aligned to the Bionano cmaps scaffolded assembly using the Arima Genomics mapping pipeline (https://github.com/ArimaGenomics/mapping_pipeline), as described on Rhie et al. (2021). The restriction enzymes used to generate each library were specified using parameters *-e GATC, GANTC* for Arima reads. The processed Hi-C alignments were then used for scaffolding with Salsa2 (Ghurye et al. 2019) using the parameters *-m yes -i 5 -p yes*. In parallel, the mitochondrial genome was assembled by the mitoVGP pipeline (Formenti et al. 2021) using the corrected PacBio reads and 10XG reads as input.

Following the scaffolding steps, primary, alternative and mitochondrial assemblies were concatenated for two rounds of nucleotide polishing. As described in Rhie et al. (2021), a first round of polishing was performed with Arrow (Chin et al. 2013) using the PacBio CLR reads, followed by two rounds of polishing using the 10XG Illumina short-reads. For the latter, reads were first aligned to the assembly with Longranger align 2.2.2 (Garrison and Marth 2012) and variants were called with FreeBayes v1.2.0 (Garrison and Marth 2012) using default options. Consensus were called with bcftools consensus (Li et al. 2009). To minimize the impact of the remaining algorithmic shortcomings, both assemblies were subjected to rigorous manual curation (Howe et al. 2021). All data generated for both of the resulting assemblies; rDerCor1 and rCheMyd1 were collated, aligned to the primary assembly and analyzed in gEVAL (Chow et al. 2016); (https://vgp-geval.sanger.ac.uk/index.html), visualizing discordances in a feature browser and issue lists. In parallel, each species' Hi-C data were mapped to the primary assembly and visualized using Juicebox (Durand et al. 2016; Dudchenko et al. 2018) and HiGlass (Kerpedjiev et al. 2018). Based on identified mis-joins, missed joins and other anomalies from genome curators, the primary assembly was corrected accordingly. A second round of curation was performed after the synteny analysis between both genomes revealed a small number of remaining anomalies.

Genome annotation

Annotation was performed as previously described (Rhie et al. 2021; Pruitt et al. 2014), using the same RNA-Seq, IsoSeq and proteins input evidence for the prediction of genes in the leatherback and green turtle. A total of 3.5 billion RNA-Seq reads from eight the green turtle tissues (blood, brain, gonads, heart, kidney, lung, spleen and thymus) and 427 million reads from three leatherback turtle tissues (blood, brain, lung and ovary) were aligned to both genomes, in addition to 144,000 leatherback and 1.9 million green turtle PacBio IsoSeq reads, and all *Sauropsida* and *Xenopus* GenBank proteins, known RefSeq *Sauropsida, Xenopus*, and human RefSeq proteins, and RefSeq model proteins for *Gopherus evgoodei* and *Mauremys reevesii*.

Transposable element analysis

Transposable elements (TEs) from the genomes of the leatherback and green turtles were identified by creating a denovo database of transposable elements using RepeatModeller2 (Flynn et al. 2020) using the module -LTRStruct for each genome. Using this database, RepeatMasker (Tarailo-Graovac and Chen 2009; Smit, Hubley, and Green 2015) was run with the additional parameters of -a -s -gccalc to calculate kimura values for all the transposable elements identified using the script *calcDivergenceFromAlign.pl* with the parameters -s and -a. An inhouse script was also used, *align_with_divHandeler.py*, to isolate the TEs flagged as Unknowns from which each representative sequence of all TE families of Unknowns was isolated. Once isolated the distribution of size and number of transposable elements was analysed for both genomes for the complete scaffolds and for the low synteny regions using the inhouse script StatsTeRegion.py (Table S5); CheckNesting.py, Size_nesting.py (Table S4); Calculate_masking_size.sh (Figure S2) and *createRepeatLandscape.pl* with the same parameters used in the first iteration, to create the TE landscape presented in Fig. S5.

Genome alignment

The genomes of the sea turtles were aligned against each other using two outgroups. For this, genome assemblies of four turtle species (leatherback turtle, green turtle, *Gopherus evgoodei* [GCA_007399415.1] and *Mauremys reevesii* [GCA_016161935.1]) were first soft-masked with RepeatMasker to reduce the total number of potential genomic anchors formed by the many matches that occur among regions of repetitive DNA. Progressive Cactus, a reference-free whole genome aligner, was used (Paten et al. 2011; Armstrong et al. 2020) to align all other genomes applying the parameter --realTimeLogging. The guide tree and divergence time used as input for Cactus were retrieved from (Thomson, Spinks, and Shaffer 2021), with branch lengths reflecting neutral substitutions per site. To obtain an alignment only for the two sea turtles the parameter --root was used, setting as root the ancestral of the two sea turtles. For the alignment among all four turtles no root was set.

Analysis of regions of low synteny

Leatherback and green turtle genomes were mapped to each other using Minimap2 and a dot plot with the mappings was generated using D-GENIES (Cabanettes and Klopp 2018) to evaluate genome synteny and identify regions that presented low identity or structural rearrangements. Specifically, windows of 20 Mb were screened by eye in the dotplot, and every region bigger than 1 Mb presenting one or more breaks in the synteny was cataloged (Table S3). Some regions smaller than 1 Mb but larger than 100,000bp that contained obvious signals of genomic rearrangements were also cataloged for future analysis. To identify if these low

syntenic regions present differences in content or nucleotide composition, they were compared to two sections of the same length immediately upstream and downstream in the chromosome. In cases where the low syntenic region was located at one of the chromosome extremities, either two upstream or downstream sections were used for comparison for all of them (Table S3). The function of the genes present on those regions were extracted using the annotation results as well as the identification of protein domains using Interproscan (Blum et al. 2021). To verify if the low synteny regions present a pattern of higher sequence duplication, the Cactus alignment was analyzed. First, the tool hal2maf from HalTools (Hickey et al. 2013) was used to convert the output of cactus to the .maf format selecting (1) green turtle as reference and (2) leatherback turtle as reference. Also, using the coordinates for the low synteny regions, coding sequences (CDS) were isolated from the genomes fasta files based on the coordinates provided by the annotation file (.gff) using GFFreads tool (Pertea and Pertea 2020). A reciprocal blast (Aubry et al. 2014) was performed between the two species and, for each low synteny region, all homologous genes that presented more than one copy for one of the two species were isolated to retrieve duplicated genes using an inhouse script.

To determine if olfactory receptor (OR) genes were more numerous in one of the species throughout the genome in addition to the differences found within RRCs, we searched the annotation for the term "olfactory". Grep searches were performed on annotation files (gff) for both sea turtle species, *M. reevesii, G. evgoodei* and *T. scripta* in order to identify and compare gene numbers between these species. ORs were considered as Class I if numbered 51-56, while the remaining ORs were considered as Class II genes. After preliminary findings showing consistent higher gene copy numbers in the green turtle, we performed multiple analyses in order to rule out the possibility of collapsed multicopy genes in the leatherback turtle assembly. Specifically, we checked gene connections based on similarity for each set of gene copies manually, and estimated the predicted number of multicopy genes based on short read (Illumina 10X data) coverage for each RRC. Both analyses showed no evidence of gene collapse in the leatherback turtle.

Gene families and gene functional analysis

To estimate the timing of gene family evolution for the olfactory receptor gene families on sea turtles we used Computational Analysis of gene Family Evolution v5 (Mendes et al. 2020) https://github.com/hahnlab/CAFE5). CAFE5 uses phylogenomics and gene family sizes to identify gene families with rapid expansions and/or contractions for all branches in a phylogeny. First, we generated a dataset containing the numbers of OR genes for a dataset containing 8 species of turtles, 4 non-turtle reptiles, 3 mammals and 1 anura species using Orthofinder v 2.5.4 (Emms and Kelly 2015, 2019). OR orthogroups were grouped based on OR class I and class II subfamilies as described previously (Vandewege et al. 2016) and identified from the human genome (Glusman et al. 2001). We generated an ultrametric phylogeny by gathering all 1:1 orthologues identified by Orthofinder. We aligned amino acid sequences from each ortholog group with MAFFT v6.864b (Katoh and Standley 2013) using default parameters and trimmed with Trimal v1.4 (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009) using the "automated1" algorithm. Then we concatenated the trimmed alignments in a supermatrix using geneSticher.py (https://github.com/ballesterus/Utensils/blob/master/geneStitcher.py) and generated a tree with IqTree v2.1.4 (Minh et al. 2020; Nguyen et al. 2015), considering each orthogroup as a partition and with 1000 bootstrap. We then calibrated the tree using r8s (Sanderson 2003) with the same known evolutionary divergences based on fossil records used by (Wang et al. 2013).

We additionally searched the genomes for known TSD-related genes. We initially searched the annotation files (gff) using gene identification strings from our gene reference list using a 'grep' search. Given that some genes have many aliases depending on the lineages they were discovered in, and their function, we additionally applied a BLAST (Camacho et al. 2009)

search using orthologous protein sequences pulled from the NCBI protein database. We used 'tblastn' (e-value =$1e^{-3}$; max_target_sequences=5; and max_hsps=10) to query the protein sequences against the genome, and where possible, pulled down sequences from the species where the gene had been previously implicated in TSD. The majority of the gene sequences were sourced from *Trachemys scripta scripta*, *Chrysemys picta belli*, and *Alligator mississippiensis* (but see Table S7). Matches were then filtered downstream such that only sequences with ≥90% identity matches were retained, and positions of matches were checked against the annotation file. Results from grep and BLAST searches were then examined and compiled to create a comprehensive list of TSD genes for each of the two genomes. To compare the position of the genes within the genome, the positions of each gene were plotted on a Circos plot using CIRCA (http://omgenomics.com/circa).

Genome-wide heterozygosity

We used the 10X Genomics paired-end reads generated for the leatherback and green turtles and aligned them back to their respective primary assembly to conduct analyses of genome-wide diversity and historical demography. To apply standard mapping and genotype calling pipelines to the data, we first removed 10X linked barcodes from the raw reads using the script 'process_10xReads.py' (Andrews et al. 2012). Reads were aligned to the reference with BWA-MEM v0.7.17 (Li 2013) using default parameters. PCR duplicates were removed and read group headers were added with Picard-Tools v2.23.2 using the MarkDuplicates and AddOrReplaceReadGroups functions, respectively (http://broadinstitute.github.io/picard). The resulting alignment files for each species were used for all downstream analyses described below.

Genome-wide heterozygosity was calculated using a sliding-window approach adapted from methods described in (Robinson et al. 2019), and using the Genome Analysis Toolkit (GATK; v4.1.8.1 (McKenna et al. 2010)). HaplotypeCaller was applied to identify and call loci in the emit reference confidence mode with base-pair resolution (-ERC BP_RESOLUTION), with the output GVCF file containing both variant and non-variant sites. Genotypes at each site were then generated from this output using GenotypeGVCFs, including at the non-variant sites. We removed unused alternate alleles from the genotypes using SelectVariants, and then filtered the VCF file based on depth of coverage (⅓× - 2× mean coverage) and genotype quality scores (MinQ = 20) at each site using an inhouse python script. We used the resulting filtered VCF file to estimate heterozygosity (π) in 100 Kb non-overlapping windows across the genome. To ensure the number of callable sites didn't influence our results, we calculated heterozygosity as the number of heterozygous sites divided by all sites that passed filtering steps, and only retained windows that contained a minimum of 80 Kb callable sites. Heterozygosity estimates for regions without a known location in the genome (i.e. unplaced scaffolds) were not included in calculations. We also estimated heterozygosity for subsets of the genome using the same methods as above, using an input BED file to specify the regions of interest. Specifically, we targeted regions that: (1) were not identified as containing repeat or low-complexity sequences (i.e. the 'masked genome', see *Transposable element analysis* section above), (2) were identified as exon regions through the annotation and (3) non-exon regions (i.e., regions not identified as exons, identified by inverting the exon region BED file using BedTools v2.29.2 (Quinlan and Hall 2010). For the windows containing exons, we examined the genes associated with regions of high diversity by extracting the annotation information for windows that had a proportion of heterozygosity higher than 3× SD above the mean. Gene lists were then run through PANTHER (Mi et al. 2021) to investigate gene ontology (GO) terms.

To directly compare heterozygosity between the two sea turtle species, we also mapped the 10X barcode removed reads to the reference genome for *Mauremys reevesii* (Liu et al. 2021) using the same methodology as described above for alignment, duplicate removal and genotype calling as described above, using scaffolds that were at least 10 Mb in length (*N*=43, ~98% of

the genome), and estimates diversity for whole-genome and exons. We then compared heterozygosity in corresponding exon windows between both species, and identified windows that had either (1) substantially higher heterozygosity in one species than the other, i.e. heterozygosity was greater than three times the mean in one species but not the other; or (2) exceptionally higher heterozygosity in both species, where heterozygosity was greater than three times the mean in both species. Following this identification, annotations of genes present in these windows were extracted and explored to determine differences between the two species.

To examine the context of the genomic diversity found in the two sea turtle species, we also directly estimated the genome-wide heterozygosity for a number of other reptile species (N=13). As the software and parameters used for genotyping can directly influence the heterozygosity estimates (see (Prasad, Lorenzen, and Westbury 2022)). We downloaded raw reads associated with reference genome assemblies from the EBI-ENA database and employed a standardized mapping and genotyping pipeline to generate comparable heterozygosity estimates. The heterozygosity pipeline is similar to that described above for the two focal species with slight alterations: if data was generated with 10X Chromium linked-reads, the first 22bp of the R1 read were trimmed using trimmomatic v0.39 (Bolger, Lohse, and Usadel 2014). Following this, paired and trimmed reads were used as input for trimmomatic with default parameters, before being aligned to the reference genome with BWA-MEM, having duplicate reads removed and read group headers added with Picard-Tools. The resulting alignment files were then used with the GATK pipeline described above, using 100 Kb windows, and only retaining scaffolds that were at least 100 Kb in length. Windows were discarded from downstream calculations if they contained fewer site calls than one standard deviation from the mean number of calls.

To determine the impact of genotype calling method, we also generated genome-wide heterozygosity using the Analysis of Next Generation Sequencing Data software (ANGSD; v0.933(Korneliussen, Albrechtsen, and Nielsen 2014)). To achieve comparable results to the GATK heterozygosity pipeline, we initially re-aligned the consensus genome around insertion-deletion (indel) sites using the RealignerTargetCreator and IndelRealigner functions included in GATK (v3.5), as this step is automatically included in the GATK analysis software (> v4.0). The indel realigned bam file was used as input for ANGSD, with site allelic frequencies calculated (-doSaf) using SamTools v1.9 (Li et al. 2009) genotype-likelihoods (-GL1), and the same depth and quality filters as those applied in the GATK pipeline applied. Site allelic frequency files were then parsed through the realSFS function in ANGSD to calculate the site frequency spectra (SFS), with the outputs used to calculate heterozygosity within 100kb windows which were generated through bedtools MakeWindows function.

Runs of homozygosity

To detect autozygosity within the genome, we used the PLINK v 1.90b6.9 SNP-based runs of homozygosity (ROH) analysis (Purcell et al. 2007). Given that the high-coverage (hcWGS) data used in these analyses was the same as that used to assemble the genome, and PLINK requires homozygous alleles at variant sites to call a ROH, we generated low- to medium-coverage (2-12×) whole genome resequenced data (WGR) from five individuals of each species to identify variant sites. Data for whole-genome resequenced individuals per species was generated through a Novaseq 6000 S4 using Illumina 150bp paired-end sequencing. We generated low-coverage data for green turtles from two populations, and medium-coverage data from three populations of leatherback turtle (Table S11). To ensure that coverage was not influencing downstream results, we also down-sampled leatherback turtle data to match green turtle data, and re-ran the ROH analysis, with our results not impacting the general qualitative patterns of ROH distributions. Importantly we note that our aim was not to present findings for

the WGR individuals as this is part of a companionate study, but to produce a SNP-list that could be used to detect ROHs within the hcWGS reference sample.

Briefly, we trimmed and aligned reads to the respective reference genomes, before removing PCR duplicates, adding read-group headers, and re-aligning around indels. These alignment files were then used with indel re-aligned files produced for the genome data, and used with ANGSD (Korneliussen, Albrechtsen, and Nielsen 2014) to generate a SNP-list in the form of a PLINK file with a posterior probability cutoff of 0.95 and a SNP p-value of $1e^{-6}$. The ANGSD-generated SNP-list containing all WGR samples and the genome sample was then run through PLINK to determine the distribution of ROHs across the genome for each individual using a minimum ROH length of 100 Kb (--homozyg-kb 100), a minimum of 20 SNPs (--homozyg-snp 20), an allowed missingness of 10 sites (--homozyg-window-missing 5), and a maximum of 3 heterozygous sites allowed per window to account for sequencing error (--homozyg-window-het 1). The PLINK outputs were then exported and analyzed using the R environment (R Core Team 2020). Only the high-coverage genome data was used for analysis. ROHs were segregated into length classes, with 'small' ROHs between 100-500 Kb in length, 'medium' ROHs, 500 Kb-1 Mb in size, and 'long' ROHs >1 Mb in length. Total aggregate lengths were calculated for each length class, and the proportion of each chromosome in ROH was calculated by dividing the aggregate length of ROHs by the total chromosome size.

Given that genotype-likelihood information is lost when running ANGSD to generate a SNP-list in the format of a PLINK file (as required for the PLINK ROH analysis), we also ran the medium-coverage leatherback turtle whole-genome resequenced samples through a GATK pipeline. Briefly, we used HaplotypeCaller on the individual data with the ERC parameter set to output one GVCF file to generate one file per sample including only variant sites. These were then combined using the GATK GenomicsDBImport function, with joint genotypes called using GenotypeGVCFs. The output VCF file, which contained variant sites for each of the five WGR samples as well as from the high-coverage reference individual, was filtered for mean depth (--min-meanDP 6, --max-meanDP 1000), as well as number of minor alleles required to call a heterozygote (--mac 3), and a minimum base quality threshold of 30 (--minQ 30). The filtered VCF file was then used with the same parameters as the ANGSD generated SNP-list in the PLINK ROH analysis function using the VCF read-in parameter (--vcf).

Demographic history

The demographic histories of leatherback and green turtles were inferred using the pairwise sequential Markovian coalescent (PSMC) (Li & Durbin 2011). To process the data for PSMC we used samtools v1.11 (Li et al. 2009) and bcftools v1.6 (Li 2011) to call variants, requiring base and mapping qualities of 30. We performed additional filtering by insert size retaining reads between 50-5000 bp, to remove potentially spurious short alignments. To mitigate the possibility of spurious heterozygotes we filtered by allele balance (AB), removing biallelic heterozygotes with AB<0.25 or AB>0.75 and filtered by repeat-masked positions. We retained the first 10 'SUPER' scaffolds, which do not include any sex-linked chromosomes as sex-determining genes are not localized to discrete sex chromosomes in sea turtles. Following protocol (Li and Durbin 2011) retained sites between a third of the average read depth (-d) and twice the average read depth (-D). We applied PSMC using the parameters -N25 -t15 -r5 -p "4+25*2+4+6", and scaled the output using a mu of 1.2*10-8 (Fitak and Johnsen 2018) and a generation time of 30 years (which is the midpoint between literature estimates for the two species). We additionally plotted the PSMC outputs using species-specific generation times for each species, with values of 14 and 42.8 for leatherback and green turtles respectively. This scaling factor produced negligible impacts on the curves for N_e, with the 30-year generation time used for all downstream tests.

To rule out that increases in N_e for the PSMC analyses for both species were an artifact of using the same individual that was sourced for genome assembly, we ran the same pipeline

for one additional individual for each species (Fig. S17). For the leatherback turtle, we aligned reads from a moderate-coverage (~13×) individual that was also used for the purposes of the ROH analysis. For the green turtle, we ran the PSMC analysis using the raw reads that were used to assemble the initial green turtle draft genome by (Wang et al. 2013). In both cases, reads were trimmed and aligned to the respective genomes following the methods described previously, before following through the PSMC pipeline used for the two focal individuals.

Genetic load

In the absence of genetic diversity, deleterious recessive alleles are more likely to be expressed, however, highly deleterious alleles should be purged from the population as they are less likely to be masked in a heterozygous state (Grossen et al. 2020). In order to examine deleterious allele accumulation and genetic load, we extracted variants from coding regions for both species using the outputs from the GATK analysis of heterozygosity within the exonic regions. Given the stringency of base and map quality (Q>20), as well as site depth filtering (⅓× < depth > 2×), all variants are considered to be reliable and of high quality. These variants were then annotated using snpEff (Cingolani et al. 2012), where each variant was designated as either 'modifier', 'low', 'moderate', or 'high' impact. Proportions of each type of variant were then compared between species. SnpEff also calculated the silent to missense ratio of variants, with higher ratios showing a higher proportion of variants that are expected to have an effect on amino acid sequences.

Extended Results

Analysis of regions of low synteny

Here we provide in-depth descriptions of gene function and copy number comparisons between the two sea turtle species found in each region of low synteny. See Tables S3 and S5 for complete details. Two regions of low identity were identified on chromosome 1 from 1 Mb to 8 Mb for the green turtle and 1 Mb to 6 Mb for the leatherback turtle for region A, and from 210.8 Mbp to 214.4 Mbp for the green turtle and 215.7 Mb to 216.85 Mb for the leatherback turtle for region B. Inside region B, an unusual string of Ns was observed for the green turtle (51.2% of the total region length). The 3.5 Mb region was analyzed together with the same length section upstream and downstream for both green and leatherback turtles. The cactus alignment detected that both species exhibited more than 4 times duplications in this region, and the duplications are at least double in base-pair lengths, compared to surrounding regions (Table S3). We further selected only duplications larger than 21, 100, and 500bp for examination, and in all the cases the pattern remained the same for the region of low identity. Additionally, there was a small increase in the amount of TEs for this region in the leatherback turtle (35;**46**;30 number of TEs in up to downstream order), but no difference in the green turtle (39;**35**;34 number of TEs in up to downstream order), possibly as a result of the high proportion of Ns in the green turtle for this region (Fig S5). Region A presented 59 genes with functions associated with Olfactory Receptors (OR) in the leatherback turtle, while the corresponding region for the green turtle presented a total of 256 OR gene copies (Table S5). The region B of chromosome 1 also presented multiple copies of three genes related to the Immune System (antigen WC1.1-like, TAPASIN and one gene containing Scavenger receptor cysteine-rich domain) for the green turtle compared to the leatherback turtle. We additionally checked for a possible association between the RRCs and TEs by comparing the RRCs with regions up- and down-stream, and found that the number of TEs was similar between these regions (Table S5). However, all large RRCs (> 1 Mb) in the green turtle that were associated with gene copy number differences had larger average TEs, potentially indicating an association of differential

activity of TEs and structural differences in associations with gene copy number variations between species.

Two regions of low synteny were found on chromosome 2, region 2A (0 - 2.2 Mbp green turtle and 0 - 2.4 Mbp on the leatherback turtle) were associated with the presence of a duplication of one gene related to sphingomyelin phosphodiesterase 5 for the green turtle. The beginning of chromosome 4 also encompassed a region of low synteny (0 - 4.5 Mbp green turtle and 0 - 3.03 Mbp leatherback turtle) where multiple copies of genes related to the immune system (erythroid membrane-associated protein/butyrophilin and major histocompatibility complex class I) and one gene containing maestro-related heat domain were found for the green turtle. In chromosome 6, two low identity regions were identified at the beginning of the chromosome sequence. The first one (6A- and 0 - 15.47 Mbp green turtle and 0 - 7.67 Mbp leatherback turtle) contained potential gene duplication for genes related to olfactory receptors, the immune system and zinc-fingers for the green turtle compared to the leatherback turtle (see details in Table S3), while the second (6B) contained one gene of the immune system (NACHT 2C LRR and PYD domains-containing protein 3) with three copies on the green turtle compared to one on the leatherback turtle. The low synteny region on chromosome 8 (8A - 61.7 - 2.7 Mbp green turtle and 63.53 - 64 Mbp leatherback turtle) included the immune system gene complement factor H with 3 copies in the green turtle and 1 in the leatherback turtle. On chromosome 11, one region of low identity (11A - 74. 2 - 79.5 Mbp green turtle and 80.0 - 80.022 Mbp leatherback turtle) had multiple copies of zinc-finger genes for the green turtle compared to the leatherback turtle. Chromosome 12 presented a large inversion in the beginning of the chromosome; however, no signs of gene duplication were found for this region (3.004 - 7.090 Mbp green turtle and 3.296 - 7.396 Mbp leatherback turtle). As was found for chromosomes 1 and 6, multiple copies of genes related to the immune system and OR were found on a region of low synteny on chromosome 13 (13A - 32.3 - 42.95 Mbp green turtle and 33.3 - 41.16 Mbp leatherback turtle), and chromosome 14 (14A - 26.5 - 44.3 Mbp green turtle and 27.6 - 40.02 Mbp leatherback turtle). While the first region of low synteny identified on chromosome 15 did not present signs of gene duplication, the second region (15B - 13.7 - 14.3 Mbp green turtle and 13.3 - 13.6 Mbp leatherback turtle) had eight copies of one gene related to immunoglobulin lambda constant 1 for the green turtle compared with one copy for the leatherback turtle. Chromosome 20 presented duplication signs for genes related to Keratin type II head, adhesion G protein-coupled receptor E1 in the low synteny region 20A (4.9 - 14.1 Mbp green turtle and 4.8 - 14.7 Mbp leatherback turtle). The low synteny region found on chromosome 21 did not present signs of gene duplication. Chromosome 23 presented one of the larger regions of low synteny (6.0 - 19.3 Mbp green turtle and 5.9 - 17.23 Mbp leatherback turtle) with multiple copies of genes from immune system, reproductive system and iron homeostasis for the green turtle compared to the leatherback turtle. Additionally, chromosome 24 displayed rearrangements that were confirmed using 10X data as biologically real (Fig. S3; 24A - 12.2 - 19.2 Mbp green turtle and 11.6 - 16.95 Mbp leatherback turtle) containing multiple copies of genes from the immune system and maintenance of the mucosal structure (IGGFC-binding protein) again for the green turtle relative to the leatherback turtle. Finally, chromosome 28 was one of the largest low synteny regions, corresponding to the entire chromosome and included the presence of multiple copies of zinc-finger genes in the green turtle. All the genes present in multiple copies for the green turtle are shown in Table S3. The low synteny regions present on chromosome 2 (2B), 3 (3A), 5 (5A and 5B), 12, 15 (15A), 21, and 26 did not contain genes or signs of gene duplication. Other functions of genes with higher copies for the green turtle within RRCs included lipid metabolism (region 20A and 24A), cornification (region 20A), response to hypoxia (region 23A), and mucus production (region 24A).

Genome diversity

Genome-wide diversity was approximately seven-times lower in the leatherback turtle compared to the green turtle, irrespective of whether repeat regions were masked in the analysis (unmasked, masked $\pi = 3.47 \times 10^{-4}$, 3.19×10^{-4} leatherback turtle and 22.3×10^{-4}, 22.2×10^{-4} green turtle; Figs. 4a & S11-13). At the chromosome level, variation was relatively evenly spread across the genome in both species (SD = 4.3×10^{-4} and 1.7×10^{-3}, respectively), but generally higher in the microchromosomes. In particular, diversity within the smallest chromosome (chromosome 28) was almost double the overall mean in both species (Figs. S12 & S13) despite containing approximately the same quantity of genes as other microchromosomes. Exons had lower levels of heterozygosity than the non-coding regions (Fig. 4a). The proportion of heterozygous sites (number of heterozygous sites/total callable sites) within 100KB non-overlapping windows across the genome ranged from 0 to 0.028 for the leatherback turtle, and from 0 to 0.061 for the green turtle. From the 21,285 and 20,709 windows that passed filtering steps, 610 (2.87%) and 1,367 (6.60%) contained zero heterozygous sites for the leatherback turtle and the green turtle respectively, suggesting diversity was lower overall in the leatherback turtle, but more evenly spread than the green turtle.

To identify genes with high diversity relative to baseline genome variation, we extracted exon-containing 100 Kb windows that had higher proportions of heterozygous sites than the mean for each species (see Methods) and identified 1,945 and 3,987 exons for the leatherback turtle and the green turtle, respectively (Table S10). Windows containing tRNA genes showed high heterozygosity for both species; however, the only specific genes observed in both species were *EPHA3* and *CHID1,* which encode an ephrin receptor and a response protein to excess calcium, respectively. Though a large proportion of the unique genes these exons comprise were with unannotated gene identifiers in both species (171 out of 302 for the leatherback turtle; 439 out of 506 for the green turtle), analysis of the annotated unique genes with PANTHER showed that the genes were involved with biological processes including development, locomotion, growth, response to stimulus and signaling (Fig. S14). The leatherback turtle also showed high diversity in genes associated with reproductive processes. Examination of the annotated molecular functions from these exons revealed many with diversity in the leatherback turtle were related to cell adhesion, transport, and binding, while in the green turtle, they were associated with olfactory reception, immunity, tumorigenesis, and zinc finger proteins (Table S10). In both species, these high diversity regions also included rRNA genes, as well as genes involved with biological processes including development, locomotion, growth, response to stimulus and signaling. The leatherback turtle also had high diversity in genes associated with reproductive processes (Fig. S14).

When aligned to a common reference (*M. reevesii*) as opposed to themselves, we found similar results, with the diversity of the green turtle generally higher than the leatherback turtle (Fig. 4c), albeit with a dampened difference between species (Table S10). In regions where diversity was high for both species (see Methods), many olfactory receptors were once again present, as were T-cell receptors, other immune-related genes (e.g. MHC related genes), maestro heat-like repeat-containing family members, and zinc finger proteins (Table S10). For regions that were only indicated to have high diversity in the leatherback turtle, the genes within these regions were linked to some olfactory receptor genes, zinc finger proteins, and genes involved with signaling. Olfactory receptor genes were present in a higher number in the regions of high diversity in the green turtle, as were many immune-related genes, including genes linked to the MHC. When compared to estimates from other non-avian reptiles generated using a standardized heterozygosity pipeline, we show that the leatherback turtle possesses very low genomic diversity (Fig. 4b), with estimates lower than even that of the well documented extinct *Chelonoidis abingdonii (Quesada et al. 2019)*. The green turtle diversity falls midway between the other species, with estimates close to that of *Gopherus evegoodei* (Rhie et al. 2021). Diversity did not correlate with the conservation status for the species examined.

Searches for genes related to the core region of the MHC

We further investigated immune genes associated with the core MHC region and found substantial differences between the leatherback turtle and the green turtle (Table S12). Out of the core set of MHC genes (Gemmell et al. 2020), 46 were present in the leatherback turtle and 39 in the green turtle, similar in number to those found in *Chrysemys picta bellii* and *Alligator mississippiensis* using the same gene set (Gemmell et al. 2020). Several genes were missing in both species, suggesting that either these genes have been lost in sea turtles, are too variable to be effectively annotated, or that this region still contains gap-rich regions. Eleven genes present in the leatherback turtle genome were absent from the green turtle, including *BAG6*, *DDX39B*, *RNF5*, and *STK19*, but only four genes that were present in the green turtle versus the leatherback turtle (*KIFC1*, *LTA*, *TAP1,* and *TAP2)*. Excluding the MHC Class I and II genes, all core MHC-related genes were found on chromosome 14, except for *C4*, which was found on chromosome 1 in both species. In the green turtle, the *ATFB6*, *NOTCH4*, and *PRRT1* genes were additionally located on an unplaced scaffold (NW_025111287.1), while these were found on chromosome 14 in the leatherback turtle. This suggests that the assembled MHC region in the green turtle genome may be partly fragmented. Examination of MHC Class I genes suggested that multiple copies were present on chromosome 14 in both species (Fig. 2d), with seven copies found in the region for the leatherback turtle and six copies found for the green turtle, with an additional copy located on another unplaced scaffold (NW_025111276.1). There were two additional copies of the MHC Class I α gene in both species that were not located within the core MHC region on chromosome 14, with a single copy located on chromosomes 4 and 5.

Supplemental Figures

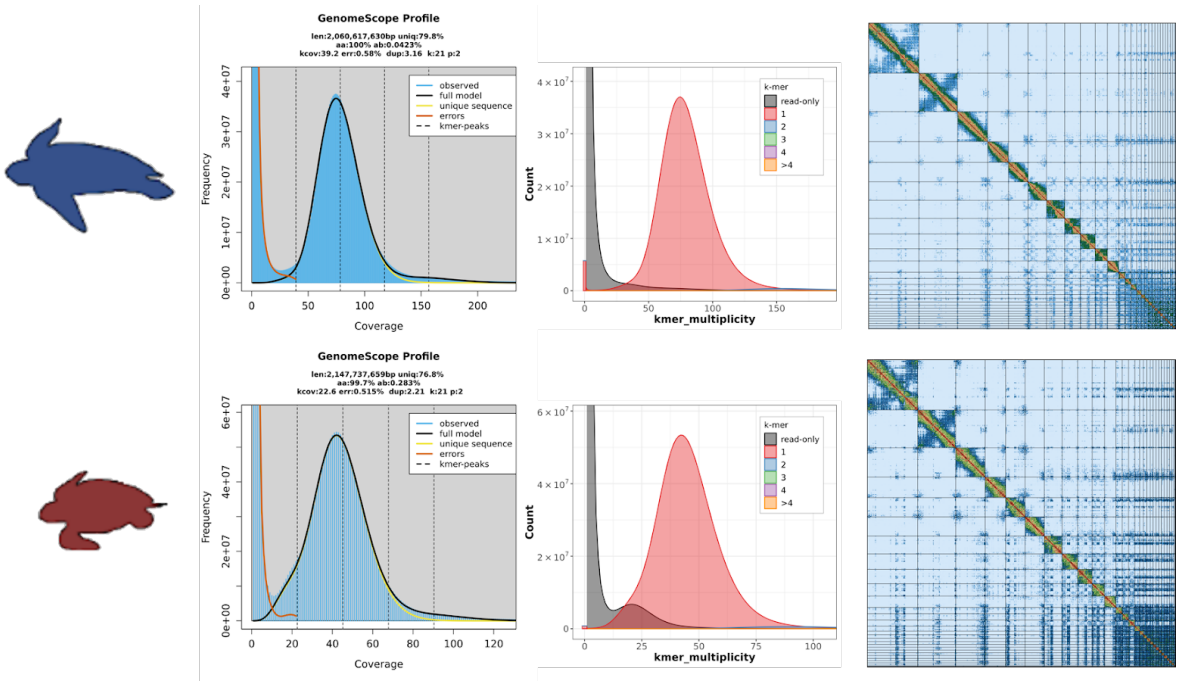


Fig. S1 | Quality control plots for the genome assemblies of *Dermochelys coriacea* (upper) and *Chelonia mydas* (lower) turtles. Plots from left to right; Genoscope profile for 21-mers collected from 10X linked reads using Meryl (https://github.com/marbl/meryl).; K-mer spectra plots for both genomes assemblies produced using KAT, showing the frequency of *k-mers* in the assembly versus the frequency of *k-mers* in the raw 10X linked reads. ; Hi-C maps contact map (Pretext https://github.com/wtsi-hpag/PretextView) for the complete assembly. Plots from left to right represent the kmer distribution profile from short reads (GenomeScope 2.0); the kmer multiplicity of reads coloured by the number of times each kmer appears in the assembly; and the contact map based on Hi-C short-read data produced using PreText.

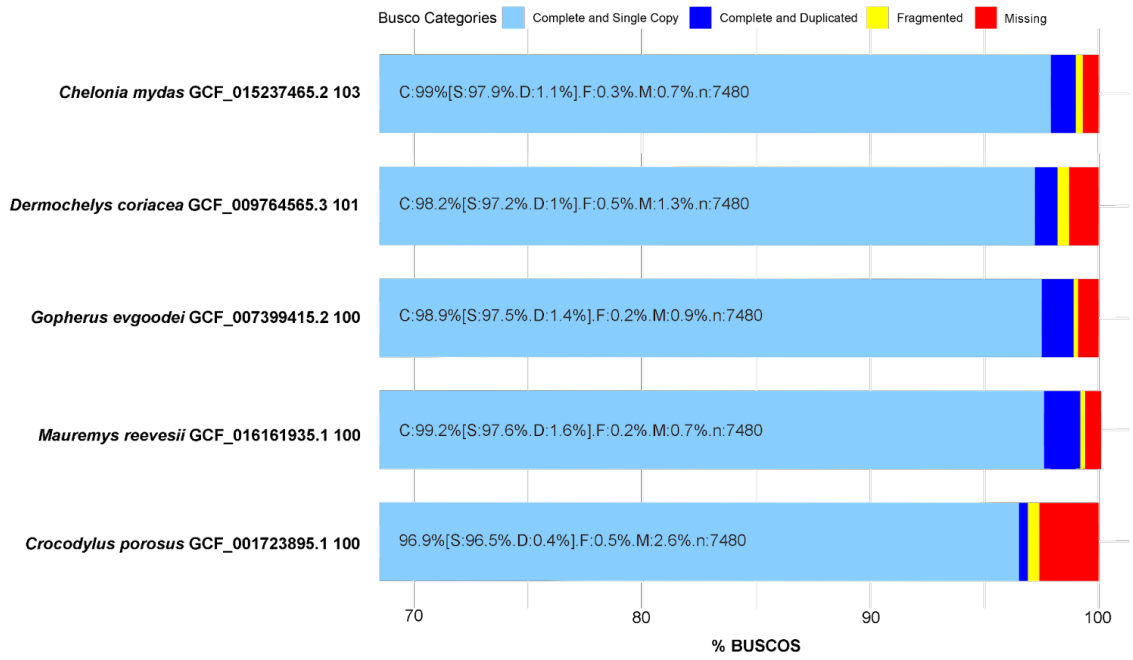
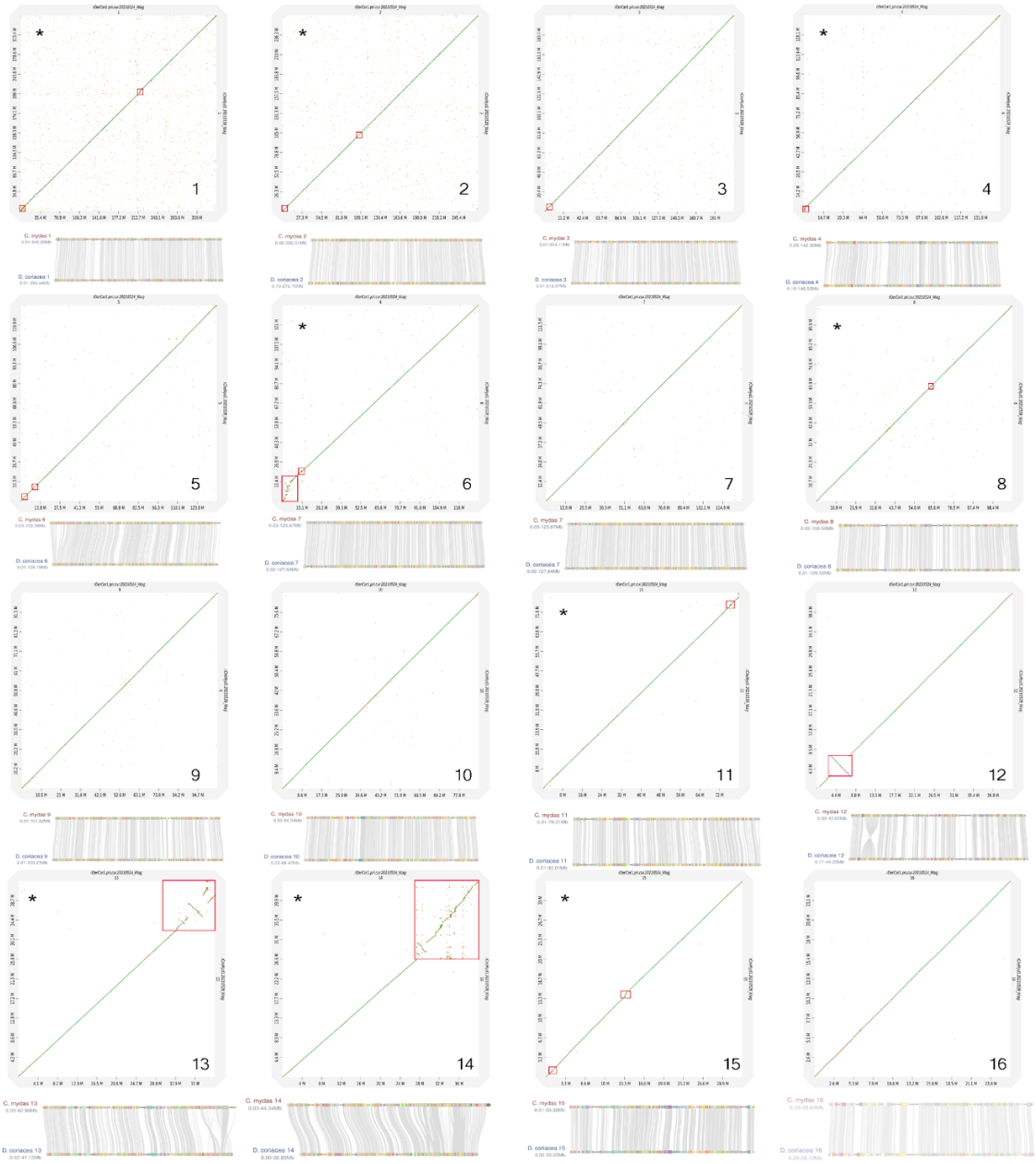


Fig. S2 | Comparison of the completeness of gene annotations, as a percentage of sauropsida_odb10 from BUSCO.

Dermochelys coriacea



Chelonia mydas

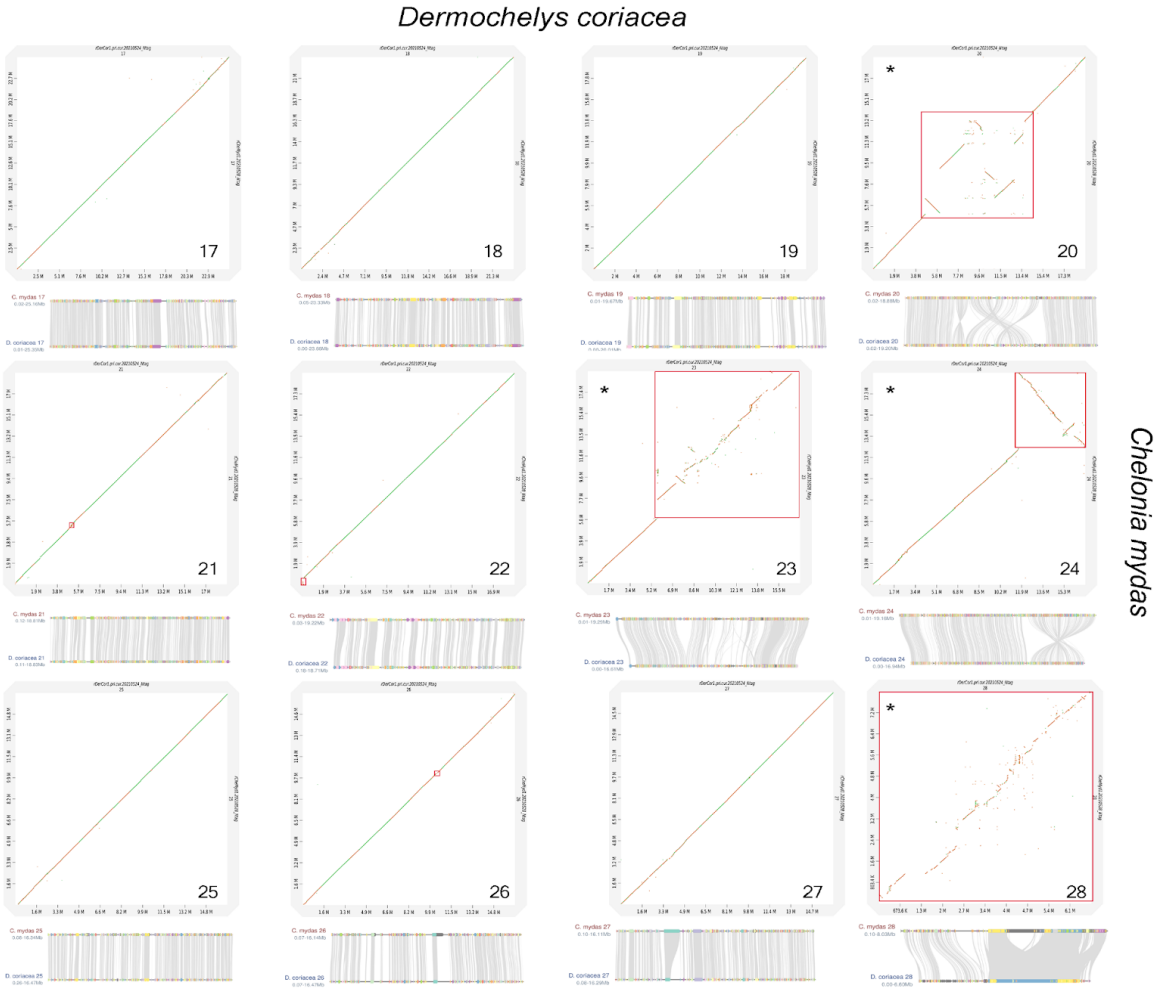


Fig. S3 | Dot plot analysis for all individual chromosomes in the leatherback turtle (*Dermochelys coriacea*) and the green turtle (*Chelonia mydas*) genomes, with identified regions of low synteny denoted by red boxes (top panel, each chromosome), and gene synteny analysis (bottom panel, each chromosome). The colored blocks with the same color in gene synteny graphs represent orthologous genes and the grey lines represent the links between them in the two species. At the genomic level, near end-to-end synteny was observed in 9 chromosomes (chromosomes: 7, 9, 10, 16, 17, 18, 19, 25, and 27), while from the remaining 19, 8 exhibited lower synteny restricted to specific sub-regions (>0.1Mbp - 3Mbp; chromosomes: 2, 3, 5, 8, 15, 21, 22, and 26), and 11 present low synteny regions larger than 3Mbp (chromosomes: 1, 4, 6, 11, 12, 13, 14, 20, 23, 24 and 28). Of the 19 chromosomes with regions of low synteny, the 13 that exhibited putative gene duplications within these regions are denoted by (*) in the upper left graph corner. The low synteny regions found on chromosomes 1, 4, 6, 8, 13, 14, 15, 20, 23, and 24 present multiple copies of genes related to immune system and/or olfactory reception in *C. mydas*. See details of region locations and compositions in Table S3.

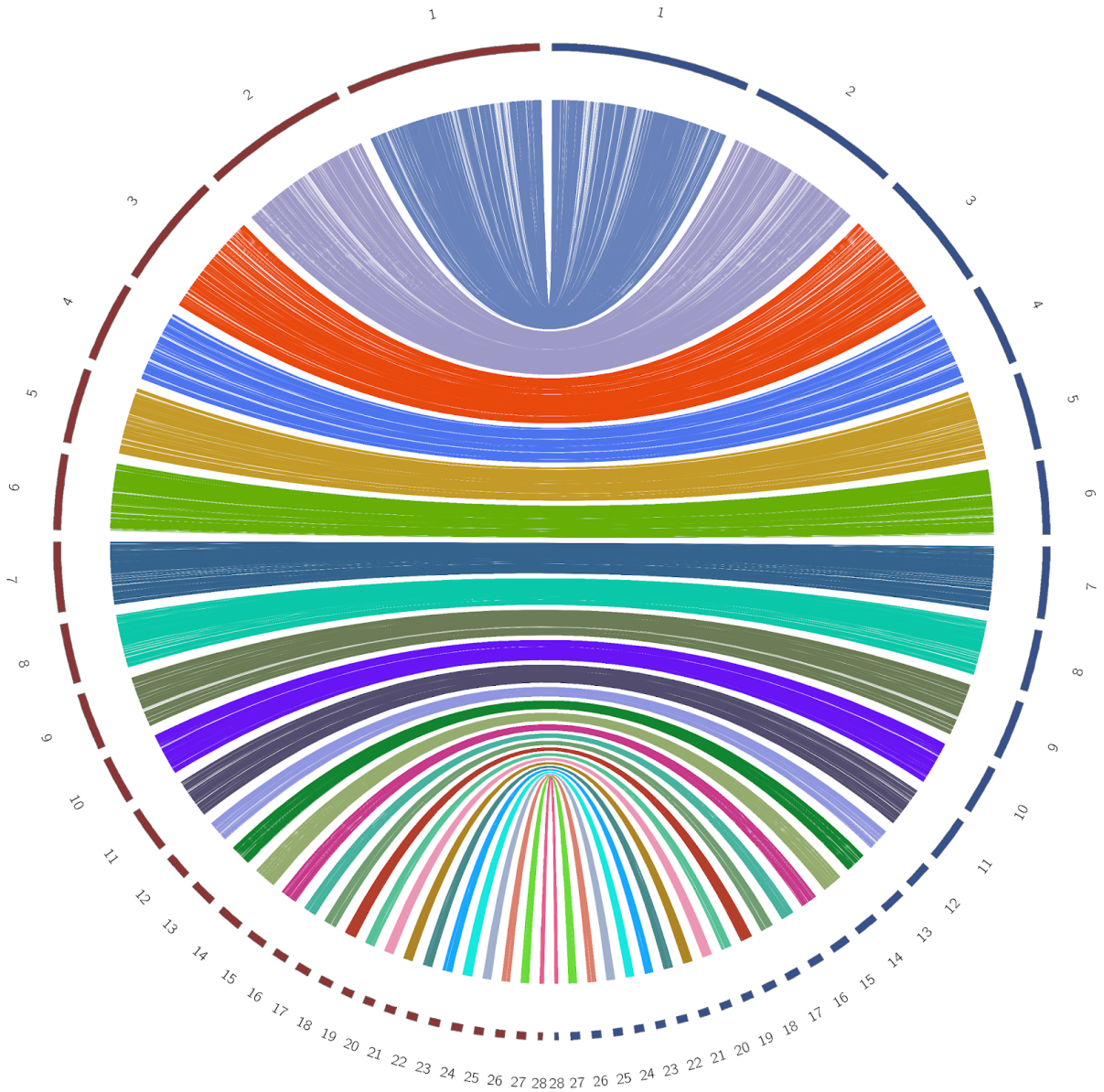


Fig. S4 | Circos plot for the genomes of the leatherback turtle (*Dermochelys coriacea*) and the green turtles (*Chelonia mydas*) showing high synteny, with the outer rings showing respective chromosome numbers for *C. mydas* (red) and *D. coriacea* (blue).

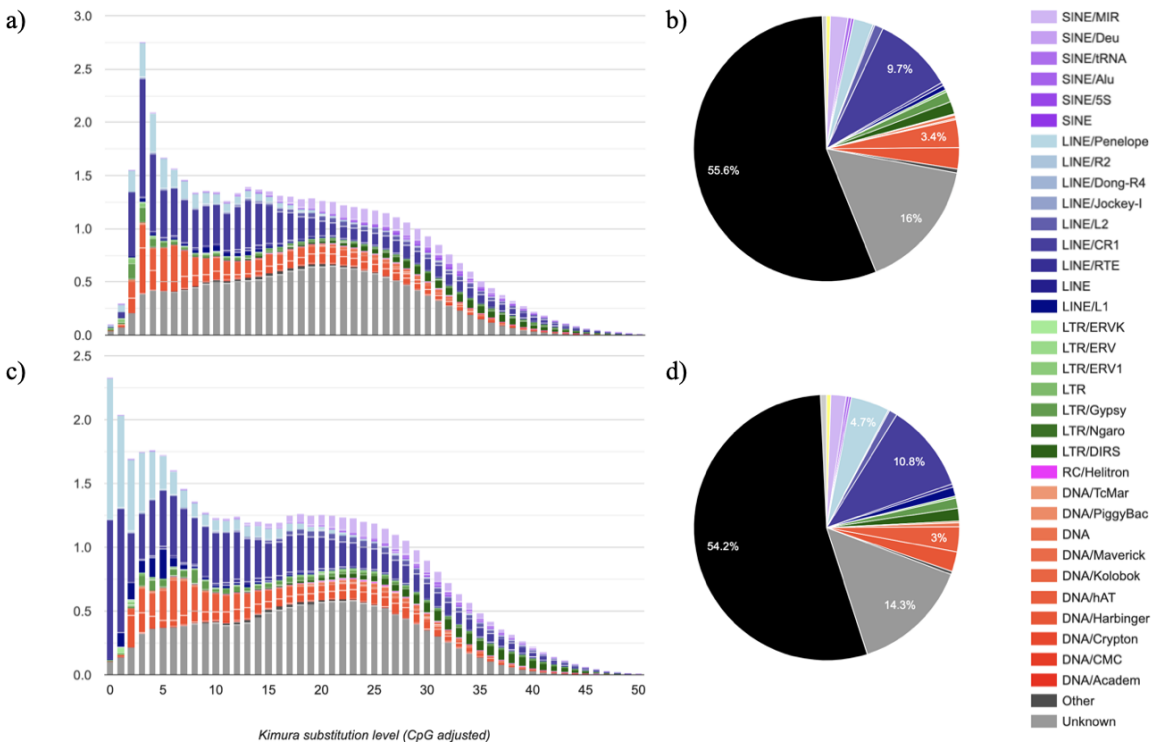
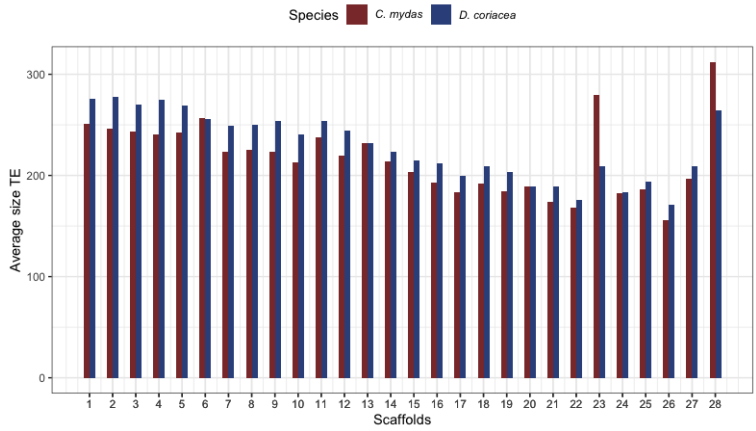
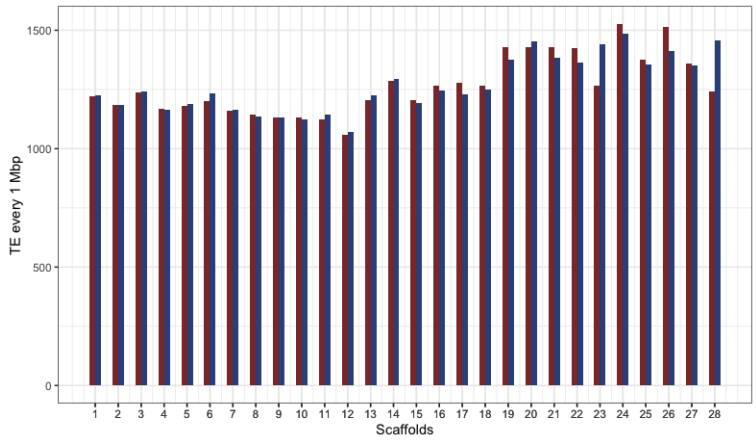


Fig. S5 | Repeat element (RE) landscape for *Chelonia mydas* (a,b) and *Dermochelys coriacea* (c,d). Colors in the stacked bar charts and pie charts correspond to the transposable elements subfamilies and Unknown REs as indicated in the key, with the proportion of the unmasked genome depicted in black in b and d. See Table S4 for details.

a)



b)



c)

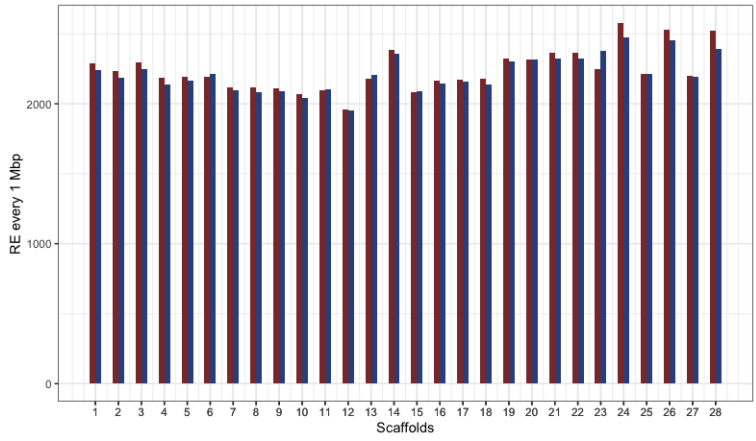


Fig. S6 | Distribution of **(a)** average size in bp of classified transposable elements (TEs), **(b)** number of TEs per 1 million bp and **(c)** number of all Repeat Elements per 1 million bp for each chromosome in *Chelonia mydas* (red) and *Dermochelys coriacea* (blue).

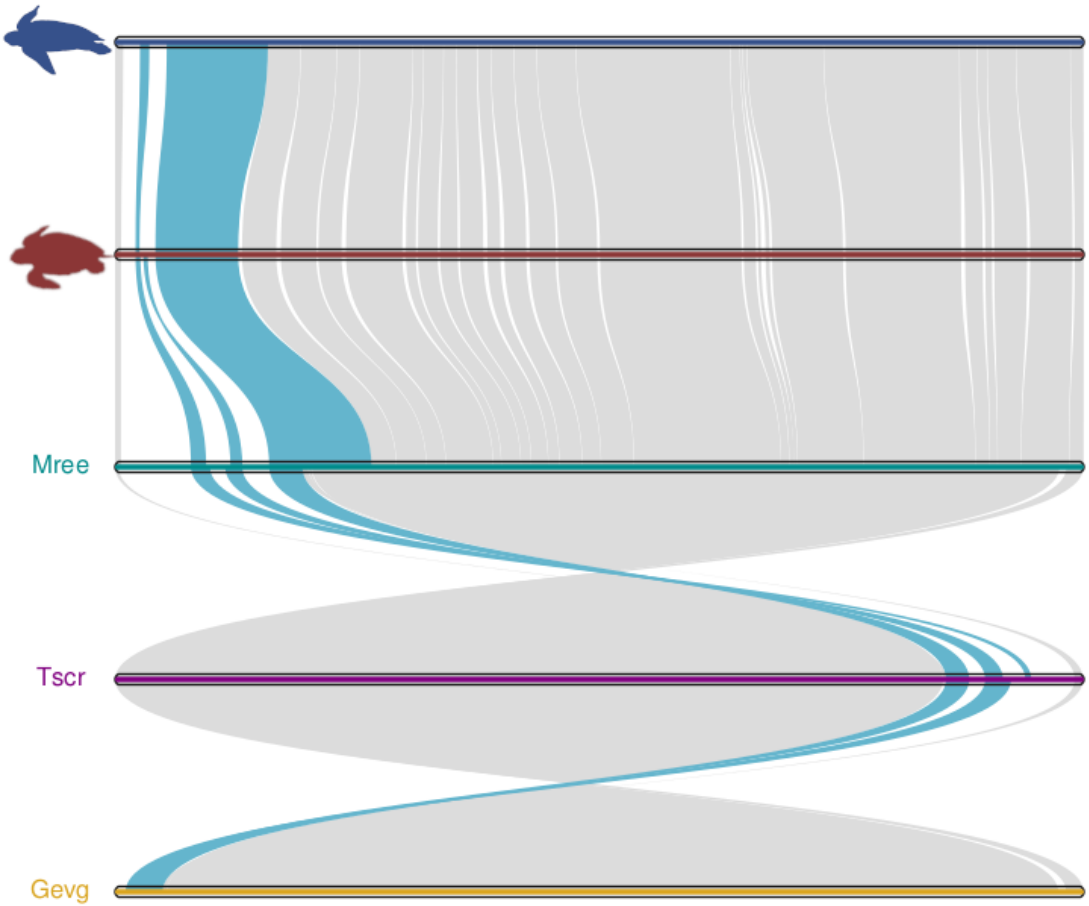


Fig. S7 | Comparison of Chromosome 1 homology across five turtle species depicting (cyan) the region with a cluster of Olfactory receptors class I. *Chelonia mydas* (red), *Dermochelys coriacea* (blue), *Mauremys reevesii* (Mree), *Trachemys scripta* (Tscr) and *Gopherus evgoodei* (Gevg).

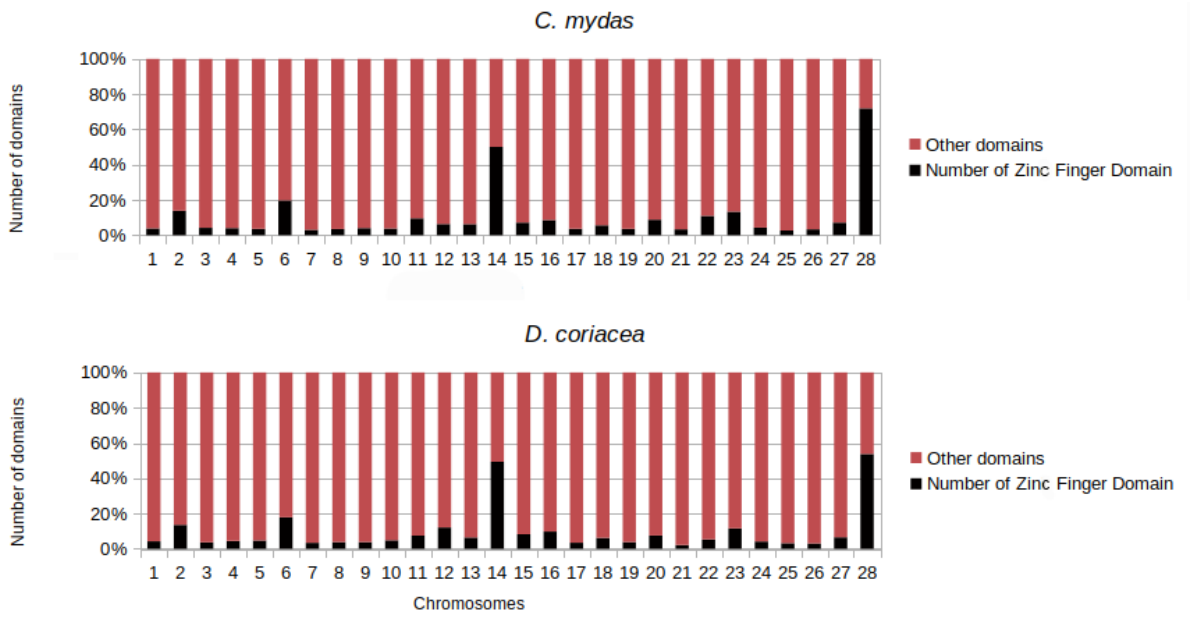


Figure S8 | Proportion of Zinc finger domains per chromosome for the green turtle (*Chelonia mydas*) and the leatherback turtle (*Dermochelys coriacea*). A concentration of Zinc finger domains can be observed in chromosomes 6, 14 and 28 for both species.

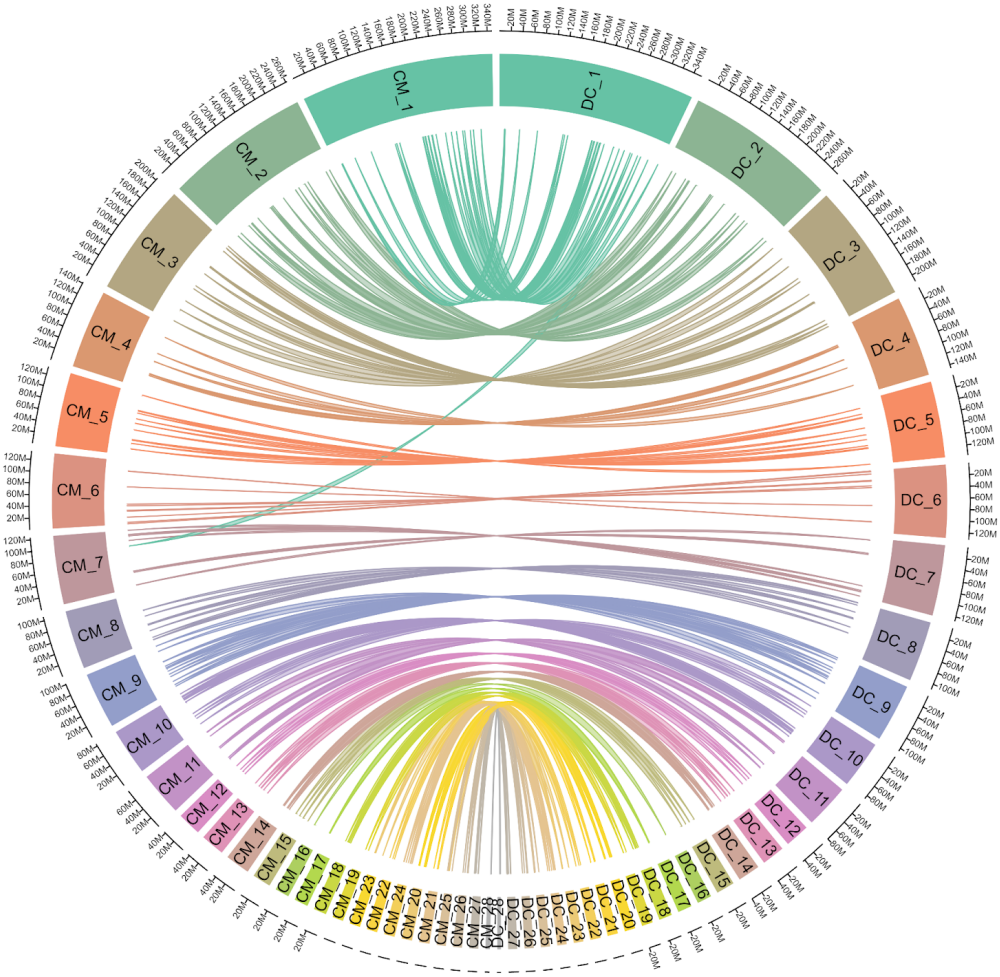


Fig S9 | Locations of 213 genes that have been implicated in temperature-dependent sex determination and that were located in the genomes of both species of sea turtle (green turtle (*Chelonia mydas*): left; leatherback turtle (*Dermochelys coriacea*)*:* right).

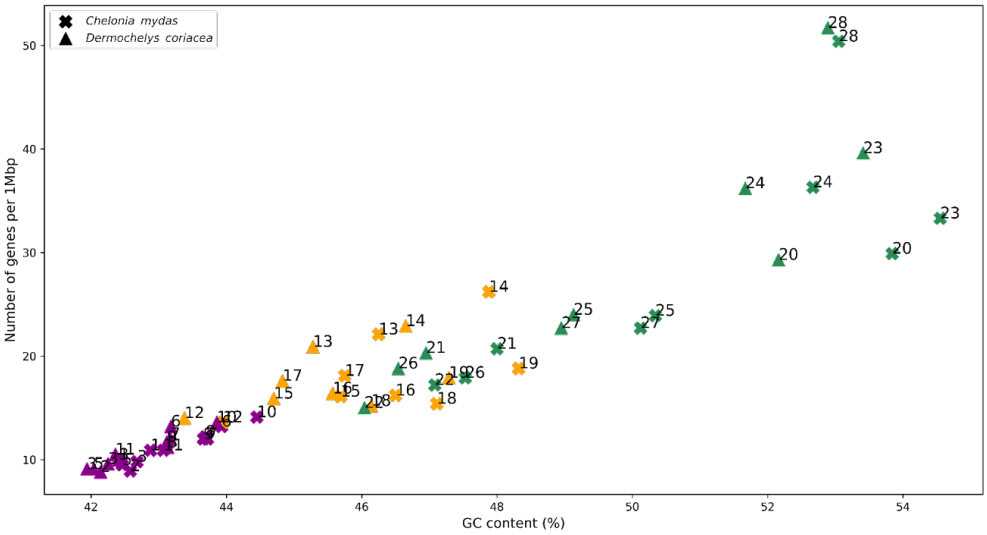


Fig S10 | Relation between number of genes per 1 Mb and GC content for *Chelonia mydas* and *Dermochelys coriacea*. Macro-chromosomes are grouped in purple, micro-chromosomes with >20 Mb in orange and micro-chromosomes with <20 Mb in *C. mydas*.

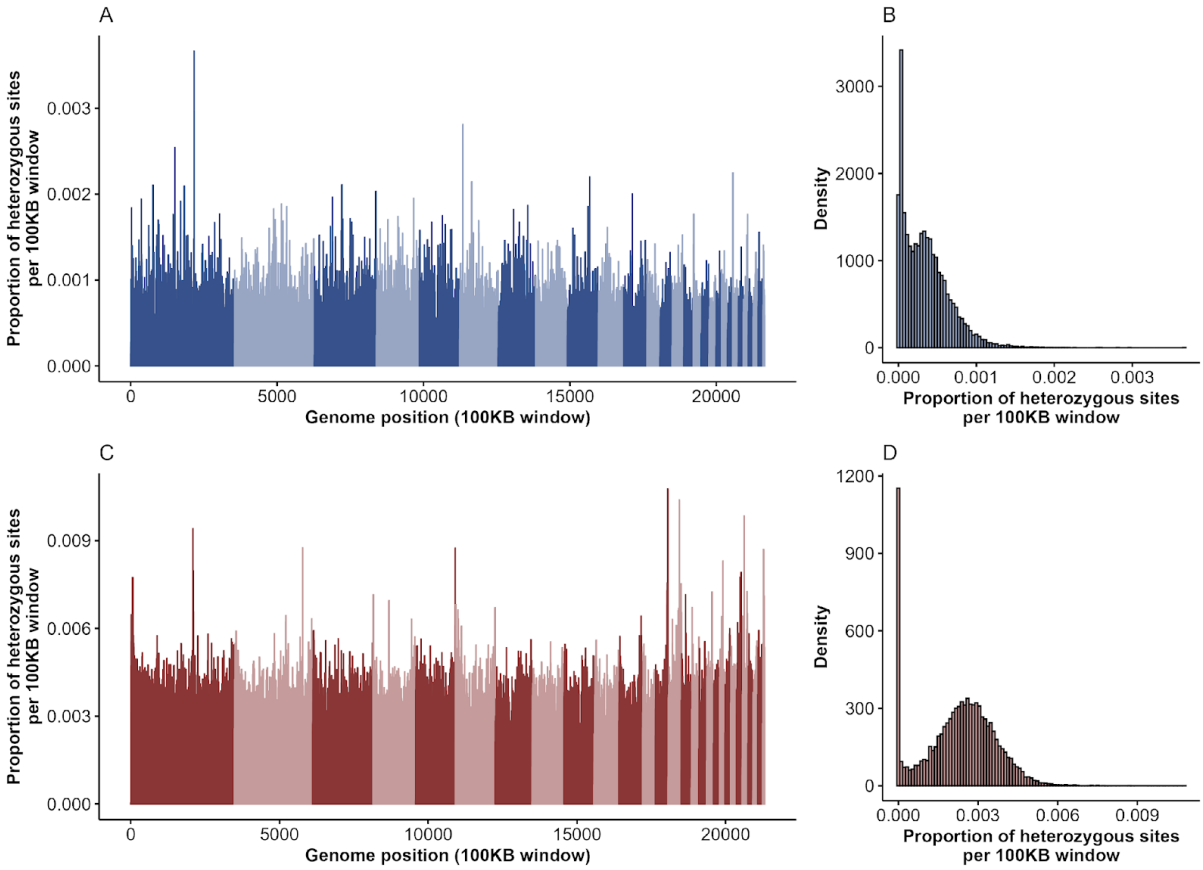


Fig. S11 | Genome-wide heterozygosity plots generated through GATK for both *Dermochelys coriacea* (A, B) and *Chelonia mydas* (C, D) turtle genome assemblies for the known 28 chromosomes. Both (A) and (C) show the proportion of heterozygous sites in 100kb windows where at least 90% of the sites were callable. Alternating colors show breaks between chromosomes. Plots (B) and (D) are histograms displaying the relative density of windows with associated heterozygous proportions. Note that the mean genome-wide heterozygosity estimates are approximately 6.5-times higher for *C. mydas*.

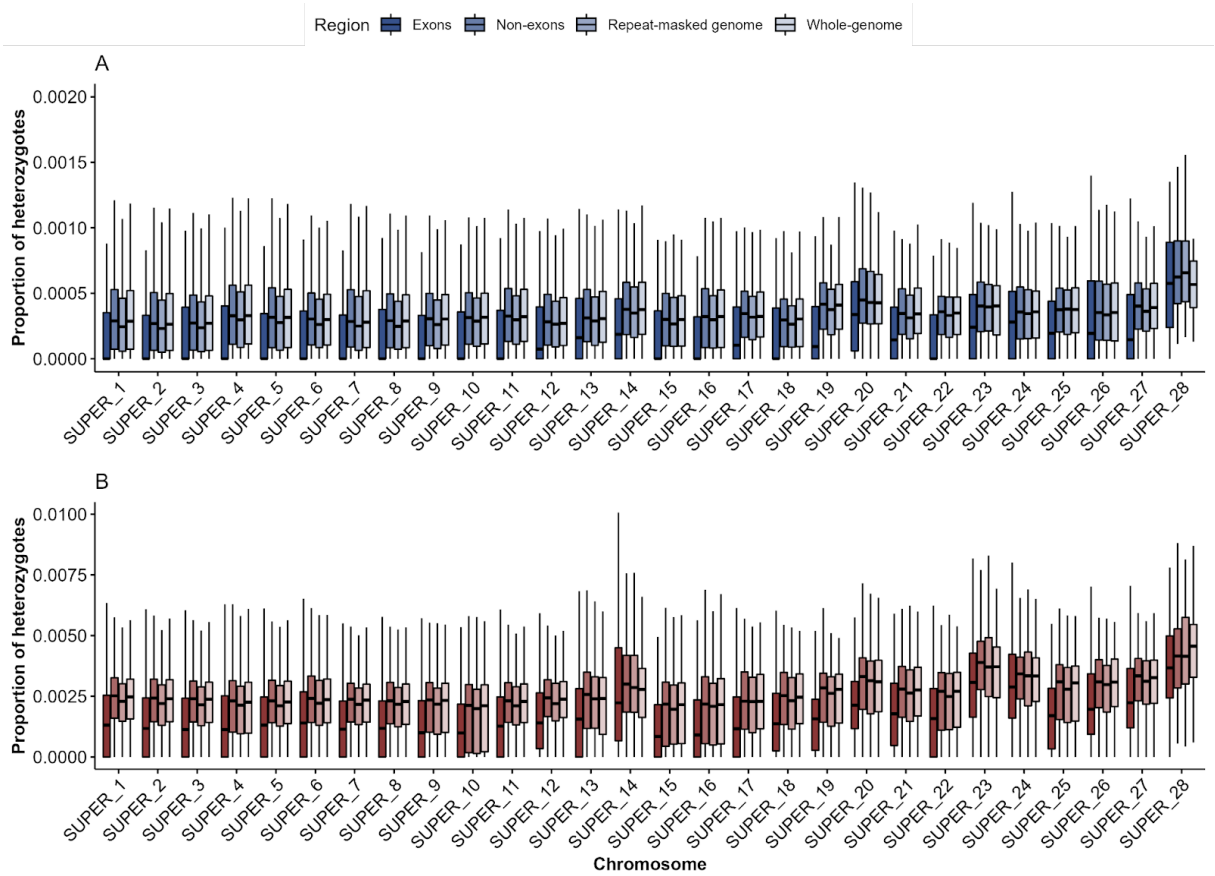


Fig. S12| Chromosome-specific estimations of diversity for whole-genome, repeat-masked, exon, and non-exon regions for both species.

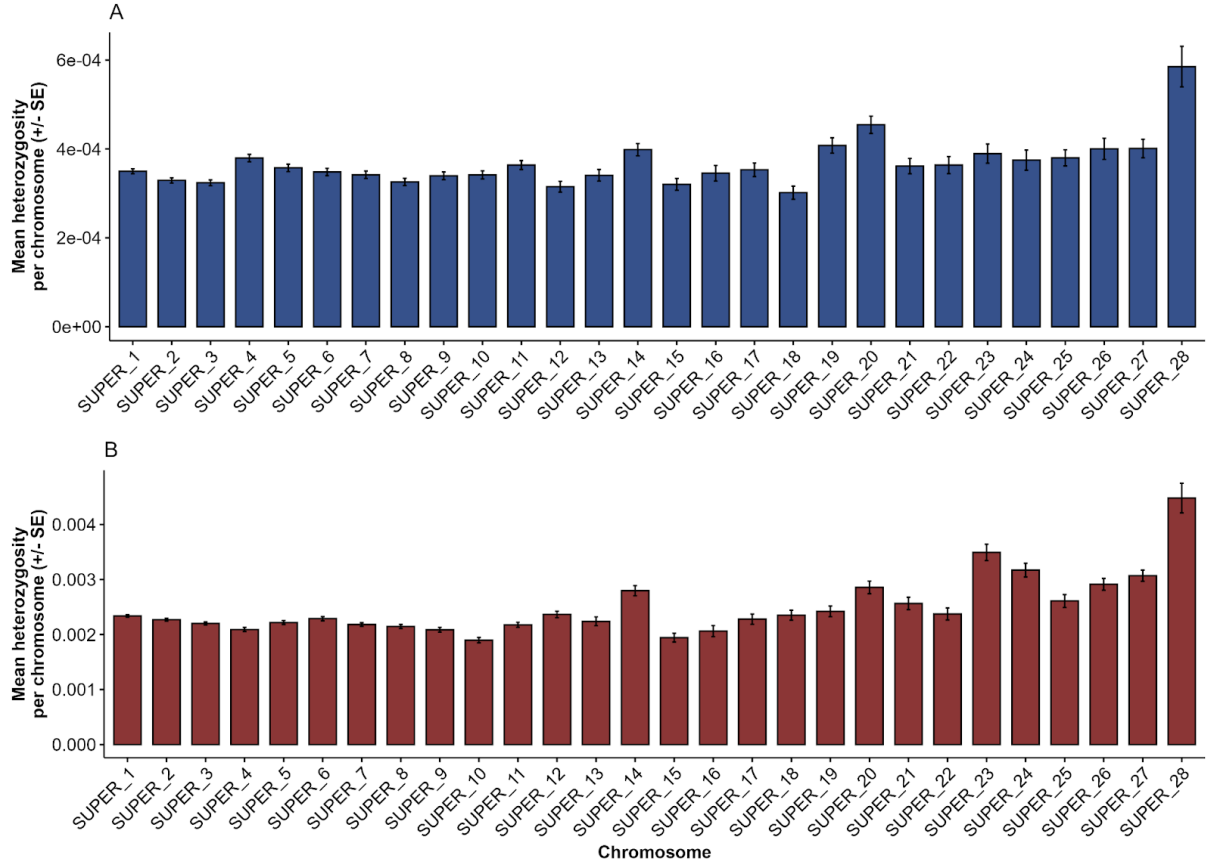


Fig. S13 | Mean heterozygosity per chromosome (+/- SE).

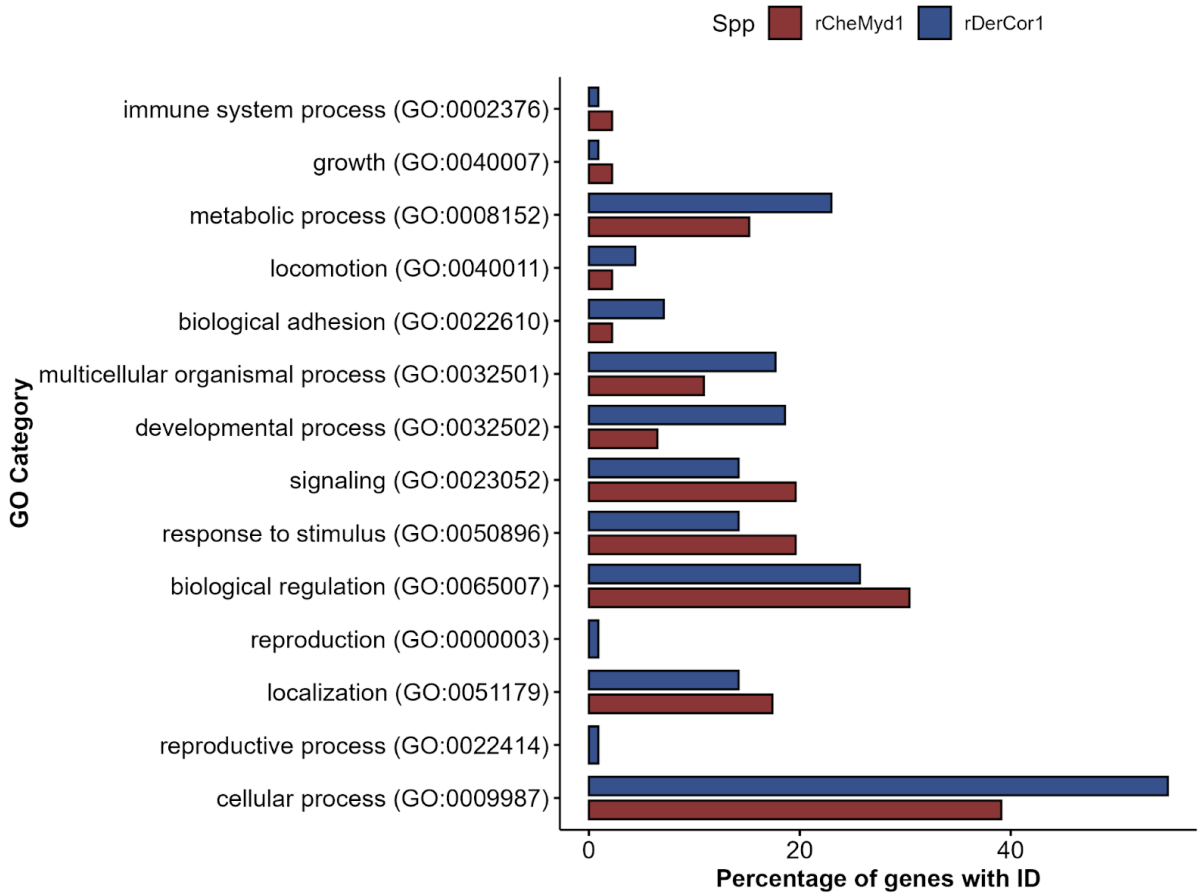


Fig. S14 | GO Biological Process Categories for genes identified with higher than average (mean + 3*SD) diversity in the leatherback turtle (*Dermochelys coriacea*) and the green turtle (*Chelonia mydas*) turtle genomes as predicted by PANTHER.

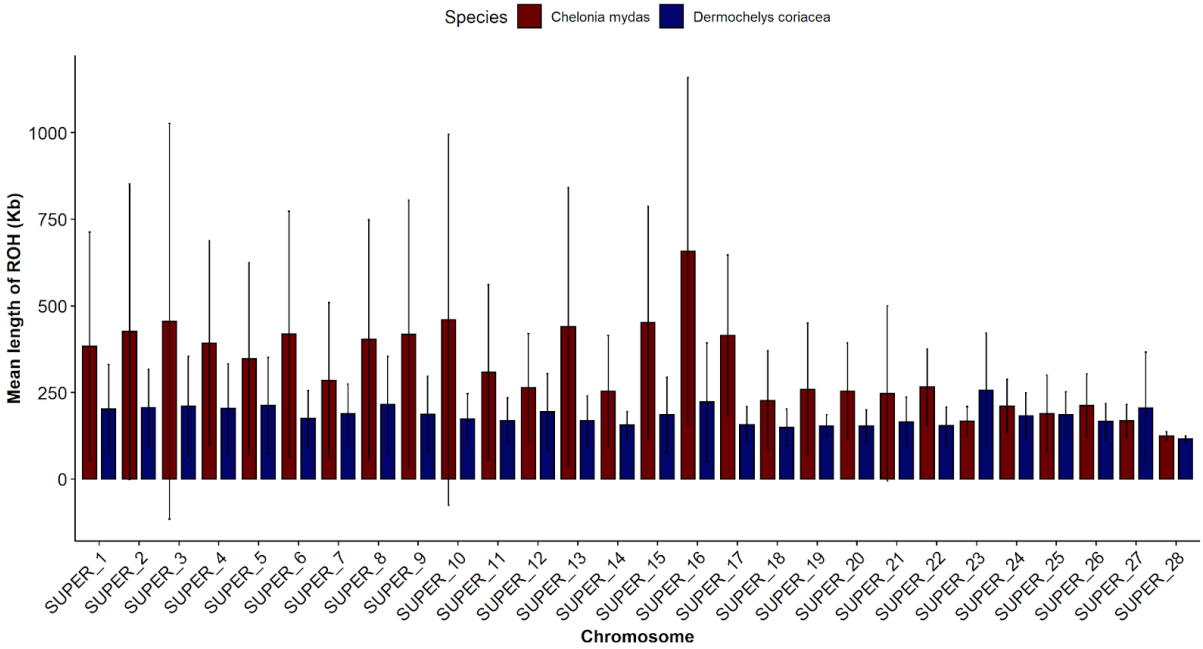


Fig. S15 | Mean length (KB) of runs of homozygosity (ROH) per chromosome for *Dermochelys coriacea* and *Chelonia mydas*.

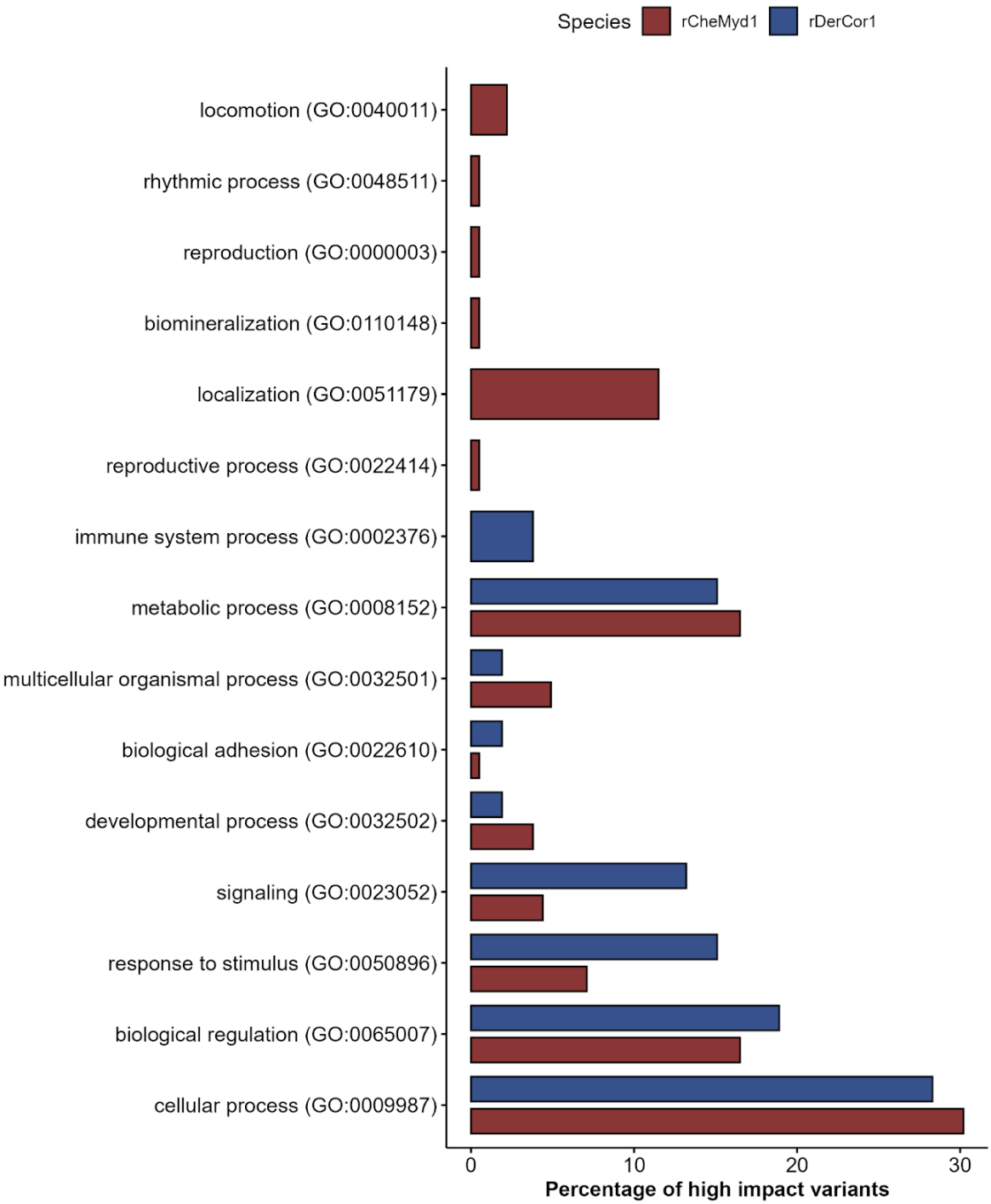


Fig. S16 | Gene ontology categories of Biological Processes predicted with PANTHER for genes with annotated gene identifiers that have putative 'high impact' variants as predicted by snpEff.

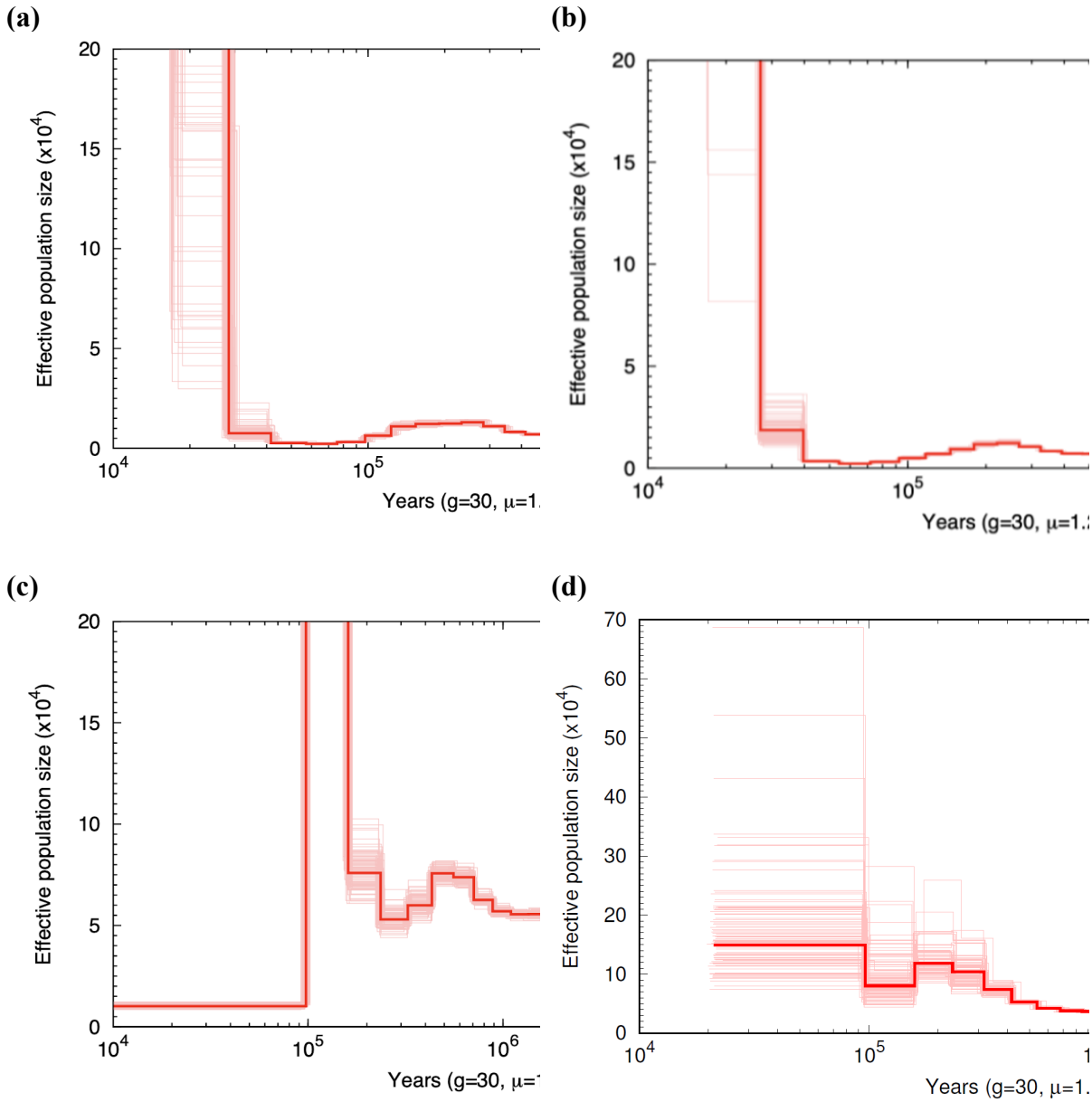
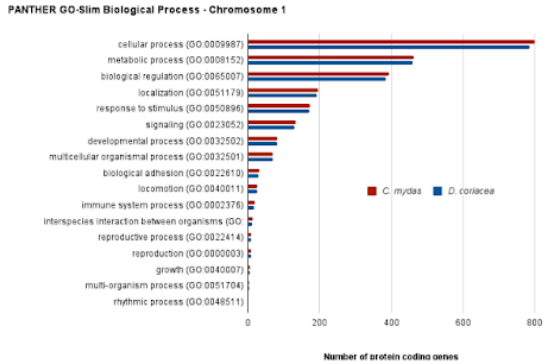
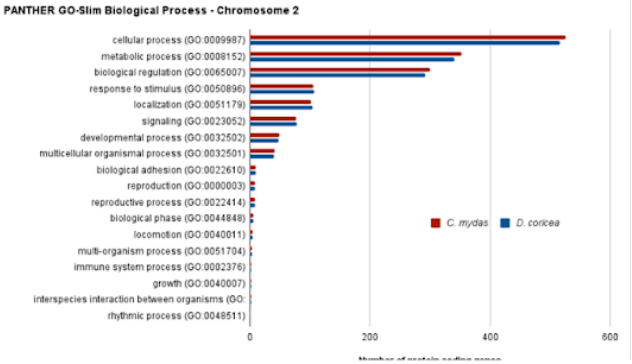
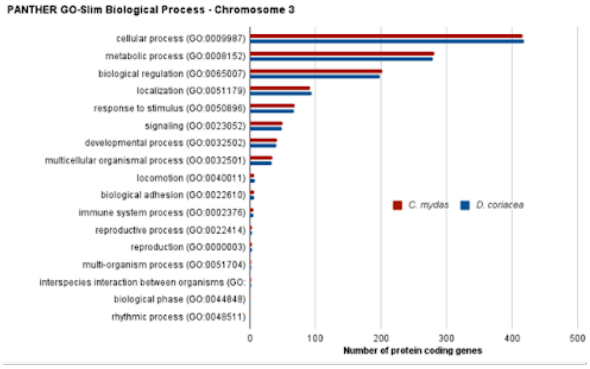
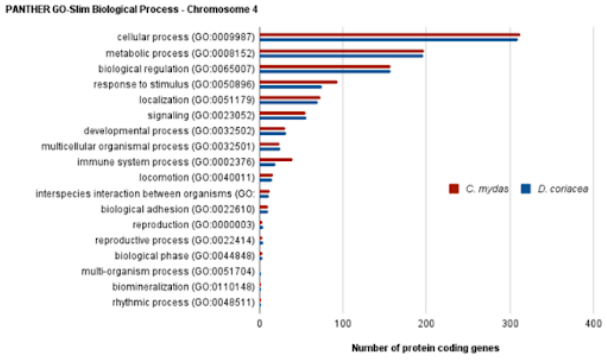


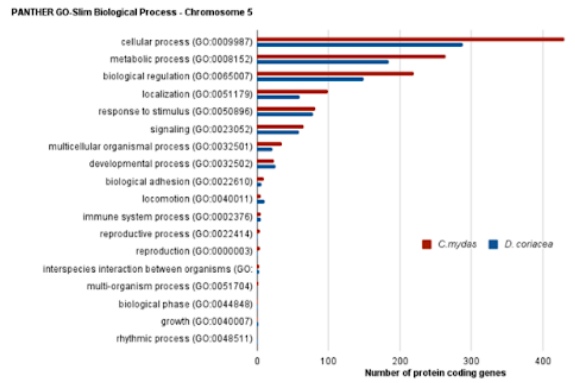
Fig. S17 | Additional PSMC plots for *Dermochelys coriacea* (a,b) and *Chelonia mydas* (c,d). Panels (a) and (c) show the bootstrapping replicates for both species with the reads from the reference individual mapped back to itself, while panels (b) and (d) show PSMC curves (with bootstrapping) for two additional individuals included to ensure observed patterns were not sample artefacts. All PSMC outputs here were generated using genome versions rDerCor1.pri.cur.20201106 and rCheMyd1.pri.cur.20200811.

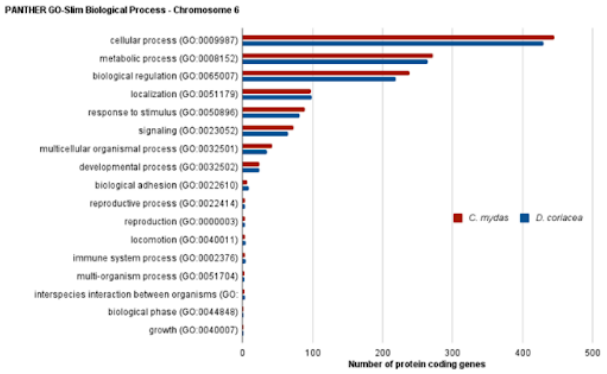

PANTHER GO-Slim Biological Process - Chromosome 1
cellular process (GO:0009987)
metabolic process (GO:0008152)
biological regulation (GO:0065007)
localization (GO:0051179)
response to stimulus (GO:0050896)
signaling (GO:0023052)
developmental process (GO:0032502)
multicellular organismal process (GO:0032501)
biological adhesion (GO:0022610)
locomotion (GO:0040011)
immune system process (GO:0002376)
interspecies interaction between organisms (GO:
reproductive process (GO:0022414)
reproduction (GO:0000003)
growth (GO:0040007)
multi-organism process (GO:0051704)
rhythmic process (GO:0048511)
C. mydas
D. coriacea
0
200
400
600
800
Number of protein coding genes

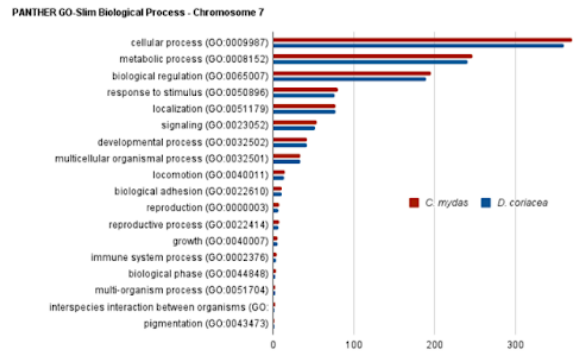

PANTHER GO-Slim Biological Process - Chromosome 2
cellular process (GO:0009987)
metabolic process (GO:0008152)
biological regulation (GO:0065007)
response to stimulus (GO:0050896)
localization (GO:0051179)
signaling (GO:0023052)
developmental process (GO:0032502)
multicellular organismal process (GO:0032501)
biological adhesion (GO:0022610)
reproduction (GO:0000003)
reproductive process (GO:0022414)
biological phase (GO:0044848)
locomotion (GO:0040011)
multi-organism process (GO:0051704)
immune system process (GO:0002376)
growth (GO:0040007)
interspecies interaction between organisms (GO:
rhythmic process (GO:0048511)
C. mydas
D. coricea
0
200
400
600

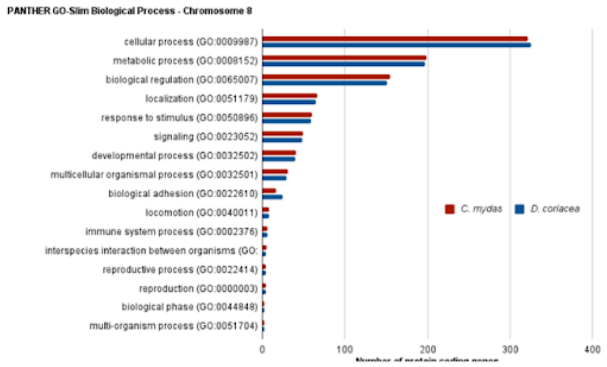

PANTHER GO-Slim Biological Process - Chromosome 3
cellular process (GO:0009987)
metabolic process (GO:0008152)
biological regulation (GO:0065007)
localization (GO:0051179)
response to stimulus (GO:0050896)
signaling (GO:0023052)
developmental process (GO:0032502)
multicellular organismal process (GO:0032501)
locomotion (GO:0040011)
biological adhesion (GO:0022610)
immune system process (GO:0002376)
reproductive process (GO:0022414)
reproduction (GO:0000003)
multi-organism process (GO:0051704)
interspecies interaction between organisms (GO:
biological phase (GO:0044848)
rhythmic process (GO:0048511)
C. mydas
D. coriacea
0
100
200
300
400
500
Number of protein coding genes

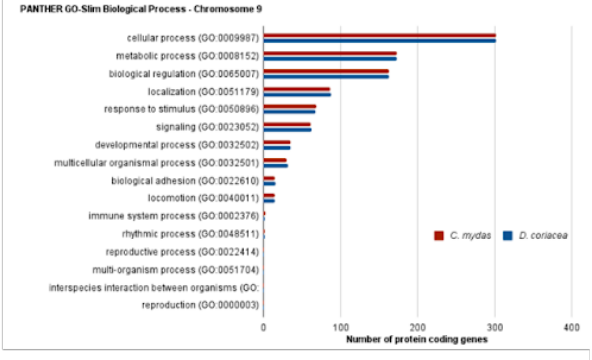

PANTHER GO-Slim Biological Process - Chromosome 4
cellular process (GO:0009987)
metabolic process (GO:0008152)
biological regulation (GO:0065007)
response to stimulus (GO:0050896)
localization (GO:0051179)
signaling (GO:0023052)
developmental process (GO:0032502)
multicellular organismal process (GO:0032501)
immune system process (GO:0002376)
locomotion (GO:0040011)
interspecies interaction between organisms (GO:
biological adhesion (GO:0022610)
reproduction (GO:0000003)
reproductive process (GO:0022414)
biological phase (GO:0044848)
multi-organism process (GO:0051704)
biomineralization (GO:0110148)
rhythmic process (GO:0048511)
C. mydas
D. coriacea
0
100
200
300
400
Number of protein coding genes

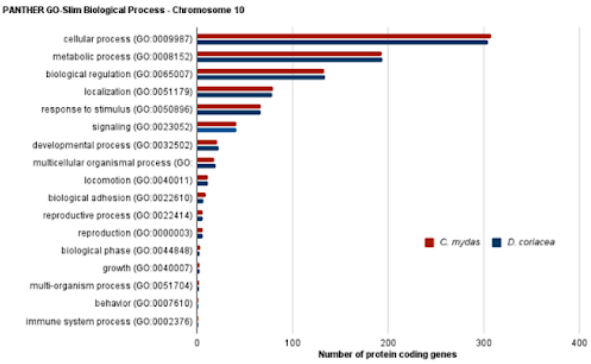

PANTHER GO-Slim Biological Process - Chromosome 5
cellular process (GO:0009987)
metabolic process (GO:0008152)
biological regulation (GO:0065007)
localization (GO:0051179)
response to stimulus (GO:0050896)
signaling (GO:0023052)
multicellular organismal process (GO:0032501)
developmental process (GO:0032502)
biological adhesion (GO:0022610)
locomotion (GO:0040011)
immune system process (GO:0002376)
reproductive process (GO:0022414)
reproduction (GO:0000003)
interspecies interaction between organisms (GO:
multi-organism process (GO:0051704)
biological phase (GO:0044848)
growth (GO:0040007)
rhythmic process (GO:0048511)
C.mydas
D. coriacea
0
100
200
300
400
Number of protein coding genes

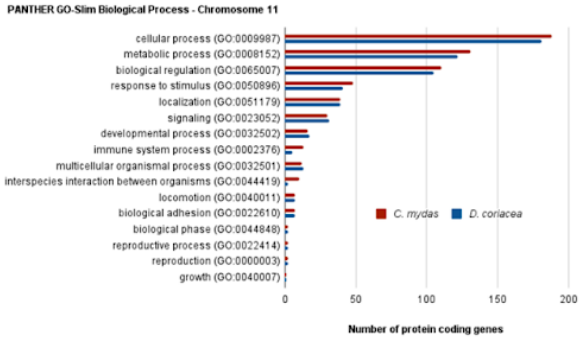

PANTHER GO-Slim Biological Process - Chromosome 6
cellular process (GO:0009987)
metabolic process (GO:0008152)
biological regulation (GO:0065007)
localization (GO:0051179)
response to stimulus (GO:0050896)
signaling (GO:0023052)
multicellular organismal process (GO:0032501)
developmental process (GO:0032502)
biological adhesion (GO:0022610)
reproductive process (GO:0022414)
reproduction (GO:0000003)
locomotion (GO:0040011)
immune system process (GO:0002376)
multi-organism process (GO:0051704)
interspecies interaction between organisms (GO:
biological phase (GO:0044848)
growth (GO:0040007)
C. mydas
D. coriacea
0
100
200
300
400
500
Number of protein coding genes

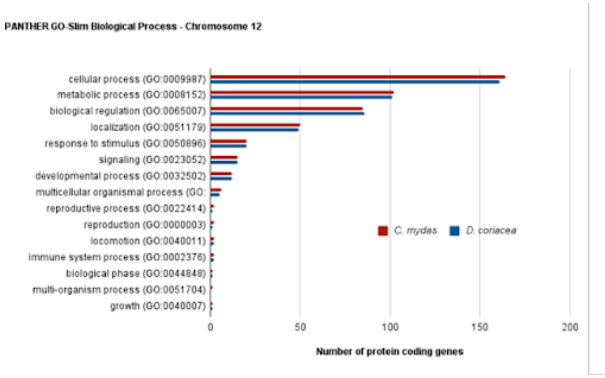

PANTHER GO-Slim Biological Process - Chromosome 7
cellular process (GO:0009987)
metabolic process (GO:0008152)
biological regulation (GO:0065007)
response to stimulus (GO:0050896)
localization (GO:0051179)
signaling (GO:0023052)
developmental process (GO:0032502)
multicellular organismal process (GO:0032501)
locomotion (GO:0040011)
biological adhesion (GO:0022610)
reproduction (GO:0000003)
reproductive process (GO:0022414)
growth (GO:0040007)
immune system process (GO:0002376)
biological phase (GO:0044848)
multi-organism process (GO:0051704)
interspecies interaction between organisms (GO:
pigmentation (GO:0043473)
C. mydas
D. coriacea
0
100
200
300

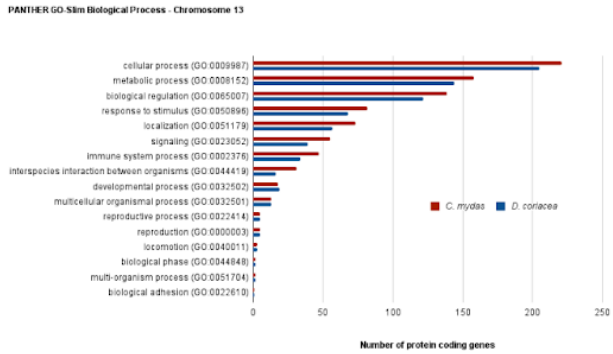

PANTHER GO-Slim Biological Process - Chromosome 8
cellular process (GO:0009987)
metabolic process (GO:0008152)
biological regulation (GO:0065007)
localization (GO:0051179)
response to stimulus (GO:0050896)
signaling (GO:0023052)
developmental process (GO:0032502)
multicellular organismal process (GO:0032501)
biological adhesion (GO:0022610)
locomotion (GO:0040011)
immune system process (GO:0002376)
interspecies interaction between organisms (GO:
reproductive process (GO:0022414)
reproduction (GO:0000003)
biological phase (GO:0044848)
multi-organism process (GO:0051704)
C. mydas
D. coriacea
0
100
200
300
400

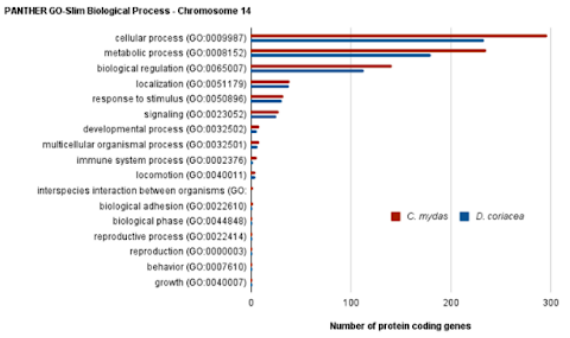

PANTHER GO-Slim Biological Process - Chromosome 9
cellular process (GO:0009987)
metabolic process (GO:0008152)
biological regulation (GO:0065007)
localization (GO:0051179)
response to stimulus (GO:0050896)
signaling (GO:0023052)
developmental process (GO:0032502)
multicellular organismal process (GO:0032501)
biological adhesion (GO:0022610)
locomotion (GO:0040011)
immune system process (GO:0002376)
rhythmic process (GO:0048511)
reproductive process (GO:0022414)
multi-organism process (GO:0051704)
interspecies interaction between organisms (GO:
reproduction (GO:0000003)
C. mydas
D. coriacea
0
100
200
300
400
Number of protein coding genes

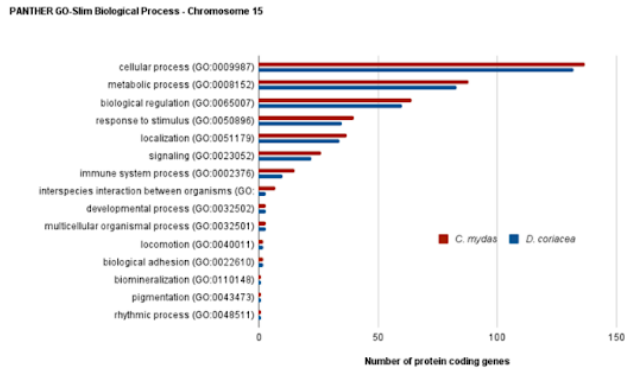

PANTHER GO-Slim Biological Process - Chromosome 10
cellular process (GO:0009987)
metabolic process (GO:0008152)
biological regulation (GO:0065007)
localization (GO:0051179)
response to stimulus (GO:0050896)
signaling (GO:0023052)
developmental process (GO:0032502)
multicellular organismal process (GO:
locomotion (GO:0040011)
biological adhesion (GO:0022610)
reproductive process (GO:0022414)
reproduction (GO:0000003)
biological phase (GO:0044848)
growth (GO:0040007)
multi-organism process (GO:0051704)
behavior (GO:0007610)
immune system process (GO:0002376)
C. mydas
D. coriacea
0
100
200
300
400
Number of protein coding genes

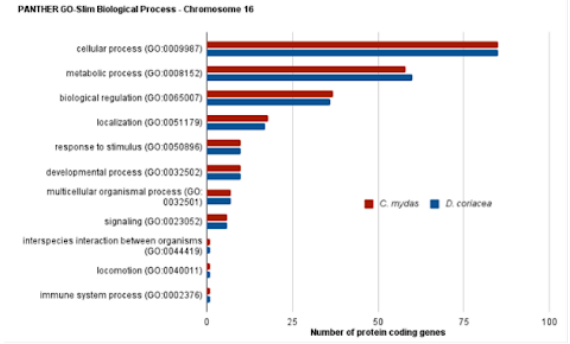

PANTHER GO-Slim Biological Process - Chromosome 11
cellular process (GO:0009987)
metabolic process (GO:0008152)
biological regulation (GO:0065007)
response to stimulus (GO:0050896)
localization (GO:0051179)
signaling (GO:0023052)
developmental process (GO:0032502)
immune system process (GO:0002376)
multicellular organismal process (GO:0032501)
interspecies interaction between organisms (GO:0044419)
locomotion (GO:0040011)
biological adhesion (GO:0022610)
biological phase (GO:0044848)
reproductive process (GO:0022414)
reproduction (GO:0000003)
growth (GO:0040007)
C. mydas
D. coriacea
0
50
100
150
200
Number of protein coding genes

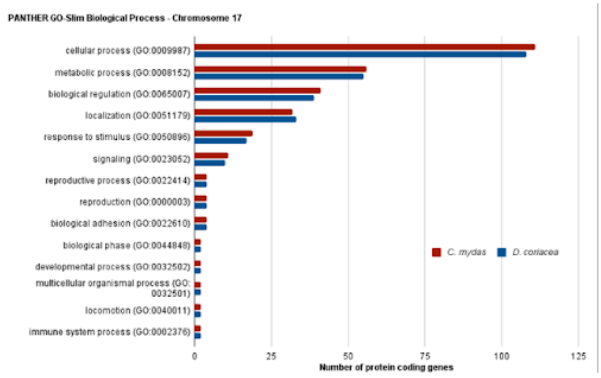

PANTHER GO-Slim Biological Process - Chromosome 12
cellular process (GO:0009987)
metabolic process (GO:0008152)
biological regulation (GO:0065007)
localization (GO:0051179)
response to stimulus (GO:0050896)
signaling (GO:0023052)
developmental process (GO:0032502)
multicellular organismal process (GO:
reproductive process (GO:0022414)
reproduction (GO:0000003)
locomotion (GO:0040011)
immune system process (GO:0002376)
biological phase (GO:0044848)
multi-organism process (GO:0051704)
growth (GO:0040007)
C. mydas
D. coriacea
0
50
100
150
200
Number of protein coding genes

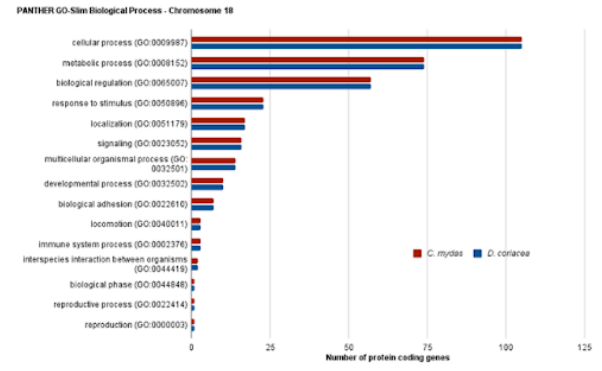

PANTHER GO-Slim Biological Process - Chromosome 13
cellular process (GO:0009987)
metabolic process (GO:0008152)
biological regulation (GO:0065007)
response to stimulus (GO:0050896)
localization (GO:0051179)
signaling (GO:0023052)
immune system process (GO:0002376)
interspecies interaction between organisms (GO:0044419)
developmental process (GO:0032502)
multicellular organismal process (GO:0032501)
reproductive process (GO:0022414)
reproduction (GO:0000003)
locomotion (GO:0040011)
biological phase (GO:0044848)
multi-organism process (GO:0051704)
biological adhesion (GO:0022610)
C. mydas
D. coriacea
0
50
100
150
200
250
Number of protein coding genes

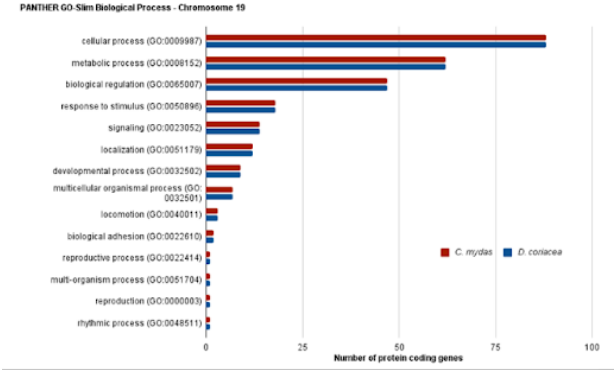

PANTHER GO-Slim Biological Process - Chromosome 14
cellular process (GO:0009987)
metabolic process (GO:0008152)
biological regulation (GO:0065007)
localization (GO:0051179)
response to stimulus (GO:0050896)
signaling (GO:0023052)
developmental process (GO:0032502)
multicellular organismal process (GO:0032501)
immune system process (GO:0002376)
locomotion (GO:0040011)
interspecies interaction between organisms (GO:
biological adhesion (GO:0022610)
biological phase (GO:0044848)
reproductive process (GO:0022414)
reproduction (GO:0000003)
behavior (GO:0007610)
growth (GO:0040007)
C. mydas
D. coriacea
0
100
200
300
Number of protein coding genes

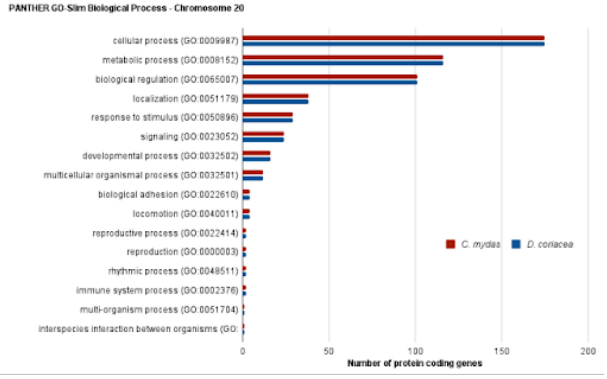

PANTHER GO-Slim Biological Process - Chromosome 15
cellular process (GO:0009987)
metabolic process (GO:0008152)
biological regulation (GO:0065007)
response to stimulus (GO:0050896)
localization (GO:0051179)
signaling (GO:0023052)
immune system process (GO:0002376)
interspecies interaction between organisms (GO:
developmental process (GO:0032502)
multicellular organismal process (GO:0032501)
locomotion (GO:0040011)
biological adhesion (GO:0022610)
biomineralization (GO:0110148)
pigmentation (GO:0043473)
rhythmic process (GO:0048511)
C. mydas
D. coriacea
0
50
100
150
Number of protein coding genes


PANTHER GO-Slim Biological Process - Chromosome 16
cellular process (GO:0009987)
metabolic process (GO:0008152)
biological regulation (GO:0065007)
localization (GO:0051179)
response to stimulus (GO:0050896)
developmental process (GO:0032502)
multicellular organismal process (GO:0032501)
signaling (GO:0023052)
interspecies interaction between organisms (GO:0044419)
locomotion (GO:0040011)
immune system process (GO:0002376)
C. mydas
D. coriacea
0
25
50
75
100
Number of protein coding genes


PANTHER GO-Slim Biological Process - Chromosome 17
cellular process (GO:0009987)
metabolic process (GO:0008152)
biological regulation (GO:0065007)
localization (GO:0051179)
response to stimulus (GO:0050896)
signaling (GO:0023052)
reproductive process (GO:0022414)
reproduction (GO:0000003)
biological adhesion (GO:0022610)
biological phase (GO:0044848)
developmental process (GO:0032502)
multicellular organismal process (GO:0032501)
locomotion (GO:0040011)
immune system process (GO:0002376)
C. mydas
D. coriacea
0
25
50
75
100
125
Number of protein coding genes


PANTHER GO-Slim Biological Process - Chromosome 18
cellular process (GO:0009987)
metabolic process (GO:0008152)
biological regulation (GO:0065007)
response to stimulus (GO:0050896)
localization (GO:0051179)
signaling (GO:0023052)
multicellular organismal process (GO:0032501)
developmental process (GO:0032502)
biological adhesion (GO:0022610)
locomotion (GO:0040011)
immune system process (GO:0002376)
interspecies interaction between organisms (GO:0044419)
biological phase (GO:0044848)
reproductive process (GO:0022414)
reproduction (GO:0000003)
C. mydas
D. coriacea
0
25
50
75
100
125
Number of protein coding genes


PANTHER GO-Slim Biological Process - Chromosome 19
cellular process (GO:0009987)
metabolic process (GO:0008152)
biological regulation (GO:0065007)
response to stimulus (GO:0050896)
signaling (GO:0023052)
localization (GO:0051179)
developmental process (GO:0032502)
multicellular organismal process (GO:0032501)
locomotion (GO:0040011)
biological adhesion (GO:0022610)
reproductive process (GO:0022414)
multi-organism process (GO:0051704)
reproduction (GO:0000003)
rhythmic process (GO:0048511)
C. mydas
D. coriacea
0
25
50
75
100
Number of protein coding genes


PANTHER GO-Slim Biological Process - Chromosome 20
cellular process (GO:0009987)
metabolic process (GO:0008152)
biological regulation (GO:0065007)
localization (GO:0051179)
response to stimulus (GO:0050896)
signaling (GO:0023052)
developmental process (GO:0032502)
multicellular organismal process (GO:0032501)
biological adhesion (GO:0022610)
locomotion (GO:0040011)
reproductive process (GO:0022414)
reproduction (GO:0000003)
rhythmic process (GO:0048511)
immune system process (GO:0002376)
multi-organism process (GO:0051704)
interspecies interaction between organisms (GO:
C. mydas
D. coriacea
0
50
100
150
200
Number of protein coding genes

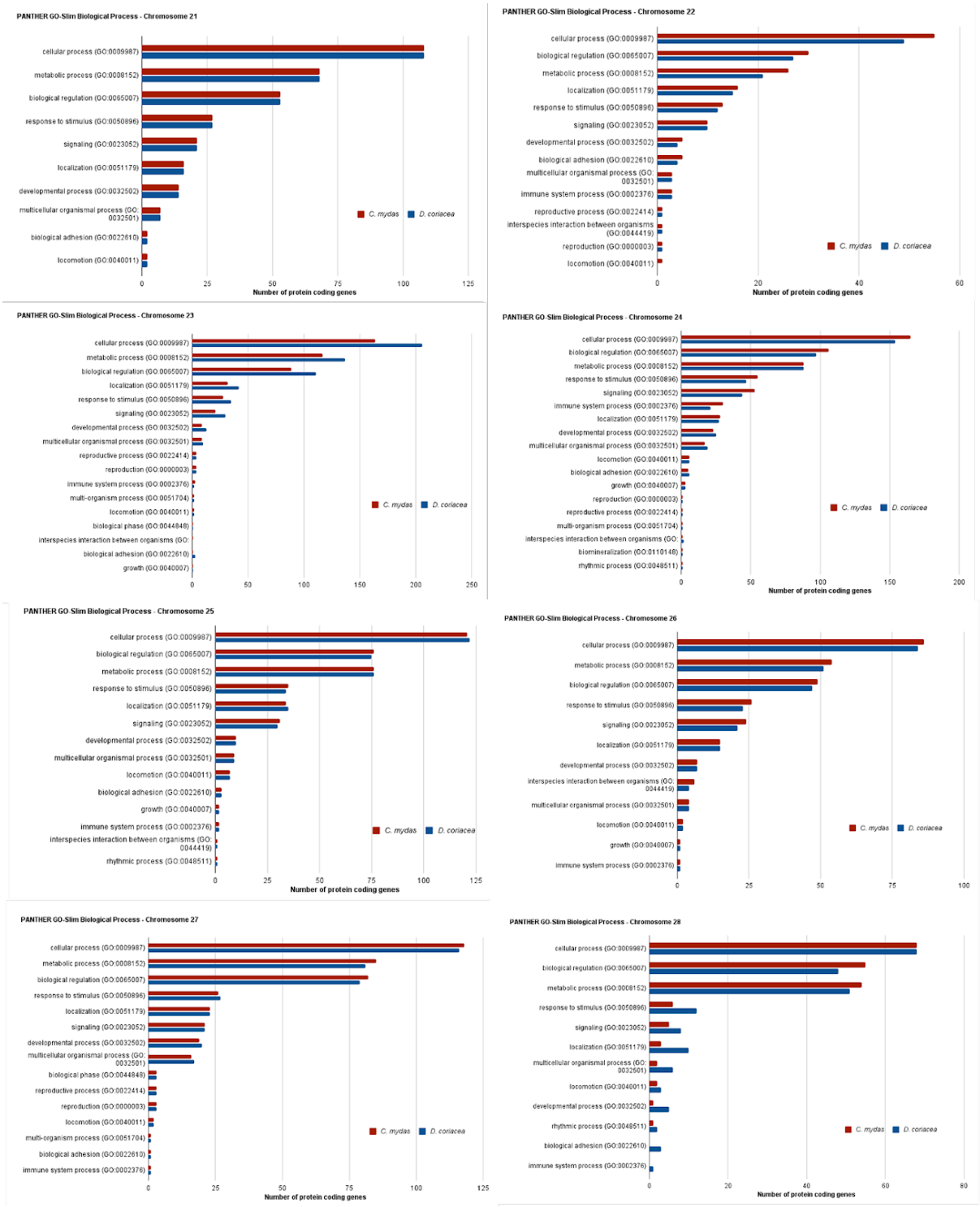


Fig. S18 | PANTHER GO-slim classification by biological process of the coding sequences present in each chromosome for *Chelonia mydas* and *Dermochelys coriacea*.

Supplementary references

Andrews, Simon, Felix Krueger, Anne Segonds-Pichon, Laura Biggins, Christel Krueger, and Steven Wingett. 2012. "FastQC: A Quality Control Tool for High Throughput Sequence Data." Babraham, UK.

Armstrong, Joel, Glenn Hickey, Mark Diekhans, Ian T. Fiddes, Adam M. Novak, Alden Deran, Qi Fang, et al. 2020. "Progressive Cactus Is a Multiple-Genome Aligner for the Thousand-Genome Era." *Nature* 587 (7833): 246–51.

Aubry, Sylvain, Steven Kelly, Britta M. C. Kümpers, Richard D. Smith-Unna, and Julian M. Hibberd. 2014. "Deep Evolutionary Comparison of Gene Expression Identifies Parallel Recruitment of Trans-Factors in Two Independent Origins of C4 Photosynthesis." *PLoS Genetics* 10 (6): e1004365.

Blum, Matthias, Hsin-Yu Chang, Sara Chuguransky, Tiago Grego, Swaathi Kandasaamy, Alex Mitchell, Gift Nuka, et al. 2021. "The InterPro Protein Families and Domains Database: 20 Years on." *Nucleic Acids Research* 49 (D1): D344–54.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics* 30 (15): 2114–20.

Cabanettes, Floréal, and Christophe Klopp. 2018. "D-GENIES: Dot Plot Large Genomes in an Interactive, Efficient and Simple Way." *PeerJ* 6 (June): e4958.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10 (December): 421.

Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón. 2009. "trimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses." *Bioinformatics* 25 (15): 1972–73.

Chin, Chen-Shan, David H. Alexander, Patrick Marks, Aaron A. Klammer, James Drake, Cheryl Heiner, Alicia Clum, et al. 2013. "Nonhybrid, Finished Microbial Genome Assemblies from Long-Read SMRT Sequencing Data." *Nature Methods* 10 (6): 563–69.

Chin, Chen-Shan, Paul Peluso, Fritz J. Sedlazeck, Maria Nattestad, Gregory T. Concepcion, Alicia Clum, Christopher Dunn, et al. 2016. "Phased Diploid Genome Assembly with Single-Molecule Real-Time Sequencing." *Nature Methods* 13 (12): 1050–54.

Chow, William, Kim Brugger, Mario Caccamo, Ian Sealy, James Torrance, and Kerstin Howe. 2016. "gEVAL — a Web-Based Browser for Evaluating Genome Assemblies." *Bioinformatics* 32 (16): 2508–10.

Cingolani, Pablo, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. 2012. "A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of *Drosophila Melanogaster* Strain w1118; Iso-2; Iso-3." *Fly* 6 (2): 80–92.

Dudchenko, Olga, Muhammad S. Shamim, Sanjit S. Batra, Neva C. Durand, Nathaniel T. Musial, Ragib Mostofa, Melanie Pham, et al. 2018. "The Juicebox Assembly Tools Module Facilitates de Novo Assembly of Mammalian Genomes with Chromosome-Length Scaffolds for under $1000." *bioRxiv*. https://doi.org/10.1101/254797.

Durand, Neva C., James T. Robinson, Muhammad S. Shamim, Ido Machol, Jill P. Mesirov, Eric S. Lander, and Erez Lieberman Aiden. 2016. "Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom." *Cell Systems* 3 (1): 99–101.

Emms, David M., and Steven Kelly. 2015. "OrthoFinder: Solving Fundamental Biases in Whole Genome Comparisons Dramatically Improves Orthogroup Inference Accuracy." *Genome Biology* 16 (August): 157.

———. 2019. "OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics." *Genome Biology* 20 (1): 238.

Fitak, Robert R., and Sönke Johnsen. 2018. "Green Sea Turtle (*Chelonia Mydas*) Population History Indicates Important Demographic Changes near the Mid-Pleistocene Transition." *Marine Biology* 165 (7): 110.

Flynn, Jullien M., Robert Hubley, Clément Goubert, Jeb Rosen, Andrew G. Clark, Cédric Feschotte, and Arian F. Smit. 2020. "RepeatModeler2: Automated Genomic Discovery of Transposable Element Families." *Proceedings of the National Academy of Sciences* 117 (17): 9451–57.

Formenti, Giulio, Arang Rhie, Jennifer Balacco, Bettina Haase, Jacquelyn Mountcastle, Olivier Fedrigo, Samara Brown, et al. 2021. "Complete Vertebrate Mitogenomes Reveal Widespread Repeats and Gene Duplications." *Genome Biology* 22 (1): 120.

Garrison, E., and G. Marth. 2012. "Haplotype-Based Variant Detection from Short-Read Sequencing. arXiv 1207.3907 [q-Bio. GN]." *Version: V9*, 9–2.

Gemmell, Neil J., Kim Rutherford, Stefan Prost, Marc Tollis, David Winter, J. Robert Macey, David L. Adelson, et al. 2020. "The Tuatara Genome Reveals Ancient Features of Amniote Evolution." *Nature* 584 (7821): 403–9.

Ghurye, Jay, Arang Rhie, Brian P. Walenz, Anthony Schmitt, Siddarth Selvaraj, Mihai Pop, Adam M. Phillippy, and Sergey Koren. 2019. "Integrating Hi-C Links with Assembly Graphs for Chromosome-Scale Assembly." *PLoS Computational Biology* 15 (8): e1007273.

Glusman, G., I. Yanai, I. Rubin, and D. Lancet. 2001. "The Complete Human Olfactory Subgenome." *Genome Research* 11 (5): 685–702.

Grossen, Christine, Frédéric Guillaume, Lukas F. Keller, and Daniel Croll. 2020. “Purging of Highly Deleterious Mutations through Severe Bottlenecks in Alpine Ibex.” *Nature Communications* 11 (1): 1001.

Guan, Dengfeng, Shane A. McCarthy, Jonathan Wood, Kerstin Howe, Yadong Wang, and Richard Durbin. 2020. “Identifying and Removing Haplotypic Duplication in Primary Genome Assemblies.” *bioRxiv*. https://doi.org/10.1101/729962.

Hickey, Glenn, Benedict Paten, Dent Earl, Daniel Zerbino, and David Haussler. 2013. “HAL: A Hierarchical Format for Storing and Analyzing Multiple Genome Alignments.” *Bioinformatics* 29 (10): 1341–42.

Howe, Kerstin, William Chow, Joanna Collins, Sarah Pelan, Damon-Lee Pointon, Ying Sims, James Torrance, Alan Tracey, and Jonathan Wood. 2021. “Significantly Improving the Quality of Genome Assemblies through Curation.” *GigaScience* 10 (1). https://doi.org/10.1093/gigascience/giaa153.

Katoh, Kazutaka, and Daron M. Standley. 2013. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.” *Molecular Biology and Evolution* 30 (4): 772–80.

Kerpedjiev, Peter, Nezar Abdennur, Fritz Lekschas, Chuck McCallum, Kasper Dinkla, Hendrik Strobelt, Jacob M. Luber, et al. 2018. “HiGlass: Web-Based Visual Exploration and Analysis of Genome Interaction Maps.” *Genome Biology* 19 (1): 125.

Korneliussen, Thorfinn Sand, Anders Albrechtsen, and Rasmus Nielsen. 2014. “ANGSD: Analysis of Next Generation Sequencing Data.” *BMC Bioinformatics* 15 (November): 356.

Li, Heng. 2011. “A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data.” *Bioinformatics* 27 (21): 2987–93.

———. 2013. “Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM.” *arXiv [q-bio.GN]*. arXiv. http://arxiv.org/abs/1303.3997.

Li, Heng, and Richard Durbin. 2011. “Inference of Human Population History from Individual Whole-Genome Sequences.” *Nature* 475 (7357): 493–96.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25 (16): 2078–79.

Liu, Jianjun, Siqi Liu, Kai Zheng, Min Tang, Liping Gu, James Young, Ziming Wang, et al. 2021. “Chromosome-Level Genome Assembly of the Chinese Three-Keeled Pond Turtle (Mauremys Reevesii) Provides Insights into Freshwater Adaptation.” *Molecular Ecology Resources*, November. https://doi.org/10.1111/1755-0998.13563.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data.” *Genome Research* 20 (9): 1297–1303.

Mendes, Fábio K., Dan Vanderpool, Ben Fulton, and Matthew W. Hahn. 2020. “CAFE 5 Models Variation in Evolutionary Rates among Gene Families.” *Bioinformatics* , December. https://doi.org/10.1093/bioinformatics/btaa1022.

Mi, Huaiyu, Dustin Ebert, Anushya Muruganujan, Caitlin Mills, Laurent-Philippe Albou, Tremayne Mushayamaha, and Paul D. Thomas. 2021. “PANTHER Version 16: A Revised Family Classification, Tree-Based Classification Tool, Enhancer Regions and Extensive API.” *Nucleic Acids Research* 49 (D1): D394–403.

Minh, Bui Quang, Heiko A. Schmidt, Olga Chernomor, Dominik Schrempf, Michael D. Woodhams, Arndt von Haeseler, and Robert Lanfear. 2020. “IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era.” *Molecular Biology and Evolution* 37 (5): 1530–34.

Nguyen, Lam-Tung, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. 2015. “IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies.” *Molecular Biology and Evolution* 32 (1): 268–74.

Ondov, Brian D., Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. 2016. “Mash: Fast Genome and Metagenome Distance Estimation Using MinHash.” *Genome Biology* 17 (1): 132.

Paten, Benedict, Mark Diekhans, Dent Earl, John St John, Jian Ma, Bernard Suh, and David Haussler. 2011. “Cactus Graphs for Genome Comparisons.” *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 18 (3): 469–81.

Pertea, Geo, and Mihaela Pertea. 2020. “GFF Utilities: GffRead and GffCompare.” *F1000Research* 9 (April). https://doi.org/10.12688/f1000research.23297.2.

Prasad, Aparna, Eline D. Lorenzen, and Michael V. Westbury. 2022. “Evaluating the Role of Reference-Genome Phylogenetic Distance on Evolutionary Inference.” *Molecular Ecology Resources* 22 (1): 45–55.

Pruitt, Kim D., Garth R. Brown, Susan M. Hiatt, Françoise Thibaud-Nissen, Alexander Astashyn, Olga Ermolaeva, Catherine M. Farrell, et al. 2014. “RefSeq: An Update on Mammalian Reference Sequences.” *Nucleic Acids Research* 42 (Database issue): D756–63.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, et al. 2007. “PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses.” *American Journal of Human Genetics* 81 (3): 559–75.

Quesada, Víctor, Sandra Freitas-Rodríguez, Joshua Miller, José G. Pérez-Silva, Zi-Feng Jiang, Washington Tapia, Olaya Santiago-Fernández, et al. 2019. "Giant Tortoise Genomes Provide Insights into Longevity and Age-Related Disease." *Nature Ecology & Evolution* 3 (1): 87–95.

Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–42.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rhie, Arang, Shane A. McCarthy, Olivier Fedrigo, Joana Damas, Giulio Formenti, Sergey Koren, Marcela Uliano-Silva, et al. 2021. "Towards Complete and Error-Free Genome Assemblies of All Vertebrate Species." *Nature* 592 (7856): 737–46.

Robinson, Jacqueline A., Jannikke Räikkönen, Leah M. Vucetich, John A. Vucetich, Rolf O. Peterson, Kirk E. Lohmueller, and Robert K. Wayne. 2019. "Genomic Signatures of Extensive Inbreeding in Isle Royale Wolves, a Population on the Threshold of Extinction." *Science Advances* 5 (5): eaau0757.

Sanderson, Michael J. 2003. "r8s: Inferring Absolute Rates of Molecular Evolution and Divergence Times in the Absence of a Molecular Clock." *Bioinformatics* 19 (2): 301–2.

Smit, A., R. Hubley, and P. Green. 2015. "RepeatMasker Open-4.0. 2013-2015http." *Repeatmasker. Org*.

Tarailo-Graovac, Maja, and Nansheng Chen. 2009. "Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences." *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* Chapter 4 (1): Unit 4.10.

Thomson, Robert C., Phillip Q. Spinks, and H. Bradley Shaffer. 2021. "A Global Phylogeny of Turtles Reveals a Burst of Climate-Associated Diversification on Continental Margins." *Proceedings of the National Academy of Sciences of the United States of America* 118 (7). https://doi.org/10.1073/pnas.2012215118.

Vandewege, Michael W., Sarah F. Mangum, Toni Gabaldón, Todd A. Castoe, David A. Ray, and Federico G. Hoffmann. 2016. "Contrasting Patterns of Evolutionary Diversification in the Olfactory Repertoires of Reptile and Bird Genomes." *Genome Biology and Evolution* 8 (3): 470–80.

Vurture, Gregory W., Fritz J. Sedlazeck, Maria Nattestad, Charles J. Underwood, Han Fang, James Gurtowski, and Michael C. Schatz. 2017. "GenomeScope: Fast Reference-Free Genome Profiling from Short Reads." *Bioinformatics* 33 (14): 2202–4.

Wang, Zhuo, Juan Pascual-Anaya, Amonida Zadissa, Wenqi Li, Yoshihito Niimura, Zhiyong Huang, Chunyi Li, et al. 2013. "The Draft Genomes of Soft-Shell Turtle and Green Sea Turtle Yield Insights into the Development and Evolution of the Turtle-Specific Body Plan." *Nature Genetics* 45 (6): 701–6.

Chapter 2

Recent expansion of *Penelope*-like retrotransposons in the leatherback turtle *Dermochelys coriacea*

Recent expansion of *Penelope*-like retrotransposons in the leatherback turtle *Dermochelys coriacea*

Tomas Carrasco-Valenzuela[1,2,3]; Luísa Marins[1]; Elisa K. S. Ramos[4], Alexander Suh[5,6]; Camila J. Mazzoni[1,2]

Under review: *Mobile DNA*

1 Berlin Center for Genomics in Biodiversity Research (BeGenDiv), Berlin, Germany

2 Evolutionary Genetics Department, Leibniz-Institut für Zoo- und Wildtierforschung (IZW), Berlin, Germany

3 Universität Potsdam, Brandenburg, Potsdam, Germany

4 Laboratory of Evolutionary Genomics. Genetics, Evolution, Immunology and Microbiology Department, State University of Campinas, Brazil.

5 School of Biological Sciences, Organisms and the Environment, University of East Anglia, NR4 7TU, Norwich, UK.

6 Department of Organismal Biology, Systematic Biology, Evolutionary Biology Centre (EBC), Science for Life Laboratory, Uppsala University, Uppsala SE-752 36, Sweden.

Abstract

Transposable elements are known to induce variation in vertebrate genomes through their diversity and number, with related species usually presenting consistency in the proportion and abundance of TE families. Despite their ancient divergence times, sea turtles *Chelonia mydas* and *Dermochelys coriacea* show high levels of overall genomic synteny and gene collinearity, but there is still a lot to explore regarding their TE panorama. In light of this, we analysed high-quality reference genomes of these species, which represent the two different extant superfamilies of sea turtles - Dermochelyidae and Cheloniidae - to explore their mobilomes and compared them with the 13 available Testudines draft genomes. In line with previous genome-wide comparisons between the two distantly related sea turtle superfamilies, our analyses showcased that turtle genomes generally share similar mobilomes. Nonetheless, we identified that the main difference between these mobilomes is a much higher proportion of *Penelope*-like Elements (PLEs) and Long Interspersed Elements (LINEs) in *D. coriacea*. Finally, we identified a new PLE subfamily of *Neptune*-1 present in *D. coriacea'*s genome, with evidence for a substantial amount of recent insertions. These results show that despite the overall slow evolutionary pace of turtle genomes, at least *D. coriacea* exhibits an active mobilome.

Introduction

One of the genomic features that are known to vary the most among vertebrates is the number and diversity of transposable elements (TEs) (Sotero-Caio et al. 2017; Tollis and Boissinot 2012). TE is an umbrella term used to describe a wide variety of mobile genetic elements that can replicate and multiply in their host's genome (Boissinot et al. 2019). TE abundance is one of the main determinants of haploid genome size variation (Margaret G. Kidwell 2002; Elliott and Gregory 2015), and the difference in TE abundance across genomes contributes indirectly to other characteristics of genomes, such as regional variations in base composition (Symonová and Suh 2019).

TEs are self-replicating genetic elements that can mobilise across the genome. The proportion of TEs varies among eukaryotic genomes, comprising around 30-60% of reptilian and mammalian genomes (Canapa et al. 2015). Despite their abundance, TE identification is a rather challenging and time-consuming process due to their complexity and the amount of data that needs to be processed and compared (Rodriguez and Makałowski 2022). In addition, TEs are extremely diverse, as they comprise multiple classes of genetic elements grouped into orders, superfamilies, families, and subfamilies, which can vary immensely in sequence, length, structure, and distribution (Wicker et al. 2007).

In eukaryotic genomes, TEs propagate in a selfish manner, being considered essentially genomic parasites (Legrand et al. 2019; Orgel and Crick 1980). TEs and their hosts are in a constant arms race where TE invasion may be counteracted by suppression of TE expression or by TE hypermutation (Skipper et al. 2013). TE families can be very old in evolutionary time and consequently accumulate mutations that would produce inactive copies, as a result of mutations or fragmentation during or after insertion, and this can be quantified using Kimura 2-parameter distance to consensus (K-value) (Kimura 1980). It has been shown that TEs may represent a major source of genetic variation in living organisms (M. G. Kidwell and Lisch 2001), and they could be a powerful source of data to compare genomes from closely related species or species with slow-paced evolution, such as some major reptilian clades (Green et al. 2014). Testudines constitute one of the reptilian clades with slow rates of nucleotide substitution compared to other vertebrates (Green et al. 2014; Avise et al. 1992) and sea turtles (superfamily Chelonioidea) have been shown to keep low levels of genetic divergence in different genome-wide analyses (Komoroske, Miller, and O'Rourke 2019; Vilaça et al. 2021; Zbinden et al. 2007; van der Zee et al. 2022; Driller, Vilaca, and Arantes 2020). Despite their ancient divergence times of 58-100 MY (Thomson, Spinks, and Shaffer 2021), *Chelonia mydas* and *Dermochelys coriacea* – representatives from the two living sea turtle families – show strikingly similar

genomic synteny and gene colinearity (Bentley et al. 2023). The high conservation levels suggest sea turtles as an excellent model group to study the evolution of TEs since speciation. Interest in turtle mobilomes emerged over 30 years ago (Endoh and Okada 1986) and has led to major contributions, such as the discovery that Short Interspersed Elements (SINEs) hijack the retropositional machinery of LINEs, achieving this by acquiring 3' sequence fragments from LINEs (Kajikawa, Ohshima, and Okada 1997). Despite these early findings, little is known about the recent dynamics of TEs in this reptilian clade.

Out of the different orders of TEs, the most abundant in reptilian genomes are LINEs (Sotero-Caio et al. 2017; Shaffer et al. 2013; Wang et al. 2013). Nonetheless, *Penelope*-like Elements (PLEs) are a particularly interesting group of group I transposon, characterised by two open reading frames (ORFs): one coding for reverse transcriptase (RT) and another for a GIY-YIG endonuclease (EN) (Evgen'ev and Arkhipova 2005; Wicker et al. 2007). Moreover, PLEs seem to have a different origin than the other retrotransposons group I elements (Wicker et al. 2007). The GIY–YIG EN domain typically associated with PLEs may have its evolutionary origins in bacterial group I introns, which are not retroelements (Stoddard 2014). PLE ENs are characteristically homing proteins because of the CCHH Zn-finger motif, with two cysteines located directly between the CIY and the YIG motifs (Arkhipova 2006). PLEs are also interesting from a phylogenetic perspective since their RT does not belong to either long terminal repeat (LTR) or LINE retrotransposon classes, but to a sister clade of telomerase reverse transcriptase (TERTs), which use a specialised RNA template to add G-rich repeats capping telomeres (Arkhipova et al. 2003). All described PLEs can be classified into two main categories: endonuclease-deficient (EN-), which are found in several kingdoms at or near telomeres, and endonuclease-containing (EN+), which use the aforementioned GIY-YIG endonuclease to transpose throughout the genome (Craig et al. 2021; Gladyshev and Arkhipova 2007). Despite the ancient origin of PLEs predating their divergence from TERTs, which are pan-eukaryotic, the phylogenetic distribution of PLEs (EN+) so far appears to be restricted to animals, with one exception of documented horizontal transfer to conifers (Lin et al. 2016). Additionally to this classification, PLEs have been subclassified into clades by the presence or absence of different ORFs (Capy 2005; Arkhipova 2006; Craig et al. 2021). The described EN+ clades are *Penelope, Poseidon, Neptunes, Hydra, Chlamys, Naiad,* and *Nematis. EN- clades include Athena* and *Coprina*, among others.

In this study, we compared the high-quality genomes (Rhie et al. 2021) from two sea turtles representing the two extant families Dermochelyidae and Cheloniidae. We analysed and explored the mobilomes of these family representatives and compared them with the 13 available Testudines assemblies. We identified that the main difference between these

mobilomes is the expansion of PLEs in *D. coriacea*. More specifically, we identified a new subfamily of PLEs present in *D. coriacea*, with evidence of recent insertions and similarities to other *Neptune* elements identified on different species.

Methods

Genomes and their raw sequencing data were retrieved from the National Center for Biotechnology Information database (NCBI: http://www.ncbi.nlm.nih.gov/) using the latest version available for each assembly. In order to assess the quality of the assemblies prior to analysis, the Genome Evaluation Pipeline (https://git.imp.fu-berlin.de/cmazzoni/GEP) was run, yielding results for analyses such as BUSCO (Seppey, Manni, and Zdobnov 2019), Sanger contig stats (*Assembly-Stats: Get Assembly Statistics from FASTA and FASTQ Files* n.d.), kmer analysis, mercury (Rhie et al. 2020) and N50 values for each assembly (Supplementary Table 1).

TEs and unclassified repeats from the testudines genome assemblies of *Emydura subglobosa*, *Podocnemis expansa*, *Carettochelys insculpta, Pelodiscus sinensis, Chelydra serpentina, C. mydas, D. coriacea, Platysternon megacephalum, Terrapene carolina triunguis, Chrysemys picta bellii, Trachemys scripta elegans, Chelonoidis abingdonii, Gopherus evgoodei, Mauremys reevesii* and *Cuora mccordi* (Bioproject Id at Supplementary Table 1) were recovered by creating a *de-novo* TE library for each genome using RepeatModeler2 (Flynn et al. 2020) and the module -LTRStruct. Using the library for each species, RepeatMasker (Tarailo-Graovac and Chen 2009; Smit, Hubley, and Green 2015) was run with the additional parameters -a -s -gccalc to calculate Kimura 2-parameter distance to consensus (K-value) with divCpGMod (Smit, Hubley, and Green 2015; Tarailo-Graovac and Chen 2009) for all the TEs identified using the script *calcDivergenceFromAlign.pl*. To recover and plot the TEs statistics, two in-house scripts were used, respectively, *align_with_divHandeler.py* and *PlotTEstats.R* (https://github.com/Tcvalenzuela/Recent-expansion-of-Penelope-like-elements-in-the-leatherback-turtle-Dermochelys-coriacea).

To improve the annotation, manual curation was performed on each sea turtle TE library, where each insertion was extended to 2000 bp on both flanks and then clustered together for examination of the characteristic component of the respective family of TE. This was done using a set of manual curation and identification as TE-Aid (Goubert et al. 2022), Repbase (Jurka et al. 2005; Kohany et al. 2006; Kapitonov and Jurka 2008), and CDD/SPARCLE (Lu et al. 2020) following the recommendations of Goubert et al. (2022).

In order to investigate the evolutionary context of a newly identified PLE from the *Neptune* family present in *D. coriacea*'s genome (*Neptune*-1_DC), we performed phylogenetic analysis using the RT of PLEs retrieved from Repbase and present in genomes from many animal species (Supplementary Table 2). The NCBI CDD database (Lu et al. 2020; Marchler-Bauer et al. 2015) was searched to identify conserved protein domains. After extraction of the

conserved protein domains identified, multiple sequence alignments were performed with PROMALS3D (Pei, Tang, and Grishin 2008; Pei and Grishin 2014; Pei, Kim, and Grishin 2008) including telomerase reverse transcriptase Protein Data Bank files (3kyl, 3du5) to assess the secondary structure of the proteins. Alignments were visualised in Jalview (Waterhouse et al. 2009), using the Clustal2 colouring scheme, and visually checked to confirm the presence of each conserved RT motif (Supplementary Figure 1). Maximum likelihood phylogenetic inference was then performed using IQ-TREE v.2.0.3 (Minh et al. 2020) and the most appropriate model of evolution was selected using ModelFinder (Kalyaanamoorthy et al. 2017). Branch support was assessed through 1,000 ultrafast bootstrap replicates (Hoang et al. 2018). Finally, the trees were visualised and edited with FigTree v1.4.4 (Rambaut 2014). To identify if any TE subfamily was significantly younger than the others, we used the confidence interval around the median (+/- 1.57 x IQR/sqrt of n) (Chambers et al. 1983).

To identify potentially active copies of TEs, we analysed the long-read transcriptome (IsoSeq) data from 3 different tissues (brain, ovaries, and lungs) of *D. coriacea* available at the NCBI database under identifiers SRR9594996, SRR9594994, and SRR9594995, respectively (Bentley et al. 2023). IsoSeq reads were mapped to the genome of *D. coriacea* (GCF_009764565.3) with minimap2 (Li 2018), applying the additional parameters *-ax splice -uf –secondary=no -C5 -O6,24 -B4.* To visually explore the genome using IGV (Thorvaldsdóttir, Robinson, and Mesirov 2013; Robinson et al. 2022), we put together the fasta file, the gff for the fasta file, the sorted bam of the mapping and a custom-made bed file that indicates scaffold, start, end, name, and K-value for each TE separated by tabs. This custom-made bed file is a simplification of the output of *align_with_divHandeler.py* where we filtered for the desired TE families, k-value, length, or any other characteristic that we could be interested in for each particular case.

Results

TE comparison between *C. mydas* and *D. coriacea*

Manual curation of the *de novo* TE library generated by RepeatModeler2 substantially reduced the number of unclassified ("Unknown") elements via their assignment to TE categories whenever possible. The genome percentages of unknown elements for *D. coriacea* and *C. mydas* were reduced from 25.64% and 24.48% to 14.5% and 16.8% , respectively (Table 1). This was performed mainly by identifying different subfamilies clustered together by RepeatMasker as one subfamily, and splitting them into the respective "real" subfamilies as described in Methods (Goubert et al. 2022).

The overall proportion of TEs belonging to the different TE orders was found to be similar in the genomes of *C. mydas* and *D. coriacea* (Table 1). The most striking difference in the TE genome composition between the two turtles refers to LINEs and PLEs, where the latter represent more than double the genome proportion in *D. coriacea* in comparison to *C. mydas* (4.70% vs. 2.34%).

Table 1. Summary statistics for the two analysed reference genome assemblies. Repeat-masked regions are summarised in main categories.

	D. coriacea			*C. mydas*		
	Number of elements	Length occupied [bp]	Percentage of the genome	Number of elements	Length occupied [bp]	Percentage of the genome
Retrotransposon	1,413,834	520,779,780	24.06	1,324,757	437,477,849	20.50
SINEs	352,713	54,104,600	2.50	398,125	61,313,780	2.87
PLEs	282,130	101,830,861	4.70	166,116	50,010,116	2.34
LINEs	962,901	394,684,017	18.23	833,888	307,190,132	14.39
LTR elements	98,220	71,991,163	3.33	92,744	68,973,937	3.23
DNA transposons	487,853	139,785,029	6.46	521,973	146,552,919	6.87
Unclassified	1,776,631	316,126,834	14.60	1,790,156	351,644,615	16.48
Small RNA	45,675	7,619,320	0.35	54,482	9,128,255	0.43

PLE expansion of *D. coriacea*

To explore differences in the accumulation of TE insertions between *C. mydas* and *D. coriacea*, the K-value of each insertion against its consensus sequence was calculated and a divergence profile generated (Figure 1). We identified that *D. coriacea* has a substantial accumulation of TE insertions with K-values between 0% and 2% (136,923 copies in total), and these younger TE subfamilies mostly belonged to LINEs and PLEs. Additionally, we explored the divergence profile of 15 Testudines representatives from 6 turtle superfamilies - Chelidae, Pelomedusoidea, Trionychoidea, Kinosternoidea, Chelonioidea, and Testudinoidea - (Supplementary Figure 2: summarised in Figure 2) of similarly between them high-quality assemblies based on BUSCO and QV scores (Supplementary Table 1). We identified that *D. coriacea* has a recent expansion of LINEs and PLEs not present in any of the other Testudines genomes analysed. *D. coriacea* was the only one for which TEs with K-values of 0% and 1% surpassed a genome proportion of 2% (i.e., 43,295 Mp; see first two columns on the TE divergency profiles plots, Supplementary Figure 2). Other turtles with a relatively high percentage of genomes composed of TEs with low K-value are *Chelydra serpentina* (0.79% or 17,836 Mb at K-value 0%) and *Cuora mccordi* (0.80% or 18,082 Mb at K-value 0%). We explored retrotransposons of the family LINE but we did not find anything as relevant there as we found in PLEs.

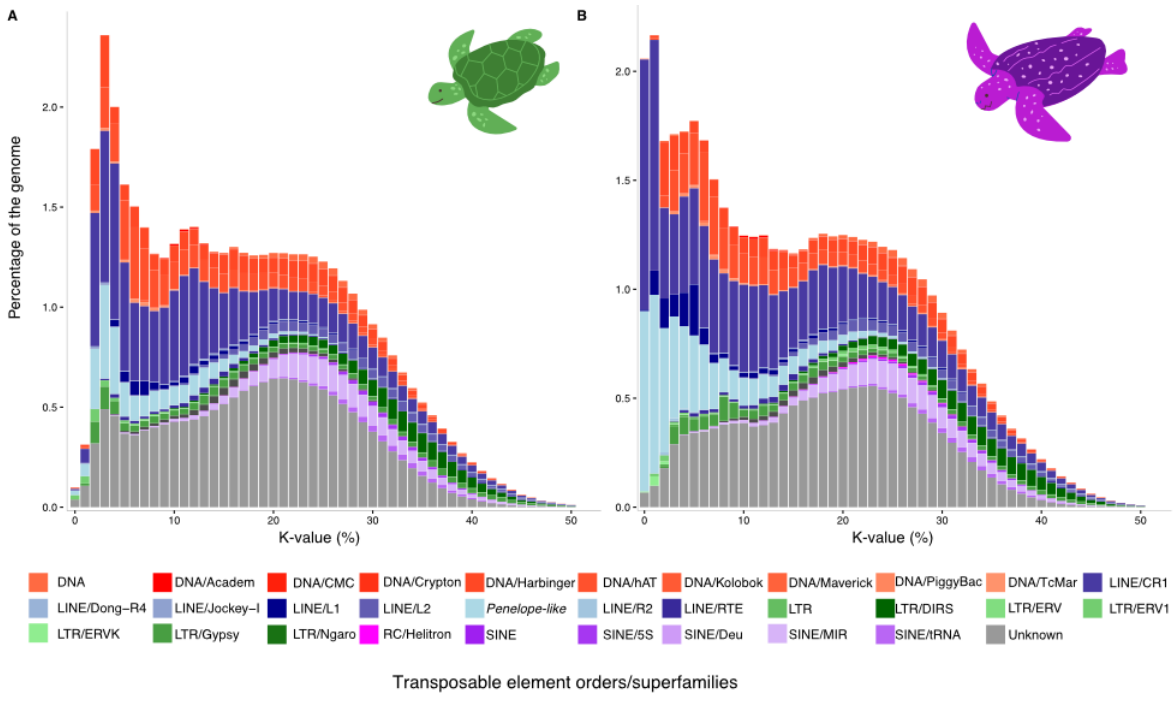


Figure 1. TE divergence profiles. The relative age of each insertion and their relative proportion in the genomes was calculated for *C. mydas* (A) and *D. coriacea* (B). On display are the main orders and superfamilies of transposable elements identified together with unclassified repeats ("Unknown", grey). The main observed

differences are in the lowest K-values (0-3%), where *D. coriacea* presents a higher proportion of their genome composed of LINE and PLEs.

Due to the expansion of PLEs observed in *D. coriacea* and absent in all other turtles included in this analysis, we explored PLE insertions in more detail by contrasting low, medium, and high K-values (low <5%; medium ≥5% and <15%; high ≥15%) to better understand the expansion of PLEs. We found that 140,928 (43.49%) PLE insertions have a low K-value, out of a total of 324,013 PLEs insertions. This proportion is close to twice the amount of low K-value PLEs found for *C. mydas* 65,931 (28.5%). Moreover, the *D. coriacea* proportion of recent PLEs is higher than all the turtles included in this analysis (Supplementary Table 3), where the average was 25,372 (18.92%) of low K-value PLE insertions.

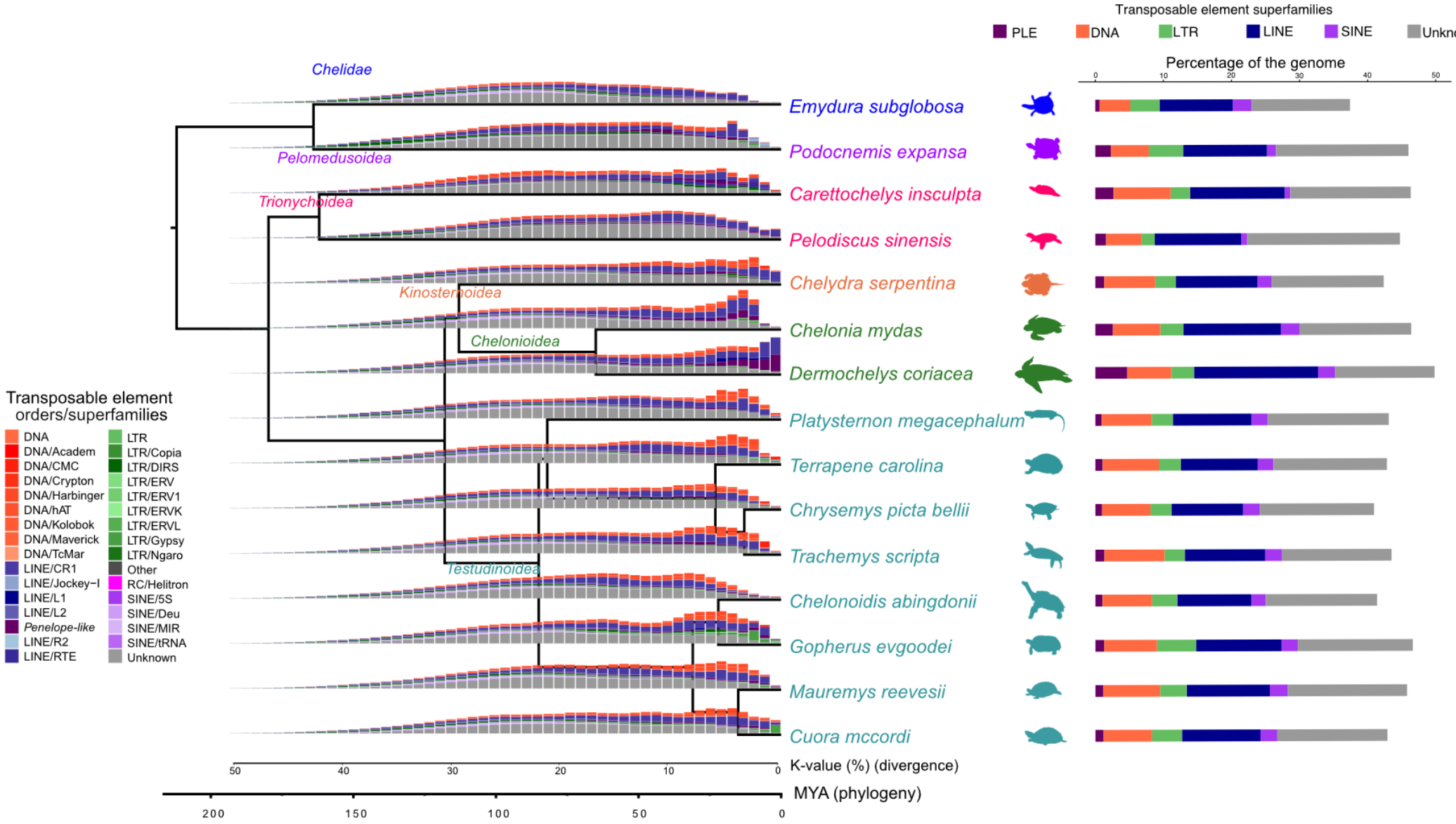


Figure 2. Transposable element content for genome assemblies from 15 different turtle species spanning the breadth of turtle diversity. TE genome proportion for 15 turtles from 6 Testudines superfamilies with their phylogeny and the divergence profile for each turtle (full plot in Supplementary Figure 2). For each turtle species, RepeatModeler and RepeatMasker were run on each genome assembly and the major TE orders and superfamilies were summarised.

Active PLE subfamily in *D. coriacea*

To explore if any particular subfamily of PLEs has been recently active in the genome of *D. coriacea*, we rely on the K-value distribution, given that an active TE will generate identical copies and thus lead to significantly lower K-values than older, non-active TEs. We identified that seven subfamilies have a distribution of divergence for their insertions with 3 out of the 4 quartiles under the overall average and significantly lower than all others using the confidence interval around the median (Supplementary Figure 3) (Chambers et al. 1983). One particular subfamily (see below) was identified with a high proportion of elements with low K-values, 140,713 (43.4 %) copies with low K-values, 131,897 (40.7 %) copies with medium K-value, and presented an average K-value significantly lower than all the other PLE subfamilies in *D. coriacea* using the confidence interval around the median (Supplementary Figure 1; Supplementary Table 3).

We explored the Repbase database and identified that this particular subfamily shows a sequence similarity of 91.38% with *Neptune*-1_CPB from *Chrysemys picta bellii*, in a segment longer than 80 bp, fulfilling the 95-80-98 rule for a separate subfamily (Flutre et al. 2011) (Figure 3C). Therefore, we decided to name it *Neptune*-1_DC. It is important to highlight that here we identified a 5' truncation on *Neptune*-1_DC, since the sequence similarity match with *Neptune*-1_CPB starts only at position 1088 bp of *Neptune*-1_DC, a phenomenon that is expected of PLEs given their transposition strategy. We identified the GIY-YIG endonuclease domains (Accessions Cdd:cd00304 and Cdd:cd10442, respectively) together with retrotranscriptase TERTs characteristic for this element and in the correct order according to previous PLE studies (Figure 3B) cataloguing *Neptune*-1_DC as an EN+ PLE (Craig et al. 2021; Evgen'ev and Arkhipova 2005; Arkhipova 2006). Through an analysis comparing expression data of the turtle's ovaries, brain, and lung tissues, we explored each genomic *Neptune*-1_DC insertion and identified expression in the three tissues independently (example case shown in Figure 3A).

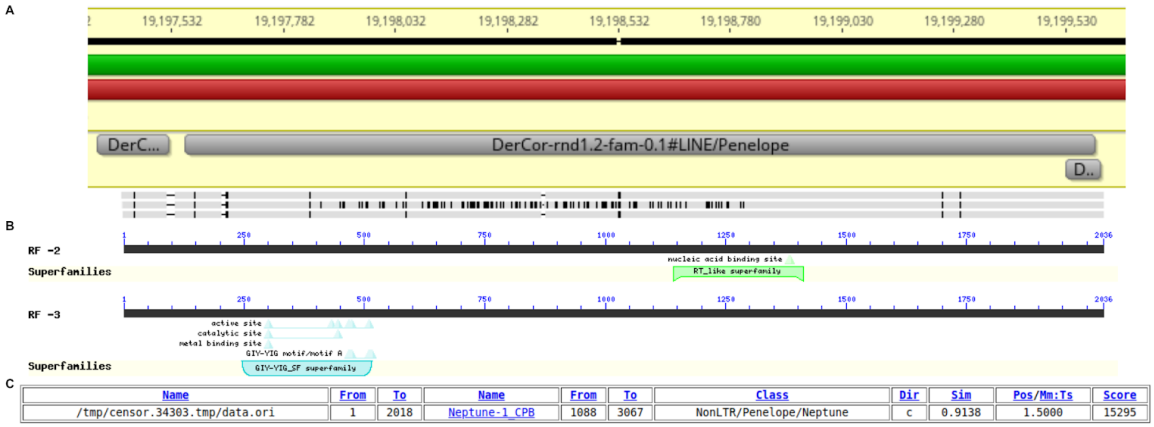


Figure 3. Characterization of a likely active *Neptune*-1_DC copy in *Dermochelys coriacea*. (A) Geneious browser view of a *Neptune*-1_DC copy in Scaffold 1 and three aligned IsoSeq reads (grey). (B) CDD/SPARCLE view of the protein domain detected from the sequence of *Neptune*-1_DC shown in (A). (C) Repbase CENSOR results of masking the consensus sequence of *Neptune*-1_DC, indicating high similarities with *Neptune*-1_CPB from *Chrysemys picta bellii*.

We generated a phylogeny of the RT domain of the newly identified *Neptune*-1_DC element and other PLEs sequences retrieved from Repbase and previous studies (Craig et al. 2021, Arkhipova, 2006) and present in other animal genomes (Figure 4). As expected, *Neptune*-1_DC clustered together with other *Neptune* elements and was more distantly related to other PLE superfamilies such as *Poseidon* and *Naiad* (Figure 4).

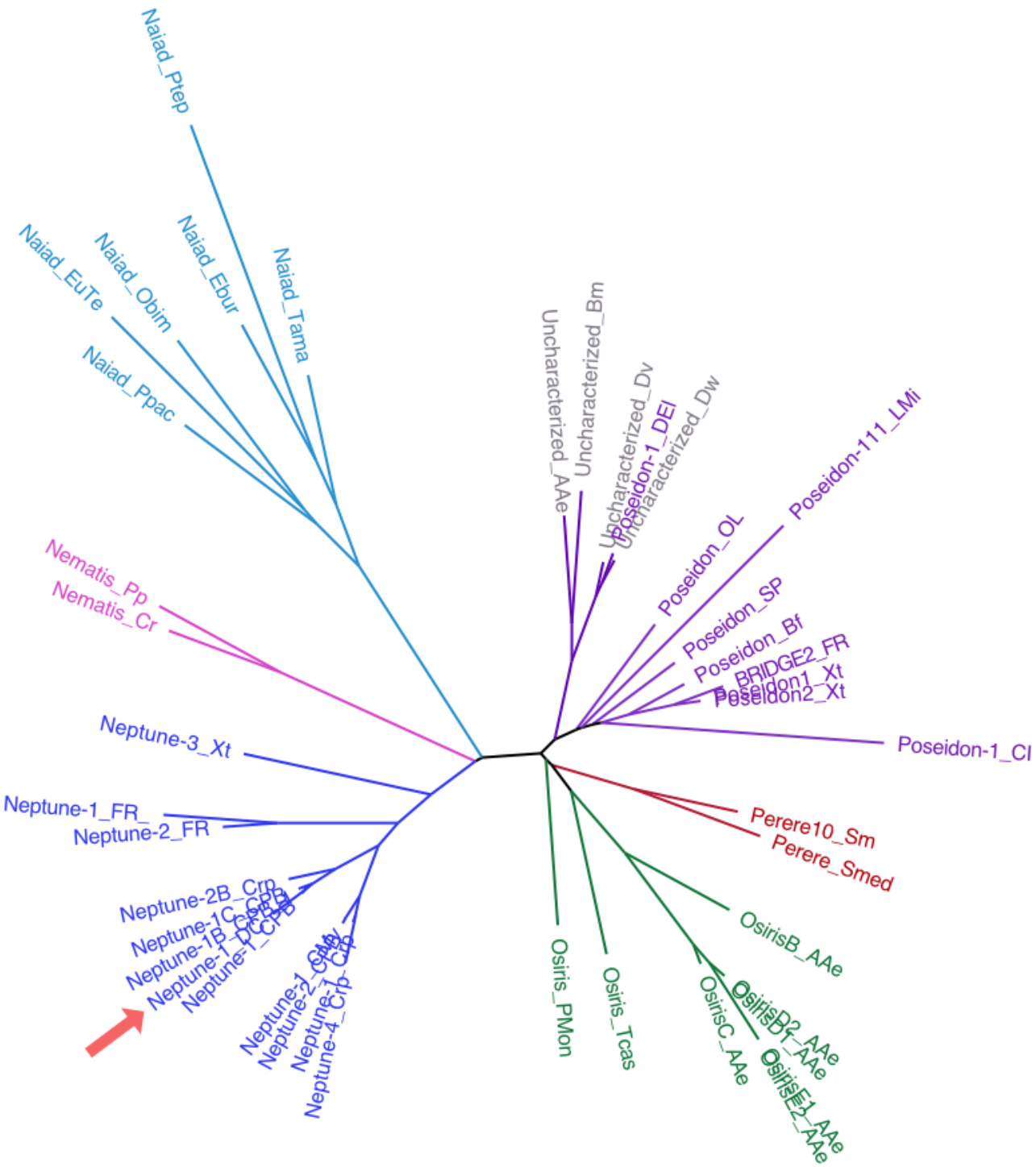


Figure 4. Maximum likelihood phylogeny of PLEs based on the amino acid sequence of the RT domain for different PLE subfamilies from several species listed in Supplementary Table 4. Sequences was obtained from Repbase. The colours indicate the different PLE superfamilies. Marked with a red arrow is the likely active PLE subfamily we identified in *D. coriacea* (*Neptune*-1_DC).

Furthermore, the novel described *Neptune*-1_DC formed a strongly supported clade with all *Neptune* PLEs from testudinian (*Chrysemys picta bellii*) and crocodilian genomes (*Crocodylus porosus*) included in the analysis. Additionally, the *Neptune* element clade shows some degree of congruence with the phylogenetic relationships between the host species, with *Neptune* elements from reptilians clustering together with those present in amphibian (*Xenopus tropicalis* - Xt) and fish species (*Takifugu rubripes* - FR). It is important to highlight that this pattern is present also in the other clades from the different PLEs included in the phylogeny of the RT present in this study, showing an astonishing level of clade diversity in PLEs RT as also described by Craig (2021).

Discussion

This study brings new insights into the transposable element (TE) dynamics within Testudines and reports a recent expansion of *Penelope*-like elements (PLEs) on an otherwise highly conserved and slow-evolving sea turtle genome.

We compared the TE composition of two extant species of sea turtles - *C. mydas* and *D. coriacea* - estimated to have shared their last common ancestor around 58-100 MYA (Wang et al. 2013; Shaffer et al. 2013; Vilaça et al. 2021; Thomson, Spinks, and Shaffer 2021). In a recent study (Bentley et al, 2023), we showed that TEs comprise a similar proportion of these species' genomes, reaching 45.79% for *D. coriacea* and 44.41% for *C. mydas* (Bentley et al, 203), values significantly higher than those reported by previous studies (close to 10% for *C. mydas*) (Wang et al. 2013; Shaffer et al. 2013; Sotero-Caio et al. 2017). However, it is important to note that previous analyses have been performed using a draft version of the species genome available at the time, something that we addressed with reference genomes in Bentley et al. (2023) and expanded upon here. Despite being an assembly based only on short-read sequencing technologies, the draft genome of *C. mydas* (Wang et al. 2013) presented high completeness and broad contiguity levels in commonly used metrics such as BUSCO and scaffold N50. Nonetheless, BUSCO scores and scaffold N50 values are not considered good indicators to assess highly repetitive regions of the genome (Peona, Blom, Xu, et al. 2021; Prost et al. 2019). As Peona et al. (2021) have shown, when compared to Illumina short reads, PacBio long reads allow the assembly of higher numbers of (young) TEs and are especially effective in the identification of novel subfamilies of TEs. Given that TE identification is highly influenced by the completeness and accuracy of the genome assemblies used (Wierzbicki et al. 2020; Bergman and Quesneville 2007; Rhie et al. 2021; Peona, Blom, Xu, et al. 2021), we believe that our analyses - based on high-quality near error-free reference genomes assembled using long reads - significantly increase the robustness of the results and represent an important advancement in the understanding of the mobilomes within Chelonioidea.

Additionally, we reinforce the importance of manual curation of the TE repeats identified by algorithms. While performing manual curation, we were able to identify several incorrect classifications, including i) consensus sequences that were annotated as belonging to a certain superfamily of TE, but were in fact a mixture of different subfamilies, ii) insertions lacking the characteristic component of the respective family and iii) multigenic families flagged as "#Unknown". These issues with classifications of TE are a known problem of currently available TE identification and classification pipelines and strategies (Goubert et al. 2022; Peona, Blom, Frankl-Vilches, et al. 2021; Galbraith et al. 2021; Boman et al. 2019).

Despite the high degree of similarity of the TE content, differences in abundance were found when comparing the divergence profiles of some TE subfamilies using K-values (Figure 1). *D. coriacea* presents 2% of the genome with younger insertions within K-values of 0-2% (Figure 1), which are not present on *C. mydas*. These insertions are mostly LINEs and PLEs, indicating a recent expansion of these elements in *D. coriacea*'s genome. LINEs are the most abundant TE order in both genomes, with CR1 as the most abundant TE superfamily and a difference of only 4% between both sea turtles (Supplementary table 2). These results are in line with previous reports that LINEs and PLEs are comparatively abundant in Testudines (Sotero-Caio et al. 2017; Shaffer et al. 2013; Wang et al. 2013) and that the CR1 LINE superfamily is dominant among amniotes (Suh 2015; Suh et al. 2014). Nonetheless, for PLEs, the difference is more accentuated: they constitute twice the proportion of the genome of *D. coriacea* compared to *C. mydas* (4.70% versus 2.34%) indicating a more accentuated expansion on the PLEs of *D. coriacea*.

The phenomenon of expansion of a TE family (or several) post speciation has been well studied on several organisms, including Arabidopsis (Slotkin et al. 2009) and tobacco (McCormick 2004), *Drosophila* (Marcillac, Grosjean, and Ferveur 2005), fish (Renaut and Bernatchez 2011; Rogers and Bernatchez 2007) among others as reviewed here (Serrato-Capuchina and Matute 2018; Mérot et al. 2020). In reptiles, a similar case of expansion of TE families as a result of speciation has been shown for snakes, comparing the Burmese python genome with a TE content of ~21% of the genome versus the pit viper with a TE content of ~45% (Castoe et al. 2011; Kumar et al. 2017; Galbraith et al. 2022). These differences have been associated mostly with the TE expansion in the pit viper, occurring after these two species diverged ~90 Mya (Galbraith et al. 2022) a similar time of divergence between *C. mydas* and *D. coriacea* (Thomson, Spinks, and Shaffer 2021).

An initial simple comparison including only *C. mydas* and *D. coriacea* would not allow us to differentiate between PLE expansion in *D. coriacea* or contraction in *C. mydas*. In order to clarify this, we expanded the TE abundance analysis to include 13 other turtle species from all 6 superfamilies of testudines (Figure 3). It was identified that *D. coriacea* indeed presents a higher proportion of the genome as PLEs, supporting the idea that this species' genome went through an expansion of PLE insertions not seen in the other analysed representatives of Testudines. Additionally, in this broader comparison, we identified that *D. coriacea* was the only species with more than 2% of the genome composed of very young or young TEs with K-values between 0-2% (Supplementary Figure 2). The only other species with high abundance of TEs within low K-values are *C. serpentina* (Kinosternoidea) and *C. mccordi* (Testudinoidea) both of them with less than half of the proportion of insertions with K-values 0-2%, compared

to *D. coriacea*. Therefore, the expansion of PLEs of *D. coriacea* is exclusive to this species' lineage and we were not able to identify any recent TE expansions of comparable scale happening in any of the other turtles analysed.

After seeing the expansion in *D. coriacea*, we searched for potentially active copies of PLEs by focusing on the lowest K-values, based on the rationale that an active TE will generate identical copies and thus lead to significantly lower K-values than older, non-active TEs. We identified 7 subfamilies of PLE with significantly lower K-values (Supplementary Figure 3) using the confidence interval around the median (Chambers et al. 1983). Furthermore, within this group of subfamilies, we identified one with K-values significantly lower than all other subfamilies. In an effort to catalogue this subfamily, we identified it as a member of the *Neptune* family. PLEs as described by Arkhipova (Evgen'ev and Arkhipova 2005) have two major groups: endonuclease positive (EN+) or negative (EN-) (Craig et al. 2021; Pyatkov et al. 2004; Gladyshev and Arkhipova 2007). In this particular case, we identified evidence of an EN, distinguished by the presence of the GIY-YIG endonuclease domain (Figure 3B), particularly the accession Cdd:cd10442, characteristic of *Neptune* PLEs. This insertion also presented a reverse transcriptase TERT domain (Figure 3B). Additionally, this subfamily showed high levels of similarities with another *Neptune* from *Chrysemys picta bellii* (Figure 3C), *Neptune*-1_CPB, with evidence for a separate subfamily in *D. coriacea* following the 95-80-98 rule(Flutre et al. 2011). Therefore here we describe the *Neptune*-1_DC as an active and recently expanded subfamily of PLEs. By analysing RNA expression data from three different tissues, we identified actively transcribed copies of *Neptune*-1_DC , raising the possibility of current activity of this element in the genome of *D. coriacea*.

In summary, we have identified that in spite of the high levels of sequence similarity and chromosome collinearity between the genomes of the two sea turtles analysed, there is a recent expansion of PLEs in *D. coriacea.* We report for the first time a likely active PLE/*Neptune* in reptiles. This is a contribution to the understanding of the dynamics of TE in slow-evolving reptiles and serves as an exemplary case of how the deceleration in the evolutionary rates of testudines can make them a unique model for studying the evolution of genome features such as TEs.

Acknowledgments

We acknowledge CONICYT-DAAD for scholarship support to T.C.-V., the São Paulo Research Foundation – FAPESP (grant #2020/10372-6) to E.K.S.R. BeGenDiv is partially funded by the German Federal Ministry of Education and Research (BMBF, Förderkennzeichen 033W034A).

References

Arkhipova, Irina R. 2006. “Distribution and Phylogeny of Penelope-like Elements in Eukaryotes.” *Systematic Biology* 55 (6): 875–85.

———. 2018. “Neutral Theory, Transposable Elements, and Eukaryotic Genome Evolution.” *Molecular Biology and Evolution* 35 (6): 1332–37.

Arkhipova, Irina R., Konstantin I. Pyatkov, Matthew Meselson, and Michael B. Evgen’ev. 2003. “Retroelements Containing Introns in Diverse Invertebrate Taxa.” *Nature Genetics* 33 (2): 123–24.

Assembly-Stats: Get Assembly Statistics from FASTA and FASTQ Files. n.d. Github. Accessed October 24, 2022. https://github.com/sanger-pathogens/assembly-stats.

Avise, J. C., B. W. Bowen, T. Lamb, A. B. Meylan, and E. Bermingham. 1992. “Mitochondrial DNA Evolution at a Turtle’s Pace: Evidence for Low Genetic Variability and Reduced Microevolutionary Rate in the Testudines.” *Molecular Biology and Evolution* 9 (3): 457–73.

Bentley, Blair P., Tomás Carrasco-Valenzuela, Elisa K. S. Ramos, Harvinder Pawar, Larissa Souza, Alana Alexander, Shreya M. Banerjee, et al. 2023. “1 Divergent Sensory and Immune Gene Evolution in Sea Turtles with Contrasting Demographic and Life 2 Histories.” *The Proceedings of the National Academy of Sciences (PNAS)*.

Bergman, Casey M., and Hadi Quesneville. 2007. “Discovering and Detecting Transposable Elements in Genome Sequences.” *Briefings in Bioinformatics* 8 (6): 382–92.

Boissinot, Stéphane, Yann Bourgeois, Joseph D. Manthey, and Robert P. Ruggiero. 2019. “The Mobilome of Reptiles: Evolution, Structure, and Function.” *Cytogenetic and Genome Research* 157 (1-2): 21–33.

Boman, Jesper, Carolina Frankl-Vilches, Michelly da Silva Dos Santos, Edivaldo H. C. de Oliveira, Manfred Gahr, and Alexander Suh. 2019. “The Genome of Blue-Capped Cordon-Bleu Uncovers Hidden Diversity of LTR Retrotransposons in Zebra Finch.” *Genes* 10 (4). https://doi.org/10.3390/genes10040301.

Canapa, Adriana, Marco Barucca, Maria A. Biscotti, Mariko Forconi, and Ettore Olmo. 2015. “Transposons, Genome Size, and Evolutionary Insights in Animals.” *Cytogenetic and Genome Research* 147 (4): 217–39.

Capy, P. 2005. “Classification and Nomenclature of Retrotransposable Elements.” *Cytogenetic and Genome Research* 110 (1-4): 457–61.

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

Chambers, John M., William S. Cleveland, Beat Kleiner, and Paul A. Tukey. 1983. “Comparing Data Distributions.” *Graphical Methods for Data Analysis* 62. https://hero.epa.gov/hero/index.cfm/reference/details/reference_id/2325101.

Craig, Rory J., Irina A. Yushenova, Fernando Rodriguez, and Irina R. Arkhipova. 2021. “An Ancient Clade of Penelope-Like Retroelements with Permuted Domains Is Present in the Green Lineage and Protists, and Dominates Many Invertebrate Genomes.” *Molecular Biology and Evolution* 38 (11): 5005–20.

Driller, M., S. T. Vilaca, and L. S. Arantes. 2020. “Optimization of ddRAD-like Data Leads to High Quality Sets of Reduced Representation Single Copy Orthologs (R2SCOs) in a Sea Turtle Multi-Species Analysis.” *bioRxiv*. https://www.biorxiv.org/content/10.1101/2020.04.03.024331v1.abstract.

Elliott, Tyler A., and T. Ryan Gregory. 2015. “What’s in a Genome? The C-Value Enigma and the Evolution of Eukaryotic Genome Content.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 370 (1678): 20140331.

Endoh, H., and N. Okada. 1986. “Total DNA Transcription in Vitro: A Procedure to Detect Highly Repetitive and Transcribable Sequences with tRNA-like Structures.” *Proceedings of the National Academy of Sciences of the United States of America* 83 (2): 251–55.

Evgen’ev, and Arkhipova. 2005. “Penelope-like Elements–a New Class of Retroelements: Distribution, Function and Possible Evolutionary Significance.” *Cytogenetic and Genome Research*. https://www.karger.com/Article/Abstract/84984.

Flutre, Timothée, Elodie Duprat, Catherine Feuillet, and Hadi Quesneville. 2011. “Considering Transposable Element Diversification in de Novo Annotation Approaches.” *PloS One* 6 (1): e16526.

Flynn, Jullien M., Robert Hubley, Clément Goubert, Jeb Rosen, Andrew G. Clark, Cédric Feschotte, and Arian F. Smit. 2020. “RepeatModeler2: Automated Genomic Discovery of Transposable Element Families.” *Proceedings of the National Academy of Sciences* 117 (17): 9451–57.

Galbraith, James D., Alastair J. Ludington, Kate L. Sanders, Timothy G. Amos, Vicki A. Thomson, Daniel Enosi Tuipulotu, Nathan Dunstan, Richard J. Edwards, Alexander Suh, and David L. Adelson. 2022. “Horizontal Transposon Transfer and Its Implications for the Ancestral Ecology of Hydrophiine Snakes.” *Genes* 13 (2). https://doi.org/10.3390/genes13020217.

Galbraith, James D., Alastair J. Ludington, Kate L. Sanders, Alexander Suh, and David L. Adelson. 2021. “Horizontal Transfer and Subsequent Explosive Expansion of a DNA Transposon in Sea Kraits (Laticauda).” *Biology Letters* 17 (9): 20210342.

Gladyshev, Eugene A., and Irina R. Arkhipova. 2007. “Telomere-Associated Endonuclease-Deficient Penelope-like Retroelements in Diverse Eukaryotes.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (22): 9352–57.

Goubert, Rory J. Craig, Agustin F. Bilat, Valentina Peona, Aaron A. Vogan, and Anna V. Protasio. 2022. “A Beginner’s Guide to Manual Curation of Transposable Elements.” *Mobile DNA* 13 (1): 7.

Green, Richard E., Edward L. Braun, Joel Armstrong, Dent Earl, Ngan Nguyen, Glenn Hickey, Michael W. Vandewege, et al. 2014. “Three Crocodilian Genomes Reveal Ancestral Patterns of Evolution among Archosaurs.” *Science* 346 (6215): 1254449.

Hoang, Diep Thi, Olga Chernomor, Arndt von Haeseler, Bui Quang Minh, and Le Sy Vinh. 2018. “UFBoot2: Improving the Ultrafast Bootstrap Approximation.” *Molecular Biology and Evolution* 35 (2): 518–22.

Jurka, J., V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. 2005. “Repbase Update, a Database of Eukaryotic Repetitive Elements.” *Cytogenetic and Genome Research* 110 (1-4): 462–67.

Kajikawa, M., K. Ohshima, and N. Okada. 1997. “Determination of the Entire Sequence of Turtle CR1: The First Open Reading Frame of the Turtle CR1 Element Encodes a Protein with a Novel Zinc Finger Motif.” *Molecular Biology and Evolution* 14 (12): 1206–17.

Kalyaanamoorthy, Subha, Bui Quang Minh, Thomas K. F. Wong, Arndt von Haeseler, and Lars S. Jermiin. 2017. “ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates.” *Nature Methods* 14 (6): 587–89.

Kapitonov, Vladimir V., and Jerzy Jurka. 2008. “A Universal Classification of Eukaryotic Transposable Elements Implemented in Repbase.” *Nature Reviews. Genetics*.

Kidwell, Margaret G. 2002. “Transposable Elements and the Evolution of Genome Size in Eukaryotes.” *Genetica* 115 (1): 49–63.

Kidwell, M. G., and D. R. Lisch. 2001. “Perspective: Transposable Elements, Parasitic DNA, and Genome Evolution.” *Evolution; International Journal of Organic Evolution* 55 (1): 1–24.

Kimura, M. 1980. “A Simple Method for Estimating Evolutionary Rates of Base Substitutions through Comparative Studies of Nucleotide Sequences.” *Journal of Molecular Evolution* 16 (2): 111–20.

Kohany, Oleksiy, Andrew J. Gentles, Lukasz Hankus, and Jerzy Jurka. 2006. “Annotation, Submission and Screening of Repetitive Elements in Repbase: RepbaseSubmitter and Censor.” *BMC Bioinformatics* 7 (October): 474.

Komoroske, L. M., M. R. Miller, and S. M. O’Rourke. 2019. “A Versatile Rapture (RAD-Capture) Platform for Genotyping Marine Turtles.” *Molecular Ecology*. https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.12980.

Kumar, Sudhir, Glen Stecher, Michael Suleski, and S. Blair Hedges. 2017. “TimeTree: A Resource for Timelines, Timetrees, and Divergence Times.” *Molecular Biology and Evolution* 34 (7): 1812–19.

Legrand, Sylvain, Thibault Caron, Florian Maumus, Sol Schvartzman, Leandro Quadrana, Eléonore Durand, Sophie Gallina, et al. 2019. “Differential Retention of Transposable Element-Derived Sequences in Outcrossing Arabidopsis Genomes.” *Mobile DNA* 10 (July): 30.

Li, Heng. 2018. “Minimap2: Pairwise Alignment for Nucleotide Sequences.” *Bioinformatics* 34 (18): 3094–3100.

Lin, Xuan, Nurul Faridi, and Claudio Casola. 2016. “An Ancient Transkingdom Horizontal Transfer of Penelope-Like Retroelements from Arthropods to Conifers.” *Genome Biology and Evolution* 8 (4): 1252–66.

Lu, Shennan, Jiyao Wang, Farideh Chitsaz, Myra K. Derbyshire, Renata C. Geer, Noreen R. Gonzales, Marc Gwadz, et al. 2020. “CDD/SPARCLE: The Conserved Domain Database in 2020.” *Nucleic Acids Research* 48 (D1): D265–68.

Marchler-Bauer, Aron, Myra K. Derbyshire, Noreen R. Gonzales, Shennan Lu, Farideh Chitsaz, Lewis Y. Geer, Renata C. Geer, et al. 2015. “CDD: NCBI’s Conserved Domain Database.” *Nucleic Acids Research* 43 (Database issue): D222–26.

Marcillac, Fabrice, Yaël Grosjean, and Jean-François Ferveur. 2005. “A Single Mutation Alters Production and Discrimination of Drosophila Sex Pheromones.” *Proceedings. Biological Sciences / The Royal Society* 272 (1560): 303–9.

McCormick, Sheila. 2004. “Control of Male Gametophyte Development.” *The Plant Cell* 16 Suppl (Suppl): S142–53.

Mérot, Claire, Rebekah A. Oomen, Anna Tigano, and Maren Wellenreuther. 2020. “A Roadmap for Understanding the Evolutionary Significance of Structural Genomic Variation.” *Trends in Ecology & Evolution* 35 (7): 561–72.

Minh, Bui Quang, Heiko A. Schmidt, Olga Chernomor, Dominik Schrempf, Michael D. Woodhams, Arndt von Haeseler, and Robert Lanfear. 2020. “IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era.” *Molecular Biology and Evolution* 37 (5): 1530–34.

Orgel, L. E., and F. H. Crick. 1980. “Selfish DNA: The Ultimate Parasite.” *Nature* 284 (5757): 604–7.

Pei, Jimin, and Nick V. Grishin. 2014. “PROMALS3D: Multiple Protein Sequence Alignment Enhanced with Evolutionary and Three-Dimensional Structural Information.” *Methods in Molecular Biology* 1079: 263–71.

Pei, Jimin, Bong-Hyun Kim, and Nick V. Grishin. 2008. “PROMALS3D: A Tool for Multiple Protein Sequence and Structure Alignments.” *Nucleic Acids Research* 36 (7): 2295–2300.

Pei, Jimin, Ming Tang, and Nick V. Grishin. 2008. “PROMALS3D Web Server for Accurate Multiple Protein Sequence and Structure Alignments.” *Nucleic Acids Research* 36 (Web Server issue): W30–34.

Peona, Valentina, Mozes P. K. Blom, Carolina Frankl-Vilches, Borja Milá, Hidayat Ashari, Christophe Thébaud, Brett W. Benz, et al. 2021. "The Hidden Structural Variability in Avian Genomes." https://www.diva-portal.org/smash/record.jsf?pid=diva2:1585441.

Peona, Valentina, Mozes P. K. Blom, Luohao Xu, Reto Burri, Shawn Sullivan, Ignas Bunikis, Ivan Liachko, et al. 2021. "Identifying the Causes and Consequences of Assembly Gaps Using a Multiplatform Genome Assembly of a Bird-of-Paradise." *Molecular Ecology Resources* 21 (1): 263–86.

Prost, Stefan, Ellie E. Armstrong, Johan Nylander, Gregg W. C. Thomas, Alexander Suh, Bent Petersen, Love Dalen, et al. 2019. "Comparative Analyses Identify Genomic Features Potentially Involved in the Evolution of Birds-of-Paradise." *GigaScience* 8 (5). https://doi.org/10.1093/gigascience/giz003.

Pyatkov, Konstantin I., Irina R. Arkhipova, Natalia V. Malkova, David J. Finnegan, and Michael B. Evgen'ev. 2004. "Reverse Transcriptase and Endonuclease Activities Encoded by Penelope-like Retroelements." *Proceedings of the National Academy of Sciences of the United States of America* 101 (41): 14719–24.

Rambaut. 2014. "FigTree v1.4.4 , a Graphical Viewer of Phylogenetic Trees." *Angle Bracket Http://tree. Bio. Ed. Ac. Uk/software/figtree …*.

Renaut, S., and L. Bernatchez. 2011. "Transcriptome-Wide Signature of Hybrid Breakdown Associated with Intrinsic Reproductive Isolation in Lake Whitefish Species Pairs (Coregonus Spp. Salmonidae)." *Heredity* 106 (6): 1003–11.

Rhie, Arang, Shane A. McCarthy, Olivier Fedrigo, Joana Damas, Giulio Formenti, Sergey Koren, Marcela Uliano-Silva, et al. 2021. "Towards Complete and Error-Free Genome Assemblies of All Vertebrate Species." *Nature* 592 (7856): 737–46.

Rhie, Arang, Brian P. Walenz, Sergey Koren, and Adam M. Phillippy. 2020. "Merqury: Reference-Free Quality, Completeness, and Phasing Assessment for Genome Assemblies." *Genome Biology* 21 (1): 245.

Robinson, James T., Helga Thorvaldsdóttir, Douglass Turner, and Jill P. Mesirov. 2022. "Igv.js: An Embeddable JavaScript Implementation of the Integrative Genomics Viewer (IGV)." *bioRxiv*. https://doi.org/10.1101/2020.05.03.075499.

Rodriguez, Matias, and Wojciech Makałowski. 2022. "Software Evaluation for de Novo Detection of Transposons." *Mobile DNA* 13 (1): 14.

Rogers, S. M., and L. Bernatchez. 2007. "The Genetic Architecture of Ecological Speciation and the Association with Signatures of Selection in Natural Lake Whitefish (Coregonus Sp. Salmonidae) Species Pairs." *Molecular Biology and Evolution* 24 (6): 1423–38.

Seppey, Mathieu, Mosè Manni, and Evgeny M. Zdobnov. 2019. "BUSCO: Assessing Genome Assembly and Annotation Completeness." *Methods in Molecular Biology* 1962: 227–45.

Serrato-Capuchina, Antonio, and Daniel R. Matute. 2018. "The Role of Transposable Elements in Speciation." *Genes* 9 (5). https://doi.org/10.3390/genes9050254.

Shaffer, H. Bradley, Patrick Minx, Daniel E. Warren, Andrew M. Shedlock, Robert C. Thomson, Nicole Valenzuela, John Abramyan, et al. 2013. "The Western Painted Turtle Genome, a Model for the Evolution of Extreme Physiological Adaptations in a Slowly Evolving Lineage." *Genome Biology* 14 (3): R28.

Skipper, Kristian Alsbjerg, Peter Refsing Andersen, Nynne Sharma, and Jacob Giehm Mikkelsen. 2013. "DNA Transposon-Based Gene Vehicles - Scenes from an Evolutionary Drive." *Journal of Biomedical Science* 20 (1): 1–23.

Slotkin, R. Keith, Matthew Vaughn, Filipe Borges, Milos Tanurdzić, Jörg D. Becker, José A. Feijó, and Robert A. Martienssen. 2009. "Epigenetic Reprogramming and Small RNA Silencing of Transposable Elements in Pollen." *Cell* 136 (3): 461–72.

Smit, A., R. Hubley, and P. Green. 2015. "RepeatMasker Open-4.0. 2013-2015http." *Repeatmasker. Org*.

Sotero-Caio, Cibele G., Roy N. Platt 2nd, Alexander Suh, and David A. Ray. 2017. "Evolution and Diversity of Transposable Elements in Vertebrate Genomes." *Genome Biology and Evolution* 9 (1): 161–77.

Stoddard, Barry L. 2014. "Homing Endonucleases from Mobile Group I Introns: Discovery to Genome Engineering." *Mobile DNA* 5 (1): 7.

Suh, Alexander. 2015. "The Specific Requirements for CR1 Retrotransposition Explain the Scarcity of Retrogenes in Birds." *Journal of Molecular Evolution* 81 (1-2): 18–20.

Suh, Alexander, Gennady Churakov, Meganathan P. Ramakodi, Roy N. Platt 2nd, Jerzy Jurka, Kenji K. Kojima, Juan Caballero, et al. 2014. "Multiple Lineages of Ancient CR1 Retroposons Shaped the Early Genome Evolution of Amniotes." *Genome Biology and Evolution* 7 (1): 205–17.

Symonová, Radka, and Alexander Suh. 2019. "Nucleotide Composition of Transposable Elements Likely Contributes to AT/GC Compositional Homogeneity of Teleost Fish Genomes." *Mobile DNA* 10 (December): 49.

Tarailo-Graovac, Maja, and Nansheng Chen. 2009. "Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences." *Current Protocols in Bioinformatics / Editorial Board, Andreas D. Baxevanis ... [et Al.]* Chapter 4 (1): Unit 4.10.

Thomson, Robert C., Phillip Q. Spinks, and H. Bradley Shaffer. 2021. "A Global Phylogeny of Turtles Reveals a Burst of Climate-Associated Diversification on Continental Margins." *Proceedings of the National Academy of Sciences of the United States of America* 118 (7). https://doi.org/10.1073/pnas.2012215118.

Thorvaldsdóttir, Helga, James T. Robinson, and Jill P. Mesirov. 2013. “Integrative Genomics Viewer (IGV): High-Performance Genomics Data Visualization and Exploration.” *Briefings in Bioinformatics* 14 (2): 178–92.

Tollis, M., and S. Boissinot. 2012. “The Evolutionary Dynamics of Transposable Elements in Eukaryote Genomes.” *Genome Dynamics* 7 (June): 68–91.

Vilaça, Sibelle Torres, Riccardo Piccinno, Omar Rota-Stabelli, Maëva Gabrielli, Andrea Benazzo, Michael Matschiner, Luciano S. Soares, Alan B. Bolten, Karen A. Bjorndal, and Giorgio Bertorelle. 2021. “Divergence and Hybridization in Sea Turtles: Inferences from Genome Data Show Evidence of Ancient Gene Flow between Species.” *Molecular Ecology* 30 (23): 6178–92.

Wang, Juan Pascual-Anaya, Amonida Zadissa, Wenqi Li, Yoshihito Niimura, Zhiyong Huang, Chunyi Li, et al. 2013. “The Draft Genomes of Soft-Shell Turtle and Green Sea Turtle Yield Insights into the Development and Evolution of the Turtle-Specific Body Plan.” *Nature Genetics* 45 (6): 701–6.

Waterhouse, Andrew M., James B. Procter, David M. A. Martin, Michèle Clamp, and Geoffrey J. Barton. 2009. “Jalview Version 2—a Multiple Sequence Alignment Editor and Analysis Workbench.” *Bioinformatics* 25 (9): 1189–91.

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

Wierzbicki, Filip, Florian Schwarz, Odontsetseg Cannalonga, and Robert Kofler. 2020. “Generating High Quality Assemblies for Genomic Analysis of Transposable Elements.” *bioRxiv*. https://doi.org/10.1101/2020.03.27.011312.

Zbinden, Judith A., Carlo R. Largiadèr, Fabio Leippert, Dimitris Margaritoulis, and Raphaël Arlettaz. 2007. “High Frequency of Multiple Paternity in the Largest Rookery of Mediterranean Loggerhead Sea Turtles.” *Molecular Ecology* 16 (17): 3703–11.

Zee, Jurjan P. van der, Marjolijn J. A. Christianen, Martine Bérubé, Mabel Nava, Sietske van der Wal, Jessica Berkel, Tadzio Bervoets, Melanie Meijer Zu Schlochtern, Leontine E. Becking, and Per J. Palsbøll. 2022. “Demographic Changes in Pleistocene Sea Turtles Were Driven by Past Sea Level Fluctuations Affecting Feeding Habitat Availability.” *Molecular Ecology* 31 (4): 1044–56.

Supplementary Material

Supplementary Table 1 List of all the genomes used on the TE analysis and their respective Genebank ID and Stats.

Species	Bioproject	Assembly Level	Assembly Stats Total Sequence Length	Contig N50	Scaffold N50	Scaffold #	GC content	QV	BUSCO C	BUSCO D	BUSCO F	BUSCO M	BUSCO SC	BUSCO TC	BUSCO V
Emydura subglobosa	PRJNA512130	Scaffold	1,986,702,056	350,989	44,759,016	22,003,978	43.03%	50.38	98.20%	1.60%	0.60%	1.20%	96.60%	3,354	5.0.0
Podocnemis expansa	PRJNA512135	Scaffold	2,447,092,617	134,824	37,101,370	47,912,441	43.60%	51.76	94.80%	1.00%	2.80%	2.40%	93.80%	3,354	4.0.6
Carettochelys insculpta	PRJNA512133	Scaffold	2,356,570,494	116,202	45,881,824	73,276,037	45.32%	54.54	92.70%	6.40%	3.60%	4.00%	86.00%	3,354	4.0.6
Pelodiscus sinensis	PRJNA221645	Scaffold	2,202,466,388	21,993	3,350,749	19,904	44.41%	34.07	96.50%	1.06%	1.51%	1.99%	95.44%	7,480	4.0.2
Chelydra serpentina	PRJNA574487	Scaffold	2,401,360,239	80,593	21,135,443	113,431	44%	48.82	95.00%	1.00%	3.60%	1.40%	94.00%	3,354	4.0.6
Chelonia Mydas	PRJNA675851	Chromosome	2,134,358,617	39,415,510	134,428,053	92	44.01%	47.70	99.01%	1.12%	0.31%	0.68%	97.89%	7,480	4.1.4
Dermochelys coriacea	PRJNA655518	Chromosome	2,164,762,090	7,029,801	137,568,771	40	43.35%	38.90	98.21%	1.03%	0.52%	1.27%	97.18%	7,480	4.1.4
Platysternon megacephalum	PRJNA479402	Scaffold	1,928,419,311	205,283	7,406,563	159,903	44.55%	38.41	98.00%	1.00%	0.80%	1.20%	97.00%	3,354	5.0.0
Terrapene carolina triunguis	PRJNA415469	Scaffold	2,571,267,249	76,614	24,249,581	131,541,780	44.27%	47.63	91.30%	9.80%	4.40%	4.30%	81.5%	3,354	4.0.6
Chrysemys picta belii	PRJNA210179	Chromosome	2,481,351,664	21,318	6,605,846	78,631	44.19%	49.65	98.20%	0.78%	1.02%	0.79%	97.42%	7,480	4.1.4
Trachemys scripta elegans	PRJNA634151	Chromosome	2,126,182,493	204,575	140,411,086	138	43.81%	50.43	97.57%	0.86%	0.63%	1.80%	96.71%	7,480	4.0.2
Chelonoidis abingdonii	PRJNA611832	Scaffold	2,300,739,315	73,186	1,277,207	10,618	43.71%	26.04	97.47%	0.99%	1.08%	1.44%	96.48%	7,480	4.0.2

Continuation table Supplementary Table 1

Gopherus evgoodei	PRJNA559383	Chromosome	2,298,547,364	13,026,736	147,425,149	382	44.14%	38.31	98.94%	1.43%	0.16%	0.90%	97.51%	7,480	4.0.2
Mauremys reevesii	PRJNA699301	Chromosome	2,035,064,793	34,524,243	139,244,951	62	44.55%	33.55	99.16%	1.59%	0.16%	0.68%	97.57%	7,480	4.0.2
Cuora mccordi	PRJNA491715	Scaffold	2,390,374,423	74,338	32,628,679	129,476,169	44.70%	50.94	91.60%	2.10%	4.40%	4.00%	89.50%	3,354	4.0.6

Supplementary table 2. Proportion of PLE lower than 5 K-values for all representatives of testudines lineage

Species	Total PLE	PLE K-value<5	[%]*
Platysternon megacephalum	98,437	17,123	17.39%
Pelodiscus sinensis	194,440	30,003	15.43%
Podocnemis expansa	258,507	25,003	9.67%
Carettochelys insculpta	312,531	42,634	13.64%
Emydura subglobosa	49,357	5,269	10.68%
Mauremys reevesii	94,041	24,117	25.65%
Gopherus evgoodei	125,045	33,071	26.45%
Chelonoidis abingdonii	108,624	26,160	24.08%
Terrapene carolina triunguis	110,600	14,841	13.42%
Cuora mccordi	125,192	35,390	28.27%
Chrysemys picta bellii	121,701	17,707	14.55%
Trachemys scripta elegans	119,010	33,145	27.85%
Chelonia mydas	231,717	66,113	28.53%
Dermochelys coriacea	324,013	140,928	43.49%

* Percentage of PLE with K-value < 5

Average without *C. mydas* and *D. coriacea*

Average Total PLE	Average PLE K-value<5	Average [%]*
143,124	25,372	18.92%

Supplementary table 3. General proportions of PLEs compare with the most abundant superfamiles of TEs

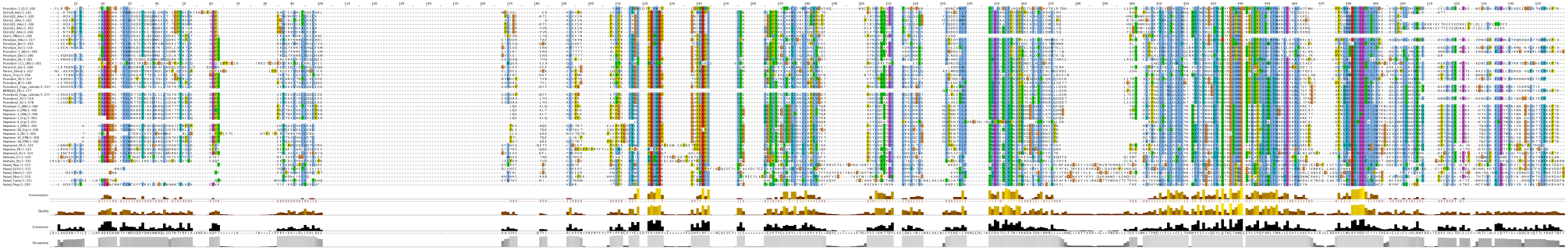
	Kimura	Total STR	%	LINEs	%	SINEs	%	PLEs	%	Neptune-1_DC	%
D. Coriacea	Low	268,003	10.25%	224,081	17.40%	836	0.17%	140,713	43.43%	24,509	70.70%
	Med	667,214	25.51%	395,975	30.75%	33,882	6.81%	131,897	40.71%	8,691	25.07%
	High	1,680,017	64.24%	667,650	51.85%	463,146	93.03%	51,403	15.86%	1,467	4.23%
	Total	2,615,234	100%	1,287,706	100.00%	497,864	100%	324,013	100.00%	34,667	100%

	Kimura	Total STR	%	LINEs	%	SINEs	%	PLEs	%
C. mydas	Low	187,468	7.48%	136,717	13.09%	702	0.12%	65,931	28.45%
	Med	626,819	25.01%	344,040	32.93%	34,741	6.11%	121,332	52.36%
	High	1,692,128	67.51%	563,949	53.98%	532,902	93.76%	44,454	19.18%
	Total	2,506,415	100%	1,044,706	100%	568,345	100%	231,717	100%

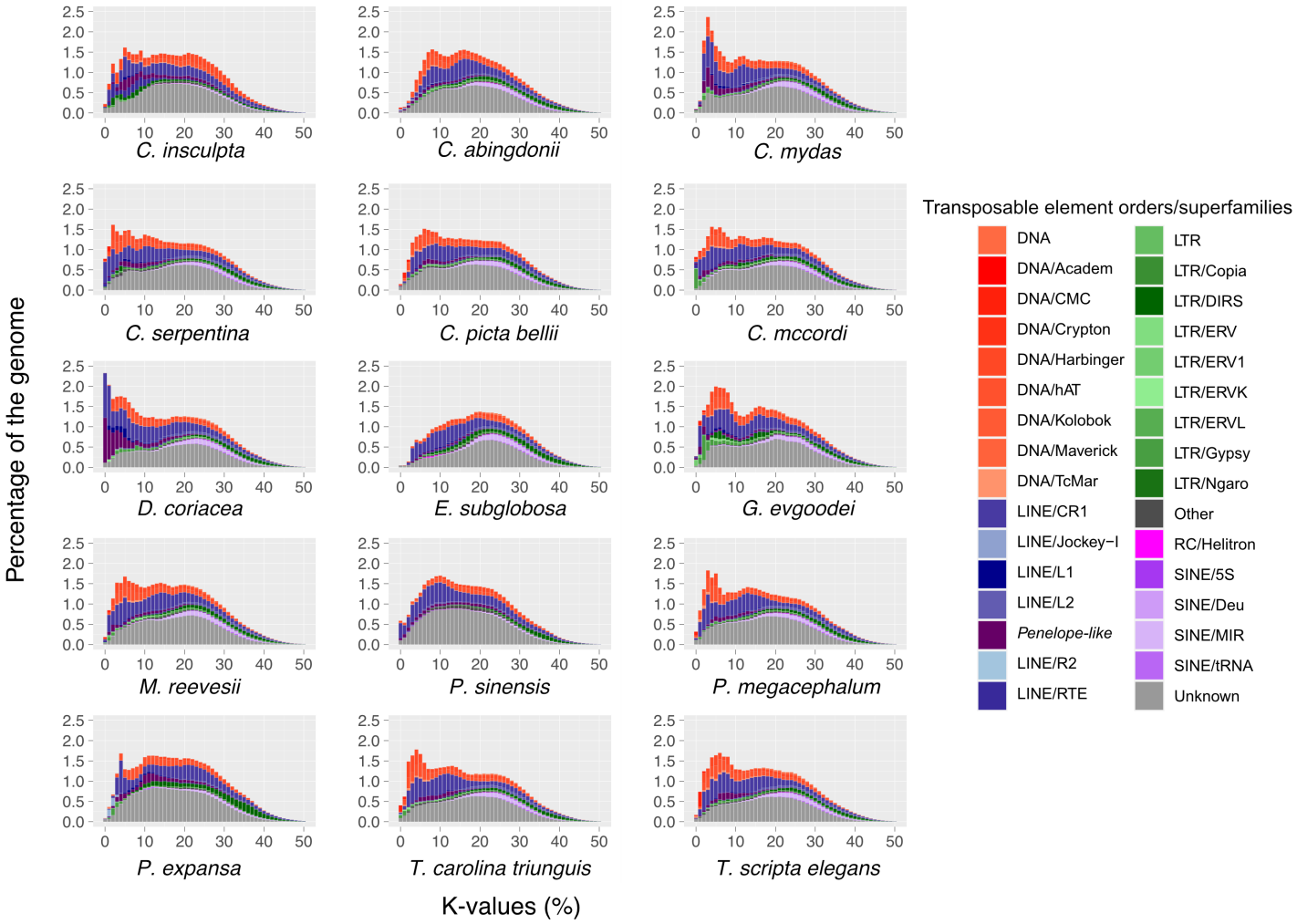
Low= [0-5[; Medium = [5-15[; High = [15-50]

Supplementary table 4. Species names and abbreviation used for phylogenetic tree.

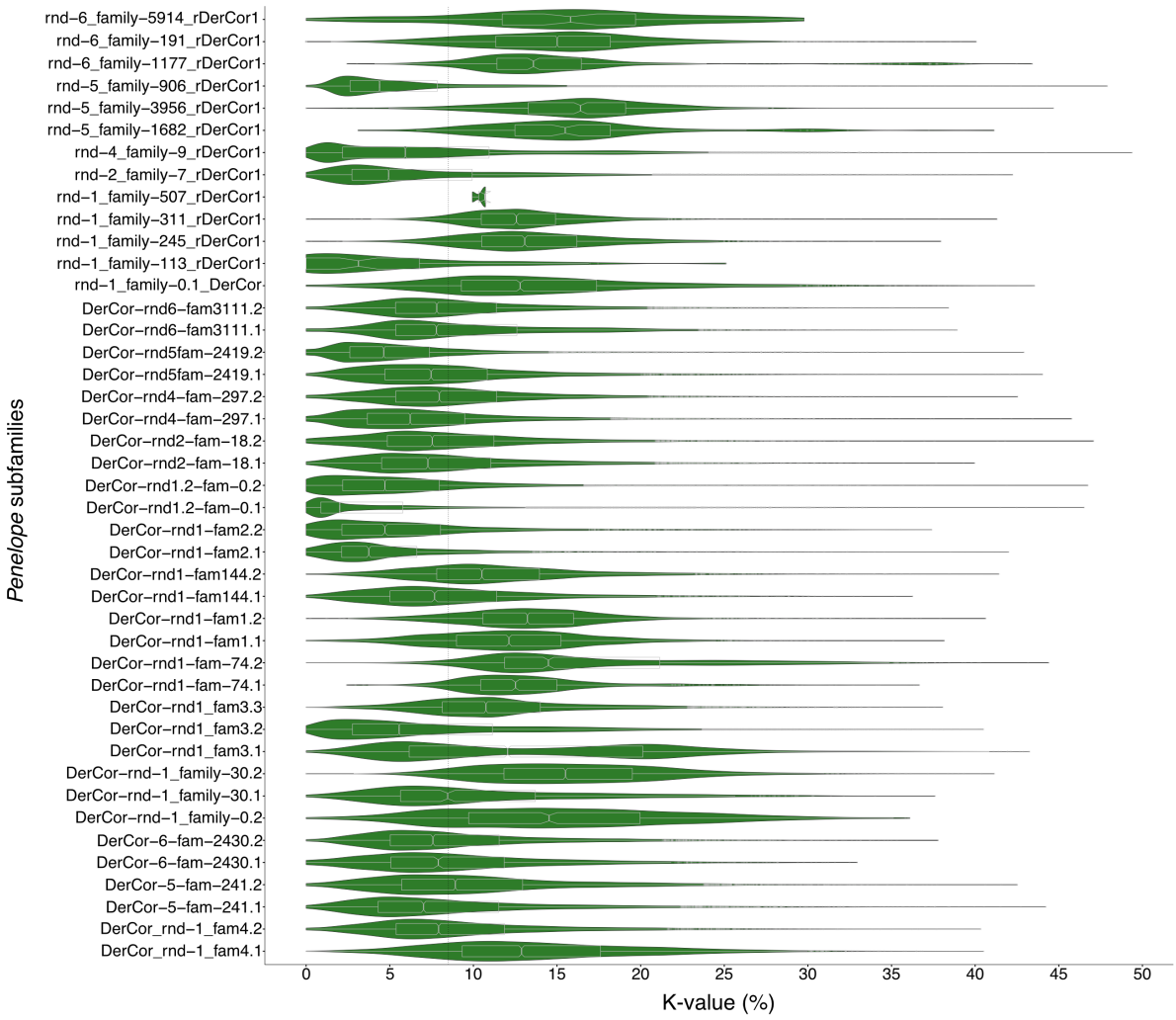
Abbreviation	Species
AAe	*Aedes aegypti*
AMi	*Alligator mississippiensis*
Bf	*Branchiostoma floridae*
Bm	*Bombyx mori*
CI	*Ciona intestinalis*
CMy	*Chelonia mydas*
CPB	*Chrysemys picta bellii*
Cr	*Caenorhabditis remanei*
Crp	*Crocodylus porosus*
DC	*Dermochelys coriacea*
DEl	*Drosophila elegans*
Dv	*Drosophila virilis*
Dw	*Drosophila willistoni*
Ebur	*Eptatretus burgeri*
EuTe	*Eulimnadia texana*
FR	*Takifugu rubripes*
LMi	*Locusta migratoria*
Obim	*Octopus bimaculoides*
OL	*Oryzias latipes*
PMon	*Penaeus monodon*
Pp	*Pristionchus pacificus*
Ppac	*Pristionchus pacificus*
Ptep	*Parasteatoda tepidariorum*
Sm	*Schistosoma japonicum*
Smed	*Schmidtea mediterranea*
SP	*Sphenodon punctatus*
Tama	*Thalassophryne amazonica*
Tcas	*Tribolium castaneum*
Xt	*Xenopus tropicalis*



Supplementary Figure 1. Sequence alignment of PLE elements for several species as listed in methods. Alignments were visualised in Jalview using the Clustal2 colouring scheme.



Supplementary Figure 2. Repeat elements divergence profiles according to RepeatMasker standart across the Testudines clade.



Supplementary Figure 3. Kimura 2-parameter distance to consensus (K-value) distribution for each PLE subfamily in *D. coriacea.* Frequency distributions are shown in violin plots (green), and quartile distributions are shown in boxplots (grey). The average K-value for all the PLEs is shown as a vertical dotted black line.

Chapter 3

Testudine-wide Transposable element exploration: A history of slow evolution and conserved genomes

Testudine-wide Transposable element exploration: A history of slow evolution and conserved genomes

Tomas Carrasco-Valenzuela[1,2,3]; Elisa K. S. Ramos[4]; Camila J. Mazzoni [1,2]

On format for submission on: *Mobile DNA*

1 Berlin Center for Genomics in Biodiversity Research (BeGenDiv), Berlin, Germany
2 Evolutionary Genetics Department, Leibniz-Institut für Zoo- und Wildtierforschung (IZW), Berlin, Germany
3 Universität Potsdam, Brandenburg, Potsdam, Germany
4 Laboratory of Evolutionary Genomics. Genetics, Evolution, Immunology and Microbiology Department, State University of Campinas, Brazil.

Abstract

Different environments offer distinct selective pressures associated to species adaptation and this process leaves behind signatures in species genomes. Transposable elements (TEs) are important agents during adaptation, changing and modulating their host genomes in a non-random way, through sequence-specific transposition, histone methylation modification and insertions dependent of gene density. This could produce gene duplication and or modification in the regulation of genes. Nonetheless, the relationship between genes and TEs has been poorly studied, especially in the focus on how TE are positionally related to genes features such as genes and exons. Here we show a comparative TE analysis on ten species from six different turtle superfamilies. We compare the general proportion of TE across different species. Additionally, we explore the interaction between TEs and gene features, describing their proportions in each species and proposing TE (*Helitron*) as an example of TEs actively duplicating gene features. The turtle species studied showed a similar level of TEs comprising around 42% of their genome content. We identify TEs interacting with different parts in a gene feature (exons and introns). We also report TEs integrating into the genome in the same proportion upstream and downstream of genes and exons. Also, we identify evidence of *Helitrons* containing a gene that could duplicate upon transposition, together with evidence of members of the same subfamily and fragments of the inserted gene in a different part of the genome. This constitutes the first attempt to describe and understand the relationship between gene features and TEs on Testudines. Our findings hard to the understanding of interactions between gene features and TE and brings a broad understanding of the Testudine clade mobilome.

Introduction

Different environments have evolutionary shaped (via selection) the differential regulation of gene expression in different species. This process can be summarised as the cooperation and coordination of different genomic elements, and according to the proximity to their gene target, they can perform a Cis or Trans regulation. Among these regulators are promoters, enhancers, silencers, and insulators, that in general are non-coding sequences whose products control gene expression. Genomic regulatory elements are classified considering their activity and could be subclassified accordingly with the necessity to be in the same orientation as the target gene (Ali, Han, and Liang 2021; Conley, Piriyapongsa, and Jordan 2008). For example, promotors are orientation-dependant elements with respect to the genes that they regulate, providing a docking site for the transcriptional machinery. In contrast, enhancers and silencers are orientation and position-independent with respect to the target genes (Franchini et al. 2011; Conley, Piriyapongsa, and Jordan 2008).

Additionally, the modification of genomes resulting from TEs activity has been suggested to modulate evolution and to facilitate species adaptation to new environments (Clément Goubert et al. 2015; Clement Goubert et al. 2017), providing modifications on the regulation of genes, genes copy duplications (Krasileva 2019; Schrader and Schmitz 2019), and horizontal transference (Galbraith et al. 2022, 2021), among others. Many studies have shown that transposable elements (TEs) can contribute to all regulatory regions as enhancers, modifying the promoter, deactivating the promoter, among others (Franchini et al. 2011; Samuelson et al. 1990; Brini, Lee, and Kinet 1993; Hambor et al. 1993). The intrinsic characteristic of TE that allows them to transpose and code their own machinery, makes them good candidates for regulating gene expression. TEs possess individual promoters, enhancers/insulators, splice sites, and terminators. Their own internal regulation allows TEs to interfere with the regulation of the sites where they transpose. As an example, LTR and LINEs -highly abundant on Testudines (Carrasco-Valenzuela in prep.) - carry Polymerase (POL) II promoters, while SINEs carry promoters for either POL III or POL II (Swergold 1990; Roy et al. 2000). This interaction of TE with the regulatory machinery is not restricted only to promoters. For example, TEs like L1 can carry antisense sequences that also interfere with the expression of genes (Speek 2001). TEs have been reported to originate conserved enhancers in vertebrates' genomes (Bejerano et al. 2006; Franchini et al. 2011; J. Wang et al. 2014).

TEs are DNA sequences with the ability to change position within a genome. TEs can be divided into two major classes (Class I and Class II) based on their mechanism of transposition. They can also be divided into subclasses based on the mechanism of

chromosomal integration. Class I elements are retrotransposons that mobilise through a 'copy-and-paste' mechanism. For these elements, an RNA intermediate is reverse-transcribed into a cDNA copy before being integrated into the genome (Boeke et al. 1985; Bourque et al. 2018). Class II elements, also known as DNA transposons, mobilise via a DNA intermediate through a 'cut-and-paste' mechanism or, in the case of Helitrons, a 'peel-and-paste' replicative mechanism involving circular DNA intermediate (Grabundzija et al. 2016; Greenblatt and Alexander Brink 1963; Rubin, Kidwell, and Bingham 1982).

Additionally, there is substantial evidence that TEs insert non-randomly in host genomes. In maize, for example, *Activator* elements transpose more frequently into linked genomic regions (Cowperthwaite et al. 2002). Additionally, *Mutator* elements target unlinked open chromatin regions near recombination spots, which tend to be close to 5' end of genes (S. Liu et al. 2009). Moreover, *P* elements in *Drosophila* have been associated with replication origins also at the 5' end of genes (Spradling, Bellen, and Hoskins 2011). This phenomenon is not restricted to sequence base regulator regions, for example, *Ty3-Gypsy* LTR retroelements can bind specific methylations on H3 of the histones to transpose exclusively to the heterochromatin and is widely found from fungi to vertebrates (Malik and Eickbush 1999). Another example of integration directed to gene-poor regions is associated with *Ty5* LTR retrotransposon. Approximately 90% of *Ty5* LTR insertions in *S. cerevisiae* are within silent mating type loci or near silent heterochromatin at telomeres (Zou and Voytas 1997; Zou et al. 1996; Zou, Wright, and Voytas 1995).

Another example of how TEs can modulate gene expression is through their abundance of interfered genes. Several TEs are able to trap genes inside them and duplicate those genes during the transposition. Examples of this process can be found broadly in bacteria (Vogan et al. 2021; Urquhart et al. 2022) and vertebrates (Morgante et al. 2005; Thomas and Pritham 2015). *Helitrons*, for example, are elements from the DNA TE group able to capture genes at RNA and DNA levels and they were reported to capture complete genes or intronless genes in several organisms such as maize (Morgante et al. 2005; Yang and Bennetzen 2009), silkworms (Han et al. 2013), rice (Sweredoski, DeRose-Wilson, and Gaut 2008), and bats (Thomas et al. 2014).

The Testudine clade is considered a good model for the study of TE dynamics (Sotero-Caio et al. 2017). However, the effort to generate good quality genomes for this clade is incipient and there is little information on TE composition for this clade, limited to a few turtle species like the western painted turtle (Shaffer et al. 2013), the Chinese softshell turtle (Z. Wang et al. 2013), the Asian yellow pond turtle (X. Liu et al. 2022), the Common Snapping Turtle (Das et al. 2020), and for sea turtles (Z. Wang et al. 2013; Bentley, Carrasco-Valenzuela,

Ramos, Pawar, Souza Arantes, et al. 2023). Recent insights from sea turtles' genomes suggest that TE activity could be related to key modifications among turtle species (Bentley, Carrasco-Valenzuela, Ramos, Pawar, Souza, et al. 2023). Therefore, studying TE evolution in the turtle clade emerges as an important strategy to comprehend how TE proportions could influence turtle evolution and diversity, as well as bring insights into the evolution of TEs in Testudines, by including information for this poorly investigated clade. Because turtle genomes present long generation times and slower mutation rates compared to mammals and most reptilians, (Janes et al. 2010), this clade provides an interesting scenario to explore the diversification of mobilomes. A comparison of TE genomes composition in turtles could answer specific questions like how TEs relate with functional genomics regions contributing to a better understanding of TE evolution.

In this study, we explored the mobilome of ten species of turtles from six different superfamilies and identified remarkable similarities in the total mobilomes of these species. The only difference that was identified pertained to the abundance of certain TE orders. Furthermore, we collected all available turtle genomes with annotations from the NCBI and examined interactions between TEs and gene records.

Firstly, we examined the relationship between TEs and genes using the gff files, and secondly, we investigated the relationship with exons. We found that despite the high number of interactions between TEs and genes, there was no evidence suggesting that TEs in the studied turtles influenced gene expression. Instead, what we discovered was that the quality of the genome assembly played a crucial role in the analysis of TEs.

Methods

Genome accessions

Genomes and their raw sequencing data for *Pelodiscus sinensis, Chelydra serpentina, Chelonia mydas, Dermochelys coriacea, Terrapene carolina triunguis, Chrysemys picta bellii, Trachemys scripta elegans, Chelonoidis abingdonii, Gopherus evgoodei, and Mauremys reevesii* were retrieved from the National Center for Biotechnology Information database (NCBI: http://www.ncbi.nlm.nih.gov/), using the latest version available for each assembly (Bioproject Id at Supplementary Table 1). The Genome Evaluation Pipeline (https://git.imp.fu-berlin.de/cmazzoni/GEP) was run to assess the quality of the assemblies prior to further analysis. The statistics such as BUSCO (Seppey, Manni, and Zdobnov 2019), Sanger contig stats (*Assembly-Stats: Get Assembly Statistics from FASTA and FASTQ Files* n.d.), kmer analysis, mercury (Rhie et al. 2020) and N50 values for each assembly were recovered for each species (Supplementary Table 1).

Transposable element analysis

To recover TEs from ten Testudines genome assemblies, a de-novo TE library was generated for each genome using RepeatModeler2 (Flynn et al. 2020) with the -LTRStruct module. RepeatMasker was then run on each species' TE library(Tarailo-Graovac and Chen 2009; Smit, Hubley, and Green 2015). To calculate Kimura 2-parameter distance to consensus (K-value) with divCpGMod, the script calcDivergenceFromAlign.pl was utilised. Two in-house scripts were created to recover (*align_with_divHandeler.py*) and plot (*PlotTEstats.R*) the TEs statistics which can be found on GitHub at https://github.com/Tcvalenzuela/Testudine-wide-Transposable-element-exploration-A-history-of-slow-evolution-and-conserved-genomes.

Gene closeness analysis

To explore the interactions between the annotated genes and the TEs, we use the in-house script AreTEsonGenes.py available at GitHub at https://github.com/Tcvalenzuela/Testudine-wide-Transposable-element-exploration-A-history-of-slow-evolution-and-conserved-genomes. Script XXX compares the genome annotation file (.gff) from the NCBI against the TE annotation file (TE-gff), which is generated with *align_with_divHandeler.py.* To explore the interaction between each annotated gene feature against the TE-gff files, script xxx compares the start and end positions for each corresponding TE against the start and end positions of the genes. Then it suggests five possible cases also described in Figure 1, flagging the occurrences accordingly:

1. **TE from inside to Upstream of the gene:** the numerical value of the start position of the TE is lower than the numerical value of the start of the gene feature but the end of the TE is lower than the end of the Gene.
2. **TE from inside to Downstream of the gene**: the start of the TE is bigger than the start of the gene feature, the end of the TE is bigger than the end of the gene feature and the start of the TE is lower than the end of the gene. In this case, we flag it as**.**
3. **Gene inside TE:** That the start of the TE is lower than the start of the gene feature and the end of the TE is bigger than the end of the gene feature..
4. **TE inside gene**: That the start of the TE is bigger than the start of the gene feature and the end of the TE is lower than the end of the gene feature.
5. **Gene exactly on TE**: That the start and end of both, gene feature and TE, are the same numbers.

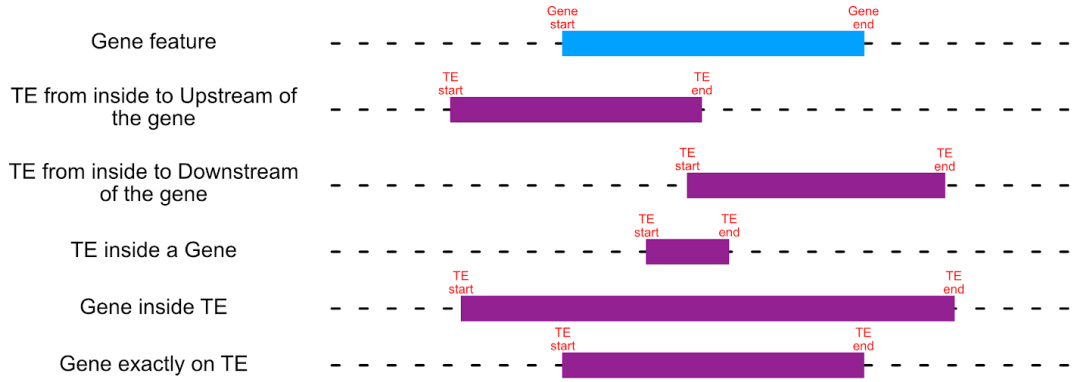


Figure 1. Diagram of the interactions between TEs and gene features showing five different categories of interaction between TEs and gene features: **TE from inside to Upstream of the gene**, **TE from inside to Downstream of the gene**, **TE inside the gene**, and **TE exactly on the gene.**

Mapping the reads of *C. Serpentina*

To validate the insertion of *Helitrons* on *C. Serpentina*. BWA2 (Vasimuddin et al. 2019), was used with default parameters to map the reads SRR10270344 against the genome GCA_018859375.1_ASM1885937v1_genomic.fna, both available at the NCBI. Subsequently, the assemble quality of this region and the positions of *Helitrons* insertions in relation to genes and exons were evaluated.

Results and Discussion

Genome quality

The turtle genomes analysed had a range of 26.04 QV (*Chelonoidis Abingdonii*) to 50.43 QV (*Trachemys scripta elegans*), with an average GC percentage of 44.06%, consistent with previous findings for this group (Bentley, Carrasco-Valenzuela, Ramos, Pawar, Souza, et al. 2023; Thomson, Spinks, and Shaffer 2021; Z. Wang et al. 2013). Out of the nine genomes analysed, four were assembled as scaffolds and five as chromosomes level. The quality of genome assemblies affects transposable element (TE) analysis, with fragmented assemblies producing a different TE profile than complete assemblies with similar QV (QV as a quality of genome assembly as Rhie (2020)). For example, the *C. serpentina* genome is more fragmented than the *C. mydas* genome, even though they have similar total sequence lengths and QV. This is evident in their assembly statistics, with *C. serpentina* having 55,422 scaffolds with 50% of the genome length contained in scaffolds equal or longer than 20,808,427bp, while *C. mydas* has only 92 total scaffolds, with 50% of the genomes length contained in scaffolds equal or longer than 134,428,053bp. It is widely recognized that more complete genome assemblies are necessary for a better detection of complete and active TEs (Peona et al. 2021; Prost et al. 2019). In particular, long-reads can play an essential role in identifying and assembling repetitive regions, as they often contain full repetitive element regions, allowing for more robust TE identification (Peona et al. 2021).

Table 1. Genome quality summary of the available assemblies on NCBI.

Species	QV	Assembly Level	Sequence Length	Contig N50	Scaffold N50	Scaffolds	GC content	Assembly version
Pelodiscus sinensis	34.07	Scaffold	2,202,466,388	21,993	3,350,749	19,904	44.41%	GCF_000230535.1_PelSin_1.0_genomic
Chelydra serpentina	49.02	Scaffold	2,064,902,118	84,752	20,808,247	55,422	44.30%	GCA_018859375.1_ASM1885937v1_genomic
Chelonia mydas	47.7	Chromosome	2,134,358,617	39,415,510	134,428,053	92	44.01%	GCF_015237465.2_rCheMyd1.pri.v2_genomic
Dermochelys coriacea	38.9	Chromosome	2,164,762,090	7,029,801	137,568,771	40	43.35%	GCF_009764565.3_rDerCor1.pri.v4_genomic
Terrapene carolina triunguis	48.07	Scaffold	2,140,043,261	80,749	28,728,777	25,686	44.16%	GCF_002925995.2_T_m_triunguis-2.0_genomic
Chrysemys picta bellii	49.65	Scaffold	2,481,351,664	21,318	6,605,846	78,631	44.19%	GCF_000241765.4_Chrysemys_picta_BioNano-3.0.4
Trachemys scripta elegans	50.43	Chromosome	2,126,182,493	204,575	140,411,086	138	43.81%	GCF_013100865.1_CAS_Tse_1.0_genomic
Chelonoidis abingdonii	26.04	Scaffold	2,300,739,315	73,186	1,277,207	10,618	43.71%	GCF_003597395.1_ASM359739v1_genomic
Gopherus evgoodei	38.31	Chromosome	2,298,547,364	13,026,736	147,425,149	382	44.14%	GCF_007399415.2_rGopEvg1_v1.p_genomic
Mauremys reevesii	33.55	Chromosome	2,035,064,793	34,524,243	139,244,951	62	44.55%	GCF_016161935.1_ASM1616193v1

Transposable elements content

TE profiles in the Testudines studied here correspond to an average of 42.72% of the genomes. The highest amount of TEs was found in *Gopherus evgoodei* at 45.33%, and the lowest was in *Chrysemys picta bellii* at 39.92% (Table 2). The Testudines clade has a very stable proportion of the genome with repetitive elements, which is expected given that turtles, after crocodilians, have the lowest heterozygosity levels among vertebrates (Green et al. 2014; Avise et al. 1992). As has been highly observed, LINEs are the most abundant TE family on Testudines (Sotero-Caio et al. 2017; Shaffer et al. 2013; Z. Wang et al. 2013).

Despite consistency in the total level of TEs, turtle families differ in the accumulation of insertions within the main TE orders and families. We compared the TE profile among the turtle species studied here and recovered different patterns for each main TE order.

The amount of Retrotransposon elements identified is directly proportional to the contiguity of the genome, with *D. coriacea* presenting the highest values while *P. sinensis* presented the lowest (Figure 2a). This dependence on genome contiguity was not identified for DNA TEs (Figure 2b). For example, highly contiguous genomes, like the sea turtles' ones, are among the species with the lowest amounts of DNA TEs. The lack of dependence of DNA TE proportions with the contiguity of the genome could suggest that this TE order is more variable among turtles, and the proportion of the insertions carry information about the natural biology or evolution of different turtle families. However, DNA TEs are more infrequent than Retrotransposon TEs in turtles' genomes, which could also contribute to the lack of a pattern found. Therefore, the fact that both sea turtles included in this study have a reduced proportion of class II TEs deserves further investigation. The total proportion of TEs also presents a pattern dependent on genome quality, showing a decrease in TE elements for genomes with lower contiguity (Figure 2c), except for the *P. sinensis* genome. Nevertheless, this turtle has the highest amount of Unclassified TE (Table 2), which could mean that the a very low contiguity genome results in a misidentification of TEs in general.

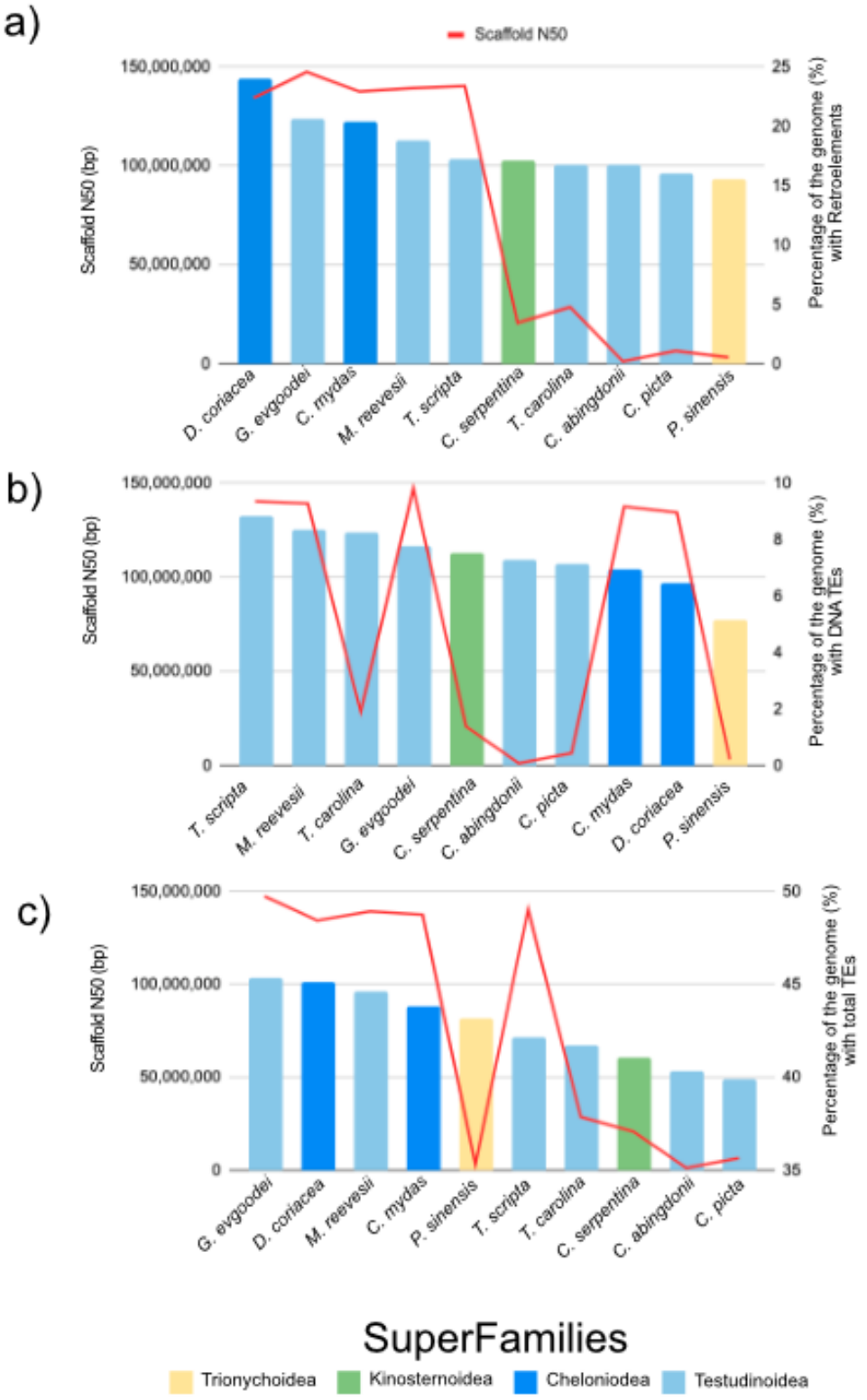


Figure 2. Relationship between proportion of the genome identified as TEs and quality of the assemblies. Shown in a), is the relationship between the contiguity of the assembly as Scaffold N50 and the proportion of the genome identified as Retroelements. In b), is the relationship of the contiguity and the proportion of the genome identified as DNA TEs. And in c), it is shown the relationship between the total identified TEs and the contiguity of the genome.

Table 2. The proportion of TEs families for the 10 genomes of the Testudine clade.

TE clasifications	*P. sinensis*	*C. serpentina*	*D. coriacea*	*C. mydas*	*T. carolina*	*C. picta*	*T. scripta*	*C.abingdonii*	*G. evgoodei*	*M. reevesii*
Retroelements	15.53%	17.07%	24.06%	20.43%	16.79%	16.04%	17.28%	16.72%	20.67%	18.77%
SINEs	0.89%	2.06%	2.5%	2.73%	2.3%	2.45%	2.49%	2.14%	2.39%	2.66%
Penelope	1.59%	1.3%	4.7%	2.59%	1.08%	0.99%	1.33%	1.05%	1.28%	1.13%
LINEs	12.68%	11.99%	18.23%	14.34%	11.31%	10.51%	11.83%	10.87%	12.52%	12.21%
LTR	1.96%	3.02%	3.33%	3.37%	3.19%	3.08%	2.95%	3.701%	5.76%	3.9%
Gypsy/DIRS1	1.92%	2.84%	2.81%	2.95%	2.78%	2.65%	2.72%	3.34%	4.93%	3.33%
Retroviral	0.03%	0.13%	0.45%	0.33%	0.35%	0.32%	0.18%	0.3%	0.78%	0.51%
DNA	5.18%	7.52%	6.46%	6.95%	8.27%	7.12%	8.84%	7.27%	7.77%	8.37%
hobo-Activator	2.36%	3.59	3.08%	3.32%	3.31%	3.39%	3.8%	3.27%	3.65%	3.78%
Tc1-IS630-Pogo	0.31%	0.13%	0.2%	0.17%	0.36%	0.35%	0.57%	0.31%	0.26%	0.32%
Tourist/Harbinger	2.22%	3.09%	2.43%	2.7%	3.66%	2.56%	3.13%	3.16%	3.05%	3.46%
Rolling-circles	0.09%	0.04%	0.1%	0.02%	0.08%	0.01%	0.09%	0.12%	0.06%	0.04%
Unclassified	22.44%	16.46%	14.6%	16.41%	16.66%	16.77%	16.05%	16.33%	16.89%	17.5%
Total	43.14%	41.05%	45.12%	43.79%	41.73%	39.92%	42.16%	40.32%	45.33%	44.64%
Small RNA	0.14%	0.54%	0.35%	0.43%	0.37%	0.4%	0.38%	0.36%	0.45%	0.046%
Simple repeats	0.38%	0.39%	0.44%	0.4%	0.4%	0.4%	0.41%	0.34%	0.37%	0.45%
Low complexity	0.06%	0.08%	0.08%	0.08%	0.07%	0.07%	0.07%	0.07%	0.07%	0.09%

Interaction between TEs and gene features

TEs x entire Genes

In regards to the way TEs interact with functional regions in the genome, we have classified them into three main categories: genes that are within TEs, TEs that are inserted in the border regions of gene features (with upstream and downstream behaving similarly), and TEs that are inside genes. For turtles, the last category has the most interactions. We have observed 1,183,543 interactions where TEs are found inside genes, with an average of 45.47% of genes containing TEs. The average K-value for TEs inside genes is 20.35% of divergence and does not vary significantly among different turtle families (as shown in Figure 3). This value is quite similar to the average K-value for all TEs, which is 20.36% of divergence (for further details, please refer to Supplementary Table 2 for mean K-value and Supplementary Table 3 for Average K-value on TEs inside genes).

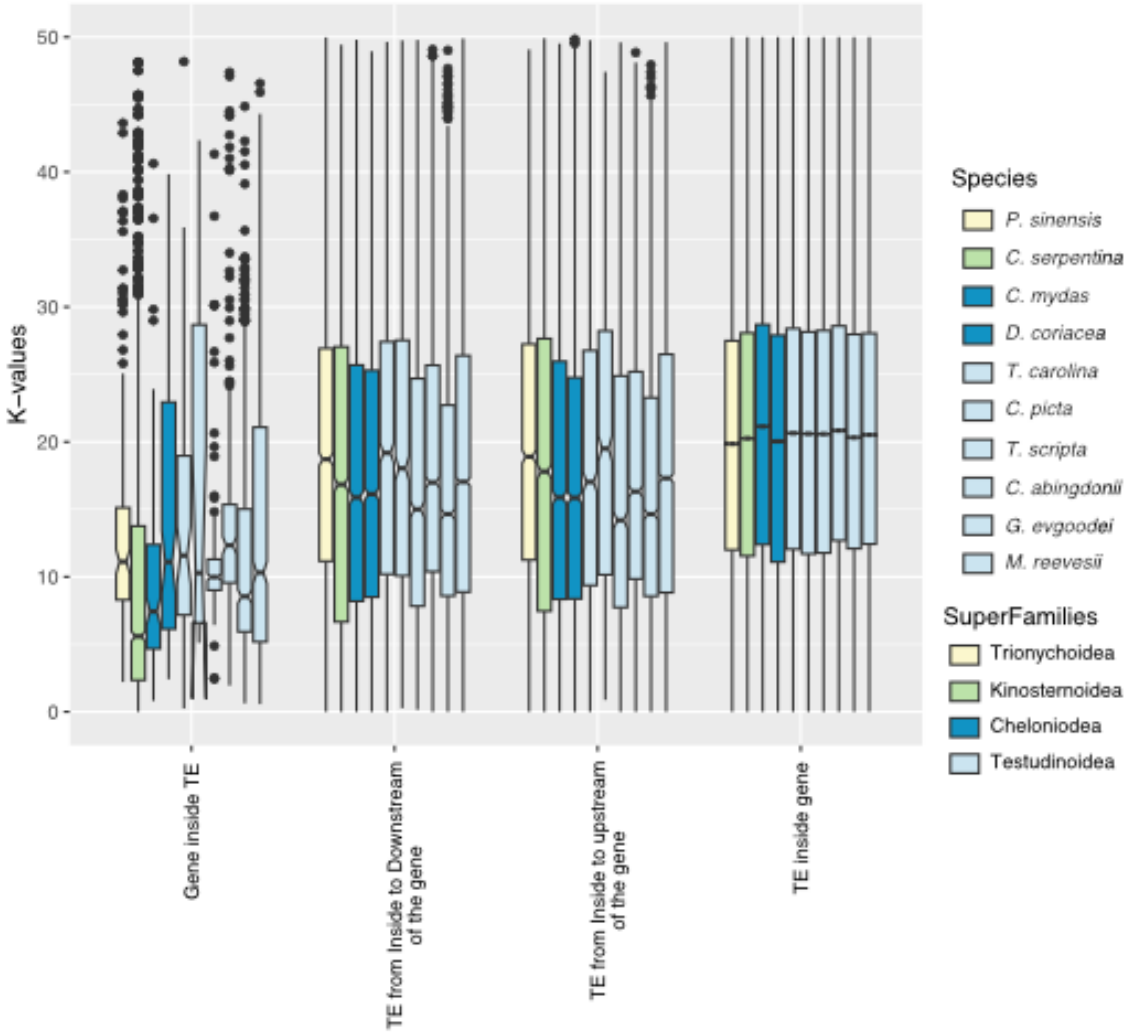


Figure 3. Kimura distance-based copy divergence analysis of transposable elements among Testudines families showing the distribution of K-values for each TEs interaction with gene features category. Turtle species appear in the order shown in the legend.

Table 3. Number of interactions of TE with genes

Type of interaction TE-gene features	***P. sinensis***	***C. serpentina***	***D. coriacea***	***C. mydas***	***T. carolina***	***C. picta***	***T. scripta***	***C. abingdonii***	***G. evgoodei***	***M. reevesii***
Genes inside TEs	142	1,085	132	154	75	14	205	201	792	926
TEs from Inside to Downstream of the gene	2,763	1,988	3,570	3,071	1,485	934	1,935	3,114	3,242	8,395
TEs from Inside to upstreamof the gene	2,660	1,885	3,658	3,197	1,535	909	1,925	3,152	3,329	8,208
TEs inside genes	1,127,477	959,917	1,517,247	1,503,182	1,143,426	228,133	1,267,553	1,170,065	1,338,756	1,579,669
Total genes	24,494	21,843	26,849	27,777	24,225	26,247	21,877	24,661	25,154	39,872
Protein-coding genes (RefSeq NCBI annotation)	19,518	20,909	18,775	19.752	22,254	21,498	18,662	19,976	20,174	22,618
Exon inside TEs	5,654	2,827	9,402	10,301	2,190	21,416	3,703	7,236	8,524	136,197
TE from Inside to downstream of the exon	7,948	2,586	12,764	14,112	3,222	19,194	5,927	7,467	10,229	102,984
TEs from Inside to upstream of the exon	7,810	2,415	12,386	14,424	3,142	19,069	5,967	7,536	10,336	102,019
TEs inside exon	28,647	6,288	65,219	117,445	13,471	27,964	30,865	28,947	41,922	129,488
Total exons	491,458	180,210	931,142	1,083,441	390,573	824,648	585,747	601,600	721,489	1,009,528

For the category of TEs interaction with the borders of genes, we have identified a similar number of upstream and downstream interactions in all the genomes analysed, regardless of their independent insertions events, as shown in Table 3. TEs are rarely randomly distributed in the genome and have the ability to target different sections of the genes (upstream and downstream) (Bourque et al. 2018). However, we did not recover more insertions in the upstream region of the gene feature when compared to the downstream regions. This pattern could suggest that TEs in turtles are not significantly targeting immediately close promoter regions of genes.

The category of "gene inside TE" varies the most among Testudines families (Figure 3). We observed that this category has the least total number of interactions and shows significantly lower k-values for *C. serpentina*'s compared to other turtles included in this analysis. The turtle superfamily Testudinoidea has a relatively similar average number of K-value for genes inside TEs. On Cheloniodea, *C. mydas* present a lower average of k-values for gene inside TEs than *D. coriacea*. This is interesting because it has been reported that *D. coriacea* presents a higher accumulation of TEs with lower k-values for the *Penelope-like* elements and LINEs in general (Carrasco-Valenzuela in prep).This results therefore implies that the few young TEs present in *C. mydas* are in average more prone to carry genes inside.

TEs x Exons

We were also interested in TE insertions occurring inside the coding regions of the genes (TEs x Exons). We recovered similar distributions of K-values across the 4 categories of interactions between TEs and exons analysed. For the category of TE interacting with downstream and upstream regions of an exon, we observed a similar pattern observed for TE interactions with the entire gene described above. The number of interactions although not exactly the same numbers are remarkably similar. For all the species analysed, the category TE inside exon is the one with more interactions (Table 3). *Mauremys reevesii*, presents substantially more exon-TE interaction than any other turtle. The number of exons in the *M. reevesii* annotation is similar to *D. coriacea*. However, *M. reevesii* present about 10 times more TEs interaction inside the exons than *D. coriacea*. Although presenting more annotated genes, a great amount of them are considered non-coding genes (see *M. reevesii* annotation report on NCBI (GCF_016161935.1)). The increased amount of TEs inside exons for this species could explain the highest amount of non-coding genes for *M. reevesii*, and the presence of TEs inside these non-coding genes could be further investigated. *M. reevesii* assemble is among the high-quality genomes for turtles, with a QV of 33.55, a total assembly with 62 scaffolds, and a

scaffold N50 of 139,244,951 bp, which does not suggest any particular indication of artefactual mistakes that could lead to these differences. After *M. reevesii*, sea turtles are the species that have more insertions of TEs inside exons, with *C. mydas* presenting almost double the amount of TEs inside exons compared to *D. coriacea*. Previous studies report that main differences between sea turtle genomes rely on multicopy gene families (Bentley, Carrasco-Valenzuela, Ramos, Pawar, Souza Arantes, et al. 2023). The functions of the genes with exons containing TEs should be investigated to understand if TE activity is related to those multicopy gene families in sea turtles.

For the category "exon inside TE", the average K-value for the TEs in *C. serpentina* is significantly lower than all the other species (Figure 4). We identified the presence of TEs from the family *Helitron* as the most frequent within this category, indicating that *C. serpentina* has significant younger insertions of Helitrons (with an average K-value of 0) within exons when compared to all the testudines analysed (Figure 5). Looking inside this group, we identified *Helitrons* carrying tRNA-Asp. The *Helitron* insertions in *C. serpentina* therefore possess signals of recent insertions while carrying tRNAs. For example, the *Helitron* on scaffold JAHGAV010000047.1 of *C. serpentina* starts at position 9,890,866 and ends at position 9,914,925 (24,059 bp). This insertion has a 4.19% K-value and presents 15 exons from a tRNA-Asp. The latter exons spam from positions 9,891,242 to 9,911,752 bp (Supplementary Table 3). Interestingly, we identify another 3 *Helitrons* from the same family at positions JAHGAV010000366.1 844,908-847,049; JAHGAV010000366.1 846,998-848,627 JAHGAV010000366.1 848,660-856,419 (Supplementary Table 3). These *Helitrons* are also carrying tRNA exons, nonetheless, they carry only a single exon each. Additionally, their lower K-value and their proximity, suggest that either the same *Helitron* was wrongly annotated as three independent insertions or that there are more recent attempts of transposition of the longest insertion (carrying 15 exons inside).

We further explored the Helitron found on scaffold JAHGAV010000047 from *C. serpentina* and mentioned above (Supplementary Table 4). To validate the insertion, we mapped back the reads to the reference genome. The average read coverage of mapping was significantly different than the surrounding regions of the insertion, with a peak of 2,000 mapping reads at the beginning of the *Helitron*. Also, we identified a breaking point of the assembly just before the peak in coverage. This suggests a contraction in the assembly at the beginning of the *Helitron*, reducing the quality of the region and bringing uncertainty to the veracity of the *Helitron* insertion. This supports the necessity of using high-quality genomes to study TE dynamics and evolution patterns, since badly assembled regions could mask the real size of the insertion, or even compromise the detection of the particular recognition domains of TEs.

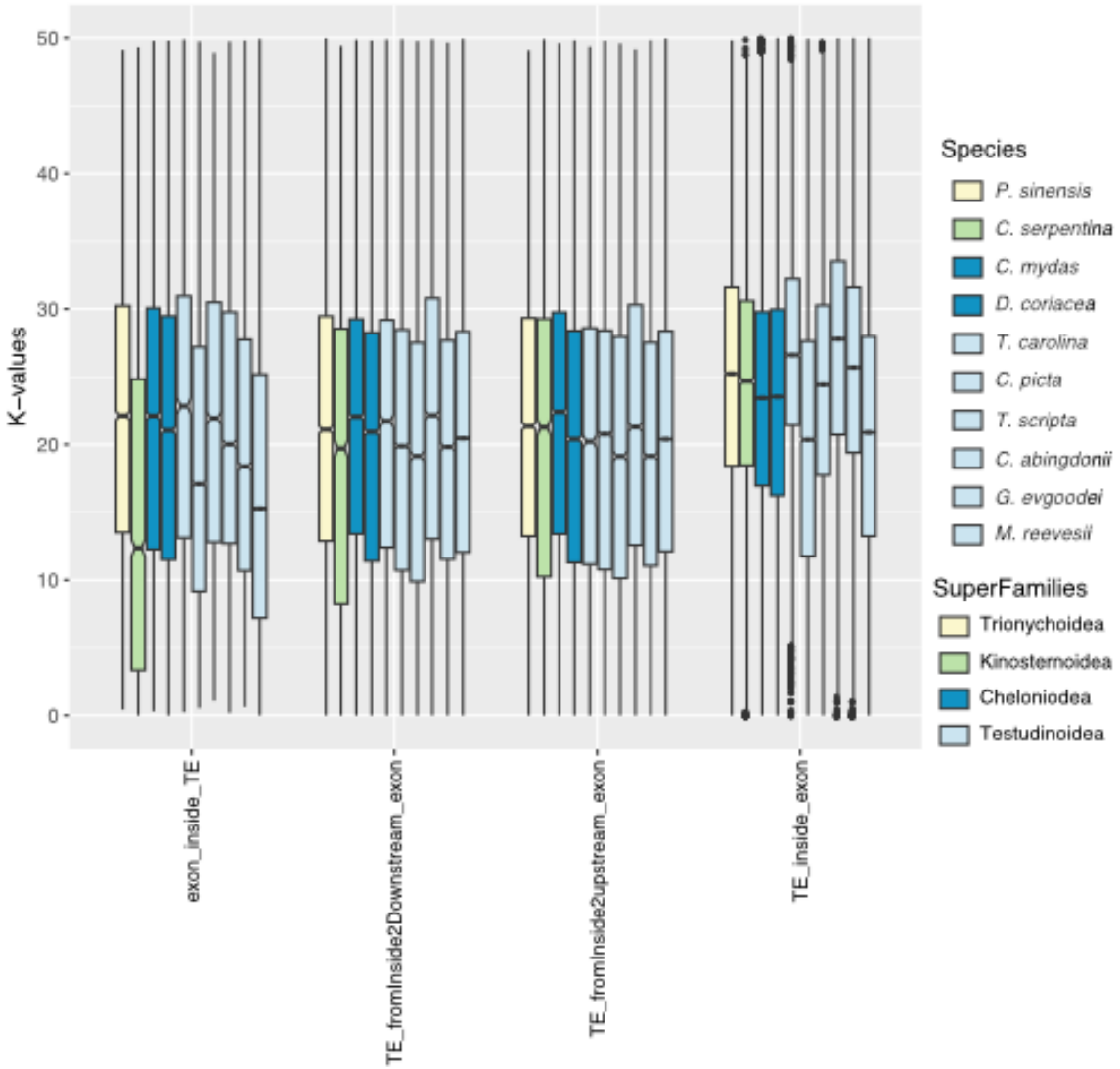


Figure 4. Kimura distance-based copy divergence analysis of transposable elements among Testudines families showing the distribution of K-values for each TEs interaction with exon features category. Turtle species appear in the order shown in the legend.

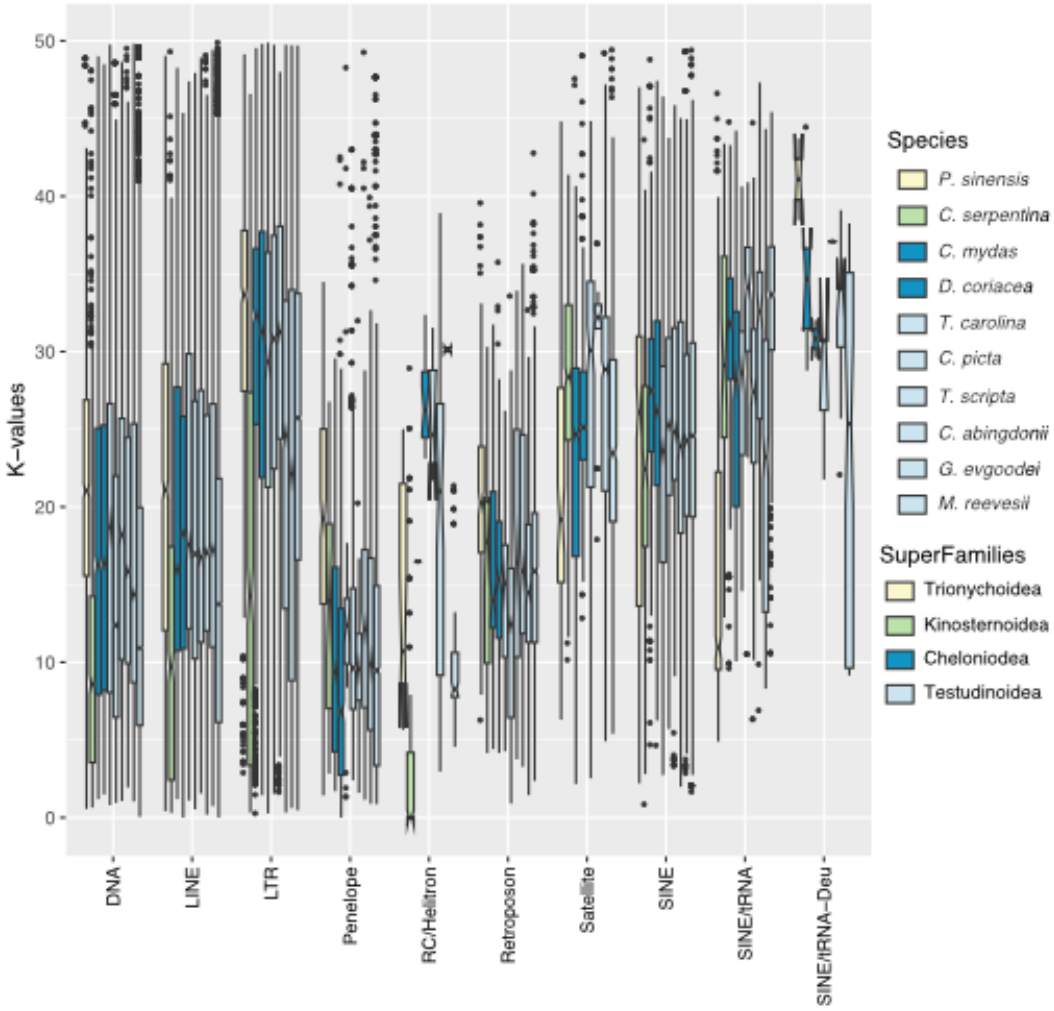


Figure 5. Interaction between exons and TEs at the category "exon inside TE" split by TE main Orders for 10 species of turtles from 4 superfamilies of the testudine clade.

In general we found substantially more TE insertions outside exons when compared to other regulatory or non-coding parts of the gene (introns). This pattern was expected, since TE-gene relationship results in an alteration in the structure and function of key genes that could be detrimental to the fitness of the individual. If the alteration is lethal, the individual will cease to exist, and therefore this transposition will not pass to the next generations (Schrader and Schmitz 2019; Mackay 1986) being purged from the population in the process of s natural selection.

Hopefully, future studies will be able to confidently identify families of TEs more associated with insertion inside genes or exons, advancing further our understanding towards TE evolution.

Conclusions

Here we show the most comprehensive comparison of TE content in turtles. Testudines, in general, have the same proportion of TEs across different turtle superfamilies, with the main differences lying inside the DNA TE order. Retrotransposon TEs proportions for this group are highly affected by the quality of the genome assemblies. As mentioned before, the analysis of TEs relies on the quality of the assemblies, and lack of contiguity and low certainty on base calls make biological conclusions out of the analysis very difficult to validate. Also, this is to our understanding the first Class-wide exploration of the relationships between TEs and genomics features. We reported that TEs transpose equally frequently in upstream and downstream regions of genes and exons. Nonetheless, to properly assess any biological meaning of a TE analysis, it is vital to have high-quality genomes. We identified no evidence of a particular family or order of TE that is more active or significantly younger and is also interacting with genes. We demonstrated that the TE proportion of the genomes for certain Orders correlates more with the quality of the genomes rather than any biological relationship of the species.

Acknowledgments

We acknowledge CONICYT-DAAD for scholarship support to T.C.-V., the São Paulo Research Foundation – FAPESP (grant #2020/10372-6) to E.K.S.R. BeGenDiv is partially funded by the German Federal Ministry of Education and Research (BMBF, Förderkennzeichen 033W034A).

References

Ali, Arsala, Kyudong Han, and Ping Liang. 2021. “Role of Transposable Elements in Gene Regulation in the Human Genome.” *Life* 11 (2). https://doi.org/10.3390/life11020118.

Assembly-Stats: Get Assembly Statistics from FASTA and FASTQ Files. n.d. Github. Accessed October 24, 2022. https://github.com/sanger-pathogens/assembly-stats.

Avise, J. C., B. W. Bowen, T. Lamb, A. B. Meylan, and E. Bermingham. 1992. “Mitochondrial DNA Evolution at a Turtle’s Pace: Evidence for Low Genetic Variability and Reduced Microevolutionary Rate in the Testudines.” *Molecular Biology and Evolution* 9 (3): 457–73.

Bejerano, Gill, Craig B. Lowe, Nadav Ahituv, Bryan King, Adam Siepel, Sofie R. Salama, Edward M. Rubin, W. James Kent, and David Haussler. 2006. “A Distal Enhancer and an Ultraconserved Exon Are Derived from a Novel Retroposon.” *Nature* 441 (7089): 87–90.

Bentley, Blair P., Tomás Carrasco-Valenzuela, Elisa K. S. Ramos, Harvinder Pawar, Larissa Souza Arantes, Alana Alexander, Shreya M. Banerjee, et al. 2023. “Divergent Sensory and Immune Gene Evolution in Sea Turtles with Contrasting Demographic and Life Histories.” *Proceedings of the National Academy of Sciences of the United States of America* 120 (7): e2201076120.

Bentley, Blair P., Tomás Carrasco-Valenzuela, Elisa K. S. Ramos, Harvinder Pawar, Larissa Souza, Alana Alexander, Shreya M. Banerjee, et al. 2023. “1 Divergent Sensory and Immune Gene Evolution in Sea Turtles with Contrasting Demographic and Life 2 Histories.” *The Proceedings of the National Academy of Sciences (PNAS)*.

Boeke, J. D., D. J. Garfinkel, C. A. Styles, and G. R. Fink. 1985. “Ty Elements Transpose through an RNA Intermediate.” *Cell* 40 (3): 491–500.

Bourque, Guillaume, Kathleen H. Burns, Mary Gehring, Vera Gorbunova, Andrei Seluanov, Molly Hammell, Michaël Imbeault, et al. 2018. “Ten Things You Should Know about Transposable Elements.” *Genome Biology* 19 (1): 1–12.

Brini, A. T., G. M. Lee, and J. P. Kinet. 1993. “Involvement of Alu Sequences in the Cell-Specific Regulation of Transcription of the Gamma Chain of Fc and T Cell Receptors.” *The Journal of Biological Chemistry* 268 (2): 1355–61.

Conley, Andrew B., Jittima Piriyapongsa, and I. King Jordan. 2008. “Retroviral Promoters in the Human Genome.” *Bioinformatics* 24 (14): 1563–67.

Cowperthwaite, Matthew, Wonkeun Park, Zhennan Xu, Xianghe Yan, Steven C. Maurais, and Hugo K. Dooner. 2002. “Use of the Transposon Ac as a Gene-Searching Engine in the Maize Genome.” *The Plant Cell* 14 (3): 713–26.

Das, Debojyoti, Sunil Kumar Singh, Jacob Bierstedt, Alyssa Erickson, Gina L. J. Galli, Dane A. Crossley 2nd, and Turk Rhen. 2020. “Draft Genome of the Common Snapping Turtle, Chelydra Serpentina, a Model for Phenotypic Plasticity in Reptiles.” *G3* 10 (12): 4299–4314.

Franchini, Lucía F., Rodrigo López-Leal, Sofía Nasif, Paula Beati, Diego M. Gelman, Malcolm J. Low, Flávio J. S. de Souza, and Marcelo Rubinstein. 2011. “Convergent Evolution of Two Mammalian Neuronal Enhancers by Sequential Exaptation of Unrelated Retroposons.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (37): 15270–75.

Galbraith, James D., Alastair J. Ludington, Kate L. Sanders, Timothy G. Amos, Vicki A. Thomson, Daniel Enosi Tuipulotu, Nathan Dunstan, Richard J. Edwards, Alexander Suh, and David L. Adelson. 2022. “Horizontal Transposon Transfer and Its Implications for the Ancestral Ecology of Hydrophiine Snakes.” *Genes* 13 (2). https://doi.org/10.3390/genes13020217.

Galbraith, James D., Alastair J. Ludington, Kate L. Sanders, Alexander Suh, and David L. Adelson. 2021. “Horizontal Transfer and Subsequent Explosive Expansion of a DNA Transposon in Sea Kraits (Laticauda).” *Biology Letters* 17 (9): 20210342.

Goubert, Clement, Helene Henri, Guillaume Minard, Claire Valiente Moro, Patrick Mavingui, Cristina Vieira, and Matthieu Boulesteix. 2017. “High-Throughput Sequencing of Transposable Element Insertions Suggests Adaptive Evolution of the Invasive Asian Tiger Mosquito towards Temperate Environments.” *Molecular Ecology* 26 (15): 3968–81.

Goubert, Clément, Laurent Modolo, Cristina Vieira, Claire ValienteMoro, Patrick Mavingui, and Matthieu Boulesteix. 2015. “De Novo Assembly and Annotation of the Asian Tiger Mosquito (Aedes Albopictus) Repeatome with dnaPipeTE from Raw Genomic Reads and Comparative Analysis with the Yellow Fever Mosquito (Aedes Aegypti).” *Genome Biology and Evolution* 7 (4): 1192–1205.

Grabundzija, Ivana, Simon A. Messing, Jainy Thomas, Rachel L. Cosby, Ilija Bilic, Csaba Miskey, Andreas Gogol-Döring, et al. 2016. “A Helitron Transposon Reconstructed from Bats Reveals a Novel Mechanism of Genome Shuffling in Eukaryotes.” *Nature Communications* 7 (March): 10716.

Greenblatt, Irwin M., and R. Alexander Brink. 1963. “Transpositions of Modulator in Maize into Divided and Undivided Chromosome Segments.” *Nature* 197 (4865): 412–13.

Green, Richard E., Edward L. Braun, Joel Armstrong, Dent Earl, Ngan Nguyen, Glenn Hickey, Michael W. Vandewege, et al. 2014. “Three Crocodilian Genomes Reveal Ancestral Patterns of Evolution among Archosaurs.” *Science* 346 (6215): 1254449.

Hambor, J. E., J. Mennone, M. E. Coon, J. H. Hanke, and P. Kavathas. 1993. “Identification and Characterization of an Alu-Containing, T-Cell-Specific Enhancer Located in the Last Intron of the Human CD8 Alpha Gene.” *Molecular and Cellular Biology* 13 (11): 7056–70.

Han, Min-Jin, Yi-Hong Shen, Meng-Shu Xu, Hong-Yu Liang, Hua-Hao Zhang, and Ze Zhang. 2013. “Identification and Evolution of the Silkworm Helitrons and Their Contribution to Transcripts.” *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes* 20 (5): 471–84.

Janes, Daniel E., Christopher L. Organ, Matthew K. Fujita, Andrew M. Shedlock, and Scott V. Edwards. 2010. “Genome Evolution in Reptilia, the Sister Group of Mammals.” *Annual Review of Genomics and Human Genetics* 11: 239–64.

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

Liu, Sanzhen, Cheng-Ting Yeh, Tieming Ji, Kai Ying, Haiyan Wu, Ho Man Tang, Yan Fu, Daniel Nettleton, and Patrick S. Schnable. 2009. “Mu Transposon Insertion Sites and Meiotic Recombination Events Co-Localize with Epigenetic Marks for Open Chromatin across the Maize Genome.” *PLoS Genetics* 5 (11): e1000733.

Liu, Xiaoli, Yakun Wang, Ju Yuan, Fang Liu, Xiaoyou Hong, Lingyun Yu, Chen Chen, et al. 2022. “Chromosome-Level Genome Assembly of Asian Yellow Pond Turtle (Mauremys Mutica) with Temperature-Dependent Sex Determination System.” *Scientific Reports* 12 (1): 7905.

Mackay, Trudy F. C. 1986. “Transposable Element-Induced Fitness Mutations in Drosophila Melanogaster.” *Genetics Research* 48 (2): 77–87.

Malik, H. S., and T. H. Eickbush. 1999. “Modular Evolution of the Integrase Domain in the Ty3/Gypsy Class of LTR Retrotransposons.” *Journal of Virology* 73 (6): 5186–90.

Morgante, Michele, Stephan Brunner, Giorgio Pea, Kevin Fengler, Andrea Zuccolo, and Antoni Rafalski. 2005. “Gene Duplication and Exon Shuffling by Helitron-like Transposons Generate Intraspecies Diversity in Maize.” *Nature Genetics* 37 (9): 997–1002.

Peona, Valentina, Mozes P. K. Blom, Luohao Xu, Reto Burri, Shawn Sullivan, Ignas Bunikis, Ivan Liachko, et al. 2021. “Identifying the Causes and Consequences of Assembly Gaps Using a Multiplatform Genome Assembly of a Bird-of-Paradise.” *Molecular Ecology Resources* 21 (1): 263–86.

Prost, Stefan, Ellie E. Armstrong, Johan Nylander, Gregg W. C. Thomas, Alexander Suh, Bent Petersen, Love Dalen, et al. 2019. “Comparative Analyses Identify Genomic Features Potentially Involved in the Evolution of Birds-of-Paradise.” *GigaScience* 8 (5). https://doi.org/10.1093/gigascience/giz003.

Rhie, Arang, Brian P. Walenz, Sergey Koren, and Adam M. Phillippy. 2020. “Merqury: Reference-Free Quality, Completeness, and Phasing Assessment for Genome Assemblies.” *Genome Biology* 21 (1): 245.

Roy, A. M., N. C. West, A. Rao, P. Adhikari, C. Alemán, A. P. Barnes, and P. L. Deininger. 2000. “Upstream Flanking Sequences and Transcription of SINEs.” *Journal of Molecular Biology* 302 (1): 17–25.

Rubin, G. M., M. G. Kidwell, and P. M. Bingham. 1982. “The Molecular Basis of P-M Hybrid Dysgenesis: The Nature of Induced Mutations.” *Cell* 29 (3): 987–94.

Samuelson, L. C., K. Wiebauer, C. M. Snow, and M. H. Meisler. 1990. “Retroviral and Pseudogene Insertion Sites Reveal the Lineage of Human Salivary and Pancreatic Amylase Genes from a Single Gene during Primate Evolution.” *Molecular and Cellular Biology* 10 (6): 2513–20.

Schrader, Lukas, and Jürgen Schmitz. 2019. “The Impact of Transposable Elements in Adaptive Evolution.” *Molecular Ecology* 28 (6): 1537–49.

Seppey, Mathieu, Mosè Manni, and Evgeny M. Zdobnov. 2019. “BUSCO: Assessing Genome Assembly and Annotation Completeness.” *Methods in Molecular Biology* 1962: 227–45.

Shaffer, H. Bradley, Patrick Minx, Daniel E. Warren, Andrew M. Shedlock, Robert C. Thomson, Nicole Valenzuela, John Abramyan, et al. 2013. “The Western Painted Turtle Genome, a Model for the Evolution of Extreme Physiological Adaptations in a Slowly Evolving Lineage.” *Genome Biology* 14 (3): R28.

Sotero-Caio, Cibele G., Roy N. Platt 2nd, Alexander Suh, and David A. Ray. 2017. “Evolution and Diversity of Transposable Elements in Vertebrate Genomes.” *Genome Biology and Evolution* 9 (1): 161–77.

Speek, M. 2001. “Antisense Promoter of Human L1 Retrotransposon Drives Transcription of Adjacent Cellular Genes.” *Molecular and Cellular Biology* 21 (6): 1973–85.

Spradling, Allan C., Hugo J. Bellen, and Roger A. Hoskins. 2011. “Drosophila P Elements Preferentially Transpose to Replication Origins.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (38): 15948–53.

Sweredoski, Michael, Leah DeRose-Wilson, and Brandon S. Gaut. 2008. “A Comparative Computational Analysis of Nonautonomous Helitron Elements between Maize and Rice.” *BMC Genomics* 9 (October): 467.

Swergold, G. D. 1990. “Identification, Characterization, and Cell Specificity of a Human LINE-1 Promoter.” *Molecular and Cellular Biology* 10 (12): 6718–29.

Thomas, Jainy, Caleb D. Phillips, Robert J. Baker, and Ellen J. Pritham. 2014. “Rolling-Circle Transposons Catalyze Genomic Innovation in a Mammalian Lineage.” *Genome Biology and Evolution* 6 (10): 2595–2610.

Thomas, Jainy, and Ellen J. Pritham. 2015. “Helitrons, the Eukaryotic Rolling-Circle Transposable Elements.” *Microbiology Spectrum* 3 (4). https://doi.org/10.1128/microbiolspec.MDNA3-0049-2014.

Thomson, Robert C., Phillip Q. Spinks, and H. Bradley Shaffer. 2021. “A Global Phylogeny of Turtles Reveals a Burst of Climate-Associated Diversification on Continental Margins.” *Proceedings of the National Academy of Sciences of the United States of America* 118 (7). https://doi.org/10.1073/pnas.2012215118.

Urquhart, A. S., A. A. Vogan, D. M. Gardiner, and A. Idnurm. 2022. “Starships Are Active Eukaryotic Transposable Elements Mobilized by a New Family of Tyrosine Recombinases.” *Biorxiv*. https://www.biorxiv.org/content/10.1101/2022.08.04.502770.abstract.

Vasimuddin, Md, Sanchit Misra, Heng Li, and Srinivas Aluru. 2019. “Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems.” In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 314–24. ieeexplore.ieee.org.

Vogan, Aaron A., S. Lorena Ament-Velásquez, Eric Bastiaans, Ola Wallerman, Sven J. Saupe, Alexander Suh, and Hanna Johannesson. 2021. “The Enterprise, a Massive Transposon Carrying Spok Meiotic Drive Genes.” *Genome Research* 31 (5): 789–98.

Wang, Jichang, Gangcai Xie, Manvendra Singh, Avazeh T. Ghanbarian, Tamás Raskó, Attila Szvetnik, Huiqiang Cai, et al. 2014. “Primate-Specific Endogenous Retrovirus-Driven Transcription Defines Naive-like Stem Cells.” *Nature* 516 (7531): 405–9.

Wang, Zhuo, Juan Pascual-Anaya, Amonida Zadissa, Wenqi Li, Yoshihito Niimura, Zhiyong Huang, Chunyi Li, et al. 2013. “The Draft Genomes of Soft-Shell Turtle and Green Sea Turtle Yield Insights into the Development and Evolution of the Turtle-Specific Body Plan.” *Nature Genetics* 45 (6): 701–6.

Yang, Lixing, and Jeffrey L. Bennetzen. 2009. “Distribution, Diversity, Evolution, and Survival of Helitrons in the Maize Genome.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (47): 19922–27.

Zou, S., N. Ke, J. M. Kim, and D. F. Voytas. 1996. “The Saccharomyces Retrotransposon Ty5 Integrates Preferentially into Regions of Silent Chromatin at the Telomeres and Mating Loci.” *Genes & Development* 10 (5): 634–45.

Zou, S., and D. F. Voytas. 1997. “Silent Chromatin Determines Target Preference of the Saccharomyces Retrotransposon Ty5.” *Proceedings of the National Academy of Sciences of the United States of America* 94 (14): 7412–16.

Zou, S., D. A. Wright, and D. F. Voytas. 1995. “The Saccharomyces Ty5 Retrotransposon Family Is Associated with Origins of DNA Replication at the Telomeres and the Silent Mating Locus HMR.” *Proceedings of the National Academy of Sciences of the United States of America* 92 (3): 920–24.

Supplementary Material

Supplementary Table 1 List of all the genomes used on the TE analysis and their respective Genebank ID and Stats.

Species	Bioproject	Assembly Level	Assembly Stats Total Sequence Length	Contig N50	Scaffold N50	Scaffold #	GC%	QV	BUSCO C	BUSCO D	BUSCO F	BUSCO M	BUSCO SC	BUSCO TCt	Annotation BUSCO database
Pelodiscus sinensis	PRJNA221645	Scaffold	2,202,466,388	21,993	3,350,749	19,904	44.41	34.07	96.50%	1.06%	1.51%	1.99%	95.44%	7,480	sauropsida_odb10
Chelydra serpentina	PRJNA574487	Scaffold	2,401,360,239	80,593	21,135,443	113,431	44.00	48.82	95.00%	1.00%	3.60%	1.40%	94.00%	3,354	vertebrata_odb10
Chelonia Mydas	PRJNA675851	Chromosome	2,134,358,617	39,415,510	134,428,053	92	44.01	47.70	99.01%	1.12%	0.31%	0.68%	97.89%	7,480	sauropsida_odb10
Dermochelys coriacea	PRJNA655518	Chromosome	2,164,762,090	7,029,801	137,568,771	40	43.35	38.90	98.21%	1.03%	0.52%	1.27%	97.18%	7,480	sauropsida_odb10
Terrapene carolina triunguis	PRJNA415469	Scaffold	2,571,267,249	76,614	24,249,581	131,541,780	44.27	47.63	91.30%	9.80%	4.40%	4.30%	81.5%	3,354	vertebrata_odb10
Chrysemys picta belii	PRJNA210179	Chromosome	2,481,351,664	21,318	6,605,846	78,631	44.19	49.65	98.20%	0.78%	1.02%	0.79%	97.42%	7,480	sauropsida_odb10
Trachemys scripta elegans	PRJNA634151	Chromosome	2,126,182,493	204,575	140,411,086	138	43.81	50.43	97.57%	0.86%	0.63%	1.80%	96.71%	7,480	sauropsida_odb10
Chelonoidis abingdonii	PRJNA611832	Scaffold	2,300,739,315	73,186	1,277,207	10,618	43.71	26.04	97.47%	0.99%	1.08%	1.44%	96.48%	7,480	sauropsida_odb10
Gopherus evgoodei	PRJNA559383	Chromosome	2,298,547,364	13,026,736	147,425,149	382	44.14	38.31	98.94%	1.43%	0.16%	0.90%	97.51%	7,480	sauropsida_odb10
Mauremys reevesii	PRJNA699301	Chromosome	2,035,064,793	34,524,243	139,244,951	62	44.55	33.55	99.16%	1.59%	0.16%	0.68%	97.57%	7,480	sauropsida_odb10

Suppementary Table 2. Average K-value for all the TEs identified for each specie.

Species	Average K-value
Pelodiscus sinensis	20.08
Chelydra serpentina	20.08
Chelonia Mydas	20.81
Dermochelys coriacea	19.83
Terrapene carolina triunguis	20.45
Chrysemys picta belii	20.26
Trachemys scripta elegans	20.33
Chelonoidis abingdonii	20.93
Gopherus evgoodei	20.36
Mauremys reevesii	20.43
General average	20.35

Suppementary Table 3. Average K-value for the TEs inside genes for each specie.

Species	Average K-value
Pelodiscus sinensis	20.08
Chelydra serpentina	20.10
Chelonia Mydas	20.82
Dermochelys coriacea	19.84
Terrapene carolina triunguis	20.45
Chrysemys picta belii	20.27
Trachemys scripta elegans	20.35
Chelonoidis abingdonii	20.94
Gopherus evgoodei	20.38
Mauremys reevesii	20.45
General average	20.37

Supplementary Table 4. Distribution of *Helitrons*, identified in *C. serpentina* within the category "Exon inside TE".

TE scaffold	Start TE	End TE	Subfamily	K-value	Size	Exon Scaffold	Start exon	End Exon	Exon ID
JAHGAV010000047.1	9,890,866	9,914,925	rnd-6_family-2979	4.19	24,059	JAHGAV010000047.1	9,891,242	9,891,313	ID=exon-G0U57_015484-1; product=tRNA-Asp
JAHGAV010000047.1	9,890,866	9,914,925	rnd-6_family-2979	4.19	24,059	JAHGAV010000047.1	9,893,791	9,893,862	ID=exon-G0U57_015485-1 product=tRNA-Asp
JAHGAV010000047.1	9,890,866	9,914,925	rnd-6_family-2979	4.19	24,059	JAHGAV010000047.1	9,895,068	9,895,139	ID=exon-G0U57_015486-1 product=tRNA-Asp
JAHGAV010000047.1	9,890,866	9,914,925	rnd-6_family-2979	4.19	24,059	JAHGAV010000047.1	9,896,346	9,896,417	ID=exon-G0U57_015487-1 product=tRNA-Asp
JAHGAV010000047.1	9,890,866	9,914,925	rnd-6_family-2979	4.19	24,059	JAHGAV010000047.1	9,897,625	9,897,696	ID=exon-G0U57_015488-1 product=tRNA-Asp
JAHGAV010000047.1	9,890,866	9,914,925	rnd-6_family-2979	4.19	24,059	JAHGAV010000047.1	9,898,903	9,898,974	ID=exon-G0U57_015489-1 product=tRNA-Asp
JAHGAV010000047.1	9,890,866	9,914,925	rnd-6_family-2979	4.19	24,059	JAHGAV010000047.1	9,900,182	9,900,253	ID=exon-G0U57_015490-1 product=tRNA-Asp
JAHGAV010000047.1	9,890,866	9,914,925	rnd-6_family-2979	4.19	24,059	JAHGAV010000047.1	9,901,461	9,901,532	ID=exon-G0U57_015491-1 product=tRNA-Asp
JAHGAV010000047.1	9,890,866	9,914,925	rnd-6_family-2979	4.19	24,059	JAHGAV010000047.1	9,902,739	9,902,810	ID=exon-G0U57_015492-1 product=tRNA-Asp
JAHGAV010000047.1	9,890,866	9,914,925	rnd-6_family-2979	4.19	24,059	JAHGAV010000047.1	9,904,018	9,904,089	ID=exon-G0U57_015493-1 product=tRNA-Asp
JAHGAV010000047.1	9,890,866	9,914,925	rnd-6_family-2979	4.19	24,059	JAHGAV010000047.1	9,905,297	9,905,368	ID=exon-G0U57_015494-1 product=tRNA-Asp
JAHGAV010000047.1	9,890,866	9,914,925	rnd-6_family-2979	4.19	24,059	JAHGAV010000047.1	9,906,575	9,906,646	ID=exon-G0U57_015495-1 product=tRNA-Asp
JAHGAV010000047.1	9,890,866	9,914,925	rnd-6_family-2979	4.19	24,059	JAHGAV010000047.1	9,907,853	9,907,924	ID=exon-G0U57_015496-1 product=tRNA-Asp
JAHGAV010000047.1	9,890,866	9,914,925	rnd-6_family-2979	4.19	24,059	JAHGAV010000047.1	9,909,132	9,909,203	ID=exon-G0U57_015497-1 product=tRNA-Asp
JAHGAV010000047.1	9,890,866	9,914,925	rnd-6_family-2979	4.19	24,059	JAHGAV010000047.1	9,911,681	9,911,752	ID=exon-G0U57_015498-1 product=tRNA-Asp

JAHGAV010000366.1	844,908	847,049	rnd-6_family-2979	3.08	2,141	JAHGAV010000366.1	846,616	846,687	ID=exon-G0U57_013304-1 product=tRNA-Asp
JAHGAV010000366.1	846,998	848,627	rnd-6_family-2979	2.6	1,629	JAHGAV010000366.1	847,366	847,437	ID=exon-G0U57_013305-1 product=tRNA-Asp
JAHGAV010000366.1	848,660	856,419	rnd-6_family-2979	1.84	7,759	JAHGAV010000366.1	852,461	852,532	ID=exon-G0U57_013306-1 product=tRNA-Asp

General discussion

The purpose of this thesis was to enhance our understanding of the evolution of transposable elements (TE) by examining their composition in the Testudines clade, an underrepresented lineage in genomic studies, investigating particularly the TE distribution and association with coding regions in these organisms. We aimed to gain insights into the relationship between transposable elements dynamics and species diversification and adaptation.

In order to achieve this, we had to generate and analyse the genomes of sea turtles, an important lineage of Testudines that lacked high-quality genomic data for investigating transposable elements (Chapter 1). We then investigated the molecular evolution of TEs in sea turtle species to identify specific changes in the transposable content in a pairwise comparison (Chapter 2). Finally, we conducted comparisons of the transposable elements proportion of the genomes in a larger number of Testudines species, including marine representatives, to explore the evolutionary pattern of transposable elements in a lineage that is distributed in diverse environments but exhibits relatively slow evolution rate in comparison to other vertebrates. We also investigated the composition of transposable elements potentially associated with functional regions of these species' genomes (Chapter 3).

In the following discussion, we present our key findings and the challenges we encountered during our research, as well as our future prospects for studying the evolution of transposable elements.

Sea turtle genomes have similar genome structure and TE composition.

The divergence of leatherback and green turtles is ancient and has resulted in species adapted to different habitats, diets, and lifestyles (Wyneken, Lohmann, and Musick 2013). In chapter 1 we showed that despite high levels of genome synteny, these two sea turtles have several regions presenting breaks in the collinearity. Additionally, we demonstrate that on microchromosomes, these sea turtles showed higher concentrations of multicopy gene families, as well as heightened nucleotide diversity and genetic distances between the species. Therefore, in chapter 1 we highlighted the potential importance of these regions as sources of variation underlying phenotypic differentiation.

Microchromosomes may have a higher adaptation value, as they accumulate variation and have a higher heterozygosity despite richer gene content. While the mechanisms driving these patterns are not well-understood, they may be related to higher recombination rates

(Rodionov 1996). As more chromosomal-level genomes become available, these findings provide a roadmap for identifying genomic regions involved in divergent evolutionary histories and the phenotypic connections of the genes within them. Further studies can be done to evaluate the prevalence of localised genomic differentiation and underlying mechanisms among other vertebrate groups.

We detected regions longer than 1 Gb with low synteny in the highly syntenic genomes of sea turtles and defined them as regions of reduced collinearity (RRCs). Our analysis revealed expansion or contraction of gene families associated with olfactory receptors, immunity, and zinc finger domains within these RRCs. However, we did not observe any correlation between the RRCs and differential accumulation of transposable elements or any specific group of transposable elements with higher concentration in the sea turtles genomes.

Sea turtles possess intricate sensory systems that allow them to detect volatile and water-soluble odorants crucial for migration, reproduction, and identifying prey, and predators (Courtney S. Endres and Lohmann 2013; C. S. Endres, Putman, and Lohmann 2009; Manton, Karr, and Ehrenfeld 1972; Kitayama et al. 2020; Courtney S. Endres et al. 2016). However, leatherback and green turtles inhabit different ecological niches and rely on distinct sensory cues. Leatherback turtles typically reside in the pelagic environment after hatching, undertaking vast horizontal and vertical migrations to locate patches of gelatinous prey (Dodge, Logan, and Lutcavage 2011). Conversely, green turtles inhabit neritic coastal and estuarine habitats as juveniles and have variable diets (Seminoff et al. 2021; Arthur, Boyle, and Limpus 2008). Our study of sea turtle genomes provided insights into the evolution of sensory and immune genes in these species. The differences in the ecological niches occupied by leatherback and green turtles have led to contrasting evolutionary paths for their olfactory receptor genes, with a greater loss of class II OR genes in the ancestral sea turtle lineage and an expansion of class I OR genes in the green turtle.

The MHC region is highly diverse and plays a vital role in the vertebrate immune response against pathogens. Greater gene copy numbers and heterozygosity within this region are associated with lower disease susceptibility (Siddle et al. 2010). While both sea turtle species have most of the core MHC-related genes, the green turtle has more copies of genes involved in adaptive and innate immunity. Pathogen prevalence and persistence are generally higher in neritic habitats than in open ocean habitats, so green turtles may be exposed to a higher diversity and load of pathogens than leatherback turtles (Escobar et al. 2015). However, research on reptilian immune systems, especially MHC genes in turtles, is limited.

The green turtle's greater immune gene diversity may reflect exposure to higher pathogen loads and diversity in neritic habitats. However, the exact relationship between

immune gene diversity and disease susceptibility or ecological adaptation in sea turtles remains unclear, and further research is needed to fully understand the role of these genes in the conservation of these species, particularly in the face of threats such as fibropapillomatosis. The availability of reference genomes will enable more accurate study of these complex gene families and advance our understanding of immune gene evolution in sea turtles.

The level of genomic diversity in a species has significant implications for their future survival and adaptive capacity, particularly in the face of rapid human-induced global change (Kardos et al. 2021). While high-quality reference genomes are not necessary for estimating genome-wide diversity, they allow for a more comprehensive examination of diversity patterns relevant to conservation. The reference genomes produced in this study reveal very low diversity in the coding regions of leatherback turtle genomes, indicating limited functional variation and potentially hindering their ability to adapt to new conditions. Leatherback turtles also exhibit lower heterozygosity compared to green turtles(Dobrynin et al. 2015; Mattila et al. 2012), which may contribute to their lower hatching success and slow population recoveries (Eckert et al. 2012). However, some species with similarly low diversity have bounced back after population declines, possibly due to purging of deleterious alleles resulting from long-term low population sizes (Robinson et al. 2018; Dussex et al. 2021; Kyriazis, Wayne, and Lohmueller 2021). The reference genomes presented in this study enable further research into these topics, clarifying the relationships between genomic diversity, genetic load, and population viability in sea turtle species to inform conservation strategies.

Patterns of diversity, genetic load, and demographic histories were generally consistent within species, but ROH analyses revealed a striking exception for the green turtle reference individual from the Mediterranean. This isolated population has suffered severe decline in the last century due to human exploitation, and our results suggest that consequent inbreeding is likely occurring, which may impact the population's recovery (Casale et al. 2018). Our study highlights the importance of understanding genomic diversity and demographic histories for conservation efforts of endangered species such as leatherback and green turtles. The low genomic diversity observed in leatherback turtles is likely a result of long-term low effective population sizes and historical bottleneck events, while higher heterozygosity and larger historical effective population sizes in green turtles reflect radiation from many refugia and frequent admixing of populations. This emphasises the significance of standing genetic variation for a species' future persistence, and the importance of deeper examination of diversity patterns within coding regions of genomes for conservation purposes. Finally, we provide insights into the impact of environmental changes on species' abundances and distributions, and emphasises the importance of using highly contiguous genomes for accurate ROH assessment

to inform conservation efforts. Overall, these findings will aid in developing better conservation strategies and help to ensure the long-term survival of these important marine species.

Leatherback turtles have recent expansion of PLE TEs.

This study presents novel findings regarding transposable element (TE) dynamics in sea turtles, highlighting a recent expansion of Penelope-like elements (PLEs) in the slow-evolving genome of leatherback turtle. The study involved a comparison of the TE composition of two sea turtle species - *C. mydas* and *D. coriacea* - estimated to have diverged around 58-100 MYA. In the first chapter, we reported that TEs constitute a similar proportion of the genomes of both species, reaching 45.79% for *D. coriacea* and 44.41% for *C. mydas*, which are significantly higher than previous estimates (approximately 10% for *C. mydas*). Nonetheless, the previous analyses like the one performed on Wang (2013), were based on draft genome versions, whereas on this thesis was sequenced, assembled and utilised high-quality reference genomes assembled using long reads, providing more comprehensive and accurate data on the mobilomes in Chelonioidea.

When comparing the TE content of the genomes of two sea turtle species, *C. mydas* and *D. coriacea*, we found differences in abundance comparing the divergence profiles and the proportion of the genomes for TE subfamilies as LINEs as PLEs (Figure 1 chapter 2). *D. coriacea*'s genome contains younger insertions within K-values of 0-2%, mainly LINEs and PLEs, which are not present in *C. mydas*. Moreover, PLEs were found to be twice as abundant as a general proportion in the genome of *D. coriacea* than in *C. mydas* (4.70% versus 2.34%). This suggests a recent expansion of these elements in the genome of *D. coriacea*.

To determine whether the higher proportion of PLEs in *D. coriacea*'s genome was due to expansion in this species or contraction in *C. mydas*, a comparison was made with 13 other turtle species from all six superfamilies of testudines. This analysis revealed that *D. coriacea* indeed had a higher proportion of the genome consisting of PLEs, indicating that its genome underwent an expansion of PLE insertions not seen in other analysed testudines. Moreover, *D. coriacea* was the only species with more than 2% of the genome composed of very young or young TEs with K-values between 0-2%, as shown in Figure 2 of chapter 2. *C. serpentina* (Kinosternoidea) and *C. mccordi* (Testudinoidea) had a lower proportion of insertions with K-values 0-2% than *D. coriacea*, suggesting that the expansion of PLEs in *D. coriacea* was exclusive to this species' lineage. No recent TE expansions of comparable scale were found in any of the other turtles analysed.

Expansion of TE families post-speciation has been extensively studied in various organisms, including Arabidopsis (Slotkin et al. 2009), tobacco (McCormick 2004), Drosophila (Marcillac, Grosjean, and Ferveur 2005), and fish (Renaut and Bernatchez 2011; Rogers and Bernatchez 2007), among others. This topic has been reviewed by Serrato-Capuchina and Matute (2018) and Mérot et al. (2020). However the fact that none other species present an expansion on TE as shown in figure 2 of chapter 2, indicates that this is an isolated phenomena happening in *D. coriacea*. As shown in chapter 1, *D. coriacea* suffered a recent bottleneck in their population. As other genetic mutations, the fixation of transposable elements (TEs) in a population is influenced not only by their fitness effects and generation time but also by demographic parameters, especially the effective population size (N_e). In populations with low N_e, TEs are more likely to be fixed by genetic drift, which can lead to their invasive fixation in the genome after genetic bottlenecks (Matzke et al. 2012).

Upon observing the PLE expansion in *D. coriacea*, we conducted a search for potentially active PLE copies by focusing on the lowest K-values. From this search, we identified 7 subfamilies of PLE with significantly lower K-values, and within this group, we found one with significantly lower K-values than all other subfamilies. To catalogue this subfamily, we identified it as a member of the *Neptune* family. PLEs can be divided into endonuclease positive (EN+) or negative (EN-) groups, and we identified evidence of an EN with the GIY-YIG endonuclease domain, characteristic of *Neptune* PLEs, in this particular subfamily. We also found a reverse transcriptase TERT domain in this subfamily. Further analysis revealed that this subfamily showed high levels of similarity with Neptune-1_CPB from *Chrysemys picta bellii*, with evidence for a separate subfamily in *D. coriacea* following the 95-80-98 rule. Therefore, we describe this subfamily as Neptune-1_DC, an active and recently expanded subfamily of PLEs. By analysing RNA expression data from three different tissues, we identified actively transcribed copies of Neptune-1_DC, suggesting that this element may still be active in the genome of *D. coriacea*.

In the second chapter, we highlight the significance of using high-quality genome assemblies to improve the accuracy of identifying and characterising transposable elements (TEs) in species, and to determine their dynamics within different lineages. The results demonstrate that sea turtle genomes are highly conserved, and sea turtle genomes contain a significant similar proportion of TEs, with few differences in specific TE family abundance between species. The discovery of a recent expansion of *Penelope-like* elements (PLEs) on leatherback within the highly conserved sea turtle clade provides new insights into the dynamics of TEs within Testudines. This finding aligns with previous research on the expansion of TE families in different organisms post-speciation (Slotkin et al. 2009). Overall, this study provides

an important contribution to the understanding of the mobilomes within Chelonioidea and advances our knowledge of TE dynamics within different lineages of species.

Influence of the reference genome quality on the probability to detect and identify TEs

As previously mentioned, *D. coriacea* has experienced long-term low effective population sizes and historical bottleneck events, which have been linked to deregulation of TE activity. In chapter 2, we identified a recent expansion of TEs in this species. Consequently, we are interested in investigating the interaction between TEs and genomic features in testudines.

We explore the correlation between TE insertion and the genomes of ten different species of Testudines. The quality of genome assemblies is known to affect transposable element (TE) analysis, with fragmented assemblies producing different TE profiles than complete assemblies with similar QV. For instance, the *C. serpentina* genome is more fragmented than the *C. mydas* genome, even though they have similar total sequence lengths and QV. To better detect complete and active TEs, more complete genome assemblies, made with long reads are required (Peona et al. 2021; Prost et al. 2019).

Although the total level of TEs is consistent, the accumulation of insertions within the main TE orders and families varies among turtle families. A comparison of the TE proportion of the genomes among the turtle species in this study showed different patterns for each main TE order. The amount of Retrotransposon elements detected is directly proportional to the quality of the genome, with *D. coriacea* having the highest values and *P. sinensis* having the lowest (Figure 2a chapter 3). However, the proportion of DNA TEs does not show a dependence on the quality of the genomes (Figure 2b chapter 3). For instance, the genomes generated in this thesis are among the species with the lowest amounts of DNA TEs. This lack of dependence of DNA TE proportions on genome quality may suggest that this TE order is more variable among turtles, and the proportion of the insertions could provide information about the natural biology or evolution of different turtle families. However, DNA TEs are less frequent than Retrotransposon TEs in turtle genomes, which may contribute to the absence of a clear pattern. Hence, the fact that both sea turtles included in this study have a reduced proportion of DNA TEs requires further investigation. The total proportion of TEs also follows a pattern that depends on genome quality, showing a decrease in TE elements for genomes with lower quality, except for *P. sinensis* genome. Nevertheless, this turtle has the highest amount of Unclassified TE (Table 2 chapter 3), which may indicate that the lower quality of the genome leads to misidentification of TEs in general.

In all genomes studied, there are similar numbers of interactions upstream and downstream of genes, regardless of independent insertion events (Table 3 chapter 3). We expected that TEs would play a role in the regulation of gene expression of the turtles, especially on *D. coriacea.* However we did not observe a higher number of insertions in the upstream region of gene features, suggesting that TEs in turtles do not significantly target promoter regions.

Among the species, *Mauremys reevesii* has the highest number of exon-TE interactions, with about 10 times more interactions than *D. coriacea*, despite having a similar number of annotated genes. *M. reevesii* has a higher number of non-coding genes, which could be due to the increased number of TEs inside exons for this species. The *M. reevesii* assembly is of high quality and does not suggest any indication of artifactual error that could lead to these differences. After *M. reevesii*, sea turtles have the highest number of TEs inside exons, with *C. mydas* having almost double the amount of TEs inside exons compared to *D. coriacea*. In chapter 1, we described that the main differences between sea turtle genomes lie in multicopy gene families, as OR and MHC. Investigating the functions of the genes with exons containing TEs may help us understand if TE activity is related to these multicopy gene families in sea turtles.

We observed a significantly higher number of TE insertions in non-coding regions, such as introns, compared to exons or other regulatory elements. This finding was not surprising, as the insertion of TEs within genes can disrupt their function and structure, potentially leading to negative impacts on the fitness of individuals. Consequently, such deleterious insertions are often purged from populations during the process of species adaptation (Schrader and Schmitz 2019; Mackay 1986). Moving forward, it would be interesting to explore whether certain families of TEs have a higher propensity for insertion within genes or exons, which could provide further insights into the evolution of TEs.

Finally, after a Testudine-wide analysis, we identified that the assembly status of the genomes affects the identification and analysis of transposable elements, and more complete assemblies are crucial for detecting active TEs. Retrotransposons are more dependent on genome quality than DNA TEs, which could suggest that the proportion of the insertions of DNA TEs in turtles' genomes in this analysis carries information about the natural biology or evolution of different turtle families. There are differences in the accumulation of insertions within TE orders and families among turtle families but not in the overall proportion of TEs. Regarding the interaction of TEs with the functional regions on the genome, TEs can affect gene expression by insertion inside genes, in the borders of gene features, or in regulatory

regions, leading to significant evolutionary consequences. Overall, the findings presented here provide a valuable resource for future studies of genome evolution and TE dynamics in turtles.

Prospects and Future Research on Testudines TE evolution.

The reference genomes for both extant sea turtle families, in addition to the insights reported here, offer an immense opportunity to conduct a wide range of fundamental and applied research that was previously unattainable. When combined with other upcoming genomes, comparative genomics analyses can shed light on the genomic basis for long-standing traits such as adaptation to saltwater, diving capacity, and long-distance natal homing among many others. By leveraging these reference genomes in conjunction with whole-genome sequencing of ancient samples, studies can determine the relationship between genomic erosion, inbreeding, and mutational load with population size, trajectories, and conservation measures in global populations. Although high-quality reference genomes are not necessary for all research goals, they are crucial for certain objectives. For example, the use of ROH metrics, that is increasingly important in species management plans, and researchers should understand how genome quality may affect their analyses and inferences. The reference genomes can also be used to develop molecular assays and amplicon panels, investigate temperature sex determination mechanisms and adaptive capacity under climate change, and assess linkages between immune genes and disease risk. Moreover, the genomes can anchor existing anonymous markers and optimise new ones for conservation-focused questions, leading to large-scale syntheses and equitable capacity building for genomics research. Therefore the necessity for high quality genomes go far beyond the boundaries of basic sciences and could have an impact on the conservation of life itself.

Also, here we describe the reduction of the population size of *D. coriacea* and how this unleash an expansion of TE on their genome. To later explore if we can catch an intervention of TEs on the genomic regulation of *D. coriacea*. Regardless, that more investigation is necessary to fully comprehend this interaction. This study provides an insight into the interaction of TE and sea turtles, with particular focus on endangered species within the Testudine clade.

Conclusion

As expected in a slow evolving clade the differences in the abundance of TE among turtles are little,with the exception of very specialised species, such as *D. coriacea*. This slow evolving pattern is even more evident compared to the differences observed inside a closely related clade of avian genomes (Kapusta and Suh 2017).

In conclusion, the comparative study of sea turtle genomes has provided valuable insights into the genomic diversity of these species, including the identification of key genomic regions and gene families that are important for phenotypic differentiation, as well as the impact of environmental changes on their populations. TEs analysis are highly susceptible to the quality of the genomes. As a response to a reduction in population size we observed an expansion of TEs on *D. coriacea*'s genome. We described the interaction between TE and genomic features as genes and exons, although no significant correlation was found, the clade-wide analysis showed once again that the quality of the genomes is of high importance in order to study the TEs abundancy on genomes.

These findings have significant implications for conservation efforts and highlight the importance of understanding the dynamics of transposable elements within different lineages. The availability of high-quality genome assemblies and manual curation of TE repeats is crucial for accurate classification and analysis of TE families. The study's results provide a foundation for further research into the evolution of genome structure and gene function in turtles and other vertebrate groups, ultimately contributing to our broader understanding of the mechanisms underlying evolutionary change and adaptation.

References

Ali, Arsala, Kyudong Han, and Ping Liang. 2021. “Role of Transposable Elements in Gene Regulation in the Human Genome.” *Life* 11 (2). https://doi.org/10.3390/life11020118.

Arthur, K. E., M. C. Boyle, and C. J. Limpus. 2008. “Ontogenetic Changes in Diet and Habitat Use in Green Sea Turtle (Chelonia Mydas) Life History.” *Marine Ecology Progress Series* 362 (June): 303–11.

Avise, J. C., B. W. Bowen, T. Lamb, A. B. Meylan, and E. Bermingham. 1992. “Mitochondrial DNA Evolution at a Turtle’s Pace: Evidence for Low Genetic Variability and Reduced Microevolutionary Rate in the Testudines.” *Molecular Biology and Evolution* 9 (3): 457–73.

Baduel, Pierre, Basile Leduque, Amandine Ignace, Isabelle Gy, José Gil Jr, Olivier Loudet, Vincent Colot, and Leandro Quadrana. 2021. “Genetic and Environmental Modulation of Transposition Shapes the Evolutionary Potential of Arabidopsis Thaliana.” *Genome Biology* 22 (1): 138.

Bentley, Blair P., Tomás Carrasco-Valenzuela, Elisa K. S. Ramos, Harvinder Pawar, Larissa Souza, Alana Alexander, Shreya M. Banerjee, et al. 2023. “1 Divergent Sensory and Immune Gene Evolution in Sea Turtles with Contrasting Demographic and Life 2 Histories.” *The Proceedings of the National Academy of Sciences (PNAS)*.

Boeke, J. D., D. J. Garfinkel, C. A. Styles, and G. R. Fink. 1985. “Ty Elements Transpose through an RNA Intermediate.” *Cell* 40 (3): 491–500.

Boissinot, Stéphane, Yann Bourgeois, Joseph D. Manthey, and Robert P. Ruggiero. 2019. “The Mobilome of Reptiles: Evolution, Structure, and Function.” *Cytogenetic and Genome Research* 157 (1-2): 21–33.

Bourque, Guillaume, Kathleen H. Burns, Mary Gehring, Vera Gorbunova, Andrei Seluanov, Molly Hammell, Michaël Imbeault, et al. 2018. “Ten Things You Should Know about Transposable Elements.” *Genome Biology* 19 (1): 1–12.

Brini, A. T., G. M. Lee, and J. P. Kinet. 1993. “Involvement of Alu Sequences in the Cell-Specific Regulation of Transcription of the Gamma Chain of Fc and T Cell Receptors.” *The Journal of Biological Chemistry* 268 (2): 1355–61.

Bruno, Melania, Mohamed Mahgoub, and Todd S. Macfarlan. 2019. “The Arms Race Between KRAB-Zinc Finger Proteins and Endogenous Retroelements and Its Impact on Mammals.” *Annual Review of Genetics* 53 (December): 393–416.

Canapa, Adriana, Marco Barucca, Maria A. Biscotti, Mariko Forconi, and Ettore Olmo. 2015. “Transposons, Genome Size, and Evolutionary Insights in Animals.” *Cytogenetic and Genome Research* 147 (4): 217–39.

Card, Daren C., W. Bryan Jennings, and Scott V. Edwards. 2023. “Genome Evolution and the Future of Phylogenomics of Non-Avian Reptiles.” *Animals : An Open Access Journal from MDPI* 13 (3). https://doi.org/10.3390/ani13030471.

Casale, P., A. C. Broderick, J. A. Camiñas, L. Cardona, C. Carreras, A. Demetropoulos, W. J. Fuller, et al. 2018. “Mediterranean Sea Turtles: Current Knowledge and Priorities for Conservation and Research.” *Endangered Species Research* 36 (August): 229–67.

Conley, Andrew B., Jittima Piriyapongsa, and I. King Jordan. 2008. “Retroviral Promoters in the Human Genome.” *Bioinformatics* 24 (14): 1563–67.

Damas, Joana, Rebecca O’Connor, Marta Farré, Vasileios Panagiotis E. Lenis, Henry J. Martell, Anjali Mandawala, Katie Fowler, et al. 2017. “Upgrading Short-Read Animal Genome Assemblies to Chromosome Level Using Comparative Genomics and a Universal Probe Set.” *Genome Research* 27 (5): 875–84.

Das, Debojyoti, Sunil Kumar Singh, Jacob Bierstedt, Alyssa Erickson, Gina L. J. Galli, Dane A. Crossley 2nd, and Turk Rhen. 2020. “Draft Genome of the Common Snapping Turtle, Chelydra Serpentina, a Model for Phenotypic Plasticity in Reptiles.” *G3* 10 (12): 4299–4314.

Davenport, John. 1997. “Temperature and the Life-History Strategies of Sea Turtles.” *Journal of Thermal Biology* 22 (6): 479–88.

Dobrynin, Pavel, Shiping Liu, Gaik Tamazian, Zijun Xiong, Andrey A. Yurchenko, Ksenia Krasheninnikova, Sergey Kliver, et al. 2015. “Genomic Legacy of the African Cheetah, Acinonyx Jubatus.” *Genome Biology* 16 (December): 277.

Dodge, Kara L., John M. Logan, and Molly E. Lutcavage. 2011. “Foraging Ecology of Leatherback Sea Turtles in the Western North Atlantic Determined through Multi-Tissue Stable Isotope Analyses.” *Marine Biology* 158 (12): 2813–24.

Driller, M., S. T. Vilaca, and L. S. Arantes. 2020. “Optimization of ddRAD-like Data Leads to High Quality Sets of Reduced Representation Single Copy Orthologs (R2SCOs) in a Sea Turtle Multi-Species Analysis.” *bioRxiv*. https://www.biorxiv.org/content/10.1101/2020.04.03.024331v1.abstract.

Dussex, Nicolas, Tom van der Valk, Hernán E. Morales, Christopher W. Wheat, David Díez-Del-Molino, Johanna von Seth, Yasmin Foster, et al. 2021. “Population Genomics of the Critically Endangered Kākāpō.” *Cell Genomics* 1 (1): 100002.

Eckert, Karen L., Bryan P. Wallace, John G. Frazier, Peter C. H. Pritchard, and Scott A. Eckert. 2012. *Synopsis of the Biological Data on the Leatherback Sea Turtle (Dermochelys Coriacea)*. Createspace Independent Pub.

Ekblom, Robert, and Jochen B. W. Wolf. 2014. “A Field Guide to Whole-Genome Sequencing, Assembly and Annotation.” *Evolutionary Applications* 7 (9): 1026–42.
Elliott, Tyler A., and T. Ryan Gregory. 2015. “What’s in a Genome? The C-Value Enigma and the Evolution of Eukaryotic Genome Content.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 370 (1678): 20140331.
Endoh, H., and N. Okada. 1986. “Total DNA Transcription in Vitro: A Procedure to Detect Highly Repetitive and Transcribable Sequences with tRNA-like Structures.” *Proceedings of the National Academy of Sciences of the United States of America* 83 (2): 251–55.
Endres, Courtney S., and Kenneth J. Lohmann. 2013. “Detection of Coastal Mud Odors by Loggerhead Sea Turtles: A Possible Mechanism for Sensing Nearby Land.” *Marine Biology* 160 (11): 2951–56.
Endres, Courtney S., Nathan F. Putman, David A. Ernst, Jessica A. Kurth, Catherine M. F. Lohmann, and Kenneth J. Lohmann. 2016. “Multi-Modal Homing in Sea Turtles: Modeling Dual Use of Geomagnetic and Chemical Cues in Island-Finding.” *Frontiers in Behavioral Neuroscience* 10 (February): 19.
Endres, C. S., N. F. Putman, and K. J. Lohmann. 2009. “Perception of Airborne Odors by Loggerhead Sea Turtles.” *The Journal of Experimental Biology* 212 (Pt 23): 3823–27.
Escobar, Luis E., Sadie J. Ryan, Anna M. Stewart-Ibarra, Julia L. Finkelstein, Christine A. King, Huijie Qiao, and Mark E. Polhemus. 2015. “A Global Map of Suitability for Coastal Vibrio Cholerae under Current and Future Climate Conditions.” *Acta Tropica* 149 (September): 202–11.
Frair, W., R. G. Ackman, and N. Mrosovsky. 1972. “Body Temperature of Dermochelys Coriacea: Warm Turtle from Cold Water.” *Science* 177 (4051): 791–93.
Franchini, Lucía F., Rodrigo López-Leal, Sofía Nasif, Paula Beati, Diego M. Gelman, Malcolm J. Low, Flávio J. S. de Souza, and Marcelo Rubinstein. 2011. “Convergent Evolution of Two Mammalian Neuronal Enhancers by Sequential Exaptation of Unrelated Retroposons.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (37): 15270–75.
Gladyshev, Eugene A., and Irina R. Arkhipova. 2007. “Telomere-Associated Endonuclease-Deficient Penelope-like Retroelements in Diverse Eukaryotes.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (22): 9352–57.
Grabundzija, Ivana, Simon A. Messing, Jainy Thomas, Rachel L. Cosby, Ilija Bilic, Csaba Miskey, Andreas Gogol-Döring, et al. 2016. “A Helitron Transposon Reconstructed from Bats Reveals a Novel Mechanism of Genome Shuffling in Eukaryotes.” *Nature Communications* 7 (March): 10716.
Greenblatt, Irwin M., and R. Alexander Brink. 1963. “Transpositions of Modulator in Maize into Divided and Undivided Chromosome Segments.” *Nature* 197 (4865): 412–13.
Green, Richard E., Edward L. Braun, Joel Armstrong, Dent Earl, Ngan Nguyen, Glenn Hickey, Michael W. Vandewege, et al. 2014. “Three Crocodilian Genomes Reveal Ancestral Patterns of Evolution among Archosaurs.” *Science* 346 (6215): 1254449.
Hambor, J. E., J. Mennone, M. E. Coon, J. H. Hanke, and P. Kavathas. 1993. “Identification and Characterization of an Alu-Containing, T-Cell-Specific Enhancer Located in the Last Intron of the Human CD8 Alpha Gene.” *Molecular and Cellular Biology* 13 (11): 7056–70.
Hirayama, Ren. 1998. “Oldest Known Sea Turtle.” *Nature* 392 (6677): 705–8.
Jacobs, Frank M. J., David Greenberg, Ngan Nguyen, Maximilian Haeussler, Adam D. Ewing, Sol Katzman, Benedict Paten, Sofie R. Salama, and David Haussler. 2014. “An Evolutionary Arms Race between KRAB Zinc-Finger Genes ZNF91/93 and SVA/L1 Retrotransposons.” *Nature* 516 (7530): 242–45.
Janes, Daniel E., Christopher L. Organ, Matthew K. Fujita, Andrew M. Shedlock, and Scott V. Edwards. 2010. “Genome Evolution in Reptilia, the Sister Group of Mammals.” *Annual Review of Genomics and Human Genetics* 11: 239–64.
Kajikawa, M., K. Ohshima, and N. Okada. 1997. “Determination of the Entire Sequence of Turtle CR1: The First Open Reading Frame of the Turtle CR1 Element Encodes a Protein with a Novel Zinc Finger Motif.” *Molecular Biology and Evolution* 14 (12): 1206–17.
Kapusta, Aurélie, and Alexander Suh. 2017. “Evolution of Bird Genomes-a Transposon’s-Eye View.” *Annals of the New York Academy of Sciences*. https://doi.org/10.1111/nyas.13295.
Kardos, Marty, Ellie E. Armstrong, Sarah W. Fitzpatrick, Samantha Hauser, Philip W. Hedrick, Joshua M. Miller, David A. Tallmon, and W. Chris Funk. 2021. “The Crucial Role of Genome-Wide Genetic Variation in Conservation.” *Proceedings of the National Academy of Sciences of the United States of America* 118 (48). https://doi.org/10.1073/pnas.2104642118.
Kelley, Joanna L., Anthony P. Brown, Nina Overgaard Therkildsen, and Andrew D. Foote. 2016. “The Life Aquatic: Advances in Marine Vertebrate Genomics.” *Nature Reviews. Genetics* 17 (9): 523–34.
Kidwell, Margaret G. 2002. “Transposable Elements and the Evolution of Genome Size in Eukaryotes.” *Genetica* 115 (1): 49–63.
Kidwell, M. G., and D. R. Lisch. 2001. “Perspective: Transposable Elements, Parasitic DNA, and Genome Evolution.” *Evolution; International Journal of Organic Evolution* 55 (1): 1–24.
Kimura, M. 1980. “A Simple Method for Estimating Evolutionary Rates of Base Substitutions through Comparative Studies of Nucleotide Sequences.” *Journal of Molecular Evolution* 16 (2): 111–20.

Kitayama, Chiyo, Yohei Yamaguchi, Satomi Kondo, Ryuta Ogawa, Yusuke K. Kawai, Mitsunori Kayano, Jumpei Tomiyasu, and Daisuke Kondoh. 2020. "Behavioral Effects of Scents from Male Mature Rathke Glands on Juvenile Green Sea Turtles (Chelonia Mydas)." *The Journal of Veterinary Medical Science / the Japanese Society of Veterinary Science* 82 (9): 1312–15.

Koepfli, Klaus-Peter, Benedict Paten, Genome 10K Community of Scientists, and Stephen J. O'Brien. 2015. "The Genome 10K Project: A Way Forward." *Annual Review of Animal Biosciences* 3: 57–111.

Komoroske, L. M., M. R. Miller, and S. M. O'Rourke. 2019. "A Versatile Rapture (RAD-Capture) Platform for Genotyping Marine Turtles." *Molecular Ecology*. https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.12980.

Kyriazis, Christopher C., Robert K. Wayne, and Kirk E. Lohmueller. 2021. "Strongly Deleterious Mutations Are a Primary Determinant of Extinction Risk due to Inbreeding Depression." *Evolution Letters* 5 (1): 33–47.

Liu, Sanzhen, Cheng-Ting Yeh, Tieming Ji, Kai Ying, Haiyan Wu, Ho Man Tang, Yan Fu, Daniel Nettleton, and Patrick S. Schnable. 2009. "Mu Transposon Insertion Sites and Meiotic Recombination Events Co-Localize with Epigenetic Marks for Open Chromatin across the Maize Genome." *PLoS Genetics* 5 (11): e1000733.

Liu, Xiaoli, Yakun Wang, Ju Yuan, Fang Liu, Xiaoyou Hong, Lingyun Yu, Chen Chen, et al. 2022. "Chromosome-Level Genome Assembly of Asian Yellow Pond Turtle (Mauremys Mutica) with Temperature-Dependent Sex Determination System." *Scientific Reports* 12 (1): 7905.

Malik, H. S., and T. H. Eickbush. 1999. "Modular Evolution of the Integrase Domain in the Ty3/Gypsy Class of LTR Retrotransposons." *Journal of Virology* 73 (6): 5186–90.

Manton, Marion, Andrew Karr, and David W. Ehrenfeld. 1972. "CHEMORECEPTION IN THE MIGRATORY SEA TURTLE, CHELONIA MYDAS." *The Biological Bulletin* 143 (1): 184–95.

Mattila, Anniina L. K., Anne Duplouy, Malla Kirjokangas, Rainer Lehtonen, Pasi Rastas, and Ilkka Hanski. 2012. "High Genetic Load in an Old Isolated Butterfly Population." *Proceedings of the National Academy of Sciences of the United States of America* 109 (37): E2496–2505.

Matzke, Andreas, Gennady Churakov, Petra Berkes, Erin M. Arms, Denise Kelsey, Jürgen Brosius, Jan Ole Kriegs, and Jürgen Schmitz. 2012. "Retroposon Insertion Patterns of Neoavian Birds: Strong Evidence for an Extensive Incomplete Lineage Sorting Era." *Molecular Biology and Evolution* 29 (6): 1497–1501.

Mérot, Claire, Rebekah A. Oomen, Anna Tigano, and Maren Wellenreuther. 2020. "A Roadmap for Understanding the Evolutionary Significance of Structural Genomic Variation." *Trends in Ecology & Evolution* 35 (7): 561–72.

Peona, Valentina, Mozes P. K. Blom, Luohao Xu, Reto Burri, Shawn Sullivan, Ignas Bunikis, Ivan Liachko, et al. 2021. "Identifying the Causes and Consequences of Assembly Gaps Using a Multiplatform Genome Assembly of a Bird-of-Paradise." *Molecular Ecology Resources* 21 (1): 263–86.

Pike, David A. 2013. "Climate Influences the Global Distribution of Sea Turtle Nesting." *Global Ecology and Biogeography: A Journal of Macroecology* 22 (5): 555–66.

Prost, Stefan, Ellie E. Armstrong, Johan Nylander, Gregg W. C. Thomas, Alexander Suh, Bent Petersen, Love Dalen, et al. 2019. "Comparative Analyses Identify Genomic Features Potentially Involved in the Evolution of Birds-of-Paradise." *GigaScience* 8 (5). https://doi.org/10.1093/gigascience/giz003.

Rhie, Arang, Shane A. McCarthy, Olivier Fedrigo, Joana Damas, Giulio Formenti, Sergey Koren, Marcela Uliano-Silva, et al. 2021. "Towards Complete and Error-Free Genome Assemblies of All Vertebrate Species." *Nature* 592 (7856): 737–46.

Robberecht, Caroline, Thierry Voet, Masoud Zamani Esteki, Beata A. Nowakowska, and Joris R. Vermeesch. 2013. "Nonallelic Homologous Recombination between Retrotransposable Elements Is a Driver of de Novo Unbalanced Translocations." *Genome Research* 23 (3): 411–18.

Robinson, Jacqueline A., Caitlin Brown, Bernard Y. Kim, Kirk E. Lohmueller, and Robert K. Wayne. 2018. "Purging of Strongly Deleterious Mutations Explains Long-Term Persistence and Absence of Inbreeding Depression in Island Foxes." *Current Biology: CB* 28 (21): 3487–94.e4.

Rodionov, A. V. 1996. "[Micro vs. macro: structural-functional organization of avian micro- and macrochromosomes]." *Genetika* 32 (5): 597–608.

Rubin, G. M., M. G. Kidwell, and P. M. Bingham. 1982. "The Molecular Basis of P-M Hybrid Dysgenesis: The Nature of Induced Mutations." *Cell* 29 (3): 987–94.

Samuelson, L. C., K. Wiebauer, C. M. Snow, and M. H. Meisler. 1990. "Retroviral and Pseudogene Insertion Sites Reveal the Lineage of Human Salivary and Pancreatic Amylase Genes from a Single Gene during Primate Evolution." *Molecular and Cellular Biology* 10 (6): 2513–20.

Seminoff, Jeffrey A., Lisa M. Komoroske, Diego Amorocho, Randall Arauz, Didiher Chacón-Chaverrí, Nelly Paz, Peter H. Dutton, et al. 2021. "Large-scale Patterns of Green Turtle Trophic Ecology in the Eastern Pacific Ocean." *Ecosphere* 12 (6). https://doi.org/10.1002/ecs2.3479.

Shaffer, H. Bradley, Evan McCartney-Melstad, Thomas J. Near, Genevieve G. Mount, and Phillip Q. Spinks. 2017. "Phylogenomic Analyses of 539 Highly Informative Loci Dates a Fully Resolved Time Tree for the Major Clades of Living Turtles (Testudines)." *Molecular Phylogenetics and Evolution* 115 (October): 7–15.

Shaffer, H. Bradley, Patrick Minx, Daniel E. Warren, Andrew M. Shedlock, Robert C. Thomson, Nicole Valenzuela, John Abramyan, et al. 2013. "The Western Painted Turtle Genome, a Model for the Evolution of Extreme Physiological Adaptations in a Slowly Evolving Lineage." *Genome Biology* 14 (3): R28.

Shahid, Saima, and R. Keith Slotkin. 2020. “The Current Revolution in Transposable Element Biology Enabled by Long Reads.” *Current Opinion in Plant Biology* 54 (April): 49–56.

Siddle, Hannah V., Jolanta Marzec, Yuanyuan Cheng, Menna Jones, and Katherine Belov. 2010. “MHC Gene Copy Number Variation in Tasmanian Devils: Implications for the Spread of a Contagious Cancer.” *Proceedings. Biological Sciences / The Royal Society* 277 (1690): 2001–6.

Skipper, Kristian Alsbjerg, Peter Refsing Andersen, Nynne Sharma, and Jacob Giehm Mikkelsen. 2013. “DNA Transposon-Based Gene Vehicles - Scenes from an Evolutionary Drive.” *Journal of Biomedical Science* 20 (1): 1–23.

Sotero-Caio, Cibele G., Roy N. Platt 2nd, Alexander Suh, and David A. Ray. 2017. “Evolution and Diversity of Transposable Elements in Vertebrate Genomes.” *Genome Biology and Evolution* 9 (1): 161–77.

Spradling, Allan C., Hugo J. Bellen, and Roger A. Hoskins. 2011. “Drosophila P Elements Preferentially Transpose to Replication Origins.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (38): 15948–53.

Symonová, Radka, and Alexander Suh. 2019. “Nucleotide Composition of Transposable Elements Likely Contributes to AT/GC Compositional Homogeneity of Teleost Fish Genomes.” *Mobile DNA* 10 (December): 49.

Thomson, Robert C., Phillip Q. Spinks, and H. Bradley Shaffer. 2021. “A Global Phylogeny of Turtles Reveals a Burst of Climate-Associated Diversification on Continental Margins.” *Proceedings of the National Academy of Sciences of the United States of America* 118 (7). https://doi.org/10.1073/pnas.2012215118.

Tollis, M., and S. Boissinot. 2012. “The Evolutionary Dynamics of Transposable Elements in Eukaryote Genomes.” *Genome Dynamics* 7 (June): 68–91.

Vilaça, Sibelle Torres, Riccardo Piccinno, Omar Rota-Stabelli, Maëva Gabrielli, Andrea Benazzo, Michael Matschiner, Luciano S. Soares, Alan B. Bolten, Karen A. Bjorndal, and Giorgio Bertorelle. 2021. “Divergence and Hybridization in Sea Turtles: Inferences from Genome Data Show Evidence of Ancient Gene Flow between Species.” *Molecular Ecology* 30 (23): 6178–92.

Wang, Zhuo, Juan Pascual-Anaya, Amonida Zadissa, Wenqi Li, Yoshihito Niimura, Zhiyong Huang, Chunyi Li, et al. 2013. “The Draft Genomes of Soft-Shell Turtle and Green Sea Turtle Yield Insights into the Development and Evolution of the Turtle-Specific Body Plan.” *Nature Genetics* 45 (6): 701–6.

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

Wyneken, Jeanette, Kenneth J. Lohmann, and John A. Musick. 2013. *The Biology of Sea Turtles, Volume III*. CRC press.

Zbinden, Judith A., Carlo R. Largiadèr, Fabio Leippert, Dimitris Margaritoulis, and Raphaël Arlettaz. 2007. “High Frequency of Multiple Paternity in the Largest Rookery of Mediterranean Loggerhead Sea Turtles.” *Molecular Ecology* 16 (17): 3703–11.

Zee, Jurjan P. van der, Marjolijn J. A. Christianen, Martine Bérubé, Mabel Nava, Sietske van der Wal, Jessica Berkel, Tadzio Bervoets, Melanie Meijer Zu Schlochtern, Leontine E. Becking, and Per J. Palsbøll. 2022. “Demographic Changes in Pleistocene Sea Turtles Were Driven by Past Sea Level Fluctuations Affecting Feeding Habitat Availability.” *Molecular Ecology* 31 (4): 1044–56.

Zou, S., N. Ke, J. M. Kim, and D. F. Voytas. 1996. “The Saccharomyces Retrotransposon Ty5 Integrates Preferentially into Regions of Silent Chromatin at the Telomeres and Mating Loci.” *Genes & Development* 10 (5): 634–45.

Zou, S., and D. F. Voytas. 1997. “Silent Chromatin Determines Target Preference of the Saccharomyces Retrotransposon Ty5.” *Proceedings of the National Academy of Sciences of the United States of America* 94 (14): 7412–16.

Zou, S., D. A. Wright, and D. F. Voytas. 1995. “The Saccharomyces Ty5 Retrotransposon Family Is Associated with Origins of DNA Replication at the Telomeres and the Silent Mating Locus HMR.” *Proceedings of the National Academy of Sciences of the United States of America* 92 (3): 920–24.

Statement of contribution to the articles:

1.- Bentley, B. P., **Carrasco-Valenzuela, T.**, Ramos, E. K., Pawar, H., Souza Arantes, L., Alexander, A., ... & Komoroske, L. M. (2023). Divergent sensory and immune gene evolution in sea turtles with contrasting demographic and life histories. *Proceedings of the National Academy of Sciences*, 120(7), e2201076120.

During the development of the project my main contribution was to assemble the high-quality genome of C. mydas at chromosome level for which I used data from four different technologies. I was also responsible for the detection and exploration of the regions of reduced collinearity. I wrote the scripts necessary to perform the gene expansion analysis, explored the TE content and contributed both to the main text writing and the conceptualization and drawing of the figures.

2.- **Carrasco-Valenzuela, T.**, Marins, L., Ramos, E., Suh, A., Mazzoni, C. (2023). Recent expansion of Penelope-like retrotransposons in the leatherback turtle Dermochelys coriacea. Submitted to Mobile DNA. Under review in *Mobile DNA.*

During the development of the project, I conceived the presented idea, developed the theory, wrote the necessary scripts performed the computations and verified the analytical methods. I also wrote the manuscript draft and designed all figures.

3.- **Carrasco-Valenzuela**,**T**., Ramos,E., Mazzoni C. (2023). Testudine-wide Transposable element exploration: A history of slow evolution and conserved genomes. In preparation (formatted for submission)

For this project too, I conceived the presented idea, developed the theory, wrote the necessary scripts, performed the computations and verified the analytical methods. I also wrote the manuscript draft and designed all figures.