THE UNIVERSITY *of* EDINBURGH

THE UNIVERSITY *of* EDINBURGH

Comparative genomics of mobulid rays using a novel West Atlantic pygmy devil ray (*Mobula hypostoma*) reference genome

Léa Soler-Clavel

Supervised by Dr Daniel Macqueen, Dr Emily Humble and Dr Manu Kumar Gundappa

MScR Genetics and Genomics

September 2024

Declaration

This project report is submitted in partial fulfilment of the requirements for the degree of MSc by Research Genetics and Genomics.

I declare that this thesis was composed by myself, that the work contained therein is my own, except where explicitly stated otherwise in the text, and that it has not been submitted, in whole or in part, for any other degree or professional qualification.

Acknowledgements

I thank my supervisors, Dan Macqueen, Emily Humble, and Manu Kumar Gundappa for being extraordinary mentors. Thank you for your support and patience, and everything I've learned this year.

Many thanks to Andy Law for his help and advice; to Zexin Jiao, Jianxuan Sun, Oliver Eve, Diego Perojil Morata, for their coding assistance; to Melissa Marr for guiding me through PSMC; to Donald Dunbar; and to the Aquaculture Labs scientists for the wealth of knowledge they shared.

Abstract

Elasmobranchs are a subclass of cartilaginous fishes comprising around 1200 species of sharks, rays and skates. A combination of low-reproductive history traits and threats such as habitat destruction, overfishing and climate change, has led to a dramatic decline in elasmobranch population numbers. In recent years, elasmobranch conservation and stock management have increasingly relied on molecular data to understand speciation, habitat use and population dynamics; simultaneously, the past decade has seen an increased recognition of elasmobranchs as an important deep-branching clade in comparative studies aimed at understanding vertebrate evolution. Therefore, a growing body of high-quality, whole-genome sequences have been assembled for several elasmobranch species across most orders, including mobulid rays.

Mobulid rays, manta and devil ray species of genus *Mobula*, are highly derived batoid fishes with unique morphological features including cephalic lobes framing a forward-facing mouth; in contrast to ancestral ray and skate lineages, who are typically adapted to benthic environments, mobulids have evolved to expand into pelagic niches. Despite their endangered status, mobulid rays remain poorly studied and understood among cartilaginous fishes, including the genomic basis for their ecological adaptations and possible susceptibilities to environmental change. Therefore, the assembly of a novel chromosome-level, high-quality reference genome for the Atlantic pygmy devil ray, *Mobula hypostoma*, provides a unique opportunity to investigate mobulid evolutionary history and adaptation, and support future conservation studies.

Whole-genome alignment of *M. hypostoma*, with its sister species oceanic manta ray (*M. birostris*), Atlantic stingray (*Hypanus sabinus*) and two skate (*Leucoraja erinacea* and *Amblyraja radiata*) reference assemblies, was carried out using Cactus. De novo repeat libraries were constructed using RepeatMasker to characterise and compare the repeat landscape of each aligned species. Synteny analysis, homology relationships and structural rearrangements between these genomes were investigated to highlight mobulid-specific mutations. Functional annotation of mobulid – and *M. hypostoma* – specific genomic rearrangements was carried out to investigate their effects on genes and phenotype.

Results showed one-to one chromosome homology was mostly maintained between the two mobulid species, reflecting their recent divergence; in contrast, more substantial rearrangements were observed between mobulids and stingray, mostly chromosomal fusions. Repeats accounted for most of the genome content for all species (60-65%), with *M. hypostoma* possessing the highest genome content across all five batoids, and the highest percentage of DNA transposons and retroelements. Genomic regions lost in the mobulid ancestor directly impacted or were located near to many candidate genes associated with vertebrate craniofacial development, including *alx3*, *fgf4*, *foxe1* among others. Further research will be needed to determine if these genomic changes were responsible

Lay Summary

Sharks and rays are an ancient group of vertebrate species, nowadays endangered by a wide array of threats such as habitat destruction, overfishing and climate change. Most sharks and rays have few pups and take longer to mature than other species, making them particularly vulnerable. In recent decades, conservation and management of these species have increasingly relied on the study of genes and whole genomes (the complete set of genetic sequences of an organism). Sharks and ray genomes have also increasingly been compared to other vertebrates to ask questions about genetic evolution. Recent improvements in technology have made the sequencing of shark and ray whole genome sequences more routine, opening up a range of potential investigations.

In contrast to older families of rays and skates, which live on the sea floor, the mobulid family (*i.e.*, devil rays, including the iconic manta rays and relatives) have evolved to inhabit open water. Although highly endangered, mobulids are less studied compared to other sharks and rays, meaning the genetic basis for their unique adaptations is largely unexplored. Therefore, the assembly of a new high-quality whole genome sequence for the West Atlantic pygmy devil ray provides the opportunity to fill key knowledge gaps in our understanding of mobulid biology and evolution.

The whole genome of the West Atlantic pygmy devil ray was compared to that of the oceanic manta ray, the Atlantic stingray and two skate species. This allowed me to identify genetic changes unique to the mobulid family. Interestingly, many genomic regions lost in the ancestor to the devil rays are genes involved in the development of facial features. Changes in genes controlling the development of the face and in particular the mouth could have helped the devil rays to adapt to a life in open water, although more research will be needed in this area.

Contents

Chapter 1. Introduction

I. Elasmobranchs are an ancient vertebrate lineage

Elasmobranchs are a subclass of cartilaginous fishes, comprising sharks, rays and skates (Kuraku, 2021). Together with the holocephalans (chimeras), they constitute class Chondrichthyes, one of the oldest vertebrate lineages and sister group to all extent bony vertebrates (Osteichthyes). Chondrichthyan divergence from Osteichthyes is estimated to have occurred 450 million years ago (Mya) (Irisarri *et al.*, 2017). Studying elasmobranchs (and chondrichthyans) therefore presents a key opportunity to understand jawed vertebrate evolution. Research on elasmobranchs is also essential to understand the evolutionary history, biological adaptations and ecology of a key vertebrate clade.

Elasmobranchs count approximately 1200-1300 species and are distributed all around the globe. They have some of the greatest diversities in morphology, body sizes and reproductive strategies among vertebrates, and occupy a range of trophic levels, from apex predators for great white sharks (*Carcharodon carcharias*), to filter feeding and benthic durophagy (Marra *et al.*, 2019). Many elasmobranchs are keystone species within their food webs, acting as top-down controls on prey population or nutrient redistributors, and are widely recognised as being both drivers and indicators of marine ecosystem health (Kuraku, 2021).

Although they have endured for over 400 million years, many, if not most elasmobranch species are now threatened by a plethora of anthropogenic stressors such as fishing pressure, bycatch, habitat destruction, climate change, among others (Dulvy *et al.*, 2014). Despite increased advocacy for conservation efforts, as of 2020 nearly half of elasmobranch species were 'data deficient' as per the International Union for the Conservation of Nature's (IUCN) records (Johri *et al.*, 2019).

In parallel, the growing interest in cartilaginous fishes to investigate vertebrate evolution, including the origin of vertebrate traits, has grown considerably (Dudgeon *et al.*, 2012; Hara *et al.*, 2018). This has led to an increase in whole-genome sequencing initiatives to establish and annotate complete elasmobranch nuclear DNA sequences, and in turn uncover clues on genome evolution, including the basis for adaptation (Pearce *et al.*, 2021). Over the past

decade, this has transformed the elasmobranch genomics landscape from a dearth of resources to a growing catalogue of assemblies of ever-improving quality, led by consortia such as Squalomix and the Vertebrate Genome Project (Rhie *et al.*, 2021; Nishimura *et al.*, 2022).

II. The current landscape of elasmobranch genomics

i. History of elasmobranch genome sequencing

Towards the end of the 2010's, the elephant shark's genome (*Callorhinchus milii*) was the first cartilaginous fish nuclear genome to be assembled (Ravi *et al.*, 2009) using next-generation sequencing techniques. For the next decade, and despite the development of high-throughput sequencing approaches and improvements in assembly tools, the elephant shark genome remained the only chondrichthyan nuclear genome available (Read *et al.*, 2017; Hara *et al.*, 2018; Supple and Shapiro, 2018). Several issues were associated with this: holocephalans and elasmobranchs diverged over 400 Mya, shortly after the cartilaginous and bony fish split, and are thus highly genetically distinct lineages. The *C. milii* genome therefore provides limited use as a reference point for elasmobranchs genomes. *C. millii* was originally chosen for its relatively small genome size (1.9 Giga base-pairs [Gb]), while chondrichthyan (Irisarri *et al.*, 2017), and in particular elasmobranch, genomes vary greatly in size within and between clades, and globally have the largest genomes among living vertebrates (Kuraku, 2021). The elephant shark's small genome size is believed to be due to erosion over time, which should have led to the loss of ancestral chondrichthyan features (Hara *et al.*, 2018). As a consequence, the exclusive use of *C. milli* in comparative studies was never sufficient to understand the evolution and biology of cartilaginous fishes as a lineage, much less elasmobranchs, which are evolutionarily separated from holocephalans by hundreds of millions of years (Redmond, Macqueen and Dooley, 2018).

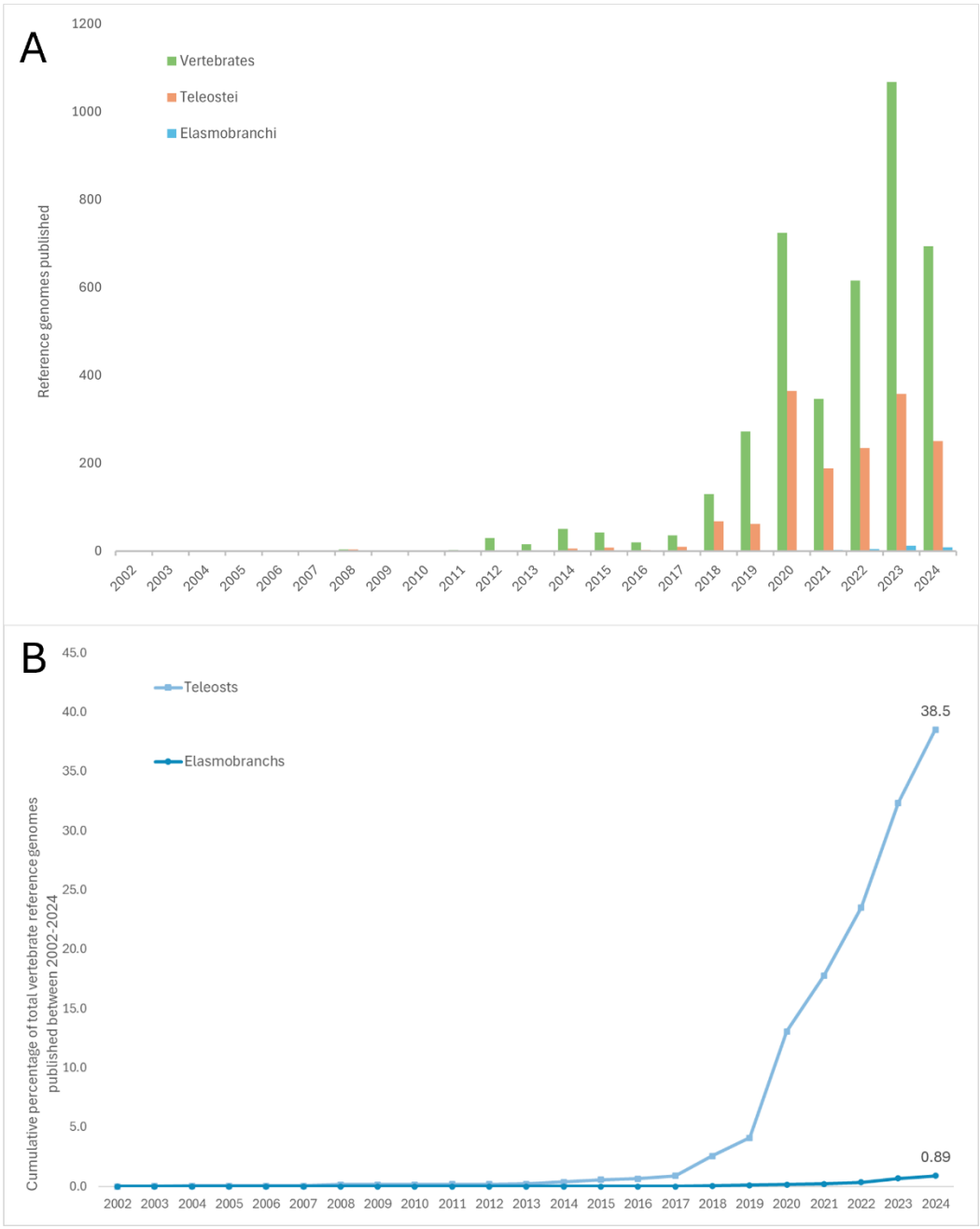


Figure 1. Yearly published elasmobranch reference genomes: **a**) number of available reference genomes for all vertebrates including teleosts and elasmobranchs, compared to teleosts and elasmobranchs only; **b**) percentage of total vertebrate reference genomes (including elasmobranchs and teleosts) published represented by each lineage.

In recent years however, elasmobranch whole-genome sequencing projects have multiplied (Hara *et al.*, 2018; Marra *et al.*, 2019; Weber *et al.*, 2020; Marlétaz *et al.*, 2023). As of July 2024, NCBI GenBank holds 35 elasmobranch reference genomes. These include 22 genomes assembled at chromosome-level, of which 11 have a RefSeq annotation record. Twenty of

these 35 reference genomes were released between January 2023 and July 2024, including 15 of the chromosome-level assemblies (Table 1), evidencing a recent ‘explosion’ of elasmobranch high quality genome resources.

Table 1: Available elasmobranch reference genomes on NCBI as of 25 August 2024

Species	Date published	Submitter	Genome size	Scaffold N50	Conting N50	Coverage	Sequence Method	GenBank accession	Reference	Annotation
***Mobula hypostoma* (West Atlantic pygmy devil ray)**	**06/01/2024**	**Centro Nacional de Análisis Genómico**	**3.5 Gb**	**152.4 Mb**	**23.3 Mb**	**100.0x**	**ONT, Illumina, OmniC**	GCA_963921235.1	**NA**	**RefSeq**
Mobula birostris (Oceanic manta ray)	22/05/2023	VGP	3.6 Gb	152.4 Mb	8.6 Mb	32.0x	PacBio Sequel II HiFi; Arima Hi-C v2	GCA_030035685.1	NA	NA
Hemiscyllium ocellatum (Epaulette shark*)*	05/11/2021	VGP	4 Gb	83.6 Mb	8.8Mb	148.6x	PacBio Sequel II CLR; Illumina NovaSeq; Bionano Genomics DLS; Arima Genomics Hi-C v1	GCA_020745735.1	Sendell-Price *et al.*, 2023	NA
Pristis pectinata (Smalltooth sawfish)	19/12/2019	VGP	2.3 Gb	101.7 Mb	17 Mb	61.9x	PacBio Sequel I CLR; Illumina NovaSeq; Arima Genomics Hi-C; Bionano Genomics DLS	GCA_009764475.2	NA	RefSeq
Rhincodon typus (Whale shark)	03/02/2022	Squalomix	2.9 Gb	70.8 Mb	48.5 kb	46.7x	Illumina Hi-Seq	GCA_021869965.1	Weber *et al.*, 2019	RefSeq
Carcharodon carcharias (Great white shark)	30/03/2021	VGP	4.3 Gb	169.9 Mb	6.5 Mb	66.8x	PacBio Sequel I CLR; Illumina NovaSeq; Arima Genomics Hi-C; Bionano Genomics DLS	GCA_017639515.1	Marra *et al.*, 2018	RefSeq
Leucoraja erinacea (Little skate)	22/02/2023	Okinawa IoST	2.2Gb	57.2Mb	5.3 Mb	150.0x	Illumina; PacBio Sequel	GCA_028641065.1	Marlétaz *et al.*, 2023	RefSeq
Scyliorhinus canicula (Small spotted catshark)	17/12/2020	Wellcome Sanger Inst	4.2Gb	198.8Mb	1.9 Mb	63.0x	PacBio data, 43x 10X Genomics Chromium data, and BioNano data	GCA_902713615.1	NA	RefSeq
Okamejei kenojei (ocellated spot skate)	12/01/2024	BGI Qingdao	2.8 Gb	74 Mb	1.7 Mb	247.8X	PacBio	GCA_035221665.1	NA	-
Chiloscyllium plagiosum (Whitespotted bamboo shark)	10/01/2019	BGI (Beijing Genomic Institute)	3.8 Gb	72.1 kb	37.2 kb	280.0x	BGIseq-500	GCA_004010195.1	NA	RefSeq
Amblyraja radiata (Thorny skate)	21/02/2020	VGP	2.6 Gb	62.1 Mb	1.5 Mb	128.2x	PacBio Sequel; Illumina NovaSeq; Arima Genomics Hi-C; Bionano Genomics DLS	GCA_010909765.2	NA	RefSeq
Squalus acanthias (Spiny dogfish)	07/07/2023	Nord Uni	3.7Gb	124.1Mb	1.7 Mb	80.0x	PacBio Sequel	GCA_030390025.1	Wagner *et al.*, 2023	Figshare
Raja brachyura (Blonde ray)	28/09/2023	Wellcome Sanger Inst	2.7 Gb	68.4 Mb	2 Mb	33.0x	PacBio, Arima2	GCA_963514005.1	NA	NA
Stegostoma tigrinum (Zebra shark)	08/08/2023	VGP	3.2Gb	79Mb	6.2 Mb	48.7x	PacBio Sequel II HiFi	GCA_030684315.1	NA	RefSeq
Prionace glauca (Blue shark)	08/04/2024	Shanghai Ocean University	3.7 Gb	109.8 Mb	5.3 Mb	35.0x	PacBio	GCA_037974335.1	NA	NA
Hypanus sabinus (Atlantic stingray)	05/06/2023	VGP	4Gb	112.3 Mb	4.3 Mb	34.2x	PacBio Sequel II HiFi; Arima Hi-C v2	GCA_030144855.1	NA	RefSeq
Mustelus asterias (starry smooth-hound)	22/08/2024	Wellcome Sanger Inst		109 Mb	1.2 Mb	45.5X	PacBio, Arima2	GCA_964213995.1	NA	NA
Narcine bancroftii (Carribbean electric ray)	06/05/2024	VGP	3.1 Gb	258.9 Mb	5 Mb	50x	PacBio Revio HiFi; Arima Hi-C v2	GCA_036971175.1	NA	NA
Cetorhinus maximus (Basking shark)	28/06/2024	Wellcome Sanger Inst	4Gb	126 Mb	2 Mb	34.0x	PacBio, Arima2	GCA_964194155.1	NA	NA
Malacoraja senta (Smooth skate)	17/04/2024	Canada's Genomic Enterprise	1.9Gb	58.7 Mb	150.1 kb	8.7X	PacBio Sequel; Illumina NovaSeq	GCA_038087875.1	NA	NA

ii. Comparative elasmobranch genomics studies

With the development of third-generation platforms that produce very long, high-accuracy sequencing reads (*i.e.*, Nanopore and PacBio) (Rhoads and Au, 2015; Wang *et al.*, 2021) and the associated drop in cost and time requirements for sequencing, the assembly of high-quality whole-genome sequences has become feasible for many non-model species, including elasmobranchs (Marra *et al.*, 2019; Pearce *et al.*, 2021; Tan *et al.*, 2021; Nishimura *et al.*, 2022). The publication of high-quality elasmobranch genome assemblies with annotations describing the location of genes and other functionally important elements, provides new opportunities to study the evolution of vertebrate genome structure and reveal the genetic basis for adaptations. Comparative genomics encapsulates the alignment and comparison of high-quality genomic sequences from different species, as well as individuals within the same species, to answer evolutionary and phenotypic questions (de Crécy-Lagard and Hanson, 2013; Nakatani *et al.*, 2021). The use of chromosome-level assemblies in such work greatly facilitates investigations of genome structural evolution, including conservation and changes in gene proximity and order (synteny) (Sacerdot *et al.*, 2018; Zhang *et al.*, 2020; Wu *et al.*, 2022; Li and Durbin, 2024).

iii. High-quality genome assemblies are becoming the norm

The achievement of high-quality reference genome assemblies hinges on several standards from sampling to bioinformatics pipelines. High-molecular weight DNA must be extracted, along with flash-frozen tissue samples to support annotation via RNA sequencing (Wong *et al.*, 2012; Pearce *et al.*, 2021). Sufficient sequencing depth, coupled with the use of long-read sequencing technologies, gives most potential to assemble full chromosomes with few gaps in the sequence (Pearce *et al.*, 2021; Kadota *et al.*, 2023). This is particularly crucial for elasmobranch genomes, as their high repeat content and large genome sizes have historically made it challenging or impossible to achieve chromosome-level assemblies (Kadota *et al.*, 2023). Improvements in long-read sequencing accuracy and assembly technologies over the past few years are now enabling near complete (*i.e.*, 'telomere-to-telomere') whole-genome

assemblies (Li and Durbin, 2024). The availability of high-quality annotations for genome assemblies is also crucial for comparative studies. Genome annotation seeks to identify sequences of functional importance within the DNA sequence, including coding genes (Weisman, Murray and Eddy, 2022). This in turn allows researchers to make inferences on changes in gene function or expression during evolution, and how this is linked to traits of interest (de Crécy-Lagard and Hanson, 2013).

iv. Comparative elasmobranch studies provide insights into vertebrate evolution

Comparative genomic investigations using elasmobranch genome sequences have shed light into vertebrate evolution and the genomic underpinnings of this clade's adaptations and successful radiation into a wide range of environments (Weisman, Murray and Eddy, 2022). A selection of important comparative studies leveraging elasmobranch genomes are briefly reviewed in the following sections.

a. A first cartilaginous fish comparative study using the elephant shark genome

In 2014, Venkatesh and colleagues conducted the first whole-genome comparative analysis of a cartilaginous fish, using their elephant shark whole-genome assembly (Venkatesh *et al.*, 2014). The researchers found lower levels of inter-chromosomal rearrangement between *C. milii* and tetrapods compared to the levels between *C. milii* and ray-finned fishes. Consistently, synteny was also well conserved between the *C. milii*, chicken and human genomes. The researchers also compared orthologous genes in *C. milii*, tetrapods and teleost fish, and found that teleost and elephant shark genes lost in tetrapods were associated with developmental processes linked to an aquatic lifestyle, such as fin and lateral line development. In addition, they found that *C. milii* retained orthologues from many gene families involved in bone formation, except the secretory calcium-binding phosphoprotein family. Through this study, the authors highlighted the genomic basis for skeletal ossification observed in the bony fish lineage, illustrating the power of comparative genomics.

b. The first elasmobranch whole-genome comparative analyses

After about a decade of reliance on the *C. milii* reference genome, Hara et al. (2018) reported high quality genome assemblies for three elasmobranchs: namely, two members of Hemiscyllidae (Orectolobiformes), the brownbanded bamboo shark (*Chiloscyllium punctatum*) and the small spotted catshark (*Scyliorhinus torazame*)(Hara *et al.*, 2018). Despite their relatively small body sizes, both species had large genome sizes (3.38 and 6.47 Gb respectively). When investigating the drivers of large genome size, the researchers found this to result from an accumulation of repetitive sequences alongside expansion of intron and intergenic regions across the genome. No expansion of gene or protein-coding sequences were found to explain genome size increases, nor any evidence of whole-genome duplication such as the event reported during early teleost evolution (Amores *et al.*, 1998; Meyer and Van De Peer, 2005; Pasquier *et al.*, 2016).

This study also used the four vertebrate Hox gene clusters (A, B, C, D), essentially involved in the development of the embryonic body plan, to illustrate the perils of drawing conclusions on genome evolution, when comparing a small number of representative species and incomplete data. In previous research relying on local assemblies of Hox cluster sequences or low-coverage short-read sequencing of elasmobranch genomes, genes from the HoxA, B and D clusters were identified, while no HoxC orthologues were recovered from *S. canicula* and *L. erinacea* (King *et al.*, 2011). In contrast, all Hox clusters, including HoxC, were identified in the *C. milii* genome (Ravi *et al.*, 2009). Researchers had therefore previously concluded that the HoxC cluster was lost in elasmobranchs (King *et al.*, 2011). However, Hara and colleagues recovered genes belonging to the HoxC cluster, changing the previous paradigm and suggesting the resolution of previous analyses were insufficient to detect these genes (Pearce *et al.*, 2021). Finally, the authors identified putative orthologues of genes of the mammal hypothalamo-pituitary axis involved in reproduction and feeding, therefore supporting an origin of these pathways in the vertebrate ancestor.

c. Genomic underpinnings of DNA maintenance and olfaction

In 2019, Marra et al. assembled a *C. carcharias* reference genome and conducted comparative analysis using several model vertebrate genomes, alongside the whale shark (*Rhincodon*

typus) and elephant shark genomes (Marra *et al.*, 2019). The authors identified several genes involved in DNA stability and repair to be under positive selection in all three cartilaginous fish genomes. Positively selected genes shared by both elasmobranch species included the p53 regulator RRS1 (Ribosome biogenesis regulatory protein homolog) and the tumour suppressor PDCD4 (Programmed cell death protein 4) involved in apoptosis pathways. The authors also revealed an enrichment of positively selected genes involved in wound healing, thereby suggesting a molecular basis for enhanced wound healing capabilities observed in sharks. Marra and colleagues' results interestingly did not recover key olfactory genes expected in apex predators such as the great white shark. Instead, their results showed multiple copies of a specific olfactory gene, V2R (Vomeronasal type 2), in both the great white and whale shark genomes (13 and 10 copies, respectively). The authors therefore proposed these multiple copies may underpin the known olfactory capabilities of predatory elasmobranchs.

d. Retracing the evolution of vertebrate immune genes

Comparative studies using elasmobranch genomes have further led to a better understanding of vertebrate immune system evolution. Several such studies have investigated the evolutionary history of the vertebrate adaptive and innate immune system, with both present in cartilaginous fishes (Tan *et al.*, 2021). Tan and colleagues (2021) investigated the evolution of the innate immune system, focusing on key pathogen recognition receptors (PRRs) involved in pathogen detection, which initiate immune responses. Using a new whale shark genome assembled using long-read sequencing, the authors found that immune-related genes were enriched among gene families that arose in the most recent common jawed vertebrate ancestor. In addition, they established differences in the evolution of distinct PRR families, with toll- and nod-like receptors having diversified extensively across vertebrate clades, while rig-like receptors have been strongly conserved across jawed vertebrate evolution (Tan *et al.*, 2021).

Two years later, Veríssimo and colleagues (2023 investigated the evolution of major histocompatibility complex (MHC) genes. The MHC is a cluster of linked genes involved in adaptive immunity, supporting the presentation of pathogen components to immune cells (Sambrook, Figueroa and Beck, 2005; The MHC Consortium, 1999). The researchers compared chromosome-level assemblies from elasmobranchs and *C. milii* for genes homologous to the

MHC. They found most major MHC genes in elasmobranchs were in close linkage, and in particular MHC Class Ia genes. The authors also highlighted that MHCIa and MHCII regions were linked in all analysed elasmobranch genomes, and were similarly linked in reedfish, lungfish and humans but not in teleosts. Therefore, the authors inferred MHCIa and MHCII linkage likely represented the ancestral state for jawed vertebrates. Furthermore, the authors found no elasmobranch MHCI and II genes outside of the region homologous to human MHC, supporting the hypothesis that all basal MHC genes are linked and separated through rearrangements in certain lineages.

e. Elucidating the evolution of vertebrate genome structure

The study of elasmobranch genomes has also shed light on the evolutionary history of vertebrate karyotype organisation, including the evolution of sex chromosomes. Yamaguchi et al. (2023) investigated chromosome organisation by comparing whale shark and zebra shark (*S. tigrinum*) chromosome level genomes (Yamaguchi *et al.*, 2023). Their work highlighted that chromosome organisation was globally maintained despite the 50 million years or so of evolutionary distance separating these species. Moreover, the level of conserved chromosome organisation between zebra shark and its much more distant relative thorny skate (*Amblyraja radiata*) was higher than between pairs of bony vertebrate genomes separated by the same evolutionary distance, suggesting a higher conservation of karyotype organisation in cartilaginous fishes. The authors also confirmed that elasmobranch sex chromosomes are XY, and characterised the X chromosomes in both shark genomes analysed; their results revealed the X chromosome bore the HoxC cluster and was homologous to human chromosome 12 and chicken chromosome 34. Yamaguchi and colleagues' research points towards a slower rate of change in elasmobranch chromosome architecture compared to other vertebrates.

Building on this work, (Wu *et al.*, 2024) recently expanded the above comparative analysis of karyotype organisation and sex chromosomes study using six high-quality elasmobranch assemblies. They confirmed the XY sex chromosome system and slow rate of chromosomal organisation evolution observed by Yamaguchi and colleagues. They found that all elasmobranchs included possessed the X chromosome, and placed its origin at a minimum of

about 181 Mya. In addition, the authors found the Y chromosome was highly eroded in shark genomes, and that gene dosage was not compensated between X and Y chromosomes. They therefore posited that elasmobranch XY chromosomes shared a common origin in sharks, although whether this is also the case for skates, which sharks diverged from around 230-240 Mya, remains to be assessed. The researchers also reconstituted the ancestral chondrichthyan karyotype and identified 18 pairs of micro-chromosomes, which they found was relatively conserved down the vertebrate evolutionary tree, albeit expanded in avian lineages.

III. Genomics and elasmobranch conservation research

Although limitations still exist to the availability of high-quality reference genome sequences for many non-model taxa, the fields of genome biology and bioinformatics have opened valuable research avenues, including for wildlife conservation research. Over the past couple of decades, advances in DNA sequencing and genome assembly technologies have enabled the study of vertebrate genomes at unprecedented scales, presenting opportunities for conservation research.

i. Population genomics

a. Identifying population structure and connectivity to define conservation units

Population genomics allows the study of population structure and connectivity by comparing the genomes of multiple individuals of a species (Supple and Shapiro, 2018; Johri *et al.*, 2019; Hohenlohe, Funk and Rajora, 2021). Population genomics is particularly helpful in marine habitats, where monitoring population size and structure at regional and global scales is often challenged by migration within and between ocean basins, vertical movement in the water column, and the remoteness of most of the open ocean (Oleksiak and Rajora Editors, no date). In the case of elasmobranchs, tag-and-release, bycatch inventories, or drone observations may provide information on species ranges and behaviours at a specific time and place, prompting further investigations, but are unlikely to support understanding of long-term habitat use, population dynamics, or social behaviour (Benestan, 2019). Using bycatch data to gain a sense of elasmobranch abundance at a local scale may hint at increases or declines in population numbers at a given time, but is insufficient to draw a precise demographic

picture, or to determine the complex dynamics driving abundance and distribution patterns (Benestan, 2019). For example, Brunjest et al., (2023) recently found that great hammerhead shark populations around Australia showed little fragmentation, supporting their management as a single stock. Population genomics therefore shows great potential in defining conservation units in elasmobranchs (Brunjes *et al.*, 2024). Genomics can therefore be used to define units of interest for conservation and management. Marine species face many local and global stressors, such as overfishing, habitat destruction, and ocean deoxygenation (among others) (Field *et al.*, 2009; Dulvy *et al.*, 2014, 2017; Lawson *et al.*, 2017). The magnitude and interaction of these threats may vary between regions and context; therefore, defining management units at too low or too high a resolution may diminish the value of conservation efforts. Over-fragmenting a population that should be managed as a unit, for example, may lead to overlooking groups that should be managed together, while conflating distinct populations together might lead to missing important variation and population mismanagement (Humble *et al.*, 2023). Phylogenomics allows to elucidate population structure through analyses of ancestry, speciation and introgression or hybridisation between species and populations, increasingly providing conservation applications at local and global scales (Segelbacher *et al.*, 2022).

b. Characterising the genetic health of a population

Genetic diversity is often used as a proxy for a population's capacity for resilience to change (Allendorf, W., Luikart and Aitken, 2013; Hohenlohe, Funk and Rajora, 2021). Population health parameters associated with genetic diversity, such as inbreeding depression and genetic drift, are tightly linked to population size and effective population size (Wang *et al.*, 2016; Hohenlohe, Funk and Rajora, 2021). Effective population size is defined as a hypothetical population size with the same rates of genetic diversity and drift as the population studied, or in practice the number of individuals potentially passing their genes to offspring (Waples *et al.*, 2022). As an example, Delaval and colleagues (2021) investigated the effective population size of blue skate populations across European waters to infer their vulnerability to extinction (Delaval *et al.*, 2022). The researchers found that although gene flow was maintained between Celtic Sea and Scottish blue skate populations, the effective population size in Scotland, like that of the Faroe Islands, was low enough to be of concern for population sustainability

(Delaval *et al.*, 2022). Population genomics approaches can therefore be used to characterise a population's vulnerability to extinction through genetic health (Hohenlohe, Funk and Rajora, 2021).

c. Historical demographic modelling

In addition to establishing recent demographic status, genomics modelling tools have been developed to reconstruct historical population levels and trends over thousands to millions of years. Historical fluctuation in effective population size can help infer past evolutionary events and changes in genetic diversity, and therefore contextualise the modern genetic health of a species (Mather, Traves and Ho, 2020; Hohenlohe, Funk and Rajora, 2021). Reconstructing demographic history also allows population dynamics and genetic diversity to be contextualised in relation to other available historic data, including environmental trends such as mean annual surface temperature, land mass movement and CO_2 levels. By comparing historical effective population levels to concurrent environmental changes, inferences can be made regarding the drivers of population declines or expansion, thereby informing predictions of future population size changes with respect to climate change and anthropogenic disturbances (Morin *et al.*, 2021; Walsh *et al.*, 2022). Some widely used approaches for demographic history modelling in a species of interest are Pairwise or Multiple Sequential Markovian Coalescent modelling (PSMC and MSMC respectively) (Li and Durbin, 2011; Mather, Traves and Ho, 2020). Markovian coalescent models reconstitute recombination events from genomic data in order to estimate past population sizes (Mather, Traves and Ho, 2020). For example, Zhao and colleagues (2022) used PSMC to identify three periods of population increase for the White Spotted Bamboo shark (Zhao *et al.*, 2021). In contrast, by overlying geological data with PSMC data, Song et al., (2023) highlighted that the ocellate spot skate effective population increase coincided with the last glacial period and peaked around the last glacial maximum (Song *et al.*, 2023). Historical elasmobranch demography study can therefore provide a better understanding of the forces that affect modern population size declines to support management.

ii. Forensics and traceability genomics tools

Another genomics approach supporting conservation and management, with strong relevance to elasmobranchs, forensics (O'Bryhim, Parsons and Lance, 2017; Feitosa *et al.*, 2018). Shark and ray species are harvested by the millions every year by artisanal and larger fishing vessels alike, often to collect fins and gill plates exported to Asian markets and consumed as delicacies or traditional medicine products (Pardo *et al.*, 2016; O'Malley *et al.*, 2017; Humble *et al.*, 2023). Forensics relies on sampling these products at different stages down the supply chain, from fisheries stalls or retail stores, and analysing them using reliable traceability tools based on genome-wide markers; emerging approaches allow this data can be used to map fisheries products to specific populations and/or basins, thereby supporting the identification of captures from illegal fishing zones, the re-evaluation of mortality rates for populations of concern, as well as the definition of regions of conservation priority for fishing regulation (Zeng *et al.*, 2016).

iii. Conservation genomics – additional applications

Whole-genome studies beyond a few selected gene loci of interest also allow researchers to identify genomic regions linked with particular traits, predict environmental niches and responses to abiotic factors, and ultimately discover phenotypes and genomic potential linked to fitness and adaptability (Layton *et al.*, 2021; Bernos *et al.*, 2023). Whole genome studies can also create insights into regulatory and genetic pathways by uncovering genomic regions functionally linked to adaptive traits (Khan *et al.*, 2016).

IV. Mobulid rays – background and significance

i. Mobulid taxonomy, distribution and ecology

a. *Mobula*, a charismatic genus

The focus of my MScR investigations, the mobulid rays, are batoids of the genus *Mobula* (Mobulidae), representing close relatives to stingray clades of the Myliobatiform order. They are highly morphologically derived, notably due to their wing-like fins as well as the characteristic cephalic lobes extending anteriorly from either side of the mouth like horns,

from which their common name of “devil rays” originates (Notabartolo di Sciara, 1987; Couturier *et al.*, 2012). Mobulids are medium- to large-sized rays reaching, depending on the species, between 2 and 7 m disc width (the maximum distance between the distal edges of the pectoral fins) (Serra-Pereira *et al.*, 2010). Highly migratory, these rays occur circumglobally in tropical to warm-temperate waters, filter-feeding on zooplankton in the upper layers of the water column (Notabartolo di Sciara, 1987; Couturier *et al.*, 2012; Stevens *et al.*, 2019).

Though under continuous debate, the current *Mobula* taxonomy recognises nine species: two species of manta rays (*Mobula birostris* and *M. alfredi*, respectively the giant Oceanic manta ray and the reef manta ray), the largest species; and seven species of devil rays, *M. mobular* (the giant devil ray), *M. tarapacana* (sicklefin devil ray), *M. thurstoni* (bentfin devil ray), *M. eregoodoo* (longhorned pygmy devil ray), *M. munkiana* (Munk’s pygmy devil ray), *M. kuhlii* (shorthorned pygmy devil ray) and *M. hypostoma* (Atlantic pygmy devil ray) (White *et al.*, 2018; Hosegood, Humble, Ogden, Bruyn, *et al.*, 2020).

Although research on mobulids has increased over the past few years, this group of rays remains relatively poorly studied compared to other elasmobranch clades (Stewart *et al.*, 2017, 2018; Notarbartolo di Sciara *et al.*, 2020; Harris *et al.*, 2024). In addition, most data collected on life history, feeding and reproduction has historically been based on *M. alfredi* due to its coastal lifestyle and therefore greater accessibility for data collection (Stewart *et al.*, 2017). Nevertheless, a slowly growing body of knowledge on mobulid ecology has been collected by researchers, which is summarised in the following section.

b. Habitat use and niche partitioning, trophic ecology and reproduction

Mobulids occur in small populations or sub-populations, distributed across their wide range (Stewart *et al.*, 2018; Harris *et al.*, 2020). Field observations, fisheries and telemetry data have shown migration patterns between aggregation sites such as seamounts, atolls, coastal areas, coral reefs and cleaning stations (Couturier *et al.*, 2012; Stewart *et al.*, 2017). Species with overlapping ranges appear to co-occur, with evidence for slight differences in habitat use preferences, for example between mostly oceanic species such as *M. birostris, M. tarapacana* and *M. thurstoni* and predominantly coastal species like *M. alfredi* (Harris *et al.*, 2024; Ozaki *et al.*, 2024). A few studies have also suggested a segregation between adults and juveniles

and ontogenic niche changes, as regularly observed in other elasmobranchs (Notabartolo di Sciara, 1987, 1988; Stewart *et al.*, 2017). This may not be systematic however, as co-occurrence of juveniles and adults has been reported for *M. alfredi* in some regions and may also be the case for other mobulid species (Harris *et al.*, 2020).

As the waters at the latitudes inhabited by mobulids are mainly oligotrophic, their habitat use appears driven by food availability (Harris *et al.*, 2020); mobulids therefore congregate in areas where zooplankton density exceeds a certain threshold (Moriarty and O'Brien, 2013; Stewart *et al.*, 2017; Armstrong *et al.*, 2021). In the Maldives, local manta ray populations were found to undertake a bi-annual migration associated with primary productivity levels, which support zooplankton abundance (Harris *et al.*, 2020). Water temperature is also a significant factor in mobulid distribution, and was proposed as the main driver of habitat use in some studies (Ozaki *et al.*, 2024). Mobulids have been observed in cooler waters coinciding with upwellings or high tides increasing local zooplankton abundance, but clearly favour warmer temperatures in the 20-26 degrees C range (Couturier *et al.*, 2012). Telemetry data for *M. alfredi* in Western Australia and *M. japanica* in the Eastern Pacific Ocean have shown a diel change in water depth occupancy, suggesting some mobulid species forage in offshore, deeper waters at night and rest in the sunlit, surface layers during the day (Croll *et al.*, 2016; Couturier *et al.*, 2018). As ectotherms, mobulid physiology depends on the water temperature, which would support a similar preference for relative warmer waters when not feeding for most mobulid species (Ozaki *et al.*, 2024). Overall, mobulids have been found to spend the majority of their life the top 50 meters of the water column (Croll *et al.*, 2016).

c. Life history and reproduction

Mobulids have some of the most conservative life history traits not only among elasmobranchs, but among all fish species (Dulvy *et al.*, 2014; Croll *et al.*, 2016). Subtle differences exist between species, but overall characteristic traits including slow-maturing individuals, slow-growing populations and low fecundity are present among all mobulids (Rambahiniarison *et al.*, 2018). Reproductive ecology studies have found that mobulids mature at about five to ten years, depending on the species (Couturier *et al.*, 2012; Rambahiniarison *et al.*, 2018). Females reach maturity later than males, as evidenced by larger

body sizes at maturity despite a similar growth rate, although pregnancy onset is usually delayed (Rambahiniarison *et al.*, 2018). Gestation is relatively long compared to other batoids and lasts about a year (Notabartolo di Sciara, 1988; Rambahiniarison *et al.*, 2018). In addition, analyses of captured mature females suggested pregnancies were often separated by two- or three-year intervals (Marshall and Bennett, 2010; Dulvy *et al.*, 2014; Rambahiniarison *et al.*, 2018).

Mobulids usually give birth to a single pup or, very rarely, twins (Rambahiniarison *et al.*, 2018). The evidence suggests this low fecundity is tied to the high metabolic demands of mobulid biology and the associated trade-off of a pregnancy, followed by the live birth of a pup between 40-50% the size of its mother (Notabartolo di Sciara, 1988; Rambahiniarison *et al.*, 2018). As a consequence, population growth rates among mobulid species were estimated to be well below 5% a year, sometimes lower (Rambahiniarison *et al.*, 2018).

ii. Taxonomy and evolutionary history

The mobulids are members of the batoids, which last shared a common ancestor with the Neoselachii, *i.e.,* sharks, about 203-229 Mya (Aschliman *et al.*, 2012). The batoid superorder comprises four orders of elasmobranchs: the Rajiformes (skates and guitarfish) (Weigmann, 2016; Misawa, Babaran and Motomura, 2023), which diverged from the other lineages between 110-147 Mya; the Torpediniformes (torpedo, thornback and electric rays) (Claeson, 2014; Moreira and de Carvalho, 2021), which separated from the remaining two orders about 99-130 Mya; and finally the Rhinopristiformes (sawfishes, wedgefishes and guitarfishes) (Jabado, 2018) and Myliobatiformes (Dunn, McEachran and Honeycutt, 2003), a sister group that diverged from the mentioned orders about 98-130 Mya (Aschliman *et al.*, 2012; Villalobos-Segura and Underwood, 2020). Myliobatiformes includes all extant stingray families, including Mobulidae.

The mobulids are considered monophyletic, as supported by several phylogenetics studies (Poortvliet *et al.*, 2015; Hosegood, Humble, Ogden, Bruyn, *et al.*, 2020; Hosegood, Humble, Ogden, de Bruyn, *et al.*, 2020; Notarbartolo Di Sciara, Stevens and Fernando, 2020) and were proposed to have diverged from *Rhinoptera*, *i.e.*, the cownose ray family, from about 87-99 Mya (Villalobos-Segura and Underwood, 2020), to as recently as 30 Mya (Poortvliet et al.,

2015). Within *Mobula*, some taxa may have begun to diverge into two new species as recently as < 5 Mya (Poortvliet *et al.*, 2015). As a relatively "young" lineage, the low genetic difference in commonly used phylogenetic markers between these species has made mobulid phylogenetic reconstruction challenging (White *et al.*, 2018; Hosegood, Humble, Ogden, Bruyn, *et al.*, 2020). Coupled with a high degree of morphological resemblance between species and historical misidentifications, this has made taxonomical resolution a challenge within *Mobula* (Notarbartolo di Sciara *et al.*, 2020).

Taxonomy within Mobulidae is still debated and has undergone several revisions, as recently as the five to six years ago. The modern redescription of genus *Mobula* dates back 1987, when nine species of devil rays (excluding manta rays) were defined after synonymisation of several species based on morphological traits and geographical range data (Notabartolo di Sciara, 1987). *Manta* was still recognised as a separate genus from *Mobula* until very recently, and comprised a single recognised manta ray species until *Manta alfredi* was revived (Marshall, Compagno and Bennett, 2009). Marshall and colleagues (2009) also raised the possibility of a third putative manta ray species in the Gulf of Mexico. In 2018, an extensive phylogenetics study based on mitogenomic data provided evidence for manta rays sitting within *Mobula,* meaning *Manta* ceased to be recognised, with two species being renamed *Mobula birostris* and *Mobula alfredi* (Cracknell *et al.*, 2018). White and colleagues also provided molecular evidence for the hypothesised Gulf of Mexico manta species (*Mobula* cf *birostris*) and synonymised *M. 27aponica* with *M. mobular*, *M. rochebrunei* with *M. hypostoma*, and *M. eregoodootenkee* with *M. kuhlii. M. eregoodootenkee* was soon after revived under the name *M. eregoodoo,* as morphometric, ecological and molecular data supported its distinction from *M. kuhlii* ((Notarbartolo di Sciara *et al.*, 2020). Genome-wide double-digest restriction site-associated sequencing (ddRAD-Seq) markers further cemented the taxonomy established by White et al. (Hosegood, Humble, Ogden, Bruyn, *et al.*, 2020). As of 2024, genus *Mobula* includes seven species of devil rays (*M. mobular, M. munkiana, M. tarapacana, M. kuhlii, M. thurstoni*, *M. eregoodoo* and *M. hypostoma)* and two manta species (*Mobula birostris* and *M. alfredi*) (Cracknell *et al.*, 2018; Hosegood, Humble, Ogden, Bruyn, *et al.*, 2020).

iii. Ecological significance, tourism value, threats and conservation status

As large-bodied filter-feeders, mobulids play an important ecological role in marine food web (Araujo *et al.*, 2020). Nutrients not fixed into biomass production are excreted back into the water, making them available for uptake by primary producers, making mobulids significant nutrient recyclers (Le Mézo *et al.*, 2022). Through horizontal and vertical movement, mobulids may also recycle nutrients back into different locations in the water column, potentially creating concentrated zones of nutrient abundance at aggregation sites (Flowers, Heithaus and Papastamatiou, 2021). When they die, mobulid carcasses provide food to deep-sea communities, in a similar albeit more limited manner to whale carcasses (Flowers, Heithaus and Papastamatiou, 2021).

As elasmobranch-watching tourism developed, mobulids, and in particular manta rays, became a widely popular species to encounter in marine wildlife tourism activities (O'Malley, Lee-Brooks and Medd, 2013; Healy *et al.*, 2020). Their large size, approachability and slow, oscillatory swimming have made them "bucket list" species to encounter while diving, snorkelling or on sea safaris (O'Malley, Lee-Brooks and Medd, 2013). Mobulids thus represent a significant portion of marine-related tourism income in countries whose waters they inhabit, such as Indonesia, Sri Lanka, or the Maldives (O'Malley, Lee-Brooks and Medd, 2013; Murray *et al.*, 2020)(O'Malley et al., 2013, Murray et al., 2020). This "live value" provides an incentive to manage mobulid stocks and adopt protection policies, by making the capture and landing of mobulids illegal, as for manta species in Mozambique (see Marine Megafauna Foundation January 2024 press release, accessed 1 August 2024) or by designing no-take marine protected areas like the Chagos Archipelago (Venables *et al.*, 2016).

iv. Threats and conservation status

Dramatic population number declines have been reported for all *Mobula* species over the past decades (Croll *et al.*, 2016; Carpenter *et al.*, 2023). Despite the inclusion of manta rays in Appendix II of the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) since 2013, alongside the adoption of a few rare regional protection laws, mobulids are still heavily caught in targeted fisheries. Targeted mobulid fisheries are largely driven by demand for gill rakers on traditional Asian medicine markets, though no medicinal

efficacy was ever demonstrated for these products (Croll *et al.*, 2016; Zeng *et al.*, 2016; Humble *et al.*, 2023). Mobulid captures have also been reported in no-take marine protected areas (MPAs), and illegal, unregulated and unreported (IUU) fishing is likely to have led to underestimation of fishing-related mortality in mobulids (Haque *et al.*, 2021; Johnston *et al.*, 2024). In some regions of South East Asia, Central and South America, mobulids are also fished for consumption, as besides the gill plates, the body is considered of relatively low value; mobulids therefore represent an affordable source of proteins for artisanal fishermen (Croll *et al.*, 2016; Lawson *et al.*, 2017). Even when not targeted, slow swimming, surface-dwelling and aggregating behaviours make mobulids extremely susceptible to accidental capture by virtually all industrial and artisanal fishing equipment, and they are regularly caught as bycatch (Croll *et al.*, 2016; Lawson *et al.*, 2017).

In addition, devil and manta rays are vulnerable to a wide array of other anthropogenic threats, including pollution, habitat destruction, and climate change (Field *et al.*, 2009; Croll *et al.*, 2016; Waller *et al.*, 2024). As filter feeders with a large mouth opening, mobulids are exposed to plastic ingestions, from large plastic litter to microplastics (Couturier *et al.*, 2012; Stewart *et al.*, 2018). Warming waters may impact their metabolic rate and increase feeding demands to the detriment of other physiological functions like reproduction, or drive mobulids to shift their range to higher latitudes (Wheeler *et al.*, 2020). Climate change has also been linked to a decoupling, or trophic mismatch, between phytoplankton and zooplankton blooms: as surface waters warm, primary production may peak earlier in the year and decrease by the time zooplankton communities would have normally grazed on the bloom (Richardson, 2008). The resulting decrease in zooplankton abundance may lead to food scarcity for mobulids and modify their distribution and aggregative behaviour.

As a result of these pressures, coupled with extremely low reproductive output and population growth rate, dramatic population declines have been reported for all mobulid species in the past decades (Dulvy *et al.*, 2014; Lawson *et al.*, 2017). Even artisanal mobulid fisheries are considered unsustainable (Pardo *et al.*, 2016). All *Mobula* species are now classified as Endangered on the International Union for Conservation of Nature and Natural Resources (IUCN, accessed 5 August 2024) Red List, except *M. alfredi* and *M. munkiana*, which are listed as Vulnerable. Several of these statuses were updated from 'Threatened' to 'Endangered' in the past 12 months alone.

v. Challenges to mobulid data collection and availability

The depletion of manta and devil ray stocks has led researchers and conservation organisations to advocate for their urgent protection. Despite this, efficient management and conservation relies on adequate biological, life history and ecological data that are still lacking for most mobulids, and in particular devil ray species (Lawson *et al.*, 2017; Stewart *et al.*, 2018; Hosegood, Humble, Ogden, Bruyn, *et al.*, 2020). Although research efforts have increased over the past fifteen years or so, devil rays remain poorly studied and understood compared to manta species (O'Malley *et al.*, 2017). Devil ray species are largely offshore dwellers, making field studies difficult (Couturier *et al.*, 2012). Deck-based handling of live individuals for morphometrics measurements and sample collection is not an ideal solution, as post-release mortality rates are high; this is notably due to mobulids being obligatory ram ventilators with a high metabolic rate and a relatively un-rigid skeleton, making them ill-suited to withstand being out of the water (Stewart et al., 2018; Poisson et al., 2014). Embryological studies are also near impossible to carry out on mobulids since they are viviparous, unlike their egg-laying relatives (Couturier *et al.*, 2012). Demographic data collection has therefore historically largely relied on opportunistic surface sightings, citizen science and fisheries data (Couturier *et al.*, 2012; Ehemann *et al.*, 2022). However, historical identification inconsistencies, persisting taxonomical uncertainties and morphological similarities between devil ray species has led to erroneous data records(Ehemann *et al.*, 2022). As a result, this persisting dearth of data has considerably slowed down management for *Mobula* species (Lawson *et al.*, 2017). However, the afore-mentioned improvement in genomics technologies and increase in molecular data availability for elasmobranchs will likely help fill some of these knowledge gaps, as demonstrated by recent mobulid population genomics studies conducted using both mitochondrial and nuclear DNA, which have helped shed light on mobulid phylogeny (Poortvliet *et al.*, 2015; White *et al.*, 2018; Hosegood, Humble, Ogden, Bruyn, *et al.*, 2020).

vi. Genomic basis of the modified mobulid body plan

The characteristic mobulid morphology is integral to their filter-feeding and pelagic lifestyle. By unfurling their cephalic lobes when feeding, mobulids funnel zooplankton prey into their

mouths and through their gill rakers (Couturier *et al.*, 2012; Stevens *et al.*, 2019). Simultaneously, wing-shaped, laterally extended pectoral fins support oscillatory swimming, along with greater lift and mobility in the water column (Hall et al., 2018). These morphological modifications have been linked to mobulid adaptation to their ecological niche and feeding mode, in contrast to benthic batoid clades (Marlétaz *et al.*, 2023).

Until recent years however, no study had investigated the evolutionary genetic basis of these modifications (Swenson *et al.*, 2018). Swenson and colleagues investigated the developmental origin of cownose ray cephalic fins and wing-like fins, as well as differentially expressed developmental genes during embryonic fin growth. The researchers established that cephalic fins extend from the anterior part of the pectoral fin and becomes a distinct developmental domain through the creation of a "notch". However, the experiments were carried out on a cownose ray, *Rhinoptera bonasus*, as a Myliobatiform proxy for mobulids (Swenson *et al.*, 2018). Rhinopterids constitute the sister lineage to *Mobula* (Villalobos-Segura and Underwood, 2020). Like mobulids, they possess laterally extended, wing-like pectoral fins and paired cephalic lobes (Hall *et al.*, 2018). However, the cownose ray craniofacial organisation is arguably closer to that of the ancestral batoid body plan: the cephalic fins are smaller and extend about as far as the rostrum, and fuse anteriorly to the mouth (Hall *et al.*, 2018). In addition, the rhinopterid mouth is located on the ventral side of the head, while mobulids have forward-facing or near-forward-facing (also defined as sub-terminal) mouths (White *et al.*, 2018; Medeiros *et al.*, 2022). Therefore, mobulid rays have clearly undergone further developmental changes driving the evolution of their characteristic morphology, which requires further investigation in mobulid representatives specifically.

Comparative genomics provides the opportunity to reveal genomic evolutionary changes underlying unique mobulid adaptations, while shedding light on the evolutionary history of batoids leading to the derived mobulid body plan. In this respect, a chromosome-level reference assembly for *M. birostris* was released in 2023, alongside other batoids within both Myliobatiformes (*e.g.*, Atlantic stingray) and more distant orders like Rajiformes (*e.g.,* the little skate and thorny skate, Table 1; Figure 2). A new high-quality reference genome assembly and annotation for a second mobulid species, *M. hypostoma* (the West Atlantic pygmy devil ray) has become available in 2024, and made accessible for my project. This assembly opens the door to the study of potential lineage-specific genomic rearrangements underlying mobulid

morphology and adaptation through comparison with the manta genome to reconstruct the ancestral mobulid genome, and through comparison with other available batoid genomes.

vii. Mobula hypostoma

M. hypostoma, the West Atlantic pygmy devil ray, is a smaller devil ray species reaching between 1 and 2 m disc width (Notabartolo di Sciara, 1987; Couturier *et al.*, 2012; Ehemann, González-González and Trites, 2017)(Ehemann et al. 2017; Couturier et al., 2012; Notabartolo di Sciara, 1987). It has been found exclusively in Atlantic waters, and occurs along the coasts of the USA, the Carribbeans, Brazil, and as far South as Argentina (Notabartolo di Sciara, 1987; Couturier *et al.*, 2012; Bucair *et al.*, 2024). In 2018, *M. rochebrunei* was reclassified as a synonym of *M. hypostoma*, and therefore records of *M. rochebrunei* along the African Atlantic coasts are likely observations of *M. hypostoma* (Cracknell *et al.*, 2018; Hosegood, Humble, Ogden, Bruyn, *et al.*, 2020). *M. hypostoma* is often observed in relatively small aggregations of ten up to (occasionally) 40 individuals (Stevens, Hawkins and Roberts, 2018; Bucair *et al.*, 2024).

Most of this species' life history traits are poorly known, although single-pup pregnancies have been reported for several specimens in line with typical conservative mobulid life-history (Bucair *et al.*, 2024). *M. hypostoma* has historically not been directly targeted by mobulid fisheries, but is regularly caught as bycatch in artisanal and industrial fishing (Medeiros *et al.*, 2022; Bucair *et al.*, 2024). As a result, the conservation status of *M. hypostoma* was recently updated from Data Deficient to Endangered (Marshall, 2022).

vi. A novel Mobula hypostoma high-quality reference genome

Leading up to this project, a *Mobula hypostoma* genome sequencing project was completed, led by my co-supervisor Dr Emily Humble in collaboration with a broader team including my supervisor, Prof. Dan Macqueen. Whole blood and tissue samples were collected from a wild individual in the Florida Panhandle, USA, by Mote Marine Laboratory collaborators. Samples were sent to the *Centro Nacional de Análisis Genómico* (CNAG) in Barcelona, Spain for DNA extraction, sequencing, genome assembly and assembly curation. Briefly, CNAG conducted long-read sequencing using Oxford Nanopore Technologies (ONT) on the PromethION

platform. The ONT reads were assembled into contigs using NextDenovo (Hu *et al.*, 2024). Assembled haplotigs were removed using purge_dups (Guan *et al.*, 2020). Short-read sequencing libraries were prepared and sequenced using an Illumina NovaSeq 6000, which were used for polishing the contigs with HyPo (Kundu, Casey and Sung, 2019). Dovetail OmniC libraries were prepared and sequenced using an Illumina NovaSeq 6000 for scaffolding contigs into a chromosome level assembly using YaHS (Zhou, McCarthy and Durbin, 2023). To annotate the finalised *M. hypostoma* genome assembly (sMobHyp1.1, GCF_963921235.1), tissue samples from heart, gonad, spleen, kidney, liver, muscle, brain, intestine and gill were collected from a deceased aquarium female and unborn pup at the Georgia Aquarium, Atlanta (USA) and sent to Novogene for RNA sequencing (RNA-Seq). 750 million paired 150bp RNA-Seq reads were generated in total, and used by NCBI to produce the RefSeq annotation used in this project (GCF_963921235.1-RS_2024_02).

V. Project aims and objectives

My overarching project aim was to gain insights into mobulid evolution through comparative genomics of batoid chromosome-level genome assemblies including two mobulid members, including the genomic basis of the highly derived morphology and pelagic lifestyle of mobulids. I further aimed to exploit the new *M. hypostoma* genome as groundwork for future conservation genomics studies of mobulid rays.

These aims were addressed through the following objectives:

- Perform whole genome alignments of the novel *M. hypostoma* reference genome alongside the chromosome-level genomes of Oceanic manta, Atlantic stingray, thorny skate and little skate (Table 1).
- Compare genome architecture, synteny, and homology between aligned batoid genomes, and characterise ancestral mobulid and *M. hypostoma* specific genome structural rearrangements.
- Conduct *de novo* repeat content characterisation of the aligned genomes and establish variation in genome repeat composition across the batoid lineage.

- Reveal genes impacted by ancestral mobulid and *M. hypostoma*-specific sequence deletions, and conduct Gene Ontology enrichment analyses to uncover potential biological functions impacted during mobulid evolution, linking this back to mobulid-specific morphology.
- Reconstruct estimated *M. hypostoma* effective population changes through time using a Pairwise Sequential Markovian Coalescent analysis to inform *M. hypostoma* conservation research.

Chapter 2. Methods

All bioinformatics tools and packages used in the below analyses, software GitHub repositories and a glossary of file formats are available at the end of this section in Tables 2 and 3, respectively. All code and scripts are available in the Appendix at the end of this report.

I. Whole-genome alignment

i. Alignment workflow

Cactus v. 2.7.2, a reference-free genome alignment tool (Armstrong *et al.*, 2020), was used for whole-genome alignment of the unmasked *M. hypostoma* genome (Accession number: GCA_963921235.1) with genome assemblies from four other batoids. After discussion with my supervisors, the choice was made to align unmasked sequences to capture the large amount of repetitive sequences characteristic of elasmobranch genomes, as well as investigating the role of inserted and duplicated repeats in the future. The unmasked reference genomes for Atlantic stingray *H. sabinus* (Accession number: GCA_030144855.1), Oceanic manta ray *Mobula birostris* (Accession number: GCA_030035685.1), Thorny skate *Amblyraja radiata* (Accession number: GCA_010909765.2) and Little skate *Leucoraja erinacea* (Accession number: GCA_028641065.1) were downloaded from NCBI GenBank. Cactus aligns multiple genomes and reconstructs ancestral genomes based on shared sequences between each pair of diverged branches. In the present alignment, four ancestors were considered: 'Anc0', the ancestral genome for included species and root of the tree; 'Anc1', the skate ancestor; 'Anc2', the shared ancestral genome between Atlantic stingray and mobulids; and 'Anc3', the mobulid ancestor (Figure 2).

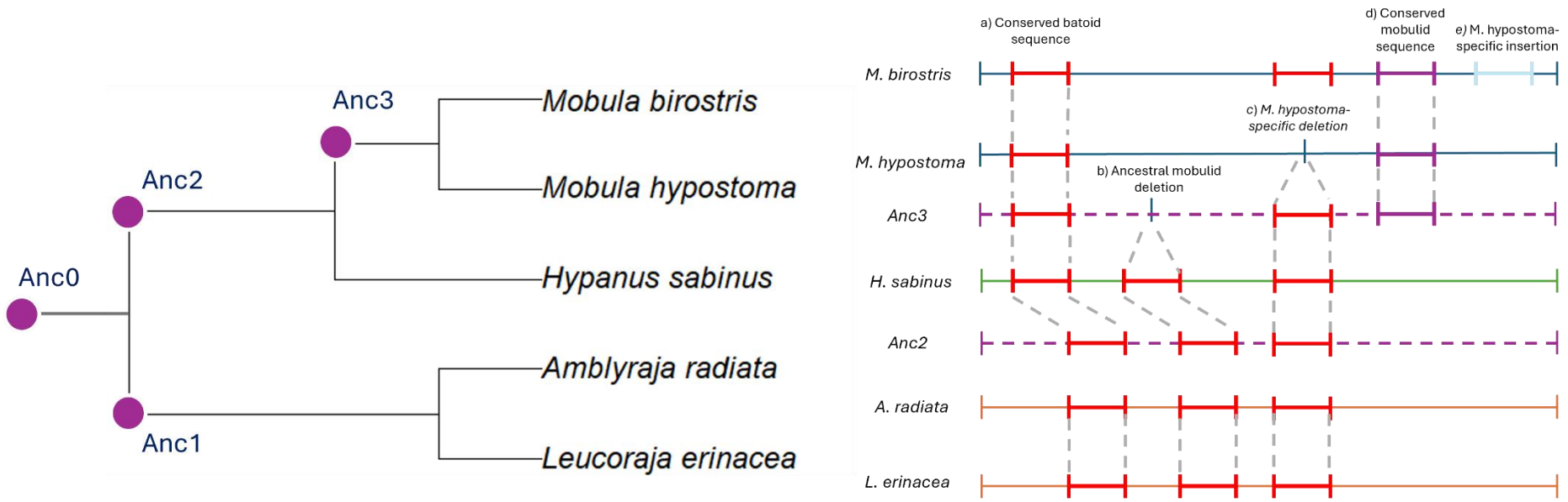


Figure 2: Aligned species and alignment schematics of a) a conserved sequence across all five species; b) a sequence conserved across skates and *H. sabinus* and lost in the mobulid ancestor; c) a sequence conserved across skates, *H. sabinus* and the mobulid ancestor, but lost in the *M. hypostoma* branch; d) a sequence inserted in the ancestral mobulid genome and conserved across mobulids; e) a *M. hypostoma*-specific inserted sequence.

Cactus was downloaded as a Singularity image file and run using Singularity v.3.5.3 on the University of Edinburgh's High Performance Computing (HPC) cluster, Eddie. Cactus requires a phylogenetic tree in Newick format as input to guide the alignment in an iterative manner. The tree is included at the top of the SeqFile, the input text file indicating to Cactus the names of each branch (species) and path to their genome FASTA files. The branch names were set to "hypostoma", "manta", "stingray", "amblyraja" and "leucoraja".

The following Newick Tree was manually generated based on batoid phylogeny and estimated evolutionary distance between each species obtained from the literature (Valsecchi *et al.*, 2005; Aschliman *et al.*, 2012; Villalobos-Segura and Underwood, 2020).

((leucoraja:12,amblyraja:12):15.5,(stingray:20,(hypostoma:10,manta:10):20):15.5);

amblyraja ./rays/GCF_010909765.2_sAmbRad1.1.pri_genomic.fna

leucoraja ./rays/GCF_028641065.1_Leri_hhj_1_genomic.fna

stingray ./rays/GCF_030144855.1_sHypSab1.hap1_genomic.fna

manta ./rays/GCA_030035685.1_sMobBir1.hap2_genomic.fna

hypostoma ./rays/sMobHyp1.curated_primary.mt.scrubbed.fa.

To run the alignment, the cactus command was called with the following options: --workDir ./workdir_cactus --maxCores 16 --maxMemory 768G --consCores 2 --consMemory 100Gi. The resulting output hal file was used in subsequent analyses using Hal Tools.

A summary of the various structural rearrangements counts and proportions (thereafter referred to as or SRs) for each branch of the alignment was extracted using HalTools, with the 'halSummarizeMutations command' ('--maxFraction' set to 0).

ii. Structural rearrangements summary

Branch-specific structural rearrangement coordinates were extracted from the Cactus alignment HAL file using the 'halSummarizeMutations' command, for *M. hypostoma*, stingray,

and the reconstructed mobulid ancestor genome (Anc3). The --maxFraction option was kept to 0 to categorise variants containing Ns as gapped variants. For each branch, two output BED files were generated: one containing insertions, inversions and transpositions; the file name for this output included the branch name, followed by the abbreviation for insertion. A second output BED file containing duplications and deletions was generated and named in the same fashion.

iii. Synteny relationships

Synteny refers to the conservation of blocks of genes and their order between two sets of chromosomes from different species (Kawashima, 2018). Synteny and more general homology relationships between chromosomes from batoid species pairs sampled from the alignment were investigated using the Cactus output. Synteny files in PSL format were extracted from the HAL alignment output for pairs of aligned species, using the 'halSynteny' command. A first file containing matching chromosome segments and their coordinates between the *M. hypostoma* and the *M. birostris* genomes was extracted by setting the query option to "--queryGenome hypostoma" and the target option to "--targetGenome manta", Matching sequence segments and their coordinates between *M. hypostoma* and *H. sabinus* chromosomes were extracted by setting options to "--queryGenome hypostoma" and "--targetGenome stingray". Finally, homologous chromosome regions between *M. hypostoma* and the thorny skate were extracted by setting "halSynteny" options to "--queryGenome hypostoma" and "--targetGenome amblyraja". The extracted PSL files were used to investigate chromosome homology and synteny between these three pairs of batoid species. The circlize (v.0.4.15) R package was used in R version 4.4.1 (Gu, 2014). Chromosome names and lengths were obtained from the NCBI GenBank records for *M. hypostoma*, *M. birostris*, *H. sabinus* and *A. radiata* to generate the chromosome track at the edge of the circular plot. Homologous chromosome segments were extracted from the PSL file visualised in Excel, and ordered by sequence size. Unplaced scaffolds were removed from the spreadsheet. The length of homologous chromosome segments was calculated by subtracting the start from the end coordinate base-pair. The target species' matching blocks were then linked to the query's to produce circos plots.

II. Repeat content landscape characterisation

Repeat content was quantified for each aligned species to compare the contribution of repeats to genome size alongside other genomic features. The proportions of major repeat families were compared between aligned species. Repeat characterisation was carried out using RepeatModeler v.2.05 and RepeatMasker v.4.1.7 software (Smit *et al*., 2015).

The Dfam-TEtools suite version 1.88 (Storer *et al.*, 2021) was downloaded as a Singularity image file to the University's HPC cluster, as it contained the RepeatModeler and RepeatMasker tools. RepeatModeler version 2.0.5 was then used to create a *de novo* repeat library for all five batoid species: first, the BuildDatabase command was used with the reference genome for each species, followed by the "nohup" command with the species' alignment name (hypostoma, stingray, manta, amblyraja or leucoraja) as the "-database" option.

As the software struggled masking the *Mobula hypostoma* genome, an alternative approach was used: the *M. hypostoma* (sMobHyp1.1) reference genome was split by chromosome and the workflow run one chromosome at a time. The output tables were then concatenated to obtain a single output file as for the other species, using the "cat" command from the command line.

III. Extraction of structural rearrangement coordinates from Cactus alignments

i. SR location flanking regions

The halBranchMutations command described in (Cactus: Branch mutations paragraph) outputs different BED files for insertions, given with respect to the query genome, and duplications/deletions, given with reference to the ancestral genome. In the case of *M. hypostoma* for example, these were the reconstructed Anc3 chromosome coordinates. To locate *M. hypostoma* genomic regions impacted by deletions and infer effects on annotated features, local sequence alignments were extracted from the whole-genome alignment for each *M. hypostoma* deletion expressed in Anc3 coordinates. The HAL output was converted into a MAF file and subset MAF files for each location were extracted using the mafExtractor command from the MafTools suite (Earl *et al.*, 2014). In order to capture the *M. hypostoma* coordinates in the subset alignments, the ins.bed file was modified so each start and end

coordinates spanned respectively one less and one additional base-pair, effectively creating a flanking or buffer genomic region (Figure 3).

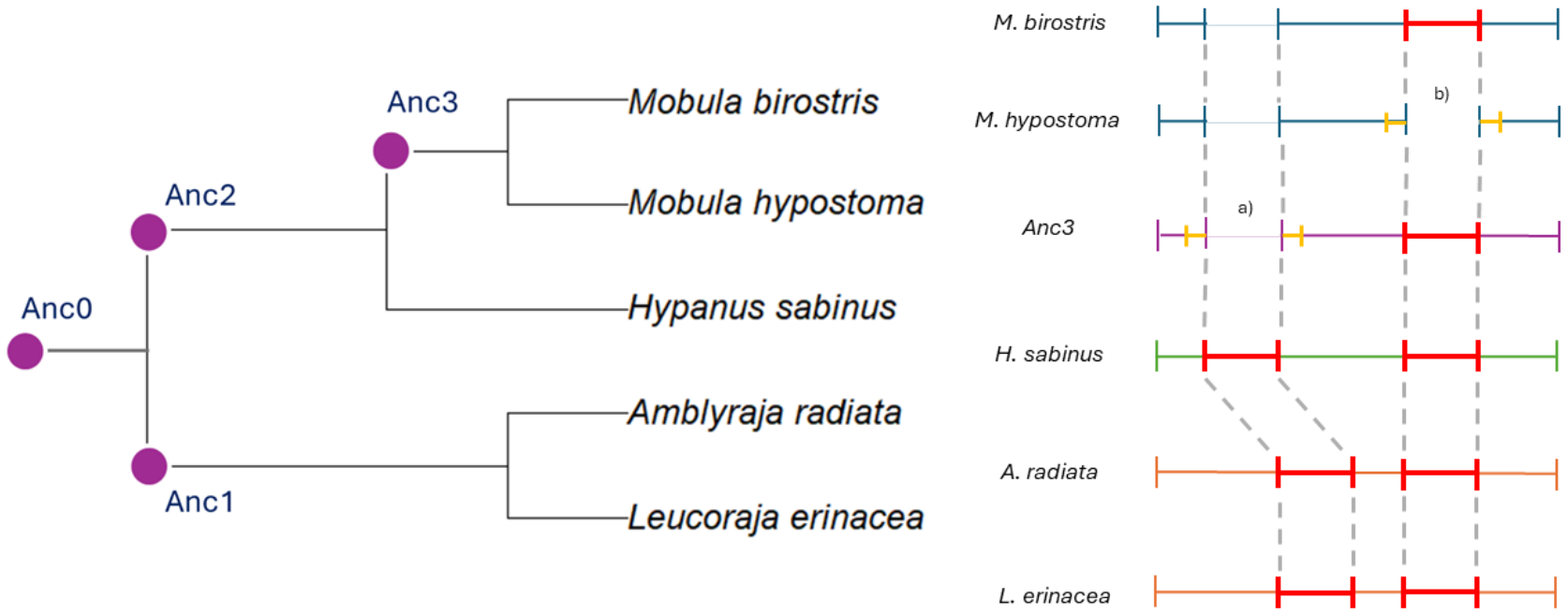


Figure 3: Schematic visualisation of a) ancestral mobulid- and b) *M. hypostoma*-specific deletions of sequences conserved in the other batoid genomes.

The modified BED files were split into subfiles, each containing a few hundred lines for easier scaling of the subalignment extraction workflow on the HPC cluster. The “mafExtractor" command from the mafTools suite was then used to parse each line of each sub BED file and extract the corresponding subalignment as a MAF file (Figure 4). An example of a subalignment MAF file is presented in Figure 5.

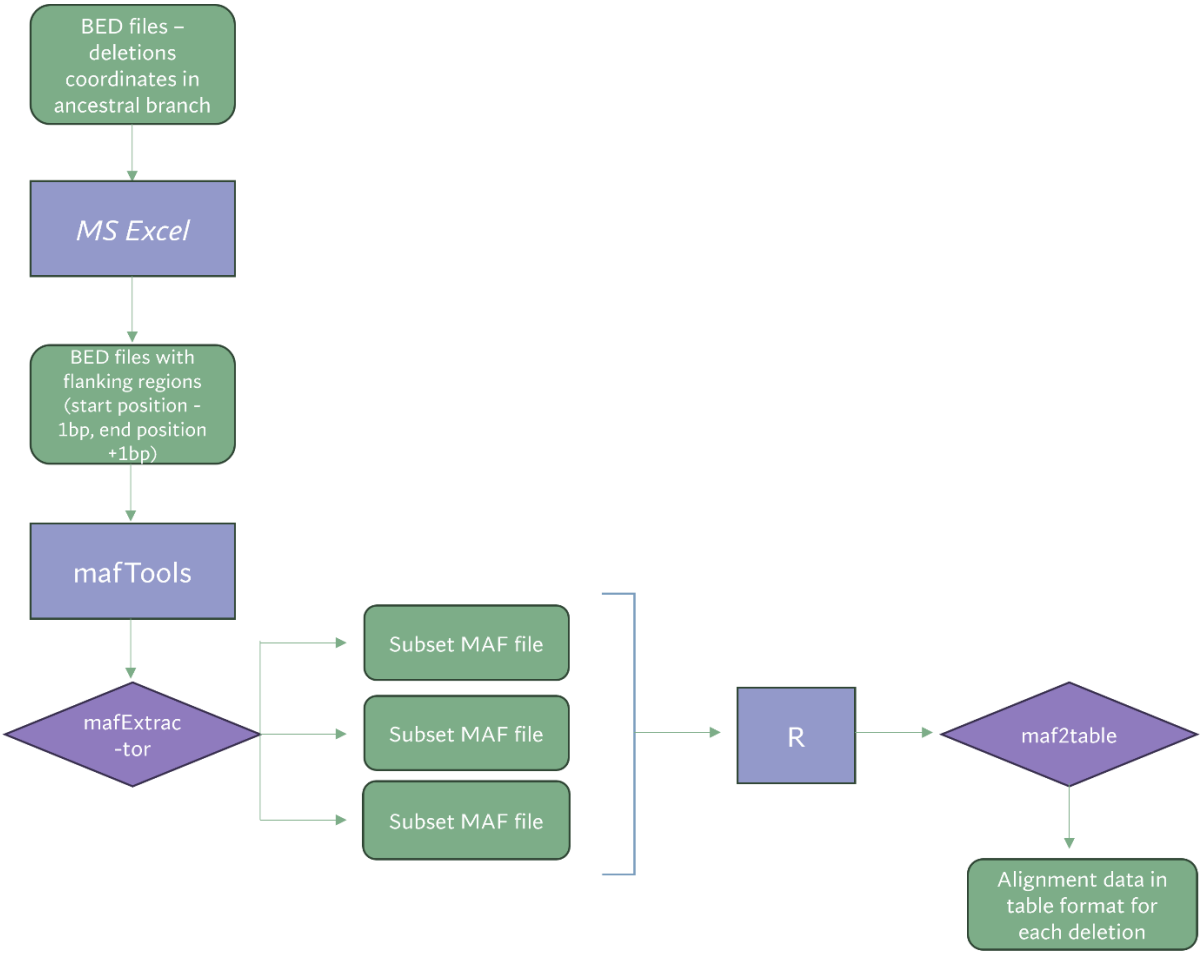


Figure 4: mafExtractor workflow**.** The start and end positions of each sequence in the ancestral ray (Anc2) and ancestral mobulid (Anc3) genome deleted in Anc3 and *M. hypostoma* respectively, were extended: one base pair was subtracted from the start coordinate, and one added to the end coordinate. Subset alignment MAF files were extracted for each deletion position using MafTools. The alignment data were transformed into a table using maf2table in R. The final dataset contained coordinates of sequences deleted in either the Anc3 or *M. hypostoma* genome and conserved in the other aligned species. The stingray coordinates were subsequently used for annotation using SnpEff.

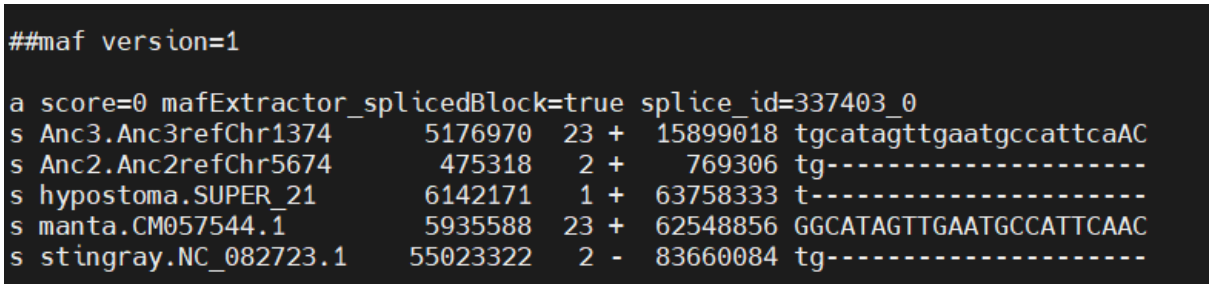


Figure 5: Example of a sub-alignment MAF file visualised on the University of Edinburgh HPC cluster.

ii. MAF subalignments concatenation

MAF subalignment files were then concatenated into a table using a custom R script provided by Dr Manu Gundappa (Figure 5). The resulting table contains the following fields: original MAF file, species of origin, chromosome of origin, cluster ID, start and end locations, size in base-pairs, range, and size-range ratio.

Due to time constraints, the decision was made to focus on mobulid- and *M. hypostoma*-specific deletions compared to stingray and the skates. Two subalignment tables were generated: one contained the coordinates of sequences conserved in the stingray and two skate genomes but lost in the mobulid ancestor (*i.e.* deletion inferred in Anc3); the other contained sequences conserved in all other aligned batoids but lost in *M. hypostoma,* and their coordinates in the other batoid genomes. The genomic location of these sequences in the stingray genome were then filtered out of each deletion table to functionally annotate ancestral mobulid and *M. hypostoma*-specific deletions, and infer their potential effects on developmental genes.

IV. Deletion annotation using SnpEff

SnpEff version 5.2c (Cingolani *et al.*, 2012) is a tool that predicts the effect of variants on genes and other features annotated in genome assemblies. As mentioned above, the deletion tables obtained using mafExtractor and maf2Table were filtered to extract the locations of ancestral batoid sequences lost in mobulids, in the stingray genome. BED files can be used as input format in SnpEff: therefore, the chromosome, start and end position fields were extracted

from the filtered tables to create new input BED files. SnpEff (Cingolani *et al.*, 2012) was used to functionally annotate the conserved sequences deleted in Anc3 (*i.e.*, mobulid ancestor) and the *M. hypostoma*-specific branch, using the Atlantic stingray genome annotation. Atlantic stingray is a non-model species, and as such needed to be added to SnpEff as a custom database. A sHypSab1.hap1 directory was added to the SnpEff data folder, to which the GTF annotation and protein fasta files were downloaded from the NCBI FTP site. An *Hypanus sabinus* entry was then added to the SnpEff config file.

Annotation of ancestral mobulid- and M. hypostoma-specific deletions

Each deletion BED file was annotated using the "SnpEff.jar" command and the "-c snpEff.config -v -nodownload -i bed" options, outputting an annotated BED file. The annotation appends to the usual BED fields and contains the genomic fragment affected (intron/exon number out of the respective total number of introns or exons for the relevant gene, gene name), and whether genes affected are coding genes or pseudogenes.

The SnpEff-annotated deletions files were manually searched for deletions affecting genes with gene ontology (GO) terms related to craniofacial, mouth and limb development which could have led to the evolution of the devil ray body plan. Genes associated with the terms "cranial skeletal system development" (GO:1904888), "roof of mouth development" (GO:0060021) and "limb development" (GO:0060173) in the Mouse Genome Informatics (MGI) database were extracted into a new dataset. The SnpEff annotation was simplified to indicate in this new dataset whether deletions affected intronic, exonic, intergenic regions or whole genes.

V. GO enrichment analyses

A GO enrichment analysis was performed to identify candidate biological functions, molecular functions and cell components over-represented among genes affected by deletions in the mobulid ancestor and the *M. hypostoma* branch. Gene names and symbols affected were extracted from the three deletion BED files described earlier (*i.e.*, sequences present in the other batoid genomes and deleted in the ancestral mobulid genome, sequences deleted in *M. hypostoma* but conserved in all other four species). The *H. sabinus* GO annotation file (GAF) was downloaded from the NCBI FTP website. Gene ID, symbols and names were extracted

from the GAF file to generate the gene information and GO tables required to create the *H. sabinus* annotation package with the AnnotationForge v.1.46.0 R package (Carson and Pagès, 2024). Once the package was created, installed and loaded, the ClusterProfiler v.4.12.6 (Wyffels *et al.*, 2014; Wu *et al.*, 2021) "enrichGO" command was run to carry out the GO analysis on each of the three files containing the extracted gene names from deletions BED files, compared to all GO annotated genes as the background. The keytype was set to "GID", and the adjusted p-value cut-off to 0.1. All of the above workflow was carried out using R v.4.4.1.

VI. Pairwise Sequential Markovian Coalescent reconstruction

As a preliminary estimation of past demographic fluctuations through geological and climatic changes, the past effective population size (N_e) for *M. hypostoma* was reconstructed using the Pairwise Sequential Markovian Coalescent (PSMC) (Li and Durbin, 2011). Assessment of effective population size through time can provide information on the historical genetic diversity of *M. hypostoma* populations, and can be compared to climatic trends and geological events to contextualise increases or decreases in N_e

Paired-end, short-read DNA sequencing data format from six flow cells, generated by CNAG and constituting of six FASTQ files, was used as input. Trim Galore v.0.6.10 (Krueger *et al.*, 2023) was used to trim the low-quality reads and adapter sequences. The trimmed reads were then aligned to the *M. hypostoma* reference genome using the Burrows-Wheeler Aligner tool v.0.7.18 (Hu, Li and Zeng, 2013) via a BWA-kit conda environment. Reads were aligned to the sMobHyp1.1 reference genome using the "bwamem" command. Unmapped reads were removed with Samtools version 1.9 (Danecek *et al.*, 2021) using the samtools view command and -bF options. Finally, the "samtools sort" command was used to sort the aligned reads. and output sorted BAM files. The six resulting BAM files were then merged using Picard software version 3.2.0 (Broad Institute, 2018), and the resulting merged and sorted BAM indexed with Samtools v.1.20. Duplicates were removed using Picard. The mapped, de-duplicated, sorted BAM file was then indexed using Samtools. "Samtools flagstat", Bedtools v.2.31.1 "genomecov" (Quinlan and Hall, 2010) and "modsdepth" (Pedersen and Quinlan, 2018) were then run on the same file to obtain the alignment and coverage statistics, which were then

visualised using MultiQC v.1.20 (Ewels *et al.*, 2016). Reads aligned to chromosomes 1 to 31 (*i.e.* excluding sex chromosomes) were extracted from the sorted and de-duplicated BAM file to generate an autosome-specific BAM file. The new BAM file was then sorted and indexed using Samtools, and statistics generated as described above. An autosome-only reference genome was also extracted and indexed using bwa index.

The consensus sequence, a necessary input file for running PSMC, was generated using BCFtools version 1.19 with the mpileup command and -C50 and -f options, and the autosome-only reference genome and sorted autosome BAM file. The PSMC v.0.6.5 workflow was then run on the consensus sequence. The ratio of population mutation rate to recombination rate, the maximum number of iterations of the Markov model and the maximum time were kept as mentioned in the tool documentation, at respectively 25, 15 and 5. To plot effective population size versus time, an estimated mutation rate of 8.1 x 10 -9 was chosen based on an estimation made for a mobulid species of a similar size (Poortvliet *et al.*, 2015). As the generation time of most mobulids is unknown, three were selected to generate different plots: 20, 25 and 29 years, the latter based on the value measured for *M. alfredi* (Couturier *et al.*, 2012).

Table 2: Software tools and packages used

Software/package	**Version**	**Github repository/website**	**Reference**
Cactus	2.7.2	https://github.com/ComparativeGenomicsToolkit/cactus/blob/v2.6.11/doc/progressive.md	Armstrong et al., 2020
halTools	2.3	https://github.com/ComparativeGenomicsToolkit/hal	Hickey et al., 2013
circlize	0.4.15	https://jokergoo.github.io/circlize_book/book/	Gu, 2014
Dfam-TEtools	1.88	https://github.com/Dfam-consortium/TETools	Storer et al., 2021
RepeatModeler	2.05	https://github.com/Dfam-consortium/RepeatModeler/blob/master/RepeatModeler	Smit et al., 2015
RepeatMasker	4.1.7	https://github.com/Dfam-consortium/RepeatMasker/issues	Smit et al., 2015
mafTools	Python 2.7	https://github.com/dentearl/mafTools	Earl et al., 2014
SnpEff	5.2c	https://pcingola.github.io/SnpEff/snpeff/running/	Cingolani et al., 2012

ClusterProfiler	4.12.6	https://guangchuangyu.github.io/software/clusterProfiler/	Yu et al., 2012
AnnotationForge	1.46.0	https://github.com/Bioconductor/AnnotationForge	Carson and Pages, 2024
Burrows-Wheeler Aligner	0.7.18	https://github.com/lh3/bwa	Li, 2013
BAMtools	2.5.2	https://github.com/pezmaster31/bamtools	Barnett et al., 2011
TrimGalore	0.6.10	https://github.com/FelixKrueger/TrimGalore	Krueger et al., 2023
Samtools	1.20	https://github.com/samtools/samtools	Li et al., 2009
MultiQC	1.23	https://github.com/MultiQC/MultiQC	Ewels et al., 2016
PSMC	0.6.5	https://github.com/lh3/psmc	Li and Durbin, 2011
Bedtools	2.31.1	https://github.com/arq5x/bedtools2	Quinlan and Hall, 2011
BCFtools	1.20	https://github.com/samtools/bcftools	Danecek et

			al., 2021
R	4.4.1	https://www.r-project.org/	R Core Team, 2021

Table 3: Glossary of file types

File type	**Description**
FASTA	DNA or RNA sequence text file written as a series of A, G, C, T letters.
HAL	Hierarchical alignment file outputted by cactus.
MAF	Mutation annotation format; text file containing aligned sequences information, their origin and their differences.
PSL	Text file containing matching seauences between two aligned species, and their coordinates.
BED	Tab-separated text file containing a list of genomic feature and their genomic coordinates.
Newick	Phylogenetic tree as a text format
SeqFile	Text file containing the Cactus alignment guiding Newick tree, followed by custom species names and their directory. Each line contains only one species, and name and path are space-delimited.
BAM	Binary alignment map; compressed format of sequence alignment maps (SAM), which contain raw sequence data mapped to a reference sequence.

Chapter 3. Results

I. M. hypostoma *genome and comparison to other batoid genomes*

i. Overview of genomes used in my project

As no description of the recently released mobulid reference genomes has been made before, the following section provides a brief comparison of the *M. hypostoma* assembly to other aligned batoid genomes for a range of relevant statistics.

The assembled reference genome for the West Atlantic pygmy devil ray *M. hypostoma* (sMobHyp1.1; GCA_963921235.1; BioProject: PRJEB71673) has a total length of 3.5 gigabases (Gb) and is composed of 34 chromosomes, with 31 autosomes and three sex chromosomes showing an X1-X2-Y pattern. According to NCBI statistics, the genome assembly has a contig N50 of 23.3 megabases (Mb), scaffold N50 of 152.4 Mb, and GC content of 43%. Contig and scaffold N50 values are defined as the sequence length of the shortest contig or scaffold at 50% that contains at least 50% of the contig or scaffolded assembly's total length, respectively. Of the 23,970 genes annotated by NCBI RefSeq, 18,916 are coding, 3,892 non-coding, while 1,061 non-transcribed pseudogenes were also predicted. For the 22,819 expressed genes, the mean/median length was 81,713/39,375 bp. The *M. hypostoma* annotation showed 94.7% BUSCO gene completeness, which refers to the percentage of benchmarking single-copy gene orthologs recovered from the RefSeq annotated proteins.

Of the five aligned batoid genomes, the Atlantic stingray *H. sabinus* has the largest assembly size, *i.e.*, (4 Gb), followed by *M. birostris* (3.6 Gbp) and *M. hypostoma.* The two skates have markedly smaller assembly sizes at 2.6 and 2.2 Gb for thorny skate *A. radiata* and little skate *L. erinacea*, respectively. *M. hypostoma* has a markedly higher contig N50 than the other species (8.6, 4.3, 1.5 and 5.3 Mb, for *M. birostris*, *H. sabinus*, *A. radiata*, and *L. erinacea*, respectively). *M. hypostoma* and *M. birostris* show the highest scaffold N50s (both 152.4 Mb), followed by *H. sabinus* (112.3 Mb), *A. radiata* (62.1 Mb) and *L. erinacea* (57.2 Mb). All annotated batoid genomes showed high BUSCO completeness, with *M. hypostoma* the highest, followed by *H. sabinus* (94.4%), *A. radiata* (92.3%) and *L. erinacea* (92.5%). No RefSeq annotation has been released for the giant oceanic manta ray *M. birostris* (sMobBir1.hap2) and therefore its predicted genome features cannot be compared to other batoids. All five batoid genomes showed similar GC content, ranging from 43% to 44.5%. While the predicted

karyotypes of the three ray species captured similar chromosome numbers (35, 34 and 33 for *H. sabinus*, *M. birostris* and *M. hypostoma*, respectively), the skates have far more predicted chromosomes: 49 for *A. radiata* and 50 for *L. erinacea*.

ii. Batoid repeat landscape

All five batoid genomes had similar predicted repeat content ranging from 60 to 65%, highest for *M. hypostoma* (Figure 6A). At a high-level, the identified repeat classes were categorised into retroelements, DNA transposons and unclassified repeats. The *M. hypostoma* genome contained the second greatest fraction of retroelements compared to the other four species at 32.62%, similar to its sister mobulid *M. birostris* (32.75%), while *H. sabinus*, *M. birostris* and *M. hypostoma* possessed 28.36%, 20.37% and 26.43 % retroelements, respectively. DNA transposons represented a negligible portion (0.5-2.89%) of repeats predicted for all five species, with *A. radiata* comprising the lowest fraction and *M. hypostoma* the highest. Tc1-IS630-Pogo DNA transposons accounted for the bulk of this repeat class (0.79-1.73% of repeats across species). Long interspersed nuclear elements (LINEs) composed the bulk of retroelements for all five genomes, highest in the mobulid species (28.54-29.64%) (Figure 6B). The main LINE family in all batoid species was L2/CR1/Rex, accounting for 12.5-22% of the genome, again with highest proportions in the two mobulid genomes. The second largest retroelement class was L1/CIN4 LINEs, although they only comprised 1.04-2.18% of batoid genome size, with the maximum proportion in *H. sabinus*. No short interspersed elements (SINEs) were detected in any of the five genomes. Long tandem repeats (LTRs) accounted for 2.66-3.67% of repeats in the three ray genomes (with a maximum in *H. sabinus*), and respectively 2.47 and 3.35% in *L. erinacea* and *A. radiata*. The Gypsy/DIRS1 family represented the bulk of LTRs for all genomes, while retroviral LTRs represented between 0.15 and 0.63% of LTRs.

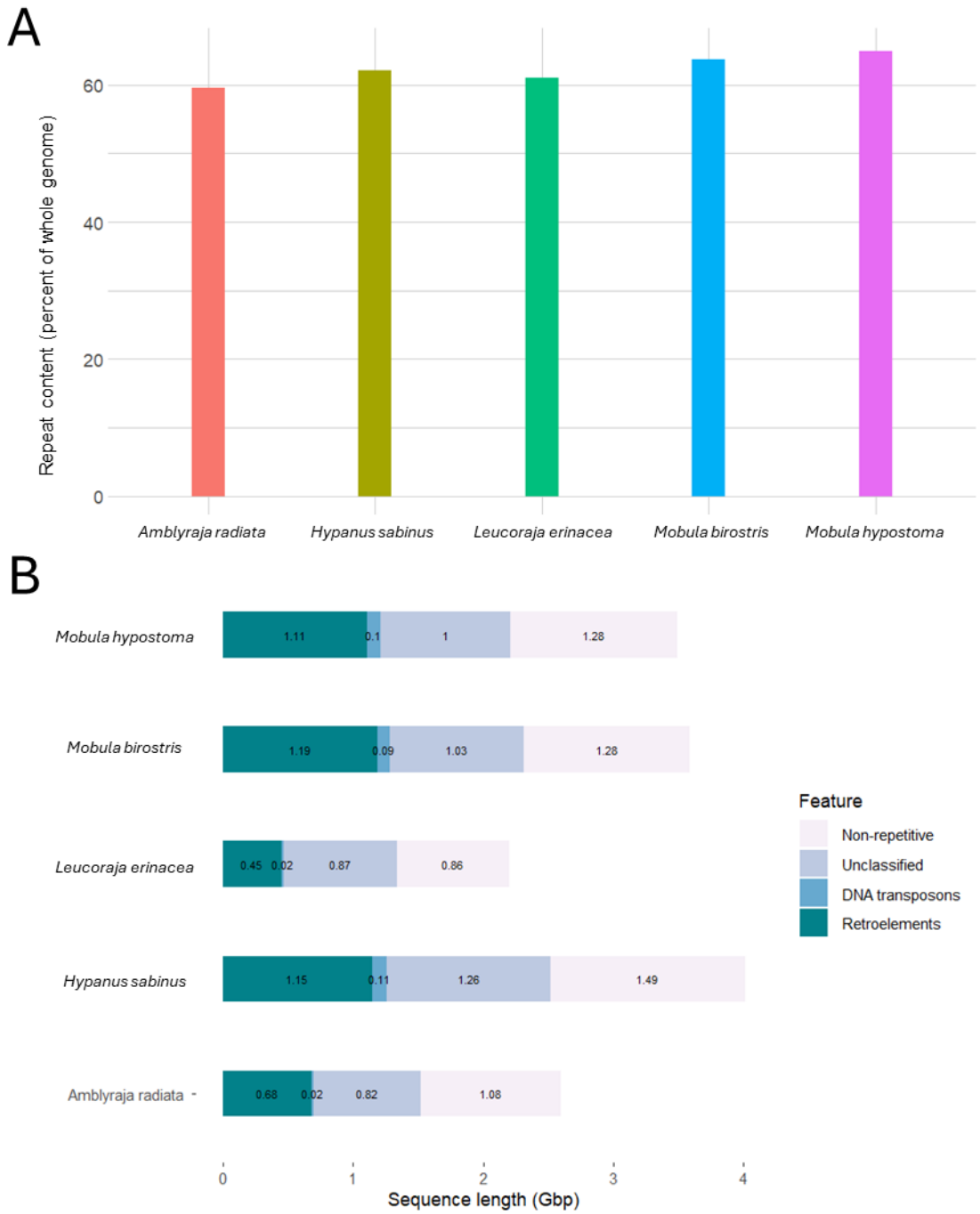


Figure 6: A) Repeat content proportion of each batoid genome used in whole genome alignment. B) Total sequence length of each major repeat family compared to total genome size for each species.

II. Pairwise synteny across batoid genomes

i. Comparison of mobulid species

Using the Cactus alignment results, one-to-one chromosome homology was largely evident for a large part between the *M. birostris* and *M. hypostoma* genomes. The first (and largest) 12 chromosomes in the *M. hypostoma* genome align across their length, with the first 12 manta chromosomes appearing conserved, with limited evidence for chromosomal rearrangements (Figure 7). The order is disrupted from *M. birostris* chromosome 13 (MB13) onwards, though most *M. birostris* chromosomes are still one-to-one homologs of a matching *M. hypostoma* chromosome. MB13 is the only chromosome with no apparent homolog in MH, with only two fragments matching short sequences in both MH3 and MH31. MB20 also breaks the expected homology order, showing instead evidence of synteny with MHX2, MHY and MH4. The *M. hypostoma* Y chromosome appears homologous to part of MB20 and MBX, while MHX1 appears to be a one-to-one homolog of MBX.

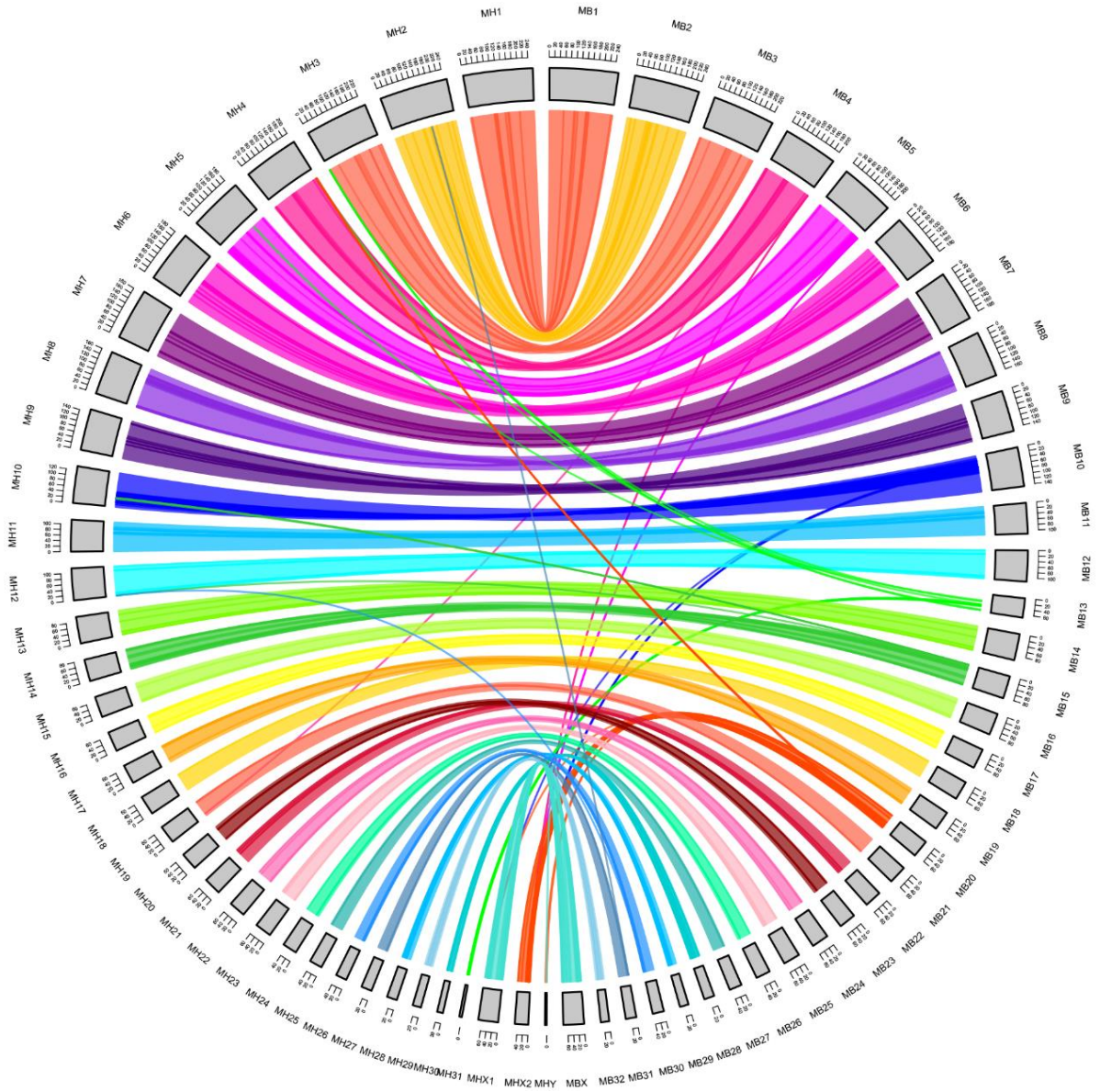


Figure 7: Circular representation of conserved synteny between chromosomes of *Mobula hypostoma* (left) and *Mobula birostris* (right). Chromosomes are represented by the innermost circular track (grey rectangular shapes). *M. hypostoma* (MH) and *M. birostris* (MB) chromosomes are respectively named MH1-31, MHX1, MHX2 and MHY, and MB1-32, MBX and MBY (outer track). Chromosome length in bp is represented by the scale bars on the middle track. Homologous regions between chromosomes of the two species are linked by coloured bands.

ii. Comparison of more distant myliobatiform genomes

More chromosomal rearrangements are evident comparing *H. sabinus* vs *M. hypostoma* (Figure 9). For example, MH7 appears to represent a fusion of HS15 and part of HS3, while MH10 appears a fusion of HS26 with part of HS8. MH3 also appears to represent a fusion of HS14 and about two-thirds of HS6. Some putative ancestral chromosome fissions appear to have resulted in new *M. hypostoma* chromosomes, with MH20 and MH27 each showing homology to non-overlapping halves of HS13. MH1 appears to have resulted from a fusion of fragments of HS1 and HS2, while the rest of HS1 appears to have combined with HS12 to form MH8. A few one-to-one chromosome homologies are maintained in the smaller *H. sabinus* and *M. hypostoma* chromosomes - including between MH15 and HS19, MH17 and HS20, MH19 and SR22, MH22 and HS23, MH25 and HS27, MH26 and HS30, MH28 and HS24, MH30 and HS31, MH31 and HS32 (Figure 8). Most of MHX1 appears homologous to HSX1, with relatively short additional matches with HSY and HS16. Similarly, MHX2 mostly corresponds to HSX2, with additional short fragments matching to HS6 and HS10. The *Mobula* Y chromosome appears to be a fusion of HSY and a fragment of HSX2.

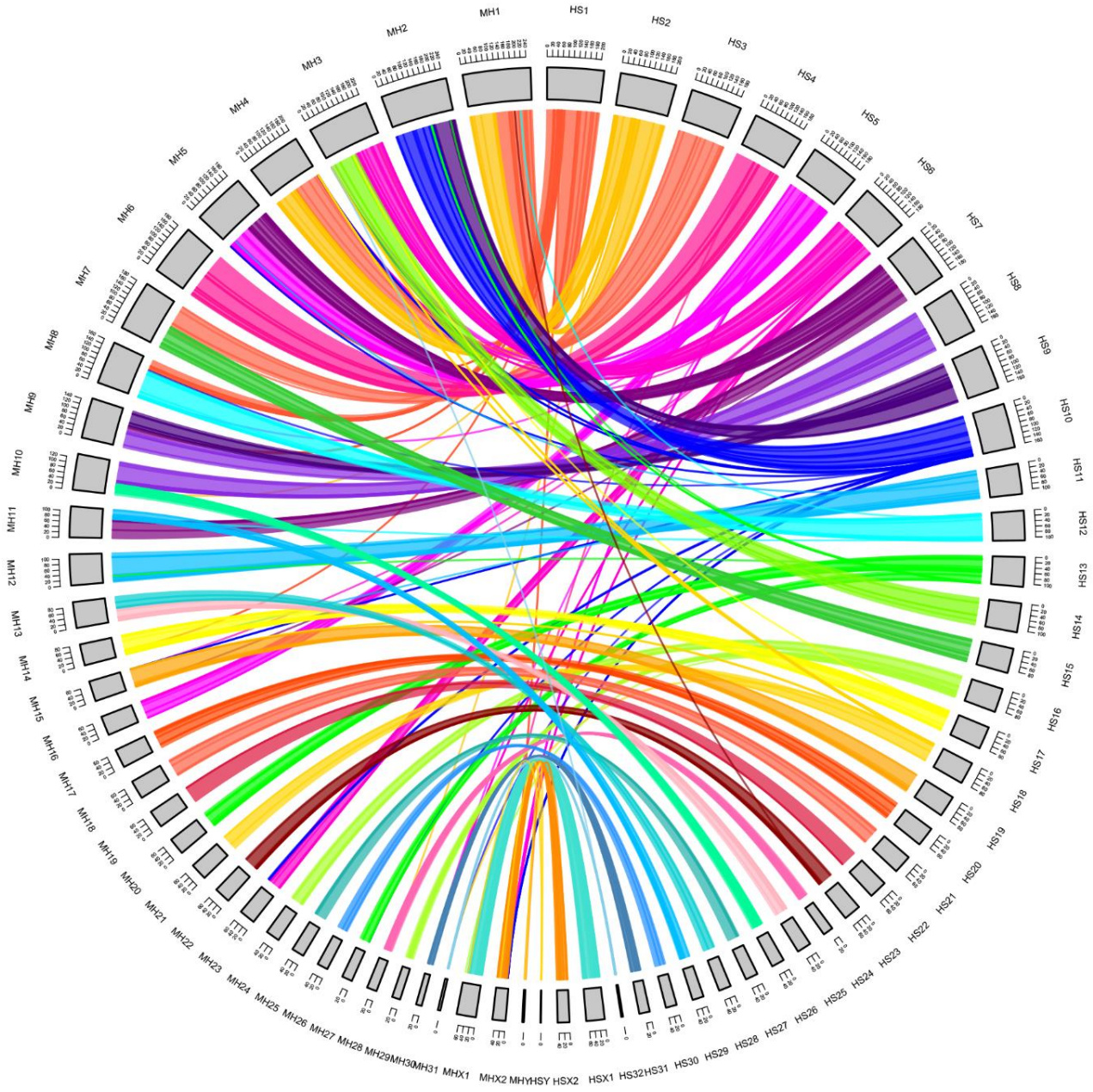


Figure 8: Circular representation of synteny and homology relationships between *Mobula hypostoma* chromosomes (left) and *Hypanus sabinus* (right). Chromosomes are represented by the innermost circular track (rectangular shapes). *M. hypostoma* (MH) and *H. sabinus* (HS) chromosomes are respectively named MH1-31, MHX1, MHX2 and MHY, and HS1-32, HSX1, HSX2 and HSY (outer track). Chromosome length in bp is represented by the scale bars on the middle track. Homologous regions between chromosomes of the two species are linked by coloured bands.

III. Genome structural rearrangements during batoid evolution

Using the Cactus results, predicted structural rearrangements (SRs - including insertions, duplications, deletions, inversions and transpositions) were extracted and characterised across the batoid tree (Figures 2 and 3). SRs were obtained for the reconstructed batoid ancestor (Anc0), the reconstructed skate ancestor (Anc1; SRs vs. Anc0), the reconstructed myliobatiform ray ancestor (Anc2; SRs vs Anc0), the reconstructed mobulid ancestor (Anc3; SRs vs Anc2), and branches for all five individual species, *i.e.* both skates (SRs vs Anc1), *H. sabinus* (SRs vs Anc2), and both mobulids (SRs vs Anc3). I focussed on SRs that occurred during myliobatiform genome evolution.

Approximately 1.5 million SRs occurred along the ancestral mobulid branch Anc3, constituting 34.7% of rearranged ancestral myliobatiform (Anc2) sequences. In contrast, 2.5 million SRs occurred along the *H. sabinus* branch after the split with mobulids, representing 27.3% of the total Anc2 sequence. SRs in Anc3 represented about 620,000 and 812,000 SRs in the *M. hypostoma* and *M. birostris* genomes, respectively.

Insertions represented the bulk of SRs in all cases, accounting for between 34-81% of all SRs. Insertions also represented the highest proportion of rearranged base pairs in all branches, ranging between 64% and 96%. The *H. sabinus* branch had the highest proportion of insertions and the *M. birostris* branch the lowest (Figure 10). The ancestral mobulid branch, Anc3, contained 26% inserted sequences (849 Mb) against 30% in the stingray branch, with 1.2 Gb of the 4.04 Gb stingray genome being inserted Anc2 sequences. Insertions account for most SRs in the Anc3 and stingray branches, reaching 81.4 and 42.4% (of all SRs), respectively (Figure 11A). The *M. hypostoma* branch contained a higher proportion of insertions than the manta branch (7% of genome length vs. 5%, respectively), with total respective inserted sequence lengths of 247 Mb and 197 Mb (Figure 10).

Duplications represented 16.8-59.21% of all SRs across the myliobatiform branches. This included 257,000 duplications in the ancestral mobulid Anc3 branch, in addition to 479,000 and 264,000 respective duplications along the *M. birostris* and *M. hypostoma* branches (Figure 9). Duplications account for about 4.5% of genome length in the *H. sabinus* branch, against only 0.8% in the ancestral mobulid branch Anc3. Duplications amounted to 2.8% (102 Mb) of the total genome length in the *M. birostris* branch. In contrast, the *M. hypostoma* branch

showed 38.4 Mb duplicated ancestral mobulid sequences, amounting to 1.1% of the total *M. hypostoma* genome length (Figure 10).

61,000 and 52,000 respective deletions were found along the *M. hypostoma* and *M. birostris* branches. About 26,000 deletions were evident along the ancestral mobulid Anc3 branch, representing a small fraction (1.7%) of all SRs (Figure 9). Similarly a small fraction of SRs in the *H. sabinus* branch were deletions (27,000/2.5 million). While deleted sequences accounted for less than 1% of total genome length in both Anc3 and *H. sabinus* branches, the ancestral mobulid branch had three times as many deletions. A similar proportion of total genome length was represented by deletions along the *M. hypostoma* and *M. birostris* branches, 0.1% for the former (3.8 Mb / 3.5 Gb) versus 0.08% for the latter (2.9 Mb / 3.6 Gb) (Figure 10).

Transpositions and inversions accounted for less than 1% of all SR counts. The Anc3 and *H. sabinus* branches showed slightly over 1 million transposed bp (0.04 and 0.03% of total Anc3 and *H. sabinus* genome length, respectively) (Figure 10C). The same branches showed about 300 inverted kb, *i.e.*, less than 0.01% of the total genome sequence (Figure 9). The *M. birostris* branch showed a slightly greater portion of transposed sequences than *M. hypostoma*, with 4.1 Mb (0.11%) versus 3 Mb (0.08%). Inversions represented a negligible fraction of both mobulid species branches.

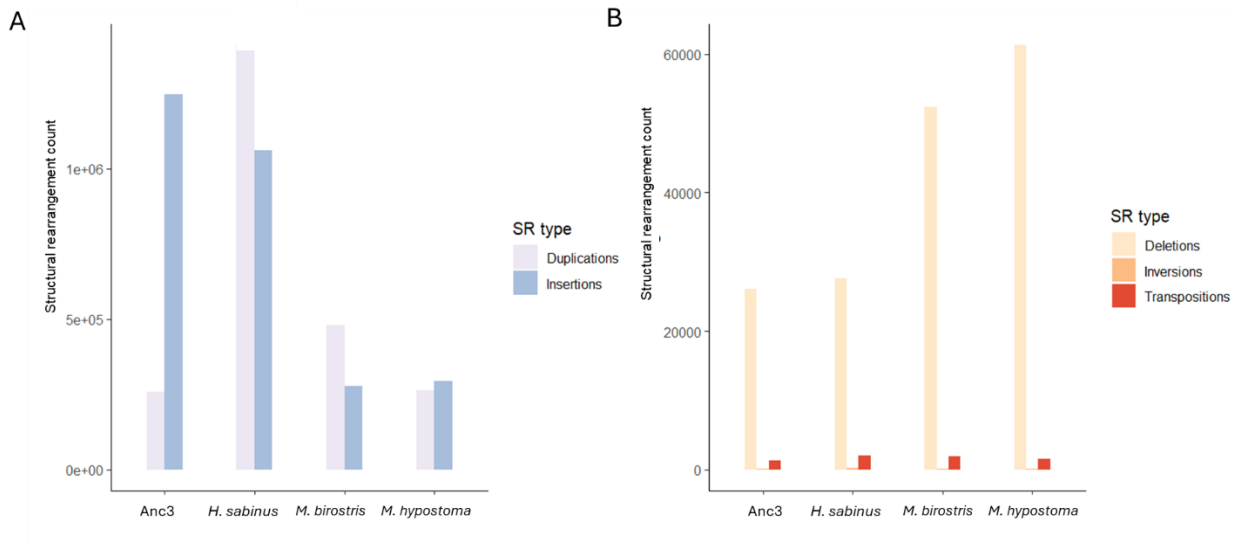
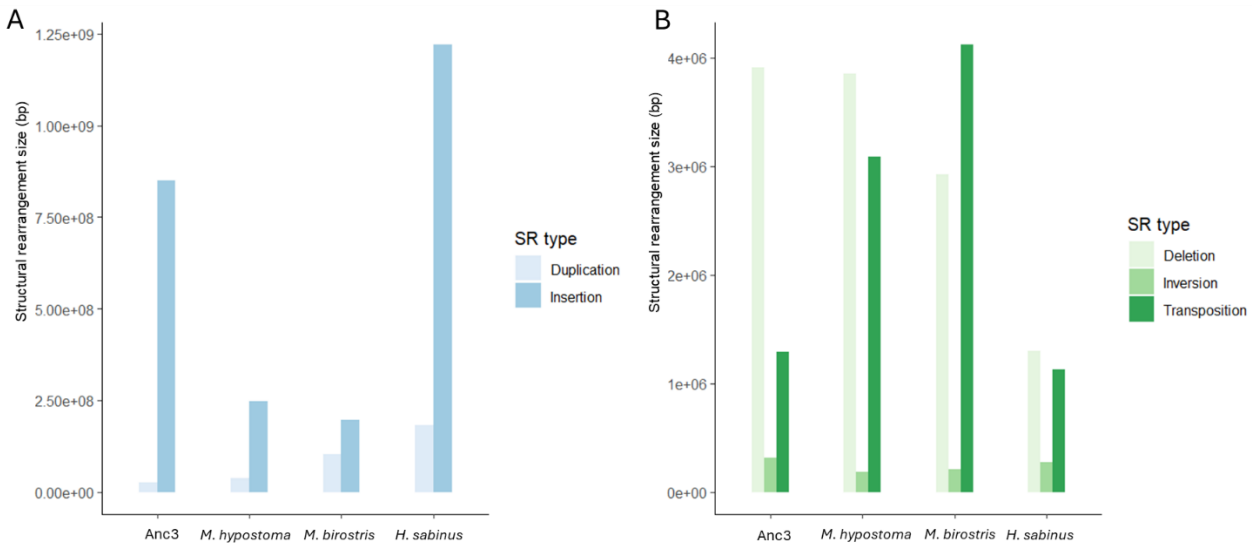


Figure 9: Total counts of each structural rearrangement (SR) in each branch. A) duplications and insertions; B) deletions, inversions, transpositions.



Figures 10: Total sequence length of each SR type in all branches. A) insertions and duplications; B) deletions, inversions and transpositions.

IV. Craniofacial and limb development genes affected by mobulid deletions

The mobulid ancestor (Anc3) branch SRs were searched to identify deletions affecting genes related to craniofacial and limb development, which I hypothesise were targeted during the evolution of the unique mobulid morphology. I focus on deletions due to the ease of interpreting their impact on annotated gene features compared to other SR types. SnpEff was used to prioritise deletions of interest in the target branches against the stingray genome annotation, representing regions that were conserved in the stingray and both skate genomes, but deleted in both mobulids. Genes were prioritised according to membership to three GO biological function terms; "craniofacial development", "roof of mouth development", and "limb development", as well as from literature searches on the genetics of vertebrate craniofacial development (Lamichhaney *et al.*, 2015; Mork and Crump, 2015).

Ninety Anc3 deletions were associated with prioritized genes according to SnpEff. Most were located in intronic sequences (Tables 4 and 5). A smaller number of deletions occurred in exons or just upstream of the first exon, potentially affecting the promoter – as for *shh* (Figure 12A), encoding sonic hedgehog – a key signalling molecule for craniofacial skeleton development in vertebrates (Dworkin *et al.*, 2016). *alx3* and *fgfr3* genes also showed deletions just upstream of their first exon (Figure 11B, C), potentially affecting the promoter region. *alx3*

regulates cellular differentiation of the forehead and nose during craniofacial development in vertebrates, and its deficiency prevents normal cranial neural tube closure in mice and humans, leading to malformations of the head and face (Lakhwani, García-Sanz and Vallejo, 2010; Mitchell *et al.*, 2021). *fgfr3* encodes a fibroblast growth factor receptor, and mutations in this gene are linked to chondrodysplasia and craniofacial malformations (Morice *et al.*, 2023). A deletion was also found in the first noncoding exon of *bpnt2* (also named *impad1*), which has been linked to craniofacial deformities (Dubail and Cormier-Daire, 2021). A deletion also affected the final, non-coding exon of *mthfd1l* (Tables 4 and 5; Figure 12D), which encodes an enzyme involved in mitochondrial formate pathways and it loss in mice causes neural tube defects (Momb *et al.*, 2013). Several other examples are further considered in the discussion section.

Table 4: Prioritised genes affected by deletions in the ancestral mobulid genome (Anc3) assigned by GO term membership

GO ID	GO term	Chromosome	Stingray start	End	Size	Gene symbol	Gene name	SnpEff annotation
GO:0060021	roof of mouth development	2	35112553	35112578	25	*dlg1b*	Discs large homolog 1	Affects introns separating various transcripts
GO:1904888;GO:0060021	cranial skeletal system development; roof of mouth development	6	24389348	24389385	37	*insig1*	Insulin-induced gene 1 protein	Intergenic region about 3 Mb upstream of gene
GO:0060021	roof of mouth development	6	26203144	26203170	26	*shha*	Sonic hedgehog a	Affects region about 1 kb upstream of first, coding exons, possibly impacting promoter region
GO:1904888	cranial skeletal system development	15	34729650	34729678	28	*sh3pxd2b*	SH3 and PX domains 2B	Affects separating two different transcripts
GO:1904888	cranial skeletal system development	2	185346434	185346459	25	*foxn3*	Forkhead box protein N3	Affects introns of several transcriots and upstream region of an alternative transcript
GO:0060021	roof of mouth development	X1	2104605	2104642	37	*fzd2*	Frizzled class receptor 2	Intergenic region, respectively 77 kb and 45 kb upstream of gene
			2134784	2134808	24			
GO:1904888	cranial skeletal system development	17	34140430	34140516	86	*irx5a*	Iroquois homeobox 5	Intergenic region about 3 kb downstream of gene
GO:1904888	cranial skeletal system development	4	42794683	42794788	105	*slc39a10*	Solute carrier family 39 member 10	Intergenic regions about 7 kb, 25 kb, 120 kb, and 140 kb upstream of gene
			42774161	42774196	35			
			42680191	42680217	26			
			42661229	42661268	39			
GO:0060021	roof of mouth development	22	54938299	54938325	26	*plekha1b*	Pleckstrin homology domain containing A1	Affects introns of various transcripts towards the 5' end of the gene
GO:0060021	roof of mouth development	6	96915067	96915145	78	*meox2a*	Mesenchyme homeobox 2	Intergenic region about 45 kb downstream of gene
GO:0060021	roof of mouth development	14	44431871	44431895	24	*pds5a*	PDS5 cohesin associated factor A	Affects introns of various Pds5a transcripts
GO:0060021	roof of mouth development	3	24222748	24222774	26	*gabrb3*	Gamma-aminobutyric acid receptor subunit beta-3	Intergenic region about 480 kb downstream of gene
GO:0060021	roof of mouth development	1	174256051	174256106	55	*chd7*	chromodomain helicase DNA-binding protein 7	Intergenic region 8 kb and 120 kb downstream of gene
			174148782	174148813	31			
			174017120	174017150	30		chromodomain helicase DNA-binding protein 7	Affects introns separating various transcripts
GO:0060021	roof of mouth development	7	75493884	75493908	24	*bnc2*	basonuclin zinc finger protein 2	Affects intron in the second half of the gene
			75289835	75289857	22			Intergenic region about 20 kb upstream of gene
GO:0060021	roof of mouth development	3	150294092	150294148	56	*lef1*	Lymphoid enhancer-binding factor 1	Intergenic region about 38k downstream of gene
GO:0060021	roof of mouth development	3	143503397	143503422	25	*msx1a*	Homeobox protein MSX-1	intergenic region closely upstream to gene, about 4 kb
			143470816	143470842	26			Intergenic region about 32 kbdownstream of gene
GO:0060021	roof of mouth development	8	51374452	51374484	32	*bcorl1*	BCL6 corepressor like 1	Intergenic region upstream of gene
GO:0060021	roof of mouth development	3	129858660	129858690	30	*intu*	Inturned planar cell polarity protein	Affects introns of various transcripts
GO:0060021	roof of mouth development	3	123817889	123817913	24	*bbs7*	Bardet–Biedl syndrome 7	Two deletions, affect separating various transcripts of gene
			123787203	123787243	40	*bcorl1*	BCL6 co-repressor-like 1	Two deletions affecting introns between several transcripts
		8	51610143	51610167	24			
			51570856	51570888	32			
GO:1904888	roof of mouth development	7	56408677	56408701	24	*frem1a*	FRAS1 related extracellular matrix 1	Intergenic region about 348 kb upstream of gene

GO:1904888; GO:0060021	cranial skeletal system development; roof of mouth development	7	57604783	57604807	24	*foxe1*	Forkhead box E1	Intergenic region about 2 kb upstream of gene
GO:1904888	roof of mouth development	7	108688532	108688556	24	*slc39a13*	Solute carrier family 39 member 13	Affects separating various transcripts of gene
GO:0060021	roof of mouth development	7	20048385	20048410	25	*hand2*	Heart and neural crest derivatives expressed 2	Intergenic region about 354 kb and 316 kb and upstream of gene
			20446445	20446469	24			
GO:1904888	roof of mouth development	2	158502146	158502175	29	*bmp4*	Bone morphogenetic protein 4	Intergenic region about 14 kb upstream of gene
GO:1904888; GO:0060021	cranial skeletal system development; roof of mouth development	5	7653542	7653568	26	*mef2cb*	Myocyte-specific enhancer factor 2C	Intergenic region about 605 kb and 523 kb upstream of gene, and two about 130 and 172 kb downstream of gene
			7590421	7590445	24			
			8481528	8482359	831			
			8672363	8673194	831			
			8526778	8526804	26			
GO:1904888	cranial skeletal system development	5	8298947	8298979	32			Affects separating various transcripts of gene
GO:0060021	cranial skeletal system development	5	64786817	64786842	25	*pax5*	Paired box 5	Two deletions affecting separating various transcripts
			65011159	65011204	45			
GO:0060021	roof of mouth development	7	155858697	155858734	37	*alx4a*	ALX homeobox 4	Intergenic region about 107 kb and 290 kb and upstream of gene
		7	155676547	155676584	37			
GO:0060021	roof of mouth development	8	162953402	162953427	25	*alx1*	ALX homeobox 1	Affects introns towards the 5' end separating Alx1 transcripts
GO:0060021	roof of mouth development	27	33820784	33820811	27	*prdm16*	PR domain containing 16	Four deletions affecting separating transcripts
			34050984	34051018	34			
			34196125	34196170	45			
			34454842	34454866	24			
GO:0060021	roof of mouth development	10	32654294	32654321	27	*runx2a*	Runt-related transcription factor 2a	Affects separating transcripts of Runx2a
GO:0060021	roof of mouth development	10	42889107	42889134	27	*osr1*	Odd-skipped-related 1	Intergenic region about 37 kb, 472 kb, 690 kb and 817 kb upstream of gene
			42998108	42998139	31			
			43216724	43216752	28			
			43344209	43344238	29			
GO:0060021	roof of mouth development	4	143562075	143562105	30	*bcor*	BCL6 corepressor	Three deletions affecting separating various transcripts of Bcor and upstream of one alternative transcript, possibly affecting promoter region; additional deletion affecting intron
			143611241	143611266	25			
			143733282	143733306	24			
			143801279	143801308	29			
GO:0060021	roof of mouth development	9	3094290	3094345	55	*ift140*	Intraflagellar transport 140	Affects separating various transcripts
GO:1904888	cranial skeletal system development	12	85226869	85226898	29	*mthfd1l*	Methylenetetrahydrofolate dehydrogenase, cyclohydrolase and formyltetrahydrofolate 1	One deletion affects final, non-coding exon of one Mthfd1l transcript; another affects separating various transcripts of Mthfd1l
			85181772	85181796	24			
GO:0060173	limb development	28	7046607	7046643	36	*aldh1a2*	Aldehyde dehydrogenase 1 family member A2	Intergenic region about 390 kb downstream of gene
GO:0060173	limb development	4	67183641	67183674	33	*col6a1-col6a2*	Collagen type 6 Alpha 1 chain	Intergenic region about 58 kb and 113 kb upstream of of Col6a2
			67128830	67128857	27			
GO:0060173	limb development	4	67096427	67096452	25			Affects intron separating various transcripts
GO:0060173	limb development	1	171389576	171390553	977	*bpnt2*	3'(2'), 5'-bisphosphate nucleotidase 2	Intergenic region about 385 kb, 367 kb, 159 kb downstream, and 816 kb upstream of gene
			171398379	171399356	977			
			171606380	171606405	25			
			171920782	171920806	24			

			171816127	171816154	27			Affects first, non-coding exon, potentially affecting transcription factor/promoter
GO:0060173	limb development	25	27631042	27631084	42	*alx3*	ALX homeobox 3	One deletion affects region just upstream of Alx3 about 2 kb, potentially affecting promoter region; another affects downstream region of Alx3 and a non-coding exon of one alternative transcript
			27681163	27681217	54			
			27722844	27722874	30			Intergenic region about 40 kb downstream of Alx3
GO:0060173	limb development	3	162764742	162764763	21	*ark2ca*	Arkadia (RNF111) C-terminal like ring finger ubiquitin ligase 2Ca	Affects intron of Ark2ca
GO:0060173	limb development	3	52525744	52525770	26	*dync2h1*	Cytoplasmic dynein 2 heavy chain 1	Two deletions affecting intron of all Dync2h1transcripts
			52444807	52444832	25			
GO:0060173	limb development	4	7330147	7330183	36	*en1a*	Engrailed homeobox 1a	Intergenic region about 61 kb downstrean of En1a
GO:0060173	limb development	22	33470331	33470355	24	*dkk1b*	Dickkopf WNT signaling pathway inhibitor 1	intergenic region about 6 kb downstream of gene
GO:0060173	limb development	16	24740566	24740593	27	*comp*	Cartilage oligomeric matrix protein	Affects downstream region of Comp transcripts
GO:0060173	limb development	8	142966239	142966265	26	*cacna1c*	Calcium voltage-gated channel subunit alpha1 C	Intergenic region 8 kb upstream of gene, possibly affecting pomoter region
GO:0060173	limb development	4	129415536	129415565	29	*dmd*	Dystrophin	19 deletions affecting intron of various Dmd transcripts
			129433821	129433850	29			
			129821368	129821391	23			
			129822101	129822158	57			
			130000129	130000173	44			
			128372576	128372601	25			
			128373618	128373669	51			
			128374217	128374244	27			
			128455400	128455425	25			
			128492113	128492169	56			
			128868211	128868250	39			
			128883534	128883560	26			
			128158881	128158916	35			
			128978191	128978227	36			
			129066458	129066512	54			
			129085396	129085423	27			
			129103098	129103140	42			
			129122813	129122855	42			
			129161565	129161589	24			

Table 5: Prioritised genes reported to play a role in craniofacial development affected by deletions in the ancestral mobulid (Anc3) genome; obtained by literature searches (Lamichhaney *et al.*, 2015; Mork and Crump, 2015; Richmond *et al.*, 2018)

Chromosome	Start	End	Size (bp)	Gene symbol	Gene name	SnpEff annotation
18	54701711	54701752	41	*wdr31*	WD repeat domain 31	Affects introns separating several transcripts
17	35882693	35882720	27	*chd9*	Chromodomain-helicase-DNA-binding protein 9	Two deletions affecting introns separating various transcripts
	36111920	36111950	30			
4	28672436	28672472	36	*hoxd3a*	Homeobox protein Hox-D3a	Affects intergenic region about 2 kb upstream of the gene
6	31630701	31630738	37	*wdr37*	WD repeat domain 37	Two deletions introns separating several transcripts
	31655982	31656012	30			
14	9097906	9097935	29	*prdm5*	PR/SET domain 5	Four deletions in intergenic region within 403 kb downstream of gene
	8971818	8971852	34			
	8703914	8703943	29			
	8461185	8461210				
5	145893189	145893212	23	*fgf14*	Fibroblast growth factor 14	Affects introns separating several transcripts and upstream region of alternative transcripts
3	163718760	163718785	25	*fgfr3*	Fibroblast growth factor receptor 3	Two deletions affecting the first, non-coding exon separating several transcripts and upstream region of other transcripts, possibly affecting promoter; another deletion affecting introns separating various transcripts
	163718582	163718606	24			
	163677345	163677369	24			
17	64262496	64262524	28	*irf8*	Interferon regulatory factor 8	Two deletions affecting intergenic region within 132 kb downstream of gene
	64333449	64333484	35			
17	61351675	61351700	25	*crispld2*	Cysteine rich secretory protein LCCL domain containing 2	Two deletions affecting intergenic region within 71 kb downstream of gene
18	74077593	74077617	24	*tbx1*	T-box protein 1	Affects downstream region of last, non-coding intron
18	18575860	18575884	24	*med27*	Mediator of RNA polymerase II transcription subunit 27	Three deletions affecting introns separating various introns across transcripts
	18824576	18824605	29			
	18914235	18914267	32			
10	75111678	75111709	31	*bmp5*	Bone morphogenetic protein 5	Two deletions, one affecting intron after the first coding exon of various transcripts, another also affecting introns separating transcripts
	74963442	74963466	24			
10	32840585	32840609	24	*supt3h*	SUPT3H SPT3 homolog, SAGA and STAGA complex component	Two deletions affecting introns separating various transcripts
	32825069	32825095	26			
X1	51945045	51945085	40	*dlx3b*	Distal-less homeobox 3	Two deletions affecting intergenic region 11 and 33 kb upstream of gene
	51923658	51923684	26			
X1	37909738	37909764	26	*med24*	Mediator of RNA polymerase II transcription subunit 24	Affects intron separating two transcripts and downstream region of two alternative transcripts
14	109103205	109103242	37	*wdr32*	WD repeat domain 32	Affects introns across various transcripts near the last coding exon
	18314970	18315013	43			

23	18556511	18556659	148	*sox9a*	SRY-box transcription factor 9a	Five deletions affecting intergenic region 6 and 248 kb upstream of gene, and 33, 175 and 242 kb downstream of gene
	18596037	18596078	41			
	18980266	18980298	32			
	18737055	18737082	27			
10	44576985	44577026	41	*wdr35*	WD repeat domain 35	Affects introns separating various transcripts
19	22263380	22263404	24	*cacna2d3a*	Calcium channel, voltage-dependent, alpha 2/delta subunit 3	Two deletions: one affects introns separating most transcripts and downstream regions of six alternative transcripts, he other intergenic region between Cacnad3a and Wnt5,
10	28004259	28004344	85	*med23*	Mediator of RNA polymerase II transcription subunit 23	Affects introns separating various transcripts
4	150780290	150780318	28	*tbx15*	T-box protein 15	Two deletions, one affects introns separating the four transcripts, another intergenic region
	151101555	151101590	35			
10	153681607	153683213	1606	*tbx16*	T-box protein 16	Two deletions, one affecting introns separating both transcripts and one affecting upstream region of the gene, possibly impacting promoter
	153827740	153829346	1606			

Figure 11: The location of example Anc3 deletions (bottom track on each panel, black vertical rectangles) in the conserved region of the stingray genome, visualised using the NCBI Genome Viewer. Genes represented by green tracks with arrows. Thin horizontal lines are introns, while the vertically elongated rectangles are exons (coding exons: dark green; non-coding exons: light green. The blue track shows RNASeq mapping coverage used in the RefSeq annotation. Purple tracks show non-coding transcripts of other neighbouring genes and should be ignored. A) Upstream to first *shha* exon, possibly affecting promoter; B) Upstream to first *alx3* exon, possibly affecting promoter; C) Within first exon of *bnpnt2*. D) Within last exon of *mthdl1*.

V. Enriched gene pathways affected by ancestral mobulid-specific deletions

14,303 *H. sabinus* annotated genes were available for the GO analysis. An adjusted p-value < 0.1 was used as the cut-off for considering a term significant. After simplification to reduce term redundancy, the five most enriched pathways were related to cell development and neurogenesis (Figure 12A). Among 60 significantly enriched terms, 16 were associated with neurogenesis and nervous system development. The most enriched ones included "generation of neurons" (GO:0048699; adjusted p < 0.0001; gene ratio = 93/1950; background ratio = 349/14303); "neurogenesis" (GO:0022008; adjusted p < 0.0001; gene ratio = 99/1950; background ratio = 389/14303); "neuron development" (GO:0048666; adjusted p < 0.0001; gene ratio = 72/1950; background ratio = 255/14303); "neuron differentiation" (GO:0030182; adjusted p < 0.0001; gene ratio = 86/1950; background ratio = 336/14303); and "neuron projection development" (GO:0031175; adjusted p < 0.0001; gene ratio = 65/1950; background ratio = 236/14303). Genes explaining these terms overlapped significantly across the various nervous system development-related terms (*e.g.*, *lhx6a*, *sema3h*, *plxnd1b*, *crtac1b*, among others) as with cell development-related terms such as GO:0048468 ("cell development"), GO:0000902 ("cell morphogenesis"), among others. Interestingly, some of the genes related to neurogenesis terms, and in particular the most enriched term globally (GO:0048699; "generation of neurons"), also overlapped with a less significant term related to multicellular organism development, *i.e.*, GO: 0050793: "regulation of developmental process" (adjusted p-value = 0.087; gene ratio = 52/1950; background ratio = 259/14303). Overlapping genes between these terms included *sema3h*, *plnxd1b*, *pacsin1b*, *sema6dl*, and *run2xa*.

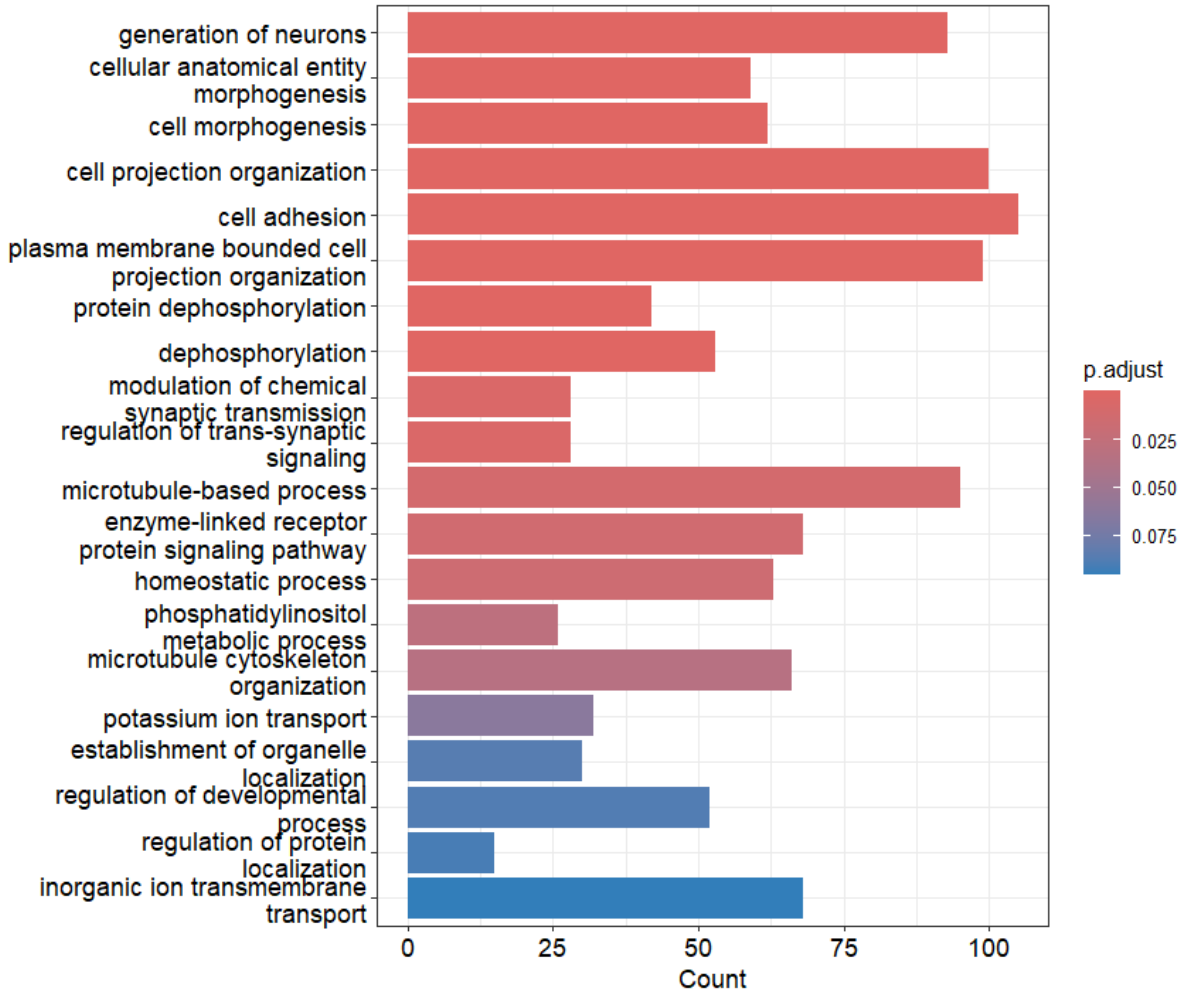


Figure 12: Enriched GO terms (biological functions) affected by ancestral mobulid-specific deletions

VI. Demographic history of M. hypostoma

The demographic history plot generated by the Pairwise Sequential Markovian Coalescent (PSMC) model using a generation time of 25 years showed two periods of effective population size decrease separated by a period of expansion (Figure 13A). Effective population size decreased between 2 and 1 Mya to about 15,000 individuals. Between 1 Mya and 300 kya, effective population size then increased to reach a peak and historical maximum at about 37-38,000 individuals. Another decrease then occurred between 300 and 100 kya to less than 5,000 individuals. The curve was identical for the plots generated with a generation time of 20 and 29 years. The population size maximum was shifted about 2 ky later, and 1 My earlier for the 20 and 29 year generation times, respectively (Figure 15B and C).

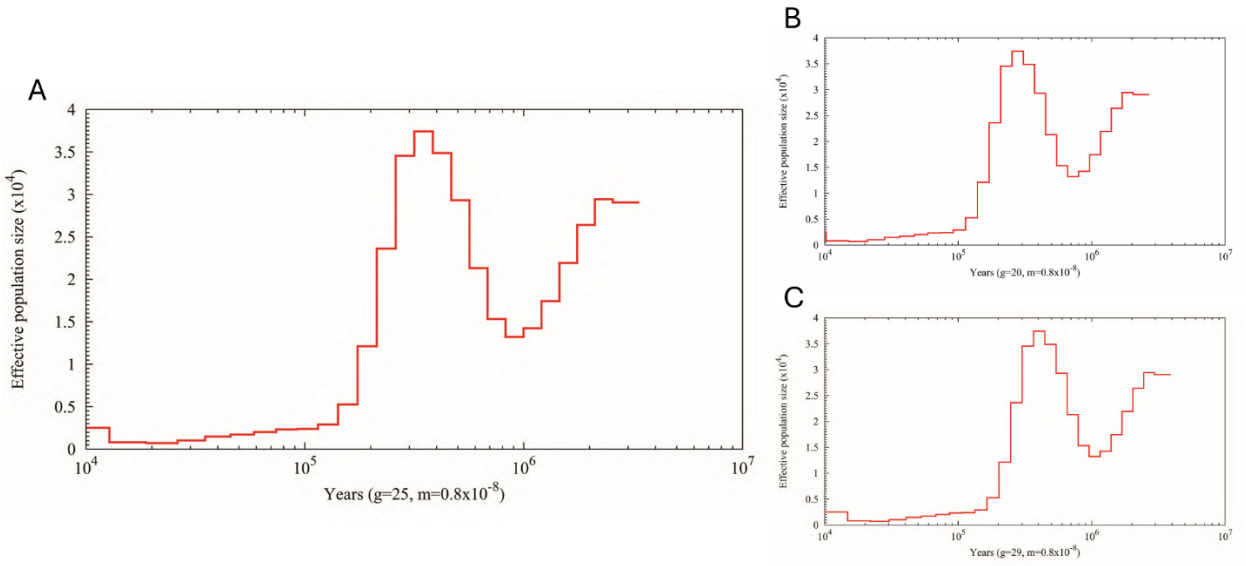


Figure 13: PSMC modelling of *M. hypostoma* demographic history. Effective population size plotted using a A) 25-year generation time, B) 20-year generation time, and C) 29-year generation time.

Chapter 4. Discussion

Aligning whole, high-quality batoid genomes including both *M. birostris* and *M. hypostoma*, allowed me to gain insight into the composition and evolution of mobulid genomes. The results of my analyses shed light onto mobulid genome architecture and reorganisation compared to *H. sabinus,* as well as the reconstructed ancestral myliobatiform and ancestral mobulid genomes. Repetitive sequences accounted for most of the mobulid genome sizes, and were proportionally more abundant compared to *H. sabinus* and the skates. Rearrangements, and in particular deletion events, were identified as potential drivers of mobulid phenotypic specialisation. A range of deletions inferred in the ancestral mobulid genome affected non-coding regions of genes shown previously to be involved in craniofacial development, which may be linked to the derived mobulid body plan. These findings are discussed further in the sections below.

I. Ancestral mobulid genome evolution

i. Repetitive sequences drive genome size variation

The five aligned batoid genomes had a predicted repeat content exceeding 60%. This exceeds values seen in teleosts and tetrapods, but sits in line with values reported for other elasmobranchs, *e.g.,* around 59% and 50% in the white and whale sharks, respectively (Hara *et al.*, 2018; Marra *et al.*, 2019). Moreover, as the two mobulid genomes contained approximately the same non-repetitive content, genome size variation between *M. birostris* and *M. hypostoma* likely results from limited differences in repetitive sequences. Retroelements formed the bulk of repetitive sequences in both mobulids. The overwhelming majority were retroelements, and in particular LINEs. The predominance of LINEs was also a feature of brownbanded bamboo shark, whale shark and cloudy catshark (Hara *et al.*, 2018).

Retroelements like LINEs have been established as a source of genomic innovation and rearrangements. By inserting themselves into genes, retroelements can disrupt both coding and non-coding regions, thereby modifying protein and regulatory sequences like promoters, or leading to the creation of new exons (Bourque, 2009). Novel genes may even arise from the proliferation of these repetitive sequences, as demonstrated for placental protein-coding

genes in mammals (Böhne *et al.*, 2008). As a result, repetitive DNA has increasingly been associated with evolutionary diversification and adaptation to new environmental conditions or niches (Sotero-Caio *et al.*, 2017). The similar amount of retroelements in both mobulid genomes and *H. sabinus* suggests a possible expansion of repeat content along the ancestral myliobatiform branch. Further investigations will be needed to determine whether specific genes have been impacted by repeat expansions during mobulid evolutionary history. In this respect, the genome alignment results I produced will make it possible to connect mobulid-specific repeat expansions with genome restructuring events, and pinpoint precise impacts of repeats on genes, which may be linked to mobulid adaptations.

ii. Evidence of karyotype reorganisation in the ancestral mobulid genome

Synteny analyses using whole genome alignment data highlighted a high degree of chromosome rearrangements between *H. sabinus* and *M. hypostoma*. In contrast, one-to-one homology was mostly maintained with a few rearranged chromosome fragments between *M. hypostoma* and *M. birostris.* This is consistent with large-scale karyotype reorganization in the mobulid ancestor, followed by more limited changes within *Mobula.* The largest twelve *M. hypostoma* chromosomes appeared to be fusions of non-overlapping *H. sabinus* chromosome halves; it can be hypothesised that the current *H. sabinus* karyotype is an approximation of the ancestral myliobatiform state, and that a large part of the ancestral mobulid genome arose from the fusion of ancestral myliobatiform chromosomes. However, further reconstruction of both ancestral karyotypes and additional pairwise synteny analyses between skate and myliobatiform genomes will be needed to ascertain the likelihood of this scenario.

Chromosome rearrangements are known to promote speciation and ecological niche adaptation (Guerrero and Kirkpatrick, 2014). Chromosome fusions in particular can bring together previously unrelated loci and create new co-evolving gene clusters with reduced recombination rates (Liu *et al.*, 2022). Given the hypothesised rearrangements of the ancestral mobulid karyotype, it can be hypothesised that genes allowing for a highly mobile, open-water lifestyle - with roles in energy generation and storage, pectoral fin muscle development, and metabolic pathways allowing for great depth variation tolerance — were rearranged into linked groups, with this arrangement being subsequently positively selected. It can be further

speculated that these advantageous genes may then have reduced constraints such as recombination and increased in frequency in the population, later possibly allowing reproductive isolation (through both reduced niche overlap and genomic divergence) and the evolution of the mobulid body plan and open-water foraging lifestyle. Further research will be needed to determine if these potentially fused chromosomes are enriched in such adaptive loci.

Chromosome architecture appeared largely conserved between the two mobulid genome assemblies. Of note however was the apparent difference in assembled sex chromosomes: the sequenced *M. hypostoma* individual, like *H. sabinus*, has two X chromosomes and a Y chromosome (X1-X2-Y), while the sequenced *M. birostris* individual has one X (homologous to *M. hypostoma* X1) and a Y chromosome. Instead, *M. hypostoma* chromosome X2 appears homologous to a region assembled on *M. birostris* chromosome 20. Based on their more distant relationship, it seems likely that the inferred X1-X2-Y system shared by stingray and *M. hypostoma* represents the ancestral myliobatiform state. Thus, assuming the assembly of *M. birostris* sex chromosomes is correct, I infer a loss of X2 in *M. birostris*, after its split from *M. hypostoma*, though the result could also be explained by an assembly artefact. Based on recent phylogenetic revision of genus *Mobula*, *M. hypostoma* is part of a clade which is sister to all other mobulid species, including the two manta species (Cracknell *et al.*, 2018). Therefore, if the lack of a second X chromosome is correct, the loss probably occurred after the *M. hypostoma* and manta branches split. Further intra-genus comparative studies of genome architecture will be needed to gain a better understanding of sex chromosome evolution in mobulids. Previous elasmobranch karyotyping studies which used chromosome level shark assemblies from various genera, revealed a highly conserved X-Y chromosome system in sharks (Yamaguchi *et al.*, 2023; Wu *et al.*, 2024). In these studies, X chromosome origin was established at about 180 mya, *i.e.*, after the Selachii and Batoidea divergence. With the increasing availability of batoid genomes however, including the *M. hypostoma* assembly, further sex chromosome comparisons should be carried out to confirm the evolutionary history of sex chromosomes in elasmobranchs. Investigating whether the shark X chromosome is homologous to a batoid X chromosome across ray and skate lineages would allow refinement of the timeline through which the X-Y, and possible myliobatiform X1-X2-Y, evolved.

An X1-X2-Y system arises when a Y chromosome fuses with an autosome, leaving a pair of X1 chromosomes and an unpaired autosome which eventually becomes an X2 (Pennell *et al.*, 2015). Based on my results, it can be speculated that this Y chromosome fusion may have occurred during batoid evolution, or later in a myliobatiform ancestor; subsequently, the X2 chromosome may have fused anew with a fragment of an autosome along the *M. birostris* branch. Sex chromosome rearrangements have been shown to participate in reproductive isolation (Pennell *et al.*, 2015). This potential autosomal fusion of ancestral mobulid chromosome X2 along the manta branch may have ultimately facilitated *M. birostris* and *M. alfredi* speciation.

iii. Ancestral mobulid structural rearrangements as a source of genomic evolution

Inferred SRs in batoid genome evolution ranged from small insertions and deletions to much larger structural variants. SRs can affect genes and their expression, through changing gene copy number, regulatory elements, deleting functional sequences, or leading to the formation of new genes; moreover, SRs can affect the 3D structure of the genome and recombination rates Merot. Such genomic changes can therefore lead to the origin and selection of new, adaptive phenotypes that may facilitate reproductive separation and speciation (Wellenreuther *et al.*, 2019; Mérot *et al.*, 2020). The ancestral mobulid genome reconstructed using the *M. hypostoma* and *M. birostris* genomes allowed me to investigate SRs that may be linked to the evolution of ancestral traits shared by both mobulids, which are derived from the ancestral state for batoids, represented in my data by conserved sequences in the Atlantic stingray and two skate genomes.

Ancestral mobulid-specific SRs were dominated by insertions. As these represented over a quarter of the reconstructed Anc3 genome length, many are likely associated with repetitive sequences, and in particular retroelements (Sotero-Caio *et al.*, 2017). As future work, insertions and repeat sequences should be intersected to differentiate inserted repeat sequences from non-repeat ones, and identify their potential impacts on genes and regulatory elements. Duplications are also likely to overlap with repetitive sequences; however, DNA transposons, *i.e.*, “copy-and-paste” repeats which would lead to duplicated repetitive sequences, represented a very small proportion of the repetitive sequences in the mobulid

genomes, and it is therefore likely DNA transposons represented a small fraction of ancestral mobulid repeats as well. Like insertions, duplications can affect gene dosage or lead to new coding and regulatory sequences (Ho, Urban and Mills, 2020; Mérot *et al.*, 2020). Further work will therefore be needed to elucidate the origin of these duplications and their effects on biology and adaptation in mobulid evolutionary history.

Inversions and transpositions can drive reproductive isolation and maintain diversification despite gene flow between co-occurring populations (McGee *et al.*, 2020; Mérot *et al.*, 2020), which may have been the case in the early stages of mobulid differentiation from other myliobatiforms. Transpositions, though evidently low in number compared to other SR types, represented over a million rearranged base pairs in the ancestral mobulid genome. Inversions in contrast accounted for a negligible portion of SRs and rearranged bases. This suggests that transpositions, along with chromosome-scale rearrangements mentioned earlier, may have played a greater role in creating a reproductive barrier between early mobulids and their myliobatiform relatives than inversions, although further research will be needed to confirm this. On the other hand, the prevalence of other SR types compared to transpositions and inversions could be a sign that this reproductive isolation was achieved predominantly through niche separation, with a shift to a mostly pelagic habitats, allowing for further selection of mobulid adaptive morphology in a second stage. To establish which evolutionary scenario is more likely, a deeper dive should be taken to study all SR types in the reconstructed ancestral mobulid genome, which was beyond the scope of my project.

Deletions can affect genes directly when they occur in coding sequences, or indirectly by affecting promoter and regulatory regions. In exons, deletions can affect the final protein product by causing a frameshift in the mRNA transcript, or completely disrupting it, leading to a defective or absent protein product (Lin *et al.*, 2017). Deletions in untranslated exons (*i.e.* 5' or 3' UTRs) could also impact which transcripts are expressed, affecting mRNA stability, expression and translation (Mayr, 2019; Ryczek, Łyś and Makałowska, 2023). Gene expression can also be affected by deletions disrupting promoter or regulatory non-coding regions (Lin *et al.*, 2017). Therefore, deletion events identified in my results provided a rather straightforward opportunity to identify potential effects on ancestral mobulid phenotypes, and in particular on development and morphology. These deletions can be hypothesised to have led to

ancestral mobulid adaptations through modified craniofacial morphology, as discussed in the following section.

II. Evolution of the mobulid craniofacial phenotype

i. Craniofacial development evolution may have led to a new, adaptive head phenotype

The majority of ancestral mobulid-specific deletions in the region of craniofacial development-related genes were located in introns or up/downstream intergenic regions, which makes it difficult to predict whether they have functional activity. Non-coding regions upstream and downstream of genes, even hundreds of kb away, often contain conserved *cis*-regulatory elements, *i.e.*, enhancers or silencers, which orchestrate specific patterns of expression (Onimaru, 2020; Rajderkar *et al.*, 2024). Enhancer disturbances in proximity to craniofacial development genes in mouse have been shown to lead to skeletal defects (Crane-Smith *et al.*, 2023). Moreover, differences in enhancer landscape were the proposed source of divergent craniofacial features between human and chimp, despite highly conserved coding sequence identity (Elias *et al.*, 2019). Some of the deletions located in intergenic regions near craniofacial development genes may therefore have affected enhancer sequences, which would have led to differential gene expression and a modified head phenotype in ancestral mobulids (Rebeiz and Tsiantis, 2017).

A few identified deletions overlapped with the promoter or UTR regions of key craniofacial development signalling and regulatory genes. Altered promoter binding may affect the transcription of these genes, with impacts on craniofacial phenotype (Ponomarenko *et al.*, 2002; Wang *et al.*, 2020). *Shh* signalling in particular, which is expressed in the fronto-nasal domain of the developing head, is known to control cranial skeleton formation as well as tissue patterning and differentiation (Adameyko and Fried, 2016). Alterations in *shh* expression can be speculated to have contributed to craniofacial changes from the basal batoid body plan. Craniofacial morphogenesis and tissue differentiation processes are also regulated by fibroblast growth factors (FGFs); a deletion in the 5'UTR region of *fgfr3*, which is known to cause cranial cartilage malformations in model vertebrate species, therefore plausibly contributed to an altered cranioskeletal morphology in the ancestral mobulid (Adameyko and Fried, 2016; Morice *et al.*, 2023). The hypothesis of ancestral mobulid craniofacial remodelling

is further supported by a deletion overlapping the promoter of *alx3*, an ALX homeobox gene responsible for regulating fronto-nasal cell differentiation during embryonic development (Lakhwani, García-Sanz and Vallejo, 2010; Mitchell *et al.*, 2021).

Further research will be required to confirm whether enhancers and promoters are functionally affected by these deletions. Chromatin immune-precipitation sequencing (ChIP-Seq) and assay for transposase accessible chromatin sequencing (ATAC-seq) of developing head tissue sampled from stingray and mobulid embryos could be used to identify and compare active enhancers and promoters at different developmental stages (Rajderkar *et al.*, 2024). Access to mobulid embryos represents a challenge however, as manta and devil rays are viviparous and sensitive to handling stress (Stewart *et al.*, 2018). A bioinformatic approach could involve mapping highly conserved non-coding elements (CNEs) and intersecting them with deletion events. CNEs are non-coding sequences with extremely low rates of change across tens or hundreds of millions of years of evolution. Many are thought to act as developmental enhancers (Onimaru, 2020; Iliopoulou *et al.*, 2023). Ultimately experiments involving embryological knockout studies of the regulatory genes targeted by the identified deletions may be needed to better understand the processes leading to the specific mobulid head shape.

However, my results are consistent with a polygenic disturbance of canonical batoid craniofacial development, eventually leading to a rearrangement of mobulid facial features and the characteristic modern devil ray phenotype.

ii. Links between neurogenesis and craniofacial development in ancestral mobulid evolution

Interestingly, the most enriched biological pathways in the ancestral mobulid-specific deletions were associated with neurogenesis and nervous system development. As mentioned in the GO analysis results section, several genes involved in neurogenesis are also involved in vertebrate embryonic development. Transcription factor *runx2* for example is involved in bone formation is Osteichthyes and has been shown to affect dorso-ventral patterning of developing zebrafish embryos (Flores *et al.*, 2008). The function of *runx2* in cartilaginous fishes is likely different from bony vertebrate orthologues, although it may still

be involved in cartilaginous skeleton development, including the chondrocranium of elasmobranchs. Deletions affecting these genes therefore may have had consequences on craniofacial development.

It is also possible that altered neurogenesis gene expression directly or indirectly led to mobulid departure from canonical batoid craniofacial morphology. Recently, researchers have reported increasing evidence of coordinated neurogenesis and craniofacial development pathways (Adameyko and Fried, 2016). Early craniofacial development is driven and regulated by neural crest cells migrating to the fronto-nasal area of the future face (Dupin, Creuzet and Le Douarin, 2006; Adameyko and Fried, 2016). After this neural crest cells-driven patterning and differentiation stage, the developing brain and peripheral nervous cells have been found to contribute to craniofacial development organisation and overall head morphology, including through signalling pathways like *shh*-led ones (Adameyko and Fried, 2016). Therefore, it can be hypothesised that altered expression of genes related to nervous system development may be intricately linked to craniofacial development. Future work could focus on neurogenesis genes-related deletions to determine which specific genes or transcription factors may have been altered. *In vivo* embryo studies using close relatives of mobulids like cownose rays (Swenson *et al.*, 2018) could then explore the effects of altered expression of these neurogenesis genes on craniofacial morphology.

iii. From altered craniofacial morphology to adaptive evolution and speciation

Based on the potential head phenotype evolution scenarios described above, hypotheses can be made about the transition between a bottom-feeding, durophagous cownose ray ancestor and a shift to pelagic habitats and zooplankton filter feeding in mobulids. Rhinopterids, i.e. cownose rays, possess mobulid features including wing-like fins and small, fused cephalic horns which are used for feeding and ventrally located (Southward *et al.*, 2004; Swenson *et al.*, 2018). These features were therefore likely shared by the common ancestor to rhinopterids and mobulids, further supporting that the mobulid ecological niche expansion may have been favoured by a modified craniofacial morphology and mouth position. The development of a forward- or near-forward-facing mouth may have driven the shift from seabed to water column feeding. This could have progressively created a reproduction barrier

between ancestral myliobatiforms and mobulids, as hypothesised previously in this discussion. Further lineage-specific adaptive – largely regulatory - mutations of mouth and feeding apparatus morphology, as well as modified constraints on cephalic lobe evolution, may subsequently have cemented mobulid speciation and led to the evolution of the typical manta and devil ray craniofacial morphology we know today.

III. Conservation research applications of the Mobula hypostoma assembly

The *M. hypostoma* reference genome used in my project also provides an opportunity for further conservation genomics analyses. As an example, the PSMC results I obtained constitute a preliminary estimation of past *M. hypostoma* effective population size, which can provide a model for population size trends as a response to past climate and oceanographic events. The reconstructed demographic history detected a dramatic decrease in effective population size between 300,000 and 100,000 years ago. This period of population decline appears to overlap with the Penultimate Gacial Period and the Last Interglacial, the latter of which lasting from 130,000 to 115,000 years before present (Bard, Antonioli and Silenzi, 2002; Margari *et al.*, 2010). The decrease in effective population size on the PSMC plot at this time may therefore reflect a decrease in ancestral mobulid population density in response to a warming climate, as well as a period of sea level rise dated between 200,000-190,000 years before present (Bard, Antonioli and Silenzi, 2002). Effective population size then appears to steadily decrease between 100,000 and 10,000 years agoto about 4,000-5,000 individuals.

In the future, recent demographic history should be investigated, for example using runs of homozygosity analyses, to gain a better understanding of *M. hypostoma* genetic health, current effective population size, inbreeding levels and overall vulnerability to extinction (Allendorf, W. , Luikart and Aitken, 2013; Wang, Santiago and Caballero, 2016). This assembly also provides a robust reference genome for future population-level comparative genomics analyses to gain a better understanding of *M. hypostoma* population structure and connectivity at global and regional scales. As the number of high-quality batoid reference assembly increases, future comparative studies, possibly including more mobulid species, could further improve our understanding of mobulid evolutionary history, as well as species-specific adaptations within the genus.

Conclusion

In this project, I successfully aligned a novel *M. hypostoma* genome with four chromosome-level batoid genome assemblies. I explored ancestral mobulid evolutionary history and identified potential sequence deletion events which may have modified craniofacial development gene expression, thereby leading to phenotype evolution and mobulid speciation. This project opened further research avenues to investigate mobulid adaptation, early vertebrate evolution, and build upon the growing elasmobranch evolutionary genomics research.

References

Adameyko, I. and Fried, K. (2016) 'The nervous system orchestrates and integrates craniofacial development: A Review', *Frontiers in Physiology*. Frontiers Media S.A., p. 185323. doi: 10.3389/fphys.2016.00049.

Allendorf, F., W. , Luikart, G. and Aitken, S. N. (2013) *Conservation and the genetics of populations (2nd ed)*. 2nd edn. Edited by John Wiley & Sons.

Amores, A. *et al.* (1998) 'Zebrafish hox clusters and vertebrate genome evolution', *Science*. American Association for the Advancement of Science, 282(5394), pp. 1711–1714. doi: 10.1126/science.282.5394.1711.

Araujo, G. *et al.* (2020) 'Changes in diving behaviour and habitat use of provisioned whale sharks: implications for management', *Scientific Reports*. Nature Publishing Group, 10(1), pp. 1–12. doi: 10.1038/s41598-020-73416-2.

Armstrong, A. O. *et al.* (2021) 'Reef manta rays forage on tidally driven, high density zooplankton patches in Hanifaru Bay, Maldives', *PeerJ*. PeerJ Inc., 9, p. e11992. doi: 10.7717/peerj.11992.

Armstrong, J. *et al.* (2020) 'Progressive Cactus is a multiple-genome aligner for the thousand-genome era', *Nature*. Nature Publishing Group, 587(7833), pp. 246–251. doi: 10.1038/s41586-020-2871-y.

Aschliman, N. C. *et al.* (2012) 'Body plan convergence in the evolution of skates and rays (Chondrichthyes: Batoidea)', *Molecular Phylogenetics and Evolution*. Academic Press, 63(1), pp. 28–42. doi: 10.1016/j.ympev.2011.12.012.

Bard, E., Antonioli, F. and Silenzi, S. (2002) 'Sea-level during the penultimate interglacial period based on a submerged stalagmite from Argentarola Cave (Italy)', *Earth and Planetary Science Letters*. Elsevier, 196(3–4), pp. 135–146. doi: 10.1016/S0012-821X(01)00600-8.

Benestan, L. (2019) 'Population Genomics Applied to Fishery Management and Conservation', in. Springer, Cham, pp. 399–421. doi: 10.1007/13836_2019_66.

Bernos, T. A. *et al.* (2023) 'Evaluating the evolutionary mechanisms maintaining alternative mating strategies in a simulated bull trout (Salvelinus confluentus) population', *Ecology and Evolution*. John Wiley & Sons, Ltd, 13(4), p. e9965. doi: 10.1002/ece3.9965.

Böhne, A. *et al.* (2008) 'Transposable elements as drivers of genomic and biological diversity in vertebrates', *Chromosome Research*. Springer, pp. 203–215. doi: 10.1007/s10577-007-1202-6.

Bourque, G. (2009) 'Transposable elements in gene regulation and in the evolution of vertebrate genomes', *Current Opinion in Genetics and Development*. Elsevier Current Trends, pp. 607–612. doi: 10.1016/j.gde.2009.10.013.

Brunjes, N. L. *et al.* (2024) 'Genomic population structure of great hammerhead sharks (Sphyrna mokarran) across the Indo-Pacific', *Marine and Freshwater Research*. CSIRO PUBLISHING, 75(6), p. NULL-NULL. doi: 10.1071/mf23236.

Bucair, N. *et al.* (2024) 'Occurrence, distribution and threats to mobulid rays in Brazil: A review and updated database', *Aquatic Conservation: Marine and Freshwater Ecosystems*. John Wiley & Sons, Ltd, 34(6). doi: 10.1002/aqc.4203.

Carpenter, M. *et al.* (2023) 'Multi-decade catches of manta rays (Mobula alfredi, M. birostris) from South Africa reveal significant decline', *Frontiers in Marine Science*. Frontiers Media S.A., 10, p. 1128819. doi: 10.3389/FMARS.2023.1128819/BIBTEX.

Cingolani, P. *et al.* (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*. Taylor & Francis, 6(2), pp. 80–92. doi: 10.4161/fly.19695.

Claeson, K. M. (2014) 'The impacts of comparative anatomy of electric rays (Batoidea: Torpediniformes) on their systematic hypotheses', *Journal of Morphology*. John Wiley & Sons, Ltd, 275(6), pp. 597–612. doi: 10.1002/jmor.20239.

Couturier, L. I. E. *et al.* (2012) 'Biology, ecology and conservation of the Mobulidae', *Journal of Fish Biology*. John Wiley & Sons, Ltd, pp. 1075–1119. doi: 10.1111/j.1095-8649.2012.03264.x.

Couturier, L. I. E. *et al.* (2018) 'Variation in occupancy and habitat use of Mobula alfredi at a major aggregation site', *Marine Ecology Progress Series*. Inter-Research, 599, pp. 125–145. doi: 10.3354/meps12610.

Cracknell, D. L. *et al.* (2018) 'Human Dimensions of Wildlife An International Journal Reviewing the role of aquaria as restorative settings: how subaquatic diversity in public aquaria can influence preferences, and human health and well-being'. doi: 10.1080/10871209.2018.1449039.

Crane-Smith, Z. *et al.* (2023) 'A non-coding insertional mutation of Grhl2 causes gene over-expression and multiple structural anomalies including cleft palate, spina bifida and encephalocele', *Human Molecular Genetics*. Oxford University Press, 32(17), pp. 2681–2692. doi: 10.1093/hmg/ddad094.

de Crécy-Lagard, V. and Hanson, A. (2013) 'Comparative Genomics', in *Brenner's Encyclopedia of Genetics: Second Edition*. Academic Press, pp. 102–105. doi: 10.1016/B978-0-12-374984-0.00299-0.

Croll, D. A. *et al.* (2016) 'Vulnerabilities and fisheries impacts: the uncertain future of manta and devil rays', *Aquatic Conservation: Marine and Freshwater Ecosystems*. John Wiley & Sons, Ltd, pp. 562–575. doi: 10.1002/aqc.2591.

Danecek, P. *et al.* (2021) 'Twelve years of SAMtools and BCFtools', *GigaScience*. Oxford Academic, 10(2), pp. 1–4. doi: 10.1093/gigascience/giab008.

Delaval, A. *et al.* (2022) 'Population and seascape genomics of a critically endangered benthic elasmobranch, the blue skate Dipturus batis', *Evolutionary Applications*. John Wiley & Sons, Ltd, 15(1), pp. 78–94. doi: 10.1111/eva.13327.

Dubail, J. and Cormier-Daire, V. (2021) 'Chondrodysplasias With Multiple Dislocations Caused by Defects in Glycosaminoglycan Synthesis', *Frontiers in Genetics*. Frontiers Media S.A., p. 642097. doi: 10.3389/fgene.2021.642097.

Dudgeon, C. L. *et al.* (2012) 'A review of the application of molecular genetics for fisheries management and conservation of sharks and rays', *Journal of Fish Biology*. John Wiley & Sons, Ltd, pp. 1789–1843. doi: 10.1111/j.1095-8649.2012.03265.x.

Dulvy, N. K. *et al.* (2014) 'Extinction risk and conservation of the world's sharks and rays', *eLife*. eLife Sciences Publications Ltd, 2014(3). doi: 10.7554/elife.00590.

Dulvy, N. K. *et al.* (2017) 'Challenges and Priorities in Shark and Ray Conservation', *Current Biology*. Cell Press, pp. R565–R572. doi: 10.1016/j.cub.2017.04.038.

Dunn, K. A., McEachran, J. D. and Honeycutt, R. L. (2003) 'Molecular phylogenetics of myliobatiform fishes (Chondrichthyes: Myliobatiformes), with comments on the effects of missing data on parsimony and likelihood', *Molecular Phylogenetics and Evolution*. Academic Press, 27(2), pp. 259–270. doi: 10.1016/S1055-7903(02)00442-6.

Dupin, E., Creuzet, S. and Le Douarin, N. M. (2006) 'The contribution of the neural crest to the vertebrate body', *Advances in Experimental Medicine and Biology*. Springer, Boston, MA, pp. 96–119. doi: 10.1007/978-0-387-46954-6_6.

Dworkin, S. *et al.* (2016) 'The role of Sonic hedgehog in craniofacial patterning, morphogenesis and cranial neural crest survival', *Journal of Developmental Biology*. MDPI Multidisciplinary Digital Publishing Institute. doi: 10.3390/jdb4030024.

Ehemann, N. *et al.* (2022) 'Manta and devil ray species occurrence and distribution in Venezuela, assessed through fishery landings and citizen science data', *Journal of Fish Biology*. John Wiley & Sons, Ltd, 101(1), pp. 213–225. doi: 10.1111/jfb.15088.

Ehemann, N. R., González-González, L. V. and Trites, A. W. (2017) 'Lesser devil rays Mobula cf. hypostoma from Venezuela are almost twice their previously reported maximum size and may be a new sub-species', *Journal of Fish Biology*. John Wiley & Sons, Ltd, 90(3), pp. 1142–1148. doi: 10.1111/jfb.13252.

Elias, M. S. *et al.* (2019) 'Functional and proteomic analysis of a full thickness filaggrin-deficient skin organoid model', *Wellcome Open Research 2019 4:134*. F1000 Research Limited, 4, p. 134. doi: 10.12688/wellcomeopenres.15405.2.

Ewels, P. *et al.* (2016) 'MultiQC: Summarize analysis results for multiple tools and samples in a single report', *Bioinformatics*. Oxford Academic, 32(19), pp. 3047–3048. doi: 10.1093/bioinformatics/btw354.

Feitosa, L. M. *et al.* (2018) 'DNA-based identification reveals illegal trade of threatened shark species in a global elasmobranch conservation hotspot', *Scientific Reports*. Nature Publishing Group, 8(1), pp. 1–11. doi: 10.1038/s41598-018-21683-5.

Field, I. C. *et al.* (2009) 'Susceptibility of sharks, rays and chimaeras to global extinction', *Advances in Marine Biology*. Academic Press, pp. 275–363. doi: 10.1016/S0065-2881(09)56004-X.

Flores, M. V. C. *et al.* (2008) 'Osteogenic transcription factor Runx2 is a maternal determinant of dorsoventral patterning in zebrafish', *Nature Cell Biology*. Nature Publishing Group, 10(3), pp. 346–352. doi: 10.1038/ncb1697.

Flowers, K. I., Heithaus, M. R. and Papastamatiou, Y. P. (2021) 'Buried in the sand: Uncovering the ecological roles and importance of rays', *Fish and Fisheries*. John Wiley & Sons, Ltd, 22(1), pp. 105–127. doi: 10.1111/faf.12508.

Guan, D. *et al.* (2020) 'Identifying and removing haplotypic duplication in primary genome assemblies', *Bioinformatics*. Oxford Academic, 36(9), pp. 2896–2898. doi: 10.1093/bioinformatics/btaa025.

Guerrero, R. F. and Kirkpatrick, M. (2014) 'LOCAL ADAPTATION AND THE EVOLUTION OF CHROMOSOME FUSIONS', *Evolution*. Oxford Academic, 68(10), pp. 2747–2756. doi: 10.1111/EVO.12481.

Hall, K. C. *et al.* (2018) 'The evolution of underwater flight: The redistribution of pectoral fin rays, in manta rays and their relatives (Myliobatidae)', *Journal of Morphology*. John Wiley & Sons, Ltd, 279(8), pp. 1155–1170. doi: 10.1002/JMOR.20837.

Haque, A. B. *et al.* (2021) 'Fishing and trade of devil rays (Mobula spp.) in the Bay of Bengal, Bangladesh: Insights from fishers' knowledge', *Aquatic Conservation: Marine and Freshwater Ecosystems*. John Wiley & Sons, Ltd, 31(6), pp. 1392–1409. doi: 10.1002/aqc.3495.

Hara, Y. *et al.* (2018) 'Shark genomes provide insights into elasmobranch evolution and the origin of vertebrates', *Nature Ecology and Evolution*. Nature Publishing Group, 2(11), pp. 1761–1771. doi: 10.1038/s41559-018-0673-5.

Harris, J. L. *et al.* (2020) 'Gone with the wind: Seasonal distribution and habitat use by the reef manta ray (Mobula alfredi) in the Maldives, implications for conservation', *Aquatic Conservation: Marine and Freshwater Ecosystems*. John Wiley & Sons, Ltd, 30(8), pp. 1649–1664. doi: 10.1002/aqc.3350.

Harris, J. L. *et al.* (2024) 'First records of the sicklefin (Mobula tarapacana), bentfin (Mobula thurstoni), and spinetail (Mobula mobular) devil rays in the Chagos Archipelago', *Journal of Fish Biology*. John Wiley & Sons, Ltd, 104(5), pp. 1628–1632. doi: 10.1111/jfb.15678.

Healy, T. J. *et al.* (2020) 'A global review of elasmobranch tourism activities, management and risk', *Marine Policy*. Pergamon, 118, p. 103964. doi: 10.1016/j.marpol.2020.103964.

Ho, S. S., Urban, A. E. and Mills, R. E. (2020) 'Structural variation in the sequencing era', *Nature Reviews Genetics*. Nature Publishing Group, pp. 171–189. doi: 10.1038/s41576-019-0180-9.

Hohenlohe, P. A., Funk, W. C. and Rajora, O. P. (2021) 'Population genomics for wildlife conservation and management', *Molecular Ecology*. John Wiley & Sons, Ltd, 30(1), pp. 62–82. doi: 10.1111/mec.15720.

Hosegood, J., Humble, E., Ogden, R., Bruyn, M. de, *et al.* (2020) 'Phylogenomics and species delimitation for effective conservation of manta and devil rays', *Molecular Ecology*, p. mec.15683. doi: 10.1111/mec.15683.

Hosegood, J., Humble, E., Ogden, R., de Bruyn, M., *et al.* (2020) 'Phylogenomics and species delimitation for effective conservation of manta and devil rays', *Molecular Ecology*. John Wiley & Sons, Ltd, 29(24), pp. 4783–4796. doi: 10.1111/mec.15683.

Hu, G., Li, J. and Zeng, G. (2013) 'Recent development in the treatment of oily sludge from petroleum industry: A review', *Journal of Hazardous Materials*, 261, pp. 470–490. doi: 10.1016/j.jhazmat.2013.07.069.

Hu, J. *et al.* (2024) 'NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads', *Genome Biology*. BioMed Central Ltd, 25(1), pp. 1–19. doi: 10.1186/S13059-024-03252-4/FIGURES/3.

Humble, E. *et al.* (2023) 'Comparative population genomics of manta rays has global implications for management', *Molecular Ecology*. John Wiley and Sons Inc. doi: 10.1111/mec.17220.

Iliopoulou, E. *et al.* (2023) 'Extensive Loss and Gain of Conserved Non-Coding Elements during Early Teleost Evolution', *bioRxiv*. Cold Spring Harbor Laboratory, p. 2023.05.23.541954. doi: 10.1101/2023.05.23.541954.

Irisarri, I. *et al.* (2017) 'Phylotranscriptomic consolidation of the jawed vertebrate timetree', *Nature Ecology and Evolution*. Nature Publishing Group, 1(9), pp. 1370–1378. doi: 10.1038/s41559-017-0240-5.

Jabado, R. W. (2018) 'The fate of the most threatened order of elasmobranchs: Shark-like batoids (Rhinopristiformes) in the Arabian Sea and adjacent waters', *Fisheries Research*. Elsevier, 204, pp. 448–457. doi: 10.1016/j.fishres.2018.03.022.

Johnston, I. A. *et al.* (2024) 'Advancing fish breeding in aquaculture through genome functional annotation', *Aquaculture*. Elsevier, p. 740589. doi: 10.1016/j.aquaculture.2024.740589.

Johri, S. *et al.* (2019) 'Taking advantage of the genomics revolution for monitoring and conservation of chondrichthyan populations', *Diversity*. Multidisciplinary Digital Publishing Institute, p. 49. doi: 10.3390/d11040049.

Kadota, M. *et al.* (2023) 'Shark and ray genome size estimation: methodological optimization for inclusive and controllable biodiversity genomics', *F1000Research*. F1000 Research Limited, 12, p. 1204. doi: 10.12688/f1000research.136385.1.

Kawashima, T. (2018) 'Comparative and evolutionary genomics', in *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*. Academic Press, pp. 257–267. doi: 10.1016/B978-0-12-809633-8.20236-7.

Khan, S. *et al.* (2016) 'Overview on the Role of Advance Genomics in Conservation Biology of Endangered Species', *International Journal of Genomics*. John Wiley & Sons, Ltd, p. 3460416. doi: 10.1155/2016/3460416.

King, B. L. *et al.* (2011) 'A natural deletion of the HoxC cluster in elasmobranch fishes', *Science*. American Association for the Advancement of Science, p. 1517. doi: 10.1126/science.1210912.

Krueger, F. *et al.* (2023) 'rimGalore: v0.6.10'. Zenodo.

Kundu, R., Casey, J. and Sung, W.-K. (2019) 'HyPo: Super Fast & Accurate Polisher for Long Read Genome Assemblies', *bioRxiv*. Cold Spring Harbor Laboratory, p. 2019.12.19.882506. doi: 10.1101/2019.12.19.882506.

Kuraku, S. (2021) 'Shark and ray genomics for disentangling their morphological diversity and vertebrate evolution', *Developmental Biology*. Academic Press, 477, pp. 262–272. doi: 10.1016/j.ydbio.2021.06.001.

Lakhwani, S., García-Sanz, P. and Vallejo, M. (2010) 'Alx3-deficient mice exhibit folic acid-resistant craniofacial midline and neural tube closure defects', *Developmental Biology*. Academic Press, 344(2), pp. 869–880. doi: 10.1016/j.ydbio.2010.06.002.

Lamichhaney, S. *et al.* (2015) 'Evolution of Darwin's finches and their beaks revealed by genome sequencing', *Nature*. Nature Publishing Group, 518(7539), pp. 371–375. doi: 10.1038/nature14181.

Lawson, J. M. *et al.* (2017) 'Sympathy for the devil: A conservation strategy for devil and manta rays', *PeerJ*. PeerJ Inc., 2017(3), p. e3027. doi: 10.7717/peerj.3027.

Layton, K. K. S. *et al.* (2021) 'Genomic evidence of past and future climate-linked loss in a migratory Arctic fish', *Nature Climate Change*. Nature Publishing Group, 11(2), pp. 158–165. doi: 10.1038/s41558-020-00959-7.

Li, H. and Durbin, R. (2011) 'Inference of human population history from individual whole-genome sequences', *Nature*. Europe PMC Funders, 475(7357), pp. 493–496. doi: 10.1038/nature10231.

Li, H. and Durbin, R. (2024) 'Genome assembly in the telomere-to-telomere era', *Nature Reviews Genetics*. Nature Publishing Group, pp. 658–670. doi: 10.1038/s41576-024-00718-w.

Lin, M. *et al.* (2017) 'Effects of short indels on protein structure and function in human genomes', *Scientific Reports*. Nature Publishing Group, 7(1), pp. 1–9. doi: 10.1038/s41598-017-09287-x.

Liu, Z. *et al.* (2022) 'Chromosomal Fusions Facilitate Adaptation to Divergent Environments in Threespine Stickleback', *Molecular Biology and Evolution*. Oxford Academic, 39(2). doi: 10.1093/molbev/msab358.

Margari, V. *et al.* (2010) 'The nature of millennial-scale climate variability during the past two glacial periods', *Nature Geoscience*. Nature Publishing Group, 3(2), pp. 127–131. doi: 10.1038/ngeo740.

Marlétaz, F. *et al.* (2023) 'The little skate genome and the evolutionary emergence of wing-like fins', *Nature*, 616(7957), pp. 495–503. doi: 10.1038/s41586-023-05868-1.

Marra, N. J. *et al.* (2019) 'White shark genome reveals ancient elasmobranch adaptations associated with wound healing and the maintenance of genome stability', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 116(10), pp. 4446–4455. doi: 10.1073/pnas.1819778116.

Marshall, A. D. and Bennett, M. B. (2010) 'Reproductive ecology of the reef manta ray Manta alfredi in southern Mozambique', *Journal of Fish Biology*. John Wiley & Sons, Ltd, 77(1), pp. 169–190. doi: 10.1111/j.1095-8649.2010.02669.x.

Marshall, A. D., Compagno, L. J. V. and Bennett, M. B. (2009) 'Redescription of the genus Manta with resurrection of Manta alfredi (Krefft, 1868) (Chondrichthyes; Myliobatoidei; Mobulidae)', *Zootaxa*, (2301), pp. 1–28. doi: 10.11646/zootaxa.2301.1.1.

Marshall, A. et al. (2022) 'Mobula hypostoma (amended version of 2019 assessment).', *The IUCN Red List of Threatened Species 2022: e.T126710128A214399766*. doi: https://dx.doi.org/10.2305/IUCN.UK.2019-3.RLTS.T126710128A896599.en.

Mather, N., Traves, S. M. and Ho, S. Y. W. (2020) 'A practical introduction to sequentially Markovian coalescent methods for estimating demographic history from genomic data', *Ecology and Evolution*. John Wiley & Sons, Ltd, pp. 579–589. doi: 10.1002/ece3.5888.

Mayr, C. (2019) 'What are 3' utrs doing?', *Cold Spring Harbor Perspectives in Biology*. Cold Spring Harbor Laboratory Press, 11(10), p. a034728. doi: 10.1101/cshperspect.a034728.

McGee, M. D. *et al.* (2020) 'The ecological and genomic basis of explosive adaptive radiation', *Nature*. Nature Publishing Group, 586(7827), pp. 75–79. doi: 10.1038/s41586-020-2652-7.

Medeiros, A. M. *et al.* (2022) 'Endangered mobulids within sustainable use protected areas of southeastern Brazil: occurrence, fisheries impact, and a new prey item', *Environmental Biology of Fishes*. Springer Science and Business Media B.V., 105(6), pp. 775–786. doi: 10.1007/s10641-022-01282-0.

Mérot, C. *et al.* (2020) 'A Roadmap for Understanding the Evolutionary Significance of Structural Genomic Variation', *Trends in Ecology and Evolution*. Elsevier Current Trends, pp. 561–572. doi: 10.1016/j.tree.2020.03.002.

Meyer, A. and Van De Peer, Y. (2005) 'From 2R to 3R: Evidence for a fish-specific genome duplication (FSGD)', *BioEssays*. John Wiley & Sons, Ltd, pp. 937–945. doi: 10.1002/bies.20293.

Le Mézo, P. *et al.* (2022) 'Global nutrient cycling by commercially targeted marine fish', *Biogeosciences*. Copernicus GmbH, 19(10), pp. 2537–2555. doi: 10.5194/bg-19-2537-2022.

Misawa, R., Babaran, R. P. and Motomura, H. (2023) 'Okamejei panayensis sp. nov., a new skate (Rajiformes: Rajidae) from the Philippines', *Ichthyological Research*. Springer, 70(1), pp. 161–176. doi: 10.1007/s10228-022-00874-1.

Mitchell, J. M. *et al.* (2021) 'The alx3 gene shapes the zebrafish neurocranium by regulating frontonasal neural crest cell differentiation timing', *Development (Cambridge)*. Company of Biologists Ltd, 148(7). doi: 10.1242/dev.197483.

Momb, J. *et al.* (2013) 'Deletion of Mthfd1l causes embryonic lethality and neural tube and craniofacial defects in mice', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 110(2), pp. 549–554. doi: 10.1073/pnas.1211199110.

Moreira, R. A. and de Carvalho, M. R. (2021) 'Phylogenetic significance of clasper morphology of electric rays (Chondrichthyes: Batoidea: Torpediniformes)', *Journal of Morphology*. John Wiley & Sons, Ltd, 282(3), pp. 438–448. doi: 10.1002/jmor.21315.

Moriarty, R. and O'Brien, T. D. (2013) 'Distribution of mesozooplankton biomass in the global ocean', *Earth System Science Data*, 5(1), pp. 45–55. doi: 10.5194/essd-5-45-2013.

Morice, A. *et al.* (2023) 'Craniofacial growth and function in achondroplasia: a multimodal 3D study on 15 patients', *Orphanet Journal of Rare Diseases*. BioMed Central Ltd, 18(1), pp. 1–13. doi: 10.1186/s13023-023-02664-y.

Morin, P. A. *et al.* (2021) 'Reference genome and demographic history of the most endangered marine mammal, the vaquita', *Molecular Ecology Resources*. John Wiley & Sons, Ltd, 21(4), pp. 1008–1020. doi: 10.1111/1755-0998.13284.

Mork, L. and Crump, G. (2015) 'Zebrafish Craniofacial Development. A Window into Early Patterning', in *Current Topics in Developmental Biology*. NIH Public Access, pp. 235–269. doi: 10.1016/bs.ctdb.2015.07.001.

Murray, A. *et al.* (2020) 'Protecting the million-dollar mantas; creating an evidence-based code of conduct for manta ray tourism interactions', *Journal of Ecotourism*. Routledge, 19(2), pp. 132–147. doi: 10.1080/14724049.2019.1659802.

Nakatani, Y. *et al.* (2021) 'Reconstruction of proto-vertebrate, proto-cyclostome and proto-gnathostome genomes provides new insights into early vertebrate evolution', *Nature Communications*, 12(1), p. 4489. doi: 10.1038/s41467-021-24573-z.

Nishimura, O. *et al.* (2022) 'Squalomix: shark and ray genome analysis consortium and its data sharing platform', *F1000Research*. F1000 Research Limited, 11, p. 1077. doi: 10.12688/f1000research.123591.1.

Notabartolo di Sciara, G. (1987) 'A revisionary study of the genus Mobula Rafinesque, 1810 (Chondrichthyes: Mobulidae) with the description of a new species', *Zoological Journal of the Linnean Society*, 91(1), pp. 1–91. doi: 10.1111/j.1096-3642.1987.tb01723.x.

Notabartolo di Sciara, G. (1988) 'Natural history of the rays of the genus Mobula in the Gulf of California', in.

Notarbartolo di Sciara, G. *et al.* (2020) 'Taxonomic status, biological notes, and conservation of the longhorned pygmy devil ray Mobula eregoodoo (Cantor, 1849)', *Aquatic Conservation: Marine and Freshwater Ecosystems*. John Wiley & Sons, Ltd, 30(1), pp. 104–122. doi: 10.1002/aqc.3230.

Notarbartolo Di Sciara, G., Stevens, G. and Fernando, D. (2020) 'The giant devil ray Mobula mobular (Bonnaterre, 1788) is not giant, but it is the only spinetail devil ray', *Marine Biodiversity Records*. BioMed Central Ltd., 13(1), pp. 1–5. doi: 10.1186/s41200-020-00187-0.

O'Bryhim, J. R., Parsons, E. C. M. and Lance, S. L. (2017) 'Forensic species identification of elasmobranch products sold in Costa Rican markets', *Fisheries Research*. Elsevier, 186, pp. 144–150. doi: 10.1016/j.fishres.2016.08.020.

O'Malley, M. p. *et al.* (2017) 'Characterization of the trade in manta and devil ray gill plates in China and South-east Asia through trader surveys', *Aquatic Conservation: Marine and Freshwater Ecosystems*. John Wiley & Sons, Ltd, 27(2), pp. 394–413. doi: 10.1002/aqc.2670.

O'Malley, M. P., Lee-Brooks, K. and Medd, H. B. (2013) 'The Global Economic Impact of Manta Ray Watching Tourism', *PLoS ONE*. Public Library of Science, 8(5), p. e65051. doi: 10.1371/journal.pone.0065051.

Oleksiak, M. F. and Rajora Editors, O. P. (no date) 'Population Genomics: Marine Organisms'. Available at: http://www.springer.com/series/13836 (Accessed: 22 August 2024).

Onimaru, K. (2020) 'The evolutionary origin of developmental enhancers in vertebrates: Insights from non-model species', *Development, Growth & Differentiation*. John Wiley & Sons, Ltd, 62(5), pp. 326–333. doi: 10.1111/DGD.12662.

Ozaki, R. *et al.* (2024) 'Evidence of environmental niche separation between threatened mobulid rays in Aotearoa New Zealand: Insights from species distribution modelling', *Journal of Biogeography*. John Wiley & Sons, Ltd, 00, pp. 1–19. doi: 10.1111/JBI.14976.

Pardo, S. A. *et al.* (2016) 'Growth, productivity, and relative extinction risk of a data-sparse devil ray', *Scientific Reports*. Nature Publishing Group, 6(1), pp. 1–10. doi: 10.1038/srep33745.

Pasquier, J. *et al.* (2016) 'Gene evolution and gene expression after whole genome duplication in fish: The PhyloFish database', *BMC Genomics*. BioMed Central Ltd., 17(1), pp. 1–10. doi: 10.1186/s12864-016-2709-z.

Pearce, J. *et al.* (2021) 'State of Shark and Ray Genomics in an Era of Extinction', *Frontiers in Marine Science*. Frontiers Media S.A., p. 744986. doi: 10.3389/fmars.2021.744986.

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

Pennell, M. W. *et al.* (2015) 'Y Fuse? Sex Chromosome Fusions in Fishes and Reptiles', *PLoS Genetics*. Public Library of Science, 11(5), p. e1005237. doi: 10.1371/journal.pgen.1005237.

Ponomarenko, J. V. *et al.* (2002) 'rSNP_Guide: An integrated database-tools system for studying SNPs and site-directed mutations in transcription factor binding sites', *Human Mutation*. John Wiley & Sons, Ltd, 20(4), pp. 239–248. doi: 10.1002/humu.10116.

Poortvliet, M. *et al.* (2015) 'A dated molecular phylogeny of manta and devil rays (Mobulidae) based on mitogenome and nuclear sequences', *Molecular Phylogenetics and Evolution*. Academic Press, 83, pp. 72–85. doi: 10.1016/j.ympev.2014.10.012.

Rajderkar, S. S. *et al.* (2024) 'Dynamic enhancer landscapes in human craniofacial development', *Nature Communications*. Nature Publishing Group, 15(1), pp. 1–18. doi: 10.1038/s41467-024-46396-4.

Rambahiniarison, J. M. *et al.* (2018) 'Life history, growth, and reproductive biology of four mobulid species in the bohol sea, Philippines', *Frontiers in Marine Science*. Frontiers Media S.A., 5(AUG), p. 373967. doi: 10.3389/fmars.2018.00269.

Ravi, V. *et al.* (2009) 'Elephant shark (Callorhinchus milii) provides insights into the evolution of Hox gene clusters in gnathostomes', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 106(38), pp. 16327–16332. doi: 10.1073/pnas.0907914106.

Read, T. D. *et al.* (2017) 'Draft sequencing and assembly of the genome of the world's largest fish, the whale shark: Rhincodon typus Smith 1828', *BMC Genomics*. BioMed Central Ltd., 18(1), pp. 1–10. doi: 10.1186/s12864-017-3926-9.

Rebeiz, M. and Tsiantis, M. (2017) 'Enhancer evolution and the origins of morphological novelty', *Current Opinion in Genetics and Development*. Elsevier Current Trends, pp. 115–123. doi: 10.1016/j.gde.2017.04.006.

Redmond, A. K., Macqueen, D. J. and Dooley, H. (2018) 'Phylotranscriptomics suggests the jawed vertebrate ancestor could generate diverse helper and regulatory T cell subsets', *BMC Evolutionary Biology*. BioMed Central, 18(1), pp. 1–19. doi: 10.1186/s12862-018-1290-2.

Rhie, A. *et al.* (2021) 'Towards complete and error-free genome assemblies of all vertebrate species', *Nature*. Nature Publishing Group, 592(7856), pp. 737–746. doi: 10.1038/s41586-021-03451-0.

Rhoads, A. and Au, K. F. (2015) 'PacBio Sequencing and Its Applications', *Genomics, Proteomics and Bioinformatics*. Oxford Academic, pp. 278–289. doi: 10.1016/j.gpb.2015.08.002.

Richardson, A. J. (2008) 'In hot water: zooplankton and climate change', *ICES Journal of Marine Science*. Oxford University Press, 65(3), pp. 279–295. doi: 10.1093/icesjms/fsn028.

Richmond, S. *et al.* (2018) 'Facial genetics: A brief overview', *Frontiers in Genetics*. Frontiers Media SA. doi: 10.3389/fgene.2018.00462.

Ryczek, N., Łyś, A. and Makałowska, I. (2023) 'The Functional Meaning of 5'UTR in Protein-Coding Genes', *International Journal of Molecular Sciences*. Multidisciplinary Digital Publishing Institute, p. 2976. doi:

10.3390/ijms24032976.

Sacerdot, C. *et al.* (2018) 'Chromosome evolution at the origin of the ancestral vertebrate genome', *Genome Biology*. BioMed Central Ltd., 19(1), pp. 1–15. doi: 10.1186/s13059-018-1559-1.

Sambrook, J. G., Figueroa, F. and Beck, S. (2005) 'A genome-wide survey of Major Histocompatibility Complex (MHC) genes and their paralogues in zebrafish', *BMC Genomics*. BioMed Central Ltd., 6(1), pp. 1–10. doi: 10.1186/1471-2164-6-152.

Segelbacher, G. *et al.* (2022) 'New developments in the field of genomic technologies and their relevance to conservation management', *Conservation Genetics*. Springer, pp. 217–242. doi: 10.1007/s10592-021-01415-5.

Serra-Pereira, B. *et al.* (2010) 'Morphometric ratios of six commercially landed species of skate from the Portuguese continental shelf, and their utility for identification', *ICES Journal of Marine Science*. Oxford Academic, 67(8), pp. 1596–1603. doi: 10.1093/icesjms/fsq056.

Song, N. *et al.* (2023) 'Genomic Characteristics of Okamejei kenojei and the Implications to Its Evolutionary Biology Study', *Marine Biotechnology*. Springer, 25(5), pp. 815–823. doi: 10.1007/s10126-023-10242-3.

Sotero-Caio, C. G. *et al.* (2017) 'Evolution and diversity of transposable elements in vertebrate genomes', *Genome Biology and Evolution*. Oxford Academic, 9(1), pp. 161–177. doi: 10.1093/gbe/evw264.

Southward, A. J. *et al.* (2004) 'Long-term oceanographic and ecological research in the western English Channel', *Advances in Marine Biology*. doi: 10.1016/S0065-2881(04)47001-1.

Stevens, G. *et al.* (2019) *GUIDE TO THE MANTA & DEVIL RAYS OF THE WORLD*, *Guide to the Manta and Devil Rays of the World*. Princeton University Press.

Stevens, G. M. W., Hawkins, J. P. and Roberts, C. M. (2018) 'Courtship and mating behaviour of manta rays Mobula alfredi and M. birostris in the Maldives', *Journal of Fish Biology*. John Wiley & Sons, Ltd, 93(2), pp. 344–359. doi: 10.1111/jfb.13768.

Stewart, J. D. *et al.* (2017) 'Trophic overlap in mobulid rays: Insights from stable isotope analysis', *Marine Ecology Progress Series*. Inter-Research, 580, pp. 131–151. doi: 10.3354/meps12304.

Stewart, J. D. *et al.* (2018) 'Research priorities to support effective Manta and Devil Ray conservation', *Frontiers in Marine Science*. Frontiers Media S.A., 5(SEP), p. 314. doi: 10.3389/fmars.2018.00314.

Supple, M. A. and Shapiro, B. (2018) 'Conservation of biodiversity in the genomics era', *Genome Biology*. BioMed Central Ltd., 19(1), pp. 1–12. doi: 10.1186/s13059-018-1520-3.

Swenson, J. D. *et al.* (2018) 'How the devil ray got its horns: The evolution and development of cephalic lobes in myliobatid stingrays (Batoidea: Myliobatidae)', *Frontiers in Ecology and Evolution*. Frontiers Media S.A., 6(NOV), p. 423904. doi: 10.3389/fevo.2018.00181.

Tan, M. *et al.* (2021) 'The whale shark genome reveals patterns of vertebrate gene family evolution', *eLife*. eLife Sciences Publications Ltd, 10. doi: 10.7554/eLife.65394.

Valsecchi, E. *et al.* (2005) 'Rapid Miocene-Pliocene dispersal and evolution of Mediterranean rajid fauna as inferred by mitochondrial gene variation', *Journal of Evolutionary Biology*. John Wiley & Sons, Ltd, 18(2), pp. 436–446. doi: 10.1111/j.1420-9101.2004.00829.x.

Venables, S. *et al.* (2016) 'Manta ray tourism management, precautionary strategies for a growing industry: A case study from the Ningaloo Marine Park, Western Australia', *Pacific Conservation Biology*. CSIRO PUBLISHING, pp. 295–300. doi: 10.1071/PC16003.

Venkatesh, B. *et al.* (2014) 'Elephant shark genome provides unique insights into gnathostome evolution', *Nature*. Nature Publishing Group, 505(7482), pp. 174–179. doi: 10.1038/nature12826.

Villalobos-Segura, E. and Underwood, C. J. (2020) 'Radiation and Divergence Times of Batoidea', *Journal of Vertebrate Paleontology*. Taylor & Francis, 40(3), p. 3. doi: 10.1080/02724634.2020.1777147.

Waller, M. J. *et al.* (2024) 'The vulnerability of sharks, skates, and rays to ocean deoxygenation: Physiological

mechanisms, behavioral responses, and ecological impacts', *Journal of Fish Biology*. John Wiley & Sons, Ltd. doi: 10.1111/jfb.15830.

Walsh, C. A. J. *et al.* (2022) 'Genomic insights into the historical and contemporary demographics of the grey reef shark', *Heredity*. Nature Publishing Group, 128(4), pp. 225–235. doi: 10.1038/s41437-022-00514-4.

Wang, J., Santiago, E. and Caballero, A. (2016) 'Prediction and estimation of effective population size', *Heredity*. Nature Publishing Group, pp. 193–206. doi: 10.1038/hdy.2016.43.

Wang, Xiangnan *et al.* (2020) 'Association of a new 99-bp indel of the CEL gene promoter region with phenotypic traits in chickens', *Scientific Reports 2020 10:1*. Nature Publishing Group, 10(1), pp. 1–12. doi: 10.1038/s41598-020-60168-2.

Wang, Xin *et al.* (2016) 'Bioremediation of marine oil pollution by Brevundimonas diminuta: effect of salinity and nutrients', *Desalination and Water Treatment*. Taylor and Francis Inc., 57(42), pp. 19768–19775. doi: 10.1080/19443994.2015.1106984.

Wang, Yunhao *et al.* (2021) 'Nanopore sequencing technology, bioinformatics and applications', *Nature Biotechnology*. Nature Publishing Group, pp. 1348–1365. doi: 10.1038/s41587-021-01108-x.

Waples, R. S. *et al.* (2022) 'Implications of Large-Effect Loci for Conservation: A Review and Case Study with Pacific Salmon', *Journal of Heredity*. Oxford Academic, pp. 121–144. doi: 10.1093/jhered/esab069.

Weber, J. A. *et al.* (2020) 'The whale shark genome reveals how genomic and physiological properties scale with body size', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 117(34), pp. 20662–20671. doi: 10.1073/pnas.1922576117.

Weigmann, S. (2016) 'Annotated checklist of the living sharks, batoids and chimaeras (Chondrichthyes) of the world, with a focus on biogeographical diversity', *Journal of Fish Biology*. John Wiley & Sons, Ltd, 88(3), pp. 837–1037. doi: 10.1111/jfb.12874.

Weisman, C. M., Murray, A. W. and Eddy, S. R. (2022) 'Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes', *Current Biology*. Cell Press, 32(12), pp. 2632-2639.e2. doi: 10.1016/j.cub.2022.04.085.

Wellenreuther, M. *et al.* (2019) 'Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification', *Molecular Ecology*. John Wiley & Sons, Ltd, 28(6), pp. 1203–1209. doi: 10.1111/mec.15066.

Wheeler, C. R. *et al.* (2020) 'Anthropogenic stressors influence reproduction and development in elasmobranch fishes', *Reviews in Fish Biology and Fisheries*. Springer, pp. 373–386. doi: 10.1007/s11160-020-09604-0.

White, W. T. *et al.* (2018) 'Phylogeny of the manta and devilrays (Chondrichthyes: Mobulidae), with an updated taxonomic arrangement for the family', *Zoological Journal of the Linnean Society*. Oxford Academic, 182(1), pp. 50–75. doi: 10.1093/zoolinnean/zlx018.

Wong, P. B. Y. *et al.* (2012) 'Tissue sampling methods and standards for vertebrate genomics', *GigaScience*. Oxford University Press, p. 8. doi: 10.1186/2047-217X-1-8.

Wu, C. S. *et al.* (2022) 'Chromosome-level genome assembly of grass carp (Ctenopharyngodon idella) provides insights into its genome evolution', *BMC Genomics*. BioMed Central Ltd, 23(1), pp. 1–14. doi: 10.1186/s12864-022-08503-x.

Wu, J. *et al.* (2024) 'Comparative genomics illuminates karyotype and sex chromosome evolution of sharks.', *Cell genomics*. Elsevier, 0(0), p. 100607. doi: 10.1016/j.xgen.2024.100607.

Wu, T. *et al.* (2021) 'clusterProfiler 4.0: A universal enrichment tool for interpreting omics data', *Innovation*. Cell Press, 2(3), p. 100141. doi: 10.1016/j.xinn.2021.100141.

Wyffels, J. *et al.* (2014) 'SkateBase, an elasmobranch genome project and collection of molecular resources for chondrichthyan fishes', *F1000Research*. F1000 Research Limited, 3, p. 191. doi: 10.12688/f1000research.4996.1.

Yamaguchi, K. *et al.* (2023) 'Elasmobranch genome sequencing reveals evolutionary trends of vertebrate

karyotype organization', *Genome Research*. Cold Spring Harbor Laboratory Press, 33(9), pp. 1527–1540. doi: 10.1101/gr.276840.122.

Zeng, Y. *et al.* (2016) 'DNA barcoding of Mobulid Ray Gill Rakers for Implementing CITES on Elasmobranch in China', *Scientific Reports*. Nature Publishing Group, 6(1), pp. 1–9. doi: 10.1038/srep37567.

Zhang, Y. *et al.* (2020) 'The White-Spotted Bamboo Shark Genome Reveals Chromosome Rearrangements and Fast-Evolving Immune Genes of Cartilaginous Fish', *iScience*. Elsevier, 23(11), p. 101754. doi: 10.1016/j.isci.2020.101754.

Zhao, J. *et al.* (2021) 'Tools for analysis and conditional deletion of subsets of sensory neurons', *Wellcome Open Research 2021 6:250*. F1000 Research Limited, 6, p. 250. doi: 10.12688/wellcomeopenres.17090.1.

Zhou, C., McCarthy, S. A. and Durbin, R. (2023) 'YaHS: yet another Hi-C scaffolding tool', *Bioinformatics*. Oxford Academic, 39(1). doi: 10.1093/bioinformatics/btac808.

Appendix

Cactus whole-genome alignment

```
#!/bin/bash
#
#
#################################################################################
#
# Setup Instructions
#
#################################################################################
#
# Grid Engine options (lines prefixed with #$)
#  job name: -N
```

```
#  use the current working directory: -cwd
#  number of cores -pe sharedmem
#  runtime limit: -l h_rt
#  memory limit: -l h_vmem
#$ -cwd
#$ -pe sharedmem 16
#$ -l h_rt=504:00:00
#$ -l h_vmem=60G
#$ -P roslin_macqueen_lab
#$ -e error.txt
#$ -o output.txt
#$ -M s2457124@ed.ac.uk
#$ -m beas

# Initialise the environment modules
. /etc/profile.d/modules.sh
# load any modules
module load roslin/singularity/3.5.3

# Make workdir

#query
# Run cactus
# output Hal is created at this step, we're naming it
singularity exec cactus.sif cactus ./rays/js ./seqFile_updt.txt
./fivespp_output.hal --workDir ./workdir_cactus --maxCores 16 --maxMemory 768G
--consCores 2 --consMemory 100Gi
```

halSynteny

```
#!/bin/bash
#
#
================================================================================
==
# Setup Instructions
#
================================================================================
==
#
# Grid Engine options (lines prefixed with #$)
#  job name: -N
#  use the current working directory: -cwd
#  number of cores -pe sharedmem
#  runtime limit: -l h_rt
#  memory limit: -l h_vmem
#$ -cwd
```

```
#$ -pe sharedmem 16
#$ -l h_rt=12:00:00
#$ -l h_vmem=30G
#$ -P roslin_macqueen_lab
#$ -e error.txt
#$ -o output.txt
#$ -M s2457124@ed.ac.uk
#$ -m beas

# Initialise the environment modules
. /etc/profile.d/modules.sh

# load any modules
module load roslin/singularity/3.5.3

# Define query genome and target genome variables (these should be entered as genome
# names found in the SeqFile
QUERY=$1
TARGET=$2
OUTPUT=$3

# halSynteny
singularity exec ./../cactus.sif halSynteny --queryGenome ${QUERY} --targetGenome ${TARGET} --maxAnchorDistance 1000000 --minBlockSize 200000 fivespp_output.hal ${OUTPUT}
```

Circlize plotting

```
# Packages needed
install.packages("circlize")
install.packages("lifecycle")
install.packages("tidyverse")
install.packages("openxlsx")
install.packages("readxl")
install.packages("scales")


library("circlize")
library("lifecycle")
library("tidyverse")
library("openxlsx")
library("readxl")
library("scales")
```

```
## set working directory
getwd()
setwd("Z:/LSC/cactus")

# Define the cytoband data. Track is the sheet with gpos50 data
cytoband.df = read.xlsx("hypostoma_manta_newsynteny.psl.xlsx", sheet = "track", colNames
= F) # replace spreadsheet name with whichever pair you're working with
cytoband.df$X1 <- as.character(cytoband.df$X1)
cytoband.df$X1 <- factor(cytoband.df$X1, levels = c(cytoband.df$X1))

cytoband.df$X1 <- factor(cytoband.df$X1)

beda <- read.xlsx("hypostoma_manta_newsynteny.psl.xlsx", sheet = "beda",colNames = F) #
make sure bed a and bed b have the same nb of rows; erase unplaced scaffolds from the
matches sheet for that

bedb <- read.xlsx("hypostoma_manta_newsynteny.psl.xlsx", sheet = "bedb",colNames = F)

## Add colours for each chromosome
chromosomes <- c( "CM057523.1", "CM057524.1", "CM057525.1", "CM057526.1",
"CM057527.1",
            "CM057528.1", "CM057529.1", "CM057530.1", "CM057531.1", "CM057532.1",
            "CM057533.1", "CM057534.1", "CM057535.1", "CM057536.1", "CM057537.1",
            "CM057538.1", "CM057539.1", "CM057540.1", "CM057541.1", "CM057542.1",
            "CM057543.1", "CM057544.1", "CM057545.1", "CM057546.1", "CM057547.1",
            "CM057548.1", "CM057549.1", "CM057550.1", "CM057551.1", "CM057552.1",
            "CM057553.1", "CM057554.1", "CM057555.1")



colours <- scales::alpha(c("#FF5733", "#FFC300", "#FF6347", "#FF1493", "#FF00FF",
"#FF00CC", "#800080", "#8A2BE2", "#4B0082", "#0000FF", "#00BFFF",
                  "#00FFFF", "#00FF00", "#7CFC00",
                  "#32CD32", "#ADFF2F", "#FFFF00", "#FFD700", "#FFA500", "#FF4500",
"#FF6347",
                  "#DC143C", "#8B0000", "#FF69B4", "#FFB6C1", "#00FA9A", "#20B2AA",
"#00CED1",
                  "#00BFFF", "#1E90FF", "#4682B4", "#87CEEB", "#40E0D0"
), alpha = 0.7)


## Define a data frame with chromosomes and colours
df <- data.frame(chromosome = chromosomes, color = colours)

# view resulting data frame
```

```
beda <- beda %>% left_join(df, by = c(X1="chromosome")) # whyyy does it need this beda pipe otherwise it gives an error? make it make sense
col.colours <- as.character(beda$color)

##plot
circos.clear()
circos.par(start.degree = 90,  gap.degree = 2, cell.padding = c(0.001, 0, 0.001, 0), canvas.xlim = c (-1.2,1.2), canvas.ylim = c(-1.2,1.2))
circos.initializeWithIdeogram(cytoband.df, plotType = NULL, chromosome.index = cytoband.df$X1) #initialises without adding anything, for full customisation

######
circos.genomicTrack(cytoband.df, ylim = c(0, 2), panel.fun = function(region, value, ...){circos.genomicText(region, value, y = 2, labels.column = 1, cex = 0.15, niceFacing = T, ...) }, track.height = 0.09, track.margin = c(0,0), bg.border = NA)

#####
circos.track(ylim=c(0,1),panel.fun = function(x, y) {
  circos.axis(h = "top" ,major.at = seq(0, round(CELL_META$xlim[2], digits = 2), by = 20000000), labels.cex = 0.20,
          labels = seq(0,280, by=20), minor.ticks = 0.5,labels.facing = "clockwise",lwd = 0.45,major.tick.length = 1.5, labels.pos.adjust = F)
  #circos.text(CELL_META$xcenter,CELL_META$cell.ylim[1] + uy(3.5, "mm"),CELL_META$sector.index, cex = 0.5, col = "black",
  #facing = "outside", niceFacing = TRUE, labels = CELL_META$sector.numeric.index )
},track.height = 0.01, track.margin = c(0,0), bg.border = NA)

#####
circos.genomicIdeogram(cytoband = cytoband.df, track.height = 0.06)

circos.genomicLink(beda, bedb, col=col.colours)

####

circos.info()
```

halBranchMutations

```
#!/bin/bash
#
#
================================================================================
==
# Setup Instructions
#
================================================================================
==
#
# Grid Engine options (lines prefixed with #$)
#  job name: -N
#  use the current working directory: -cwd
#  number of cores -pe sharedmem
#  runtime limit: -l h_rt
#  memory limit: -l h_vmem
#$ -cwd
#$ -pe sharedmem 16
#$ -l h_rt=168:00:00
#$ -l h_vmem=30G
#$ -P roslin_macqueen_lab
#$ -e error.txt
#$ -M s2457124@ed.ac.uk
#$ -m beas

# Initialise the environment modules
. /etc/profile.d/modules.sh

# load any modules
module load roslin/singularity/3.5.3

#Define the branch we're extracting mutations for as argument 1
# This should be the species name as in the SeqFile

BRANCH=$1

# Define REF_FILE (the insertion bed file in the branch species' coordinates) and PARENT_FILE
(the duplications and
# deletions file in the ancestor's coordinates) as arguments 2 and 3 when submitting the job
# These should be in a form like refspecies_ins.bed parent_dup.bed

REF_FILE=$2
PARENT_FILE=$3

singularity exec ./../cactus.sif halBranchMutations fivespp_output.hal ${BRANCH} --refFile
${REF_FILE} --parentFile ${PARENT_FILE} --maxNFraction 0
```

RepeatModeler

```
#!/bin/bash
#
#
================================================================================
==
# Setup Instructions
#
================================================================================
==
#
# Grid Engine options (lines prefixed with #$)
#  job name: -N
#  use the current working directory: -cwd
#  number of cores -pe sharedmem
#  runtime limit: -l h_rt
#  memory limit: -l h_vmem
#$ -cwd
#$ -pe sharedmem 16
#$ -l h_rt=336:00:00
#$ -l h_vmem=50G
#$ -P roslin_macqueen_lab
#$ -e error.txt
#$ -o hypostomamodeler.txt
#$ -M s2457124@ed.ac.uk
#$ -m beas

# Initialise the environment modules
. /etc/profile.d/modules.sh

# load any modules
module load roslin/singularity/3.5.3

if [ ! -e ${PWD}/.singularity ]; then mkdir ${PWD}/.singularity; fi
export SINGULARITY_TMPDIR=$PWD/.singularity
export SINGULARITY_CACHEDIR=$PWD/.singularity

singularity build dfam-tetools-latest.sif docker://dfam/tetools:latest

# Make database
singularity exec dfam-tetools-latest.sif BuildDatabase -name hypostoma
sMobHyp1.curated_primary.mt.scrubbed.fa

# Run RepeatModeler
singularity exec dfam-tetools-latest.sif nohup RepeatModeler -database hypostoma
```

RepeatMasker

```
#!/bin/bash
#
#
================================================================================
==
# Setup Instructions
#
================================================================================
==
#
# Grid Engine options (lines prefixed with #$)
#  job name: -N
#  use the current working directory: -cwd
#  number of cores -pe sharedmem
#  runtime limit: -l h_rt
#  memory limit: -l h_vmem
#$ -cwd
#$ -pe sharedmem 16
#$ -l h_rt=336:00:00
#$ -l h_vmem=50G
#$ -P roslin_macqueen_lab
#$ -e error.txt
#$ -o output.txt
#$ -M s2457124@ed.ac.uk
#$ -m beas

# Initialise the environment modules
. /etc/profile.d/modules.sh

# load any modules
module load roslin/singularity/3.5.3

# run RepeatMasker
singularity exec dfam-tetools-latest.sif RepeatMasker -lib hypostoma-families.fa
sMobHyp1.curated_primary.mt.scrubbed.fa
```

Splitting files for mafExtractor processing

```
#!/bin/bash
#
#
# ============================================================================
==
# Setup Instructions
#
# ============================================================================
==
#
# Grid Engine options (lines prefixed with #$)
#  job name: -N
#  use the current working directory: -cwd
#  number of cores -pe sharedmem
#  runtime limit: -l h_rt
#  memory limit: -l h_vmem
#$ -cwd
#$ -pe sharedmem 8
#$ -l h_rt=12:00:00
#$ -l h_vmem=30G
#$ -P roslin_macqueen_lab
#$ -e error2.txt
#$ -o output.txt
#$ -M s2457124@ed.ac.uk
#$ -m beas

# Initialise the environment modules
. /etc/profile.d/modules.sh

split -n l/600 -a 3  anc3_dels_plusone.bed split_anc3_dels/anc3_subset_dels_plusone
```

mafExtractor

```
#!/bin/bash
#
#
# ============================================================================
==
# Setup Instructions
#
# ============================================================================
==
#
# Grid Engine options (lines prefixed with #$)
#  job name: -N
#  use the current working directory: -cwd
#  number of cores -pe sharedmem
#  runtime limit: -l h_rt
#  memory limit: -l h_vmem
#$ -cwd
#$ -pe sharedmem 1
#$ -l rl9
#$ -l h_rt=336:00:00
#$ -l h_vmem=30G
#$ -P roslin_macqueen_lab
#$ -M s2457124@ed.ac.uk
#$ -m beas
#$ -t 1-600
#$ -tc 16

# Initialise the environment modules
. /etc/profile.d/modules.sh

# cd to split Anc3 dels directory
cd split_anc3_dels/

#Define variables
INPUT=$(ls anc3_subset_dels_plusonea* | awk "NR == $SGE_TASK_ID")
PREFI="$(basename $INPUT)"

echo "$INPUT"
echo "$PREFI"

# extract the deletions from each maf file into respective directories
while read -r a b c;
do
/exports/cmvm/eddie/eb/groups/macqueen_lab/manu/mkg/softwares/mafTools/bin/mafEx
tractor \
-m ./../fivespp_output_anc2ref.maf -s Anc2."$a" --start $b \
```

```
--stop $c > ./../anc3_dels_plusone_extractions/anc3_del_"$a"_"$b"_"$c".maf; done <${PREFI}
```

maf2table

```
###load libraries
library(tidyverse)
library(fuzzyjoin)
library(openxlsx)
library(tidygenomics)
message("loaded libraries")

#####set working directory
setwd(paste("/exports/cmvm/eddie/eb/groups/macqueen_lab/Lea/cactus/rays_plus_skates
/maf_anc2AsRef/anc3_dels_plusone_extractions")) ## set this according to your eddie folder
containing maf files Lea
getwd()

##### function to read maf as table

MAF2tbl <- function(mafFile){
  x <- readLines(mafFile)

  aLines <- grepl("^a",x)
  sLines <- grepl("^s",x)

  tibble(idx=cumsum(aLines),txt=x) %>%
    filter(sLines) %>%
    separate(txt, into=c("s","src","start","size","strand","srcSize","text"),sep = "[ \t]+") %>%
    select(-s) %>%
    mutate_at(vars(start,size,srcSize),as.integer)
}

###### read maf files using the function above
Files <- list.files(pattern = "*.maf",full.names = TRUE)
maflist <- lapply(Files,MAF2tbl) ### convert all maf files to tables
names(maflist) <- Files
message("maffiles read into table")

###### function to overlay atac peaks on to your maf table data
curate_maf<- function(maftable){
  maftable %>%
    select(c("src","start","size","strand","srcSize"))  %>%
    separate(src, into=c("species", "chromosome"), sep="\\.") %>%
    mutate(end= ifelse(strand=="+", start+size,
                       ifelse(strand=="-",start+size,0))) %>%
    mutate(pos_start=ifelse(strand=="+",start,
                            ifelse(strand=="-",srcSize - end,0))) %>%
    mutate (pos_end=ifelse(strand=="+",end,
                           ifelse(strand=="-",srcSize - start,0))) %>%
    select(c("species","chromosome","strand","pos_start","pos_end","size")) %>%
```

```
   filter(size>3) %>%
   genome_cluster(.,by  =  c("chromosome",  "pos_start",  "pos_end"),max_distance=1000)
%>%
   distinct(pos_start,pos_end, .keep_all = TRUE) %>%
   group_by(species,chromosome,cluster_id) %>%
   summarise(start = min(pos_start, na.rm=TRUE),end = max(pos_end, na.rm=TRUE), size =
sum(size)) %>%
   ungroup %>%
   mutate(range = end-start) %>%
   group_by(species,chromosome) %>% top_n(1, size)%>% ungroup %>%
   group_by(species) %>% top_n(1, size) %>% ungroup %>%
   mutate(size_range_ratio=size/range)
}


###run the curate maf function
cur_maf<-lapply(maflist,curate_maf)
message("maffiles curated")

AllDat_maf <- bind_rows(cur_maf, .id = "Origin") %>%
  mutate(Origin = basename(Origin))
message("all maf files curated and combined")

## write output into an excel sheet
#write.xlsx(AllDat_maf,"maf_files_summary_tabular.xlsx")
#message("maffiles overlayed with atac and written to excel file")

# Write the table to a tab-separated text file
write.table(AllDat_maf, "maf_files_anc3dels_summary_tabular.txt", sep = "\t", row.names =
FALSE, quote = FALSE)
message("maffiles written to tab-separated text file")
```

SnpEff

```
#!/bin/bash
#
#
================================================================================
==
# Setup Instructions
#
================================================================================
==
#
# Grid Engine options (lines prefixed with #$)
#  job name: -N
#  use the current working directory: -cwd
#  number of cores -pe sharedmem
#  runtime limit: -l h_rt
#  memory limit: -l h_vmem
#$ -cwd
#$ -pe sharedmem 8
#$ -l h_rt=12:00:00
#$ -l h_vmem=40G
#$ -P roslin_macqueen_lab
#$ -e error.txt
#$ -o output.txt
#$ -M s2457124@ed.ac.uk
#$ -m beas

# Initialise the environment modules
. /etc/profile.d/modules.sh

# load any modules
module load roslin/openjdk/11.0.2

# Define REF (reference genome) and BED (the BED file to annotate) variables
REF=sHypSab1.hap1
BED="mh_dels_conserved_stingray_manta_mapped.bed"
OUTPUT="mh_dels_conserved_manta_stingray"

# Annotate chromosome rearrangements/recombination BED files using SnpEff

java -Xmx8g -jar snpEff.jar -c snpEff.config -v -nodownload -i bed sHypSab1.hap1 ${BED} >
"${OUTPUT}_annotated.bed"
```

PSMC

###Trim reads

```bash
#!/bin/bash
#
#
# ============================================================================
==
# Setup Instructions
#
# ============================================================================
==
#
# Grid Engine options (lines prefixed with #$)
#  job name: -N
#  use the current working directory: -cwd
#  number of cores -pe sharedmem
#  runtime limit: -l h_rt
#  memory limit: -l h_vmem
#$ -cwd
#$ -pe sharedmem 8
#$ -l h_rt=10:00:00
#$ -l h_vmem=30G
#$ -P roslin_macqueen_lab
#$ -M s2457124@ed.ac.uk
#$ -m beas
#$ -t 1-12 #change this value to however many fastqs you have

#initialise modules and conda
. /etc/profile.d/modules.sh

#load modules and conda env
module load roslin/conda/23.7.2
source /exports/applications/apps/community/roslin/conda/4.9.1/etc/profile.d/conda.sh
source activate trim_galore

#define dirs
target_dir=/exports/cmvm/eddie/eb/groups/macqueen_lab/Lea/psmc/bwa/reads
sample_list=/exports/cmvm/eddie/eb/groups/macqueen_lab/Lea/psmc/bwa/file_names.txt
#text file with sample file names per row

#get filelist
base=`sed -n "$SGE_TASK_ID"p $sample_list | awk '{print $1}'`
R1=`sed -n "$SGE_TASK_ID"p $sample_list | awk '{print $2}'`
R2=`sed -n "$SGE_TASK_ID"p $sample_list | awk '{print $3}'`
```

```
trim_galore --paired --fastqc -q 30 -j 4 $R1 $R2

#remove original files
#rm $R1 $R2 #careful with this :)

/exports/cmvm/eddie/eb/groupas/macqueen_lab/Lea/psmc/bwa/reads/H3G5WDSX7_3_378UDI-idt-UMI_1.fastq
```

BWA kit mem alignment

```bash
#!/bin/sh
#$ -N runBWA
#$ -cwd
#$ -l h_rt=12:00:00
#$ -l h_vmem=8G
#$ -P  roslin_macqueen_lab #add in priority if you have
#$ -t 1:6 # amend to how many samples you have
#$ -pe sharedmem 8
#$ -M s2457124@ed.ac.uk
#$ -m beas

# Initialise the environment modules
. /etc/profile.d/modules.sh

# load modules and environments
module load roslin/samtools/1.9 # originally was 1.16 but that version isn't on Eddie
module load roslin/conda/23.7.2
source activate bwakit_env

#define vars
#bwapath=/exports/cmvm/eddie/eb/groups/macqueen_lab/Lea/anaconda/envs/bwakit_env #or add to PATH
OUT_DIR=/exports/cmvm/eddie/eb/groups/macqueen_lab/Lea/psmc/bwa
SAMPLE_LIST=/exports/cmvm/eddie/eb/groups/macqueen_lab/Lea/psmc/bwa/trimmed_reads_list.txt #list of file names, with sample ID in first column e.g sampleID,sampleID.trimmed.R1.fastq.gz,sampleID.trimmed.R1.fastq.gz (for example)
REFGEN=/exports/cmvm/eddie/eb/groups/macqueen_lab/Lea/psmc/bwa/sMobHyp1.curated_primary.mt.scrubbed.fa # Mobula hypostoma genome

#get sample lists
base=$(sed -n "${SGE_TASK_ID}p" $SAMPLE_LIST | awk '{print $1}')

#process
echo Processing sample: ${base} on $HOSTNAME
```

```
#align
run-bwamem -d -t 21 -o ${base}_$SGE_TASK_ID.aln.bam -HR "@RG\tID:${base}\tSM:${base}"
$REFGEN $OUT_DIR/trimmed_reads/${base}_378UDI-idt-UMI_1_val_1.fq
$OUT_DIR/trimmed_reads/${base}_378UDI-idt-UMI_2_val_2.fq | sh

#remove unmapped reads
samtools view -b -F 4 ${base}_$SGE_TASK_ID.aln.bam  > ${base}_$SGE_TASK_ID.aln.map.bam

#for some reason bwa.kit sort gives an error so do separately
samtools sort ${base}_$SGE_TASK_ID.aln.map.bam -o ${base}_$SGE_TASK_ID.sorted.bam

#best practice: things likely to change between jobs go in the env variable; global definitions
at top get capitals for variables
```

Merging bams

```bash
############################################################
#>>>>>>>>>>STEP 3. Merge your bam files<<<<<<<<<<<<<<<
############################################################

#!/bin/sh
# SGE options (lines prefixed with #$)
#$ -N runMergeBams.sh
#$ -cwd
#$ -l h_rt=36:00:00
#$ -l h_vmem=8G
#$ -pe sharedmem 4
#$ -P  roslin_macqueen_lab #add in priority if you have

#Jobscript to merge 6 bam files  - need to make into an array

#Initialise the Modules environment
. /etc/profile.d/modules.sh

#Load modules
module add java
module load roslin/samtools/1.9

#Define dirs
TMP_DIR=/exports/cmvm/eddie/eb/groups/macqueen_lab/Lea/psmc/bwa/tmp #scratch
TARGET_DIR=/exports/cmvm/eddie/eb/groups/macqueen_lab/Lea/psmc/bwa
PICARD=/exports/cmvm/eddie/eb/groups/ogden_grp/software/picard/picard.jar
#merge
```

```
java -Xmx10g -jar $PICARD MergeSamFiles -I $TARGET_DIR/H3G5WDSX7_1_1.sorted.bam \ -I $TARGET_DIR/H3G5WDSX7_1_2.sorted.bam \
   -I $TARGET_DIR/H3G5WDSX7_1_3.sorted.bam \ -I $TARGET_DIR/H3G5WDSX7_1_4.sorted.bam \ -I $TARGET_DIR/HFYM3DSX7_1_5.sorted.bam \
   -I $TARGET_DIR/HFYM3DSX7_2_6.sorted.bam \
   -O $TARGET_DIR/mh_reads_merged.map.sort.bam \
   -TMP_DIR $TMP_DIR

#index
samtools index $TARGET_DIR/mh_reads_merged.map.sort.bam
```

Remove duplicates and run mapstats
```
################################################################################
#
#>>>>>>>>>>STEP 4. Remove duplicates and calculate statistics<<<<<<<<<<<<<<<
################################################################################
#

#!/bin/sh
# SGE options (lines prefixed with #$)
#$ -N runmarkdups.sh
#$ -cwd
#$ -l h_rt=36:00:00
#$ -l h_vmem=16G
#$ -pe sharedmem 8
#$ -P  roslin_macqueen_lab #add in priority if you have
#$ -M s2457124@ed.ac.uk
#$ -m beas

# Initialise the environment modules
. /etc/profile.d/modules.sh
#source /exports/applications/apps/community/roslin/conda/4.9.1/etc/profile.d/conda.sh

#load modules
module load roslin/samtools/1.9
module load igmm/apps/BEDTools/2.27.1
module load anaconda/2024
module load roslin/openjdk/18.0.2
source activate mapstats_env

#define dirs
OUT_DIR=/exports/cmvm/eddie/eb/groups/macqueen_lab/Lea/psmc/bwa/mapstats
PICARD=/exports/cmvm/eddie/eb/groups/macqueen_lab/Lea/psmc/bwa/picard/picard.jar
TMP_DIR=/exports/cmvm/eddie/eb/groups/macqueen_lab/Lea/psmc/bwa/tmp
```

#get sample lists
base=mh_reads_merged

#mark dups
java -Xmx10g -jar $PICARD MarkDuplicates -I ${base}.map.sort.dedup.bam -O ${base}.map.sort.dedup.bam -REMOVE_DUPLICATES true -METRICS_FILE ${base}.metrics.txt -TMP_DIR ${TMP_DIR}

#index
samtools index -@ 4 ${base}.map.sort.dedup.bam #change to match file name

#getstats - sometimes multiqc errors out - can run in cwd with code below
samtools flagstat ${base}.map.sort.dedup.bam > ${base}.flagstat.txt
bedtools genomecov -ibam ${base}.map.sort.dedup.bam > ${base}.cov.txt
mosdepth -n --fast-mode -t 4 -Q30 ${base} ${base}.map.sort.dedup.bam
singularity exec multiqc-1.20.sif multiqc .

New autosome-only ref genome

```bash
#!/bin/sh
# Grid Engine options
#$ -N run_new_ref_gen
#$ -cwd
#$ -l h_rt=10:00:00
#$ -l h_vmem=16G
#$ -pe sharedmem 8
#$ -P roslin_macqueen_lab
#$ -M s2457124@ed.ac.uk
#$ -m beas

# Initialise the environment modules
. /etc/profile.d/modules.sh

#load modules
module load igmm/apps/samtools/1.16.1
module load roslin/bwa/2.1.0

#input var
input_file="sMobHyp1.curated_primary.mt.scrubbed.fa"

#output var
output_file="sMobHyp1.curated_primary.auto.fa"

#create empty output file
>"output_file"
```

```
#extract chrs to new file
for i in {1..31}
do
   chromosome="SUPER_$i"
   # Extract the chromosome and append to the output file
   samtools faidx "$input_file" "$chromosome" >> "$output_file"
done

#index
bwa index sMobHyp1.curated_primary.auto.fa
samtools faidx sMobHyp1.curated_primary.auto.fa
```

Extract autosomes from merged bams

```bash
#!/bin/sh
# SGE options (lines prefixed with #$)
#$ -N runmarkdups.sh
#$ -cwd
#$ -l h_rt=36:00:00
#$ -l h_vmem=16G
#$ -pe sharedmem 8
#$ -P  roslin_macqueen_lab #add in priority if you have
#$ -M s2457124@ed.ac.uk
#$ -m beas

#Initialise the environment modules
. /etc/profile.d/modules.sh
#source /exports/applications/apps/community/roslin/conda/4.9.1/etc/profile.d/conda.sh

#load modules
module load roslin/samtools/1.9
module load igmm/apps/BEDTools/2.27.1
module load anaconda/2024
module load roslin/openjdk/18.0.2
source activate mapstats_env

TARGET_DIR=/exports/cmvm/eddie/eb/groups/macqueen_lab/Lea/psmc/bwa
PICARD=/exports/cmvm/eddie/eb/groups/macqueen_lab/Lea/psmc/bwa/picard/picard.jar
BASE=mh_reads_merged

#extract autosomes
samtools view -b ${base}.merged.map.sorted.dedup.bam $(for i in {1..31}; do echo -n "SUPER_$i "; done) > $TARGET_DIR/${base}_autosomes.bam
```

```
#sort and index the new BAM file
samtools sort $TARGET_DIR/${base}_autosomes.bam > $TARGET_DIR/${base}_autosomes.sort.bam
samtools index  $TARGET_DIR/${base}_autosomes.sort.bam

#get stats for autosomes only
samtools flagstat $TARGET_DIR/${base}_autosomes.sort.bam > $TARGET_DIR/${base}_autosomes.sort.flagstat
bedtools genomecov -ibam $TARGET_DIR/${base}_autosomes.sort.bam > $TARGET_DIR/${base}_autosomes.sort.cov.txt
mosdepth -n -t 4 -Q30 $TARGET_DIR/${base}_autosomes.sort.bam $TARGET_DIR/${base}_autosomes.sort.bam

```

Consensus sequence and psmc

```
#!/bin/sh
# Grid Engine options
#$ -N psmc
#$ -cwd
#$ -l h_rt=10:00:00
#$ -l h_vmem=16G
#$ -pe sharedmem 8
#$ -P roslin_macqueen_lab

# Initialise the environment modules
. /etc/profile.d/modules.sh
source /exports/applications/apps/community/roslin/conda/4.9.1/etc/profile.d/conda.sh
conda activate psmc_env

#load eddie modules
module load igmm/apps/samtools/1.10  #note if using new version the mpileup options have changed
module load igmm/apps/bcftools/1.19

#extract chromosome 1 from bam
#samtools view -b Goat-39507187.ARS1_2.sorted.dedup.nocont.bam NC_030808.1 > chr1_goat_psmc.bam
#samtools index chr1_goat_psmc.bam

#define variables
INFILE=/exports/cmvm/eddie/eb/groups/macqueen_lab/Lea/psmc/bwa/mh_reads_autosomes.sort.bam
REFGEN=/exports/cmvm/eddie/eb/groups/macqueen_lab/Lea/psmc/bwa/sMobHyp1.curated_primary.auto.fa

#quick conseq - doesn't work for pmsc
```

#samtools consensus $infile | gzip > cons.fq.gz

#conseq bcftools
bcftools mpileup -C50 -f $REFGEN $INFILE | bcftools call -c | vcfutils.pl vcf2fq -d 3 -D 50 | gzip > mh_bam_auto.cons.fq.gz

#run program
fq2psmcfa -q30 mh_bam_auto.cons.fq.gz > mh_bam_auto.psmcfa #takes consensus sequence, filters on Q20 and redirects to output file

psmc -N25 -t15 -r5 -p "4+25*2+4+6" -o mh_bam_auto.psmc mh_bam_auto.psmcfa #involed main PSMC program, -N25 sets ratio of pop mutation rate to recomb rate to 25,

#-t15 species max number of iterations for the Markov model, -r5 set max time, -p

species time intervals, -o specifies outfile

psmc2history.pl mh_bam_auto.psmc | history2ms.pl > ms-cmd.sh

psmc_plot.pl -R -u 8.1e-09 -g 25 -p diploid mh_bam_auto.psmc #add mutation rate