Summer 7-28-2023

Annotation of Non-Model Species' Genomes

Taiya Jarva

Annotation of Non-Model Species' Genomes

by

Taiya Jarva

A Master's Thesis
Submitted to the Graduate School,
the College of Arts and Sciences
and the School of Biological, Environmental, and Earth Sciences
at The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Master of Science

Approved by:

Alex Flynt, Committee Chair
Nicole Phillips
Dmitri Mavrodi

July 2023

Published by the Graduate School

THE UNIVERSITY OF SOUTHERN MISSISSIPPI

ABSTRACT

The innovations in high throughput sequencing technologies in recent decades has allowed unprecedented examination and characterization of the genetic make-up of both model and non-model species, which has led to a surge in the use of genomics in fields which were previously considered unfeasible. These advances have greatly expanded the realm of possibilities in the fields of ecology and conservation. It is now possible to the identification of large cohorts of genetic markers, including single nucleotide polymorphisms (SNPs) and larger structural variants, as well as signatures of selection and local adaptation. Markers can be used to identify species, define population structure, and assess genetic health. In addition, researchers can examine unique features of genes related to the health of a threatened species, such as genes involved in immune function, reproduction, environmental response, as well as evolutionary trends and niche adaptations.

Recent developments in sequencing and software also allow researchers to examine the noncoding- “ome”, providing a glimpse into gene regulation, developmental pathways, and response to viral sequences and mobile elements. Characterization of the biogenesis pathways of noncoding RNAs has facilitated the development of RNA interference strategies which are increasingly being used in therapeutics, agriculture, and pest control. The focus of this project is to use available sequencing technology and computational methods to annotate the genome and small RNA pathways of non-model organisms.

ACKNOWLEDGMENTS

I would first like to thank my advisors and committee members, Dr. Alex Flynt and Dr. Nicole Phillips, for giving me the opportunity to achieve my goals and for helping me to become a better researcher, writer, and programmer. Words cannot express my appreciation. Thanks to Dr. Dmitri Mavrodi for his helpfulness and providing much-needed resources for writing grant proposals, as well as being a great committee member. Second, I'd like to thank the current and former members of FlyntOS for their support and advice, and friendship – Dr. Olga Mavrodi, Farid, Beatriz, Sweta, Yulica, Iyannu, Jay, and Cory. They made graduate school a lot more enjoyable and were always willing to give advice. I would also like to thank several other people who have either collaborated with me on projects, provided technical support, manuscript comments, or letters of recommendation to me throughout my time in the graduate program: Dr. Mac Alford, Annmarie Fearing, Dr. Kevin Feldheim, Dr. Brian Olsen, Mrs. Kelly Vera, Dr. Gregg Poulakis, Dr. Gavin Naylor, and Dr. John Perry. Lastly, I could not have succeeded in this program without the tireless support and encouragement of friends and family- my (now) husband, Chris; close friends, Baine, Sarah, and Emmy; my parents, Tedd and Lisa; and my grandfather, Brian.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BAM	Binary Alignment Mapping
BLAST	Basic Local Alignment Search Tool
BLASTP	Basic Local Alignment Search Tool Protein
BLAT	BLAST-like Alignment Tool
BPAs	Bisphenol A
BUSCO	Benchmarking Universal Single Copy Orthologs
DNA	Deoxyribonucleic Acid
ESA	Endangered Species Act
ESTs	Expressed Sequence Tags
FGFs	Fibroblast Growth Factors
GC	Guanine Cytosine
GO	Gene Ontology
GRAVY	Grand Average Value of Hydrophathicity
HPC	High Performance Compute Cluster
INBRE	IDeA Network of Biomedical Research Excellence
IUCN	International Union For The Conservation of Nature
MAPK	Mitogen-Activated Protein Kinase

NCBI	National Center For Biotechnology Information
NMFS	National Marine Fisheries Service
NOAA	National Oceanographic and Atmospheric Administration
PAHs	Polycyclic Aromatic Hydrocarbons
PCA	Principal Components Analysis
PCDDs	Polychlorinated Dibenzodioxins
PCR	Polymerase Chain Reaction
PSGs	Positively Selected Genes
RNA	Ribonucleic Acid
RNAi	Ribonucleic Acid Interference
RTK	Receptor Tyrosine Kinase
STL	Stretched Total Length
US	United States
UCSC	University of California Santa Cruz
VGP	Vertebrate Genome Project
dNTP	di-nucleotide tri-phosphate
miRNA	micro-ribonucleic acid
piRNA	piwi-interacting ribonucleic acid
siRNA	short interfering ribonucleic acid

CHAPTER I - Genomics and Transcriptomics For Conservation Efforts of The Smalltooth Sawfish, Pristis Pectinata

1.1 Background

1.1.1 Genomics in Conservation

Due to the continuously decreasing cost and improvement of sequencing technologies, whole genome and transcriptome data is increasingly being used for conservation purposes. Genomic sequences provide genetic markers and polymorphisms which can be used to assign relatedness, population structure, and gene flow, and the amount of data returned by current sequencing methods allows an unprecedented level of genome coverage. Concurrently, RNA sequencing provides information about gene expression, response to environment, and adaptation, which can be used to identify genetic factors which may be related to an organism's unique biology. RNA can be obtained relatively unobtrusively from many species using small tissue samples and can be used to annotate complete coding genes, provide information about gene expression, as well as novel isoforms. The benefits provided by the sequencing revolution, in conjunction with the development of new computational methods to assemble and analyze massive amounts of data has led to the initiation of large cooperative projects whose goals range from sequencing all vertebrate species on earth (https://vertebrategenomesproject.org), 10,000 species of plants (https://db.cngb.org/10kp, Cheng et al., 2018) to sequencing every eukaryotic genome on earth (https://www.earthbiogenome.org/). The utility of genome assemblies and

transcriptome data to aid conservation efforts and the preservation of biodiversity has been demonstrated in dozens of species to date.

1.1.2 Genome Annotation in the Era of High Throughput Sequencing

Despite the increasing ease and falling costs of whole genome and transcriptome sequencing, functional annotation of non-model genomes is still a daunting task which often requires multiple approaches. Complexities in the genomes of non-model organisms include differences in exon and intron structure, alternative codon usage frequencies, differences in GC content of coding genes, polymorphic genes, transcript isoforms, pseudogenes, and distinguishing orthologs from paralogs. A commonly used method of annotation is *ab initio* gene prediction using tools such as Augustus, which relies on signals from known coding genes to train models to predict genes directly from DNA sequences (Stanke & Waack, 2003). In contrast, other tools such as GeMoMa utilize sequence similarity to previously annotated genes (homology) (Keilwagen et al., 2016). Software which incorporates both gene predictions as well as evidence from expressed sequences (ESTs) or RNAseq data provide the most sensitive and accurate annotations to date, however, the high sequencing depth achieved by modern sequencers ensure that gene annotation remains a highly complex and evolving process.

1.1.3 Conservation Status of The Smalltooth Sawfish.

The smalltooth sawfish, *Pristis pectinata*, is one of only five extant and highly threatened species of sawfishes which are known for their long, tooth-covered rostral

saw. Out of the five extant species in this group, *Pristis pectinata* has experienced the greatest decline in its geographical range and abundance. They were considered highly abundant until the mid-nineteenth century, historically ranging from Brazil to the U.S. Atlantic coast, but today their core population is restricted to southwest Florida (Dulvy et al., 2016). For this reason, the U.S. population is assessed as endangered under the Endangered Species Act, and globally they are considered Critically Endangered by the International Union for the Conservation of Nature (Carlson et al., 2013). Their decline has primarily been caused by habitat degradation from development and both direct and untargeted overfishing (bycatch). Though recent research suggests that the core population may be recovering, there are numerous criteria for the de-listing of this species that have not yet been met (Lehman et al., 2022; Wiley & Brame, 2018). Furthermore, recovery and population growth in this species, as in other sharks and rays, is limited by late maturity and low reproductive fecundity (Kyne et al., 2021).

1.1.4 Sawfish Biology

Smalltooth sawfish are members of the Batoid (ray) family, and they possess morphological features of both sharks and rays. They are dorsoventrally flattened from rostrum to pectoral fin, yet their dorsal and caudal fins are similar to those of sharks. They are ovoviviparous, meaning that they hatch eggs internally and produce a yolk that sustains the young until they give birth. They are reported to have a biennial reproductive cycle with litter sizes ranging from 7-14 pups (Brame et al., 2019; Feldheim et al., 2017). Interestingly, they are among the several known species of sharks and rays which are

capable of facultative parthenogenesis, or asexual reproduction, though the environmental conditions that encourage this form of reproduction are poorly understood (Dudgeon et al., 2017; Feldheim et al., 2017; Fields et al., 2015; Harmon et al., 2016; Wyffels et al., 2021). Smalltooth sawfish utilize shallow coastal mangroves and river estuaries as nurseries and throughout their juvenile stage but become increasingly mobile as they mature and may be found in depths up to 73m (Wiley & Simpfendorfer, 2010). Though they tend to remain near the nurseries of their birth for much of their juvenile stages, adults may migrate to northern waters in response to seasonal temperature changes or possibly for reproduction (Graham et al., 2021).

Sawfish use their distinctive rostral saw, which comprises around 20% of their total length, for feeding and defense. The rostrum is lines with dozens of dermal denticle-derived teeth, and they are known to use their saws to slash prey in the water. The number of rostral teeth is fixed during early development, and as such can be used to distinguish Pristid species (Thorson, 1973). Sexual dimorphism is seen in the number of teeth on the rostral saw in some species of sawfish, including *Pristis pectinata*, with males having more teeth on average than females (Wiley et al., 2008). Interestingly, elongated rostrums with teeth are found only two extant families of chondrichthyans, the Pristiophorids, or sawsharks, and Pristids, the sawfishes. The dermis of the saw contains electrosensory organs called the ampullae of Lorenzini which aid in prey detection. Elongated rostrums appear to provide an evolutionary advantage, as they have evolved independently at least five times in chondrichthyans (Welten et al., 2015). Despite the

apparent advantages they provide, however, they have also been a major cause of decline, becoming easily entangled in fishing gear and causing mortality.

The ampullae lie at the end of electrically conductive ion-rich jelly-filled canals and may contain tens to thousands of sensory cells that detect weak (< 0.5 μV/cm) bioelectric signals in the water, allowing detection of prey or mates (Wueringer et al., 2012). Electrosensory physiology is unique by clade and often reflects differences in ecological habitat or feeding strategy. For example, the freshwater sawfish *Pristis microdon* has twice as many ampullary pores as marine sawfish (Wueringer et al., 2012). The spatial and topological distribution of ampullary pores is correlated with feeding strategy; sawfish target free-swimming prey in the water column and thus have a higher dorsal concentration. In contrast, other rays trap prey using their bodies while repositioning their mouths for feeding, and thus have a higher ventral density (Wueringer et al., 2012).

Upon electrical stimulation, sensory cells are depolarized, causing intracellular calcium influx through L-type voltage-gated calcium channels (VGCC), followed by repolarization and potassium efflux from calcium-activated potassium channels. This causes a change in post-synaptic potential at the basal surface of the cell and neurotransmitter release to afferent nerves. In the skate *Leucoraja erinacea*, gene expression in ampullary cells showed that depolarization of the apical membrane is controlled by $Ca_v1.3$, while repolarization is mediated by the large potassium (BK) channel (Bellono, Leitch, & Julius, 2017; Bennet & Obara, 1986; King et al., 2015). These channels are differentially modulated by β or γ accessory subunits which may

provide oscillatory responses to heightened or prolonged sensory input (Bellono, Leitch, & Julius, 2018). In sharks, the fast-acting voltage-gated “Shaker” potassium channel (Kv1.3) localizes with VGCC to regulate responses to repetitive sensory stimulation; however, in skates, only Kv1.1 and Kv1.5 have been found, and their function in electrosensing has not been elucidated (Bellono, Leitch, & Julius, 2018).

1.2 Gene Expression and Evolution in the Smalltooth Sawfish, Pristis Pectinata

Taiya M. Jarva[1], Nicole M. Phillips[1], Cory Von Eiff[1], Gregg R. Poulakis[2], Gavin Naylor[3], Kevin A. Feldheim[4], Alex S. Flynt[1]

[1]School of Biological, Environmental, and Earth Sciences. The University of Southern Mississippi, 118 College Drive, Hattiesburg, MS, 39401
[2]Charlotte Harbor Field Laboratory, Fish and Wildlife Research Institute, Florida Fish and Wildlife Conservation Commission, 585 Prineville St., Port Charlotte, FL 33954
[3]Florida Program for Shark Research, University of Florida, Gainesville, FL 32611
[4]Pritzker Laboratory for Molecular Systematics and Evolution, The Field Museum, 1400 S. Lake Shore Drive, Chicago, IL 60605

This chapter has been deposited in the BioRxiv pre-print server and is awaiting acceptance to be reviewed for a journal.

1.2.1 Summary

Sawfishes (Pristidae) are large, highly threatened rays named for their tooth-studded rostrum, which is used for prey sensing and capture. Of all five species, the smalltooth sawfish, *Pristis pectinata*, has experienced the greatest decline in range, currently found in only ~20% of its historic range. To better understand the genetic underpinnings of these taxonomically and morphologically unique animals, we collected transcriptomic data from several tissue types, mapped them to the recently completed reference genome and contrasted the patterns observed with comparable data from other elasmobranchs. Evidence of positive selection was detected in 79 genes in *P. pectinata*, several of which are involved in growth factor/receptor tyrosine kinase signaling and

specification of organ symmetry, suggesting a role in morphogenesis. Data acquired also allow for examination of the molecular components of *P. pectinata* electrosensory systems, which are highly developed in sawfishes and have likely been influential in their evolutionary success.

As meso- to apex-level predators, chondrichthyans (sharks, rays, and chimaeras) play essential roles in the health of marine ecosystems by increasing biodiversity, buffering against invasive species, decreasing transmission of diseases, and mitigating the effects of climate change (Ferretti et al., 2010; Ritchie et al., 2012) However, many chondrichthyan populations are in decline, primarily due to overfishing, with more than one third estimated to be threatened with extinction (Dulvy et al., 2021). Sawfishes belong to one of the most threatened families, with all five species assessed as Endangered or Critically Endangered on the International Union for the Conservation of Nature (IUCN) Red List of Threatened Species (Carlson et al., 2013). Sawfishes are notable for their tooth-studded rostrum, which is used to detect, acquire, and manipulate prey (Poulakis et al., 2017; Wueringer, 2012). This rostrum also makes them especially susceptible to entanglement in fishing gear, which has precipitated a global decline in their numbers over the last century (Brame et al., 2019; Carlson et al., 2013; Dulvy et al., 2016).

Bycatch in fisheries and habitat degradation continue to pose the greatest threats to sawfishes, including in ‘stronghold’ locations such as the United States and Australia

(NMFS 2009)(Graham et al., 2022). Fisher education and trade bans on sawfish and their parts have been used to mitigate sawfish declines (Wiley & Brame, 2018). While such strategies are essential, additional preventative initiatives are needed as sawfish populations continue to decline globally (Yan et al., 2021). Recent research aimed at supporting development of deterrent technology found that sawfish will react to electric field stimuli, but behavioral responses were not consistent and were considered insufficient to avoid fisheries gear entanglement (Abrantes et al., 2021).

The electrosensing abilities of sawfishes are currently poorly understood, and little is known about the underlying the molecular mechanisms of electroreception and environmental biosensing that underpin their behaviors. The aims of this study were to collect transcriptome sequences from different tissues of the Critically Endangered smalltooth sawfish, *Pristis pectinata*, map the data to the recently completed high resolution genome assembly, and compare patterns of gene expression and sequence evolution with those derived from other elasmobranchs to identify unique components of genomic architecture. Of all sawfishes, *P. pectinata* has experienced the most severe decline in range and is present in less than 20% of its former range in the Atlantic Ocean (Brame et al., 2019; Dulvy et al., 2016). Viable populations are currently restricted to Florida in the U.S. and western portions of The Bahamas and, like all sawfishes, reducing fisheries interactions is a top conservation priority (Carlson et al., 2013). The gene set collected in this study provides insight into evolutionary patterns and the mechanisms of electrosensing in *P. pectinata* and serve as critical first steps for future work.

1.3 Results

1.3.1 High Quality *Pristis Pectinata* Gene Set

In collaboration with the Vertebrate Genome Project, GN produced a high-quality reference genome assembly for an adult female *P. pectinata* that had died at the Ripley's Aquarium in Myrtle Beach, South Carolina (GCA_009764475.1). The assembly was based on 60x PACBIO long read sequencing, BioNano, Hi-C and Illumina short read data (scaffold N50 101.7M; Contig N50=17M; 99.61% of the data assigned to 48 chromosomes. Genome size 2.27Gb; http://vertebrategenomesproject.org). However, while the assembly had high contiguity, there were limited EST data associated with the original genome assembly (Salzberg, 2019). As a result, approximately 40% of expected orthologs were absent from the predicted gene dataset (Fig. 1.1E)(Simão et al., 2015). To address this issue and maximize utility of the *P. pectinata* genome, RNA was sequenced from tissues collected from a juvenile female (828 mm stretch total length) collected by GP under ESA Permit No. 21043. Datasets were generated from brain, kidney, liver, ovary, and skin tissues fixed in RNAlater.

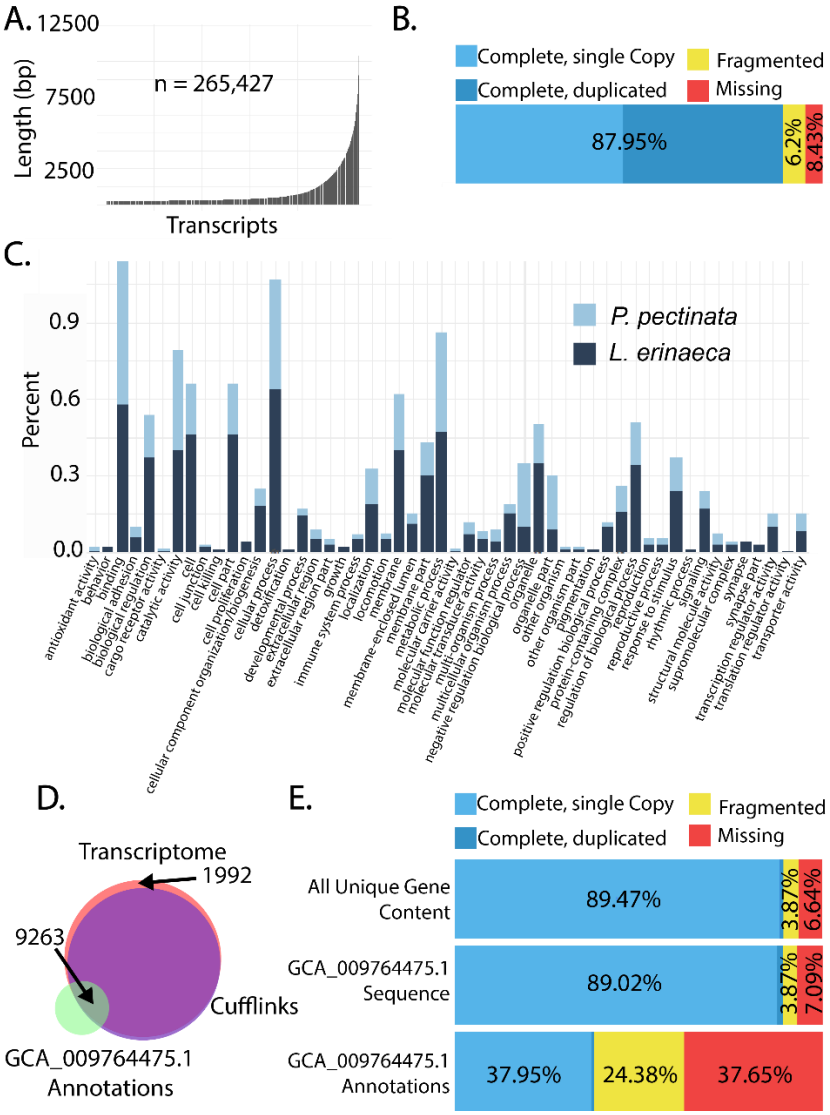


Figure 1.1 *Establishment of a near complete gene set for the smalltooth sawfish, Pristis pectinata.*

A) Log length distribution of transcripts in assembled transcriptome with total number of transcripts (n = 265,427). **B)** Outcome of Benchmarking Universal Single Copy Ortholog (BUSCO) analysis to assess completeness of assembled transcriptome. **C)** WeGO (Ye et al., 2018) gene ontology analysis of P. pectinata and little skate, Leucoraja erinacea, transcriptomes. Genes are classified according to biological process, molecular function, or cellular component and plotted by percent of genes related to meaningful biological activities. P. pectinata has a higher percentage of genes related to DNA binding, negative regulation of biological processes, organelle part, and antioxidant activity, while L. erinacea has more genes related to cellular processes, response to stimulus, signaling, and catalytic activity. **D)** Venn diagram (Hulsen et al., 2008) showing overlap of the number of transcripts provided with the public genome (GCA_009764475.1, green), the Cufflinks

assisted annotation of the public genome (purple), and transcripts in the de novo assembled transcriptome (pink). **E)** BUSCO assessment for all unique gene content from combined transcriptome and genome recourses, and annotations associated with GCA_009764475.1.

Genome annotation facilitated by RNA-seq alignments yielded substantially more transcripts (159,014), than the 19,597 genes predicted from the genome alone. RNA-seq data were also assembled into a *de novo* transcriptome of 265,427 transcripts with 31.9% being >500 bp (Fig. 1.1A-B). Substantial improvement was seen in the contribution of the *de novo* transcriptome and RNA-seq guided annotation, with approximately 2000 genes exclusive to the *de novo* transcriptome. Less redundancy was seen in the dataset after intersecting *de novo* transcripts with predicted genes from the genome, with 9,233 transcripts shared between datasets (Fig. 1.1D). Combining the reference genome assembly predictions, RNA-seq assisted annotations, and a *de novo* transcriptome resulted in a greatly enhanced gene set representation (Fig. 1.1E). In the *de novo* transcriptome alone, less than 10% of expected orthologs were absent. In the combined gene set, the percentage of missing genes was reduced to only 6.6% (Table 1.2).

Table 1.1

	Smalltooth sawfish, *Pristis pectinata*	Chain catshark *Scyliorhinus retifer*	Little skate, *Leucoraja erinacea*	Indonesian coelacanth, *Latimeria menadoensis*,	Australian ghostshark, *Callorhinchus milii*

Table 1.1 continued

No. transcripts in dataset	*175,569*	*107,231*	*103,996*	*66,138*	*92,334*
Accession	PRJNA864825	GEO: GSM6439958	GEO: GSM6439957	GAPS010661368	GEO: GSM6439959
%BUSCO complete	89.47% C 3.87% F	56.6% C 22.1% F	60.4% C 20.4% F	40.9% C 18.9% F	47.3% C 26.9% F

Table 1.1 *Summary of Transcriptomes Used For Positive Selection*

Number of transcripts, NCBI accession numbers, and percentage of complete (C) and fragmented (F) orthologs via BUSCO analysis of transcriptomes for all taxa used in positive selection analyses.

To further validate the *P. pectinata* gene set, gene ontology (GO) terms were assigned and compared to the transcriptome of the little skate, *Leucoraja erinacea* (King et al., 2011) (Fig. 1.1C). Distributions of high-level GO terms were similar between *P. pectinata* and *L. erinacea*, suggesting gene content in the *de novo* transcriptome represents what is observed in related taxa. However, differences were noted, such as a higher percentage of genes involved in antioxidant activity and DNA binding in *P. pectinata* versus a higher percentage of genes in *L. erinacea* related to signaling, response to stimulus, metabolic process, and catalytic activity. This may reflect a difference in juvenile and adult tissues sampled for *P. pectinata* relative to the embryonic tissue-derived *L. erinacea* transcriptome. The near-complete *P. pectinata* gene set enables

characterization of unique genetics in this species, which was not possible with predicted annotations offered by genome sequence approaches alone.

1.3.2 Unique Genetic Features of *Pristis Pectinata*

To identify positively selected genes (PSGs) in *P. pectinata*, the combined dataset from *P. pectinata* and transcriptomes of four other fish species were assigned to orthologous gene groups with Orthofinder (Emms & Kelly, 2015). The species included were the Australian ghostshark, *Callorhinchus milii*, chain catshark, *Scyliorhinus retifer*, and *L. erinacea*, with the Indonesian coelacanth, *Latimeria menadoensis*, serving as an outgroup (Fig. 1.2A). *Pristis pectinata* had the second highest percentage of genes (~51%) assigned to an orthogroup and the highest percentage of species-specific genes, likely due to the substantially greater completeness of the assembly relative to the other species (Sup. Fig. 1). The resulting 3,116 genes were tested for branch-specific episodic selection using aBSREL, revealing 79 PSGs in *P. pectinata* (Smith et al., 2015) (Supplementary Methods, Table 1.3).

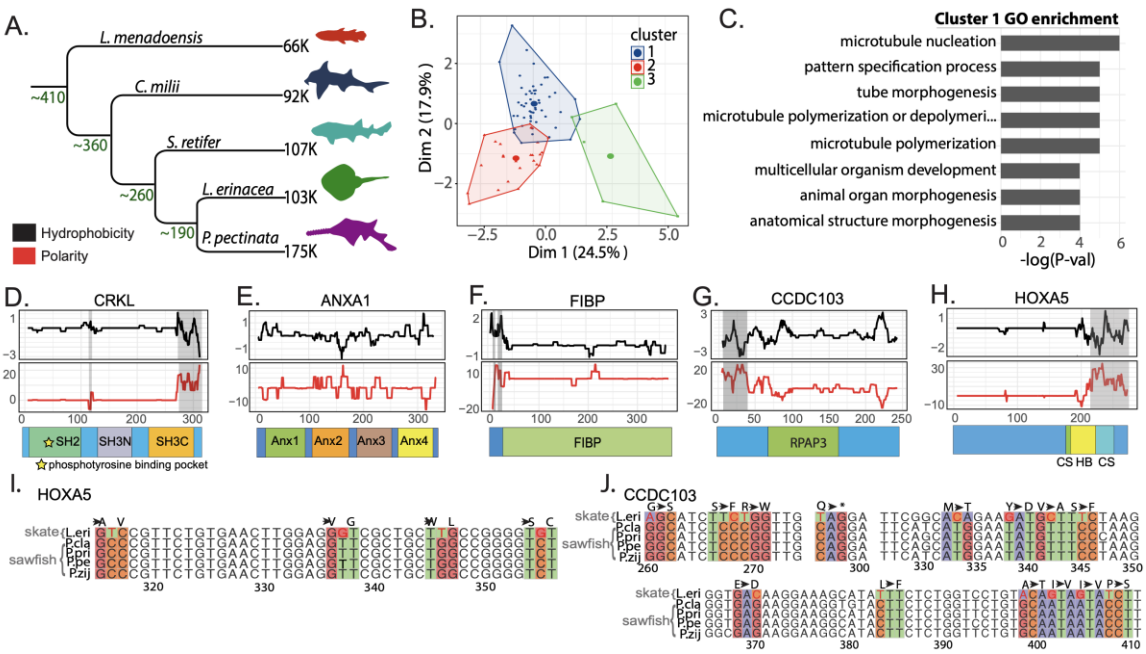


Figure 1.2 *Smalltooth sawfish, Pristis pectinata, genes under positive selection.*

A) Phylogeny of taxa used in positive selection analysis with approximate divergence times shown at each node and number of transcripts in each transcriptome noted at each terminal branch. Species included: *Latimeria menadoensis*, *Callorhinchus milii*, *Scyliorhinus retifer*, *Leucoraja erinacea*, and *P. pectinata*. **B)** Results of k-means clustering analysis of selected genes using omega value, percent of sites, change in GRAVY value relative to little skate, *L. erinacea*, and change in expression relative to *S. retifer*. **C)** Top 8 most enriched gene ontology (GO) terms related to biological processes from genes which were grouped into Cluster 1 by PCA analysis plotted by -log(p-value). **D-H)** Changes in hydrophobicity and polarity of protein sequence relative to *L. erinacea.* Sequence alignment gaps are shown by shaded regions. Functional protein domains retrieved from InterproScan and literature are shown as colored boxes below each plot. **I)** Sequence alignment of PCR-amplified region of interest from HoxA5 between *Pristis* sawfishes and *L. erinacea* with codons of interest colored by nucleotide. **J)** Alignment of region of interest from amplification of CCDC103.Alignments are colored by nucleotide.

Table 1.2

Species	Unique genes	Total genes
Smalltooth sawfish, *Pristis pectinata*	79	96
Chain catshark, *Scyliorhinus retifer*	26	38
Australian ghostshark, *Callorhinchus milii*	76	95
Little skate, *Leucoraja erinacea*	40	51
Indonesian coelacanth, *Latimeria menadoensis*	49	60

Table 1.2 *Genes Under Selection By Species.*

Number of unique genes and total number of genes found under selection in aBSREL analysis by species.

To cluster *P. pectinata* PSGs into groups which may share functional relationships or similar selection pressures, omega values from aBSREL, percent of sites under selection, gene expression rank relative to *S. retifer* orthologs, and the difference in Grand Average of Hydropathy (GRAVY) between *P. pectinata* and *L. erinacea* orthologs were collected for Principal Components Analysis (Sup. File 1). K-means clustering identified three groups containing 49, 25, and 5 genes (Fig. 1.2B). Generally, changes in GRAVY and omega values were highly correlated, with GRAVY values being the major contribution to variance (Fig. 1.3A). Between tissues, the highest correlation was seen between skin, kidney, and brain, while liver and ovary showed little correlation. Gene ontology enrichment analysis by TopGO of PSGs in Cluster 1 versus all genes analyzed

showed enrichment in functions such as multicellular organism development, animal organ morphogenesis, and anatomical structure morphogenesis (Alexa & Rahnenfuhrer J, 2022) (Fig. 1.2C). In comparison, Cluster 2 genes were related to regulation of response to biotic stimulus, regulation of defense response, and proteasomal protein catabolic process (Fig. 1.3B). Interestingly, Cluster 1 contained multiple genes implicated in developmental processes such as fibroblast growth factor/receptor tyrosine kinase (FGF/RTK) and mitogen activated protein kinase (MAPK) signaling, suggesting that *P. pectinata* PSGs share functional or evolutionary similarities. The altered FGF and MAPK signaling that may result from these diverged genes implies possible relevance to *P. pectinata*-specific morphogenesis.

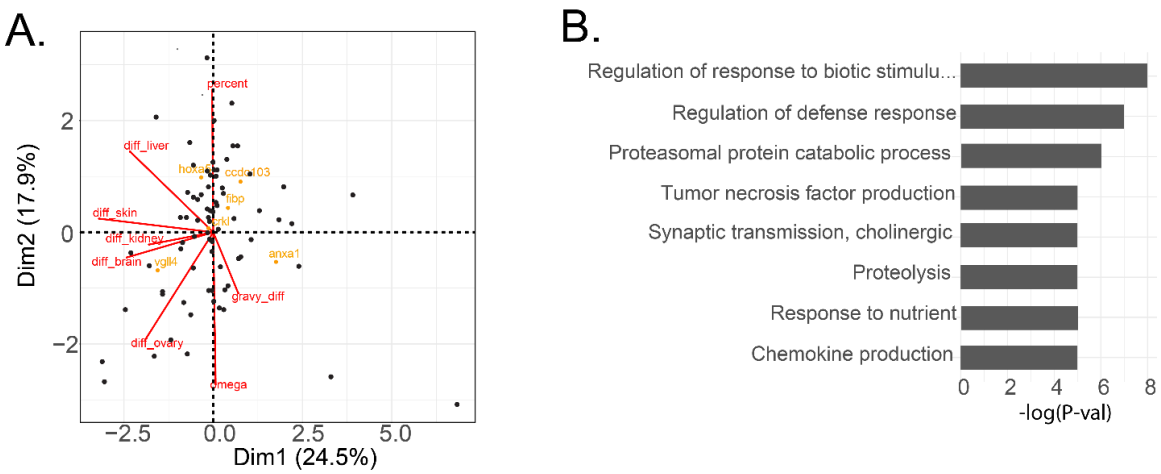


Figure 1.3 *Smalltooth sawfish, Pristis pectinata, genes under selection.*

A) PCA of genes under selection using omega value, percent of sites, change in GRAVY value relative to little skate, *Leucoraja erinacea*, and change in expression relative to chain catshark, *Scyliorhinus retifer*. Genes of interest from Cluster 1 are shown in yellow, variables are shown in red. **B)** Top 8 most enriched gene ontology terms related to biological processes from genes grouped into Cluster 2 by PCA analysis plotted by -log(p-value).

Comparing biochemical properties such as hydrophobicity and polarity of developmental genes with the closest available taxon, *L. erinacea*, revealed functional changes in *P. pectinata* in three notable genes: *Crk-l*, an integrator of multiple signaling pathways; Annexin A1 (*Anxa1*), a modulator of FGF ligands and RAS, and FGF1 intracellular binding protein (*Fibp*), an EGF/MAPK modulator (Fig. 1.2D-H) (Balasubramanian & Zhang, 2016; Katoh & Katoh, 2006). CRK-L had decreased hydrophobicity in the Src Homology 2 (SH2) domain where tyrosine phospho-proteins like growth factor receptors bind and decreased hydrophobicity in its SH3 domain that associates with RAC1 or RAS (Antoku & Mayer, 2009) (Fig. 1.2D). ANXA1 had changes in all four Annexin repeat domains, which upon Ca^{2+} binding display phosopho-sites (Fig. 1.2E). Annexins activate MAPK signaling either through growth factors, or they can be directly phosphorylated by RTKs (Babbin et al., 2006). FIBP, which binds FGF1, showed fluctuations in both hydrophobicity and polarity in the annotated FIBP domain, though little is known about the mechanisms of action of this protein (Thauvin-Robinet et al., 2016) (Fig. 1.2F). Selection in genes such as these, which are potentially involved in body patterning and growth factor signaling suggest a role related to rostral development. A prime example is *Crk-l*, which has been shown to cause craniofacial defects in mice through disruption of neural crest development or through the retinoic acid/*Tbx1* regulatory network (Guris et al., 2006; Newbern et al., 2008; Yutzey, 2010).

Another major function of Cluster 1 genes is microtubule biology with two genes being significant (*Haus2* and *Tubgcp4*). Shared evolutionary trajectory with developmental regulators suggests a similar functionality in sawfish biology, such as through cilia-mediated signaling. One example is *Ccdc103*, which is necessary for ciliogenesis. CCDC103 had numerous changes in polarity and hydrophobicity in its Rpap3 domain, which binds the axonemes of cilia (Figs. 1.2G). This is necessary for outer dynein arm attachment, and thus changes in this domain suggest functional differences which do not appear to be a result of gene duplication and subsequent divergence (King & Patel-King, 2015). Lastly, a homeobox transcription factor involved in segmentation and body patterning, *HoxA5*, was found. The protein sequence had changes in polarity and hydrophobicity near the conserved site (residues 183–188) and the beginning of the DNA-binding homeobox domain, despite being truncated relative to *L. erinacea*. These changes suggest modified interactions between HOXA5 and its cofactors and/or targets (Fig. 1.2H).

To support whether changes in PSGs were specific to *P. pectinata* or shared with other sawfishes, sequencing of genomic DNA in functional domains was performed using samples from three other *Pristis* sawfishes: dwarf sawfish (*P. clavata*), green sawfish (*P. zijsron*), and largetooth sawfish (*P. pristis*). *Ccdc103* and *HoxA5* were chosen as they demonstrated clear chemical differences between *P. pectinata* and *L. erinacea* but were conserved enough in flanking regions of functional domains for amplification. Sequence alignments revealed that the ratios of non-synonymous to synonymous nucleotide

substitutions in both genes were lower among sawfishes than when comparing sawfishes to *L. erinacea*. At *HoxA5* conserved sites, all five substitutions were shared among sawfishes but were non-synonymous with *L. erinacea*, with amino acid property changes at two sites (Fig. 1.2L, Fig. 1.5). A similar pattern was seen in *Ccdc103* among sawfishes. In the Parp3 domain of *Ccdc103*, 18/21 changes were non-synonymous between sawfishes compared to *L. erinacea*, and eight had an amino property change (Fig. 1.2J, Fig. 1.4). Fewer changes were seen among sawfishes; 10/13 were non-synonymous, and seven of these included chemical changes. Given the functions of PSGs and that the ratios of non-synonymous substitutions were lower among *Pristis* sawfishes compared to *L. erinacea*, this supports that these changes could be related to themes found in Cluster 1 genes. As conservation of the biochemical changes occurred at the base of the *Pristis* branch and conserved in all species this leads to further credence for a role in sawfish-specific morphological development.

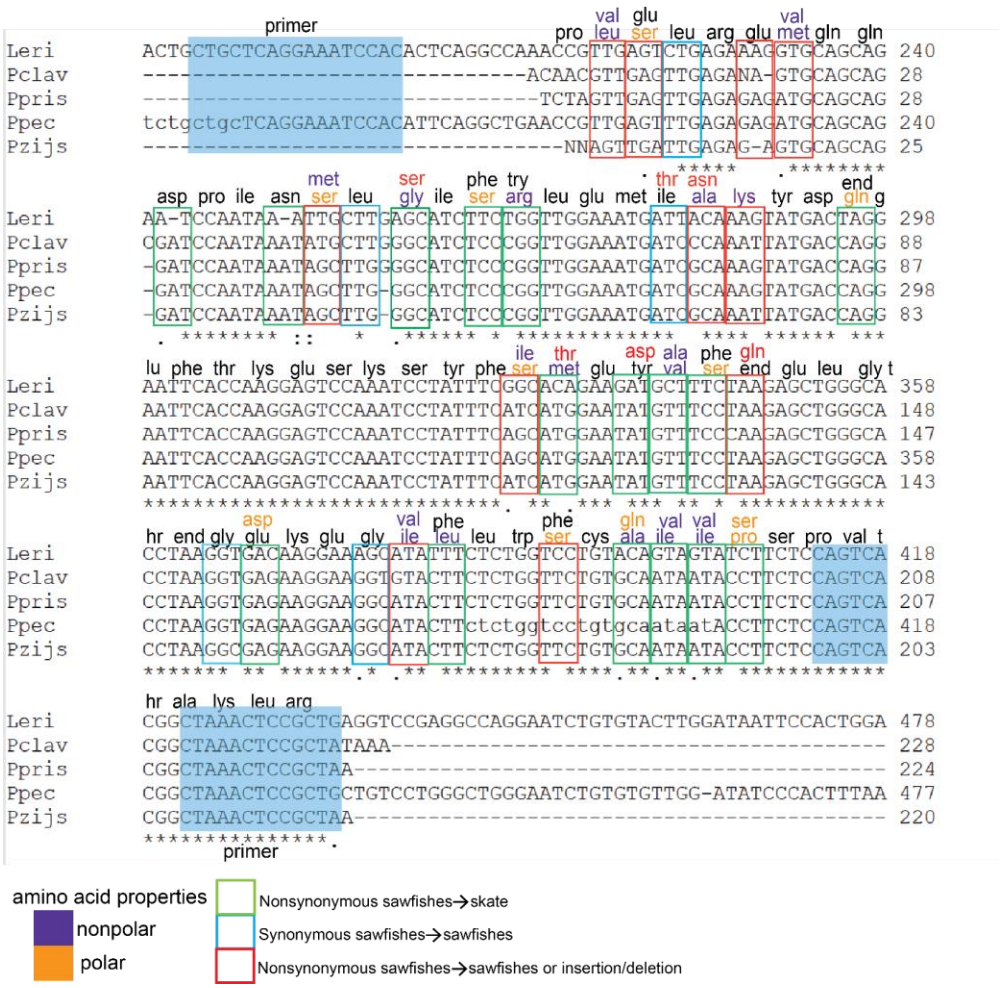


Figure 1.4 *Annotated Multiple Sequence Alignment of Ccdc103.*

Annotation of substitutions in *Ccdc103* between dwarf, green, largetooth, and smalltooth sawfishes (*Pristis clavata*, *P. zijsron*, *P. pristis*, *P. pectinata*, respectively) and little skate, *Leucoraja erinacea*. Changes in amino acids are noted above aligned nucleotide sequences. Primer sequences are shaded in blue.

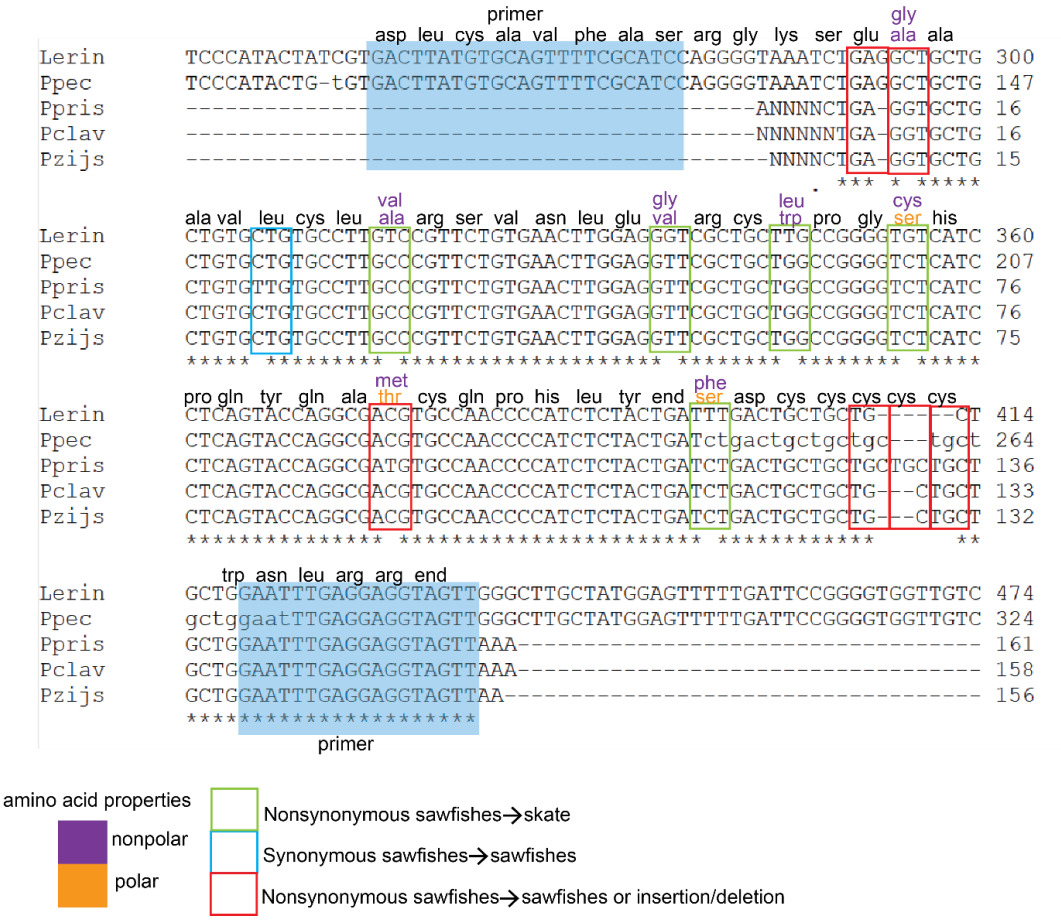


Figure 1.5 *Annotated Multiple Sequence Alignment of HoxA5.*

Annotation of substitutions in *HoxA5* between dwarf, green, largetooth, and smalltooth sawfishes (*Pristis clavata*, *P. zijsron*, *P. pristis*, *P. pectinata*, respectively) and little skate, *Leucoraja erinacea*. Changes in amino acids are noted above aligned nucleotide sequences. Primer sequences are shaded in blue.

1.3.3 Characterization of Electosensory Genes

In sawfishes, the rostrum provides an exaggerated platform for placement of electrosensing organs called ampullae of Lorenzini. Initial sensing occurs through voltage-gated calcium channels (VGCC) produced from *Cacnα1D* in both rays and sharks (Bellono, Leitch, & Julius, 2017). Following influx of Ca2+ ions, K+ efflux

occurs leading to membrane depolarization and neurotransmission. Different K+ channels participate in rays (BK-channels, BK-α) and sharks (Shaker-type, Kv). Using the *P. pectinata* transcriptome collection described above, orthologs of all channels involved were identified to understand mediators of electrosensing in this species.

First, all VGCC subunit transcriptomes were assessed for expression in rostral skin (Fig. 1.6A). Six VGCC-related genes were expressed, including a *Cacnα1D* ortholog. Others were channel accessory subunits (*Cacnα2δ, Cacnb2,* and *Cacnb4*), or a *Cacnα1C*-type VGCC, which have not been previously implicated in electrosensing chondrichthyans. A *Cacnb2* ortholog was found in the transcriptome but was not expressed in rostral skin. Thus, as expected, the *Cacnα1D* ortholog is likely the primary mediator of electrosensing. However, *P. pectinata Cacnα1D* is missing a 91-residue segment that would interact with intracellular effectors which is present in other rays (Fig. 1.6B) (Bellono, Leitch, & Julius, 2017). A portion of this region is also deleted in other chondrichthyans such as *C. mili* and the whale shark, *Rhincodon typus*, indicating that it is a site of functional novelty. Together, these results suggest that substantial changes occur during the initial step in the ability of *P. pectinata* to detect the bioelectric signal of prey.

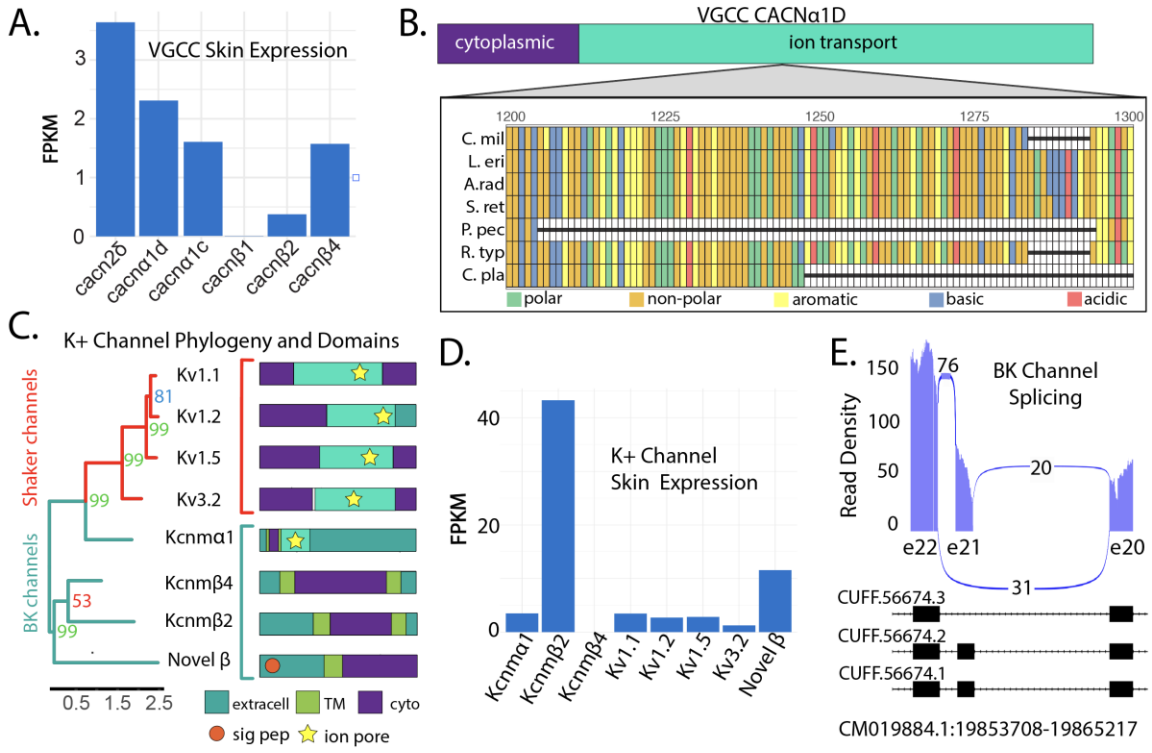


Figure 1.6 *Smalltooth sawfish, Pristis pectinata, electroreception machinery.*

A) FPKM normalized expression of voltage gated calcium channel α and β subunits in rostral skin. **B)** Multiple sequence alignment of Cacnα1D depicting missing 91-residue segment in P. pectinata ion channel domain relative to other elasmobranchs. Species included: *Callorhinchus milii* (C. mil), *Leucoraja erinacea* (L. eri), *Amblyraja radiata* (A. rad), *Scyliorhinus retifer* (S. ret), *P. pectinata* (P. pec), *Rhinchodon typus* (R. typ), and *Chiloscyllium plagiosum* (C. pla). Alignments are colored by amino acid chemistry. **C)** Phylogenetic construction of potassium channels and subunits with bootstrap values noted (left). Sequences were aligned with ClustalW and tree was constructed by RAxML with 1000 bootstraps. Shaker channel sequences are highlighted in red, while BK channels are highlighted with light blue. Annotated protein domains of each sequence obtained using Interpro Scan (Right). EC is extracellular, TM, transmembrane, cyto is cytoplasm, and pore is ion channel pore. **D)** Normalized expression of potassium channels potentially involved electrosensing in rostral skin. Kcnm- transcripts correspond to BK channel subunits, Kv are voltage-gated Shaker channels. Novel β refers to the uncharacterized BK subunit found in transcriptome data. **E)** GGSashimi plot of alternatively spliced BK channel transcripts in all P. pectinata tissues.

Transcriptome expression confirms that the BK channel transcript with highest expression in the rostral skin retains exon 21 (exon 29 in L. erinacea). Next, sequences of eight orthologs of Ca^{2+} responsive K^{+} channel were identified and compared (Fig. 1.6C). Clustering showed both major groups, BK (α and b) and Shaker types, were present in the gene set. A BK-α ortholog with clear pore motif was found (*Kcnmα1*) along with three accessory BK-β subunits, one of which was designated as novel as it could not be definitively paired with an ortholog. The four Shaker-type channels also exhibited the requisite pore domain residues. Two of these channels have been observed in *L. erinacea*, but they have only been demonstrated to work in conjunction with *Cacnα1D* orthologs in ampullae of the chain catshark, *S. retifer*, though they have also been found to be enriched in ampullae of the American paddlefish, *Polyodon spathula* (Bellono, Leitch, & Julius, 2018; Clusin et al., 2019) (Modrell, et al., 2017).

All eight channel components are expressed in rostral skin, except for the *P. pectinata* β-4 ortholog (Kcnmb4) (Fig. 1.6D). Absence of this inhibitory subunit may contribute to the heightened electrosensing seen in sawfishes. Further differences were seen in expression levels of β-2 subunits. In *L. erinacea*, expression of β- subunits in electrosensory cells is 1000- to 10000-fold lower than α-subunit expression, yet in *P. pectinata* tissue, a BK β-2 is expressed nearly 10-fold to BK-α, and the novel subunit nearly 2-fold (Clusin et al., 2019). Together, this suggests *P. pectinata* has altered BK-channel physiology. Another unusual observation was expression of all four shaker-type

channels, which differs from *L. erinacea* where only two, Kv1.1 and Kv1.5, are found (Clusin et al., 2019).

Other aspects of electrosensing machinery in *P. pectinata* showed potentially divergent biology. Several inserts and deletions were found in *Kcnmα1* that may alter its function (Fig. 1.7A-C). The ion pore-containing BK α-subunits can differentially associate with β- subunits to modulate membrane repolarization to alter activation rates or affect calcium sensitivity, drawing into question the role of the novel β-subunit in the overall performance of the system (Brenner et al., 2000; Gonzalez-Perez & Lingle, 2019; Li & Yan, 2016). Another difference found in *Kcnmα1* was the inclusion of exon 21 (known as exon 29 in *L. erinacea*) (Fig. 1.7E). Alternative splicing at exon 21 has been identified in BK channel ampullary transcripts of *L. erinacea* and in auditory hair cells of chick cochlea, though the effect on electrosensing is not known (Benjamin. King et al., 2015). The inclusion of this exon appears to be variable among elasmobranchs, being present in *P. pectinata*, thorny skate, *Amblyraja radiata*, and *L. erinacea* adult ampullae, but absent in *S. retifer*, white-spotted bamboo shark, *Chiloscyllium plagiosum*, and *L. erinacea* embryonic sequence (Fig. 1.7C).

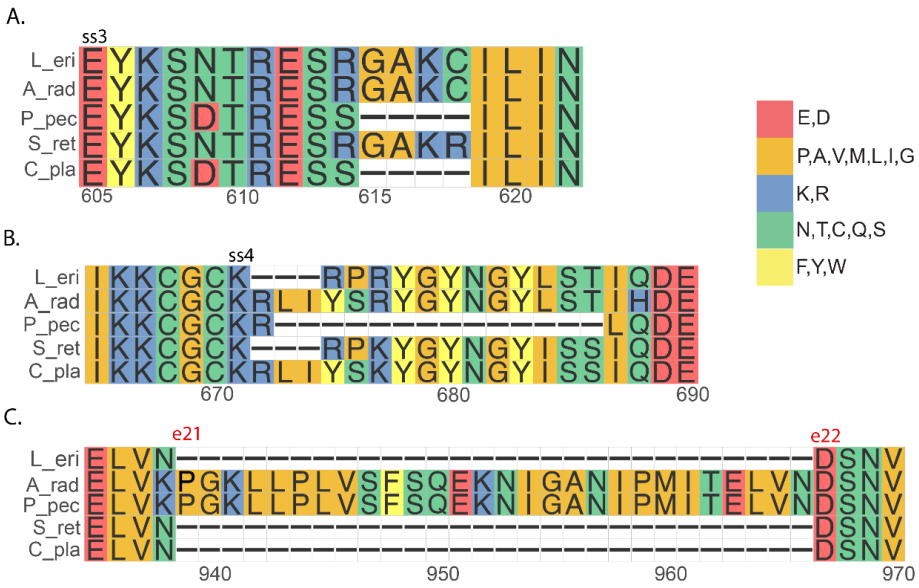


Figure 1.7 *Partial Multiple Sequence Alignments of BK alpha K+ Channel.*

Insertions, deletions, and splice sites identified in Pristis pectinata (P_pec) BK alpha K+ channel relative to Leuroraja erinacea (L_eri), Amblyraja radiata (A_rad), Scyliorhinus retifer (S_ret) and Chiloscyllium plagiosum (C_pla). Ss = splice site. e = exon. Plot colors indicate amino acid chemical properties.

1.4 Discussion

1.4.1 High Quality *Pristis Pectinata* Gene Set

Here, fundamental changes in developmental genes were found in *P. pectinata*, which may influence morphogenesis of the rostrum. Elongated rostral structures have evolved in chondrichthyans at least five times (Welten et al., 2015). Sawfishes (Batoidea: Pristidae) and sawsharks (Selachii: Pristiophoridae) are the only extant families that possess toothed rostra, but similar structures are also found in the fossil record, including two species of chimaeras, Squaloraja *polyspondyla* and *Acanthorhina jaekel* (Holocephali), Sclerorhynchoidei (Rajiformes), and *Bandringa rayi*. (Elasmobranchii) (Greenfield, 2021; Sallan & Coates, 2014; Welten et al., 2015). Convergent evolution of

elongated rostral structures suggests an evolutionary advantage in prey detection, likely through heightened bioelectric sensing, which may provide sensory information at night, at depth, or in murky shallow waters (Wueringer et al., 2012). The cohort of genes uniquely under selection in *P. pectinata* may provide a framework for the genetic changes that underpin the morphogenetic origins of saw-like rostral structures. Primarily, this appears to be through reshaping the signaling environment, as was observed in this study, with changes in FGF/MAPK mediators. Simultaneously, positive selection was also seen in *HoxA5*, a transcription factor that establishes regions along the dorsoventral axis. However, in vertebrate development, cluster 5 Hox genes are typically expressed in the somites which eventually become the upper thorax and organs, and thus changes in this gene in *P. pectinata* may be related to differences in pectoral or synarcual morphology, and not related to the rostrum.

A consequence of potentially altered FGF/MAPK signaling in *P. pectinata* is divergent cellular response to environmental stressors such as changes in salinity, temperature, and dissolved oxygen. In Florida, juveniles have an affinity for warm (>30°C), brackish waters (18–30) with high dissolved oxygen levels (>6 mg/L).(G. R. Poulakis et al., 2011) Cold water temperatures (<8–12°C) are known to alter habitat use and cause mortality(Feldheim et al., 2017b; Poulakis et al., 2011), while the physiological impacts of hypoxia and low salinities are unknown. As *P. pectinata* begins to recover, populations should re-establish in historical habitats. While there is evidence that such re-expansions have begun in the northern Gulf of Mexico and in the southern Indian River

Lagoon, Florida (Lehman et al., 2022) (G. R. Poulakis unpublished data), water quality issues and fluctuating environmental conditions, caused in part by freshwater diversions that modify the hydrology of estuaries and cause algal blooms and hypoxic conditions, may pose a risk to sawfish health and viability (Lehman et al., 2022).

The genes involved in this environmental signaling can be affected by marine pollutants including polycyclic aromatic hydrocarbons (PAHs), heavy metals, bisphenol-A's (BPAs), brominated flame retardants (BFRs), and polychlorinated benzodioxins (PCBs). For example, *Crk-l* expression is affected by BPAs, BFRs, and PCBs, and *Anxa1* is upregulated in response to PAH exposure in sea turtles as a response to increased production of reactive oxygen species (Cocci et al., 2019). Cadmium, PCBs, and other persistent organic pollutants can bioaccumulate in elasmobranch tissues, and may disrupt critical processes such as metabolism, immune function, and reproduction, though pollutant levels in sawfish tissues have not been reported (Hamers et al., 2006; Martins et al., 2021; Tiktak et al., 2020). Future research should identify contaminant levels in sawfish tissues and document sources of these pollutants in habitats where *P. pectinata* reside and may be re-establishing. Characterizing direct physiological effects of pollutants in sawfishes via *ex situ* experiments is not feasible; however, a potential circumvention of this issue could be to identify biomarkers in *P. pectinata* which are correlated with toxicological risks (Cullen et al., 2019).

A top conservation priority for *P. pectinata*, and all sawfishes, is to reduce injuries and mortalities in fisheries, especially trawl fisheries. Understanding the basis of sawfish electrosensing at a molecular level may support the development of more effective deterrent technologies that exploit this sensing modality. In future studies, molecular-level responses to stimuli, such as electric fields, should be studied to refine optimal experimental conditions, and used in parallel with aquarium trials to elicit avoidance responses (Abrantes et al., 2021). Analysis of electrosensory genes revealed numerous differences in functional domains and expression of channels and subunits which were previously not implicated in electrosensing in chondrichthyans. As basal batoids, sawfish may possess unique electrosensing biology that leverages both types of K+ channels found in sharks and other rays. Considering ray-type electrosensing alone, elevated expression of multiple previously uncharacterized subunits suggests *P. pectinata* may have substantially altered BK-channel physiology. Altogether, these results suggest unique physiology of ion channels and subunits in *P. pectinata*, and their identification could allow tailored bycatch-reduction technology. Developing this technology for trawls is a priority as shrimp trawls have been identified as still having the highest bycatch risk for sawfishes in the 'stronghold' nations of the U.S. and Australia (Graham et al., 2022).

In addition to providing an annotation of the publicly available genome, this study highlights the potential value of genomic approaches to conservation efforts, particularly through the identification and characterization of electrosensory genes and subunits and genes that could be used to reduce bycatch. Global sawfish recovery requires aggressive

conservation planning that uses novel methodological approaches to support modern, high-tech solutions. Results from this work suggest that the impacts of pollutants need to be more deeply investigated and that physiological responses may differ between sawfishes and existing model organisms. These data will also support the ability to build experimental systems to test channel behavior in a controlled *in vitro* setting that can be used to assess and quantify the performance of sawfish electrosensing to facilitate development of behavior-modifying technology. Together, these data and insights provide the foundation to support key future research, with the goal of supporting global recovery of imperiled sawfishes.

1.5 Supplementary Methods

1.5.1 Sample Collection

RNAs were collected from the brain, liver, kidneys, ovary, and skin tissues of a juvenile female *P. pectinata*. Tissues were preserved in RNAlater and stored at -80°C until use. Samples from each tissue were homogenized and extracted with the TRIzol method. Concentrations and purities of RNA extracts were measured by nanodrop and bioanalyzer 2100 (Table 1.3).

Table 1.3

Tissue	Concentration (ng/ul)	Purity (A260/A280)	RIN
Brain	174.7	1.83	6.7
Liver	962.4	2.01	6.9
Ovary	1660.7	2.03	7.0
Kidney	1150.0	2.04	7.3
Skin	273.9	1.98	7.5

Table 1.3 *Concentrations and purities of smalltooth sawfish, Pristis pectinata, RNA by tissue.*

Concentrations and purities obtained by nanodrop and Bioanalyzer. For RNA, chemical purity is indicated by A260/A280 of 2.0. RNA integrity number (RIN) ranges from 1 to 10, where 10 is intact and 1 is degraded.

1.5.2 Transcriptome Sequencing And Assembly

The transcriptome was assembled using Trinity on the Magnolia High Performance Computing (HPC) cluster (Grabherr et al., 2011). Blast2GO was used to assign gene ontology terms to transcripts, and WeGO was used for comparison between *P. pectinata* and *L. erinacea* (Gotz et al., 2008; Ye et al., 2018). Completeness was assessed with BUSCO in transcriptome mode against the most recent vertebrate lineage, odb10. To annotate the genome with coding sequences and to capture a more complete gene set for positive selection analysis, the published *P. pectinata* genome from an adult female (GCA_009764475.2) was used in addition to the assembled transcriptome.

1.5.3 Genome Annotation For Positive Selection Analysis Gene Set

Gene content exclusive to the transcriptome was isolated using STAR aligner with default parameters to align reads from all tissue samples to the genome (Dobin et al., 2013). Coverage over the genome was assessed with BEDtools and reads with low coverage (< 5 reads) were then aligned to the assembled transcriptome using Hisat2 with the –dta-cufflinks option enabled (Kim et al., 2015; Quinlan & Hall, 2010). Cufflinks was used to map RNA reads to coordinates in the genome, which were then intersected,

excluding overlapping transcripts, with Augustus-predicted gene sequence coordinates for the genome using BEDtools Intersect (Trapnell et al., 2010). All BAM and SAM file conversion and sorting was performed with SAMtools (Li et al., 2009). Unique gene sequences from the transcriptome and genome were concatenated into one dataset for further analysis and BUSCO annotation output was used to remove redundant sequences and assess completeness. To obtain transcript expression, each tissue library was individually mapped to the constructed transcriptome using Bowtie2, and SAMtools Idxstats was used to quantify mapped reads (Langmead & Salzberg, 2012).

1.5.4 Positive Selection In Transcriptome

Open reading frames were predicted from all input transcriptomes using TransDecoder with default parameters (Hass, 2022). All coding sequences were concatenated into one dataset and Orthofinder was executed with default parameters to cluster orthologous gene groups (Emms & Kelly, 2015) Sequences in each gene family were annotated using eggNOG with Diamond mode enabled (Huerta-Cepas et al., 2017). Annotations were also used to split gene families into paralogous groups using custom scripts. For positive selection, only groups that retained at least one sequence from each taxon were retained for analysis. Transcripts were discarded if their length was more than 100 amino acids shorter or longer than the average length of the gene, and only genes with Pal2Nal alignments longer than 20 amino acids were kept, excluding trees with insufficient branch lengths for analysis. ABSREL, which identifies episodic selection in individual branches, was used to analyze each group of orthologous genes (Smith et al.,

2015). Mafft v7.475 was used for protein alignment, FastTree 2.1.10 for tree construction, and Pal2Nal v14 with the –nogap option to provide gap-free codon-based nucleotide alignments (Katoh & Standley, 2013; Price et al., 2009; Suyama et al., 2006). Output JSON files were parsed with custom scripts using the JsonLite package (Ooms, 2014). To examine potential effects of substitutions in genes of interest under selection, chemical properties were compared between protein sequences from *P. pectinata* and *L. erinacea* using Expasy ProtScale (Gasteiger et al., 2005). The Kyte & Doolittle (1982) scale was used for hydrophobicity and the Zimmerman scale for polarity. After manual gap correction, scale values at each residue for *P. pectinata* were subtracted from *L. erinacea* values and the change plotted by residue. Domains of each protein were obtained from Interpro Scan. Analysis of PSGs of interest using DAMBE found no significant substitution saturation in any species from any alignment, indicating that the species are not too diverged to obtain meaningful positive selection results (Xia, 2018).

For PCA analysis, grand average values of hydropathicity (GRAVY) for each *P. pectinata* and *L. erinacea* protein were obtained from Expasy ProtParam. If there were two sequences in an orthogroup for *L. erinacea*, the closest aligning sequence was selected for each *P. pectinata* gene from a multiple sequence alignment, and the difference was taken between the values. Expression of homologous genes under selection in *P. pectinata* was compared to similar tissues from *S. retifer* by aligning publicly available single-end RNAseq libraries from brain, liver, kidney, ovary, and skin to the longest *S. retifer* homolog sequence (Table 1.4). Alignment was performed with

Hisat2 and quantified with SAMtools Idxstats (Kim et al., 2015; Li et al., 2009). Alignment statistics for each tissue type can be found in Table 1.5. Expression values for each gene were normalized to TPM and subtracted from *P. pectinata* values. Principal component analysis was performed using the factoextra R package and visualized with ggpubr (Kassambra, 2020) (Supplementary File 1).

Table 1.4

Sample	Accession code	Run
Adult 1 kidney	SAMD00098998	DRR111789
Adult 1 medulla	SAMD00098989	DRR111780
Adult 1 ovary	SAMD00098994	DRR111785
Adult 1 liver	SAMD00098997	DRR111788
Adult 3 skin	SAMD00099043	DRR111834

Table 1.4 *Summary of Datasets Used From Chain Catshark.*

NCBI sample names, accession codes, and run identifiers for the chain catshark, *Scyliorhinus retifer,* expression data used in PCA analysis.

Table 1.5

Sample	Total reads	Unaligned	Aligned once	Multimapping	Overall rate
A1 Kidney	10208584	99.41%	0.58%	0.01%	0.59%
A1 Medulla	8936204	99.32%	0.67%	0.01%	0.68%
A1 Ovary	8892600	99.30%	0.69%	0.01%	0.70%
A1 Liver	7162306	99.61%	0.38%	0.01%	0.39%
A3 Skin	9125803	99.35%	0.63%	0.02%	0.65%

Table 1.5 *Alignment statistics From Chain Catshark Data used for PCA analysis*.

Number and percent of reads mapped to genes orthologous to the smalltooth sawfish, *Pristis pectinata,* PSGs per sample for the chain catshark, *Scyliorhinus*

1.5.5 Conservation of Changes in Sawfishes

To determine whether signals of positive selection in genes of interest for *P. pectinata* were also present in other sawfish species, nucleotide changes in the three other *Pristis* sawfishes were examined for *HoxA5* and *Ccdc103*. Primers were designed to amplify divergent homologous functional domains between *P. pectinata* and *L. erinacea*. Primer sequences used were: *HoxA5* forward: 5’GACTTATGTGCAGTTTTCGCATCCA 3’*; HoxA5* reverse: 3’AACTACCTCCTCAAATTC 5’; *Ccdc103* forward: 5’CTGCTGCTCAGGAAATCCAC 3’; *Ccdc103* reverse: 3’AGCGGAGTTTAGCCGTGACTG 5’.

DNA samples from *P. clavata*, *P. zijsron*, and *P. pristis* were mixed with Phire Hot Start II DNA polymerase (ThermoFisher), dNTPs, deionized water, and primers for either *Ccdc103* or *HoxA5*, followed by PCR amplification using a Mastercycler Pro and electrophoresis apparatus. The reactions were amplified for 35 cycles at 98°C for 30s/5s, 53°C/58.5°C for 15s, and 72°C/72°C for 1min/1 min for *HoxA5* and *Ccdc103* respectively. The agarose gel bands were purified with a GeneJET Plasmid MiniPrep (ThermoFisher) and sent to Eurofins Genomics for sequencing. ApE (RRID:SCR_014266) was used to translate DNA sequences into the appropriate reading frame, and Clustal Omega was used to align sequences to *P. pectinata* and *L. erinacea* (Sievers et al., 2011). BLAT from the UCSC genome browser and the *P. pectinata* genome were used to verify amplification of the correct target sequence. Alignments were manually examined between species to identify conserved non-synonymous changes among sawfishes that were not found in *L. erinacea*.

1.5.6 Electrosensory Genes

Full coding sequences for electrosensing genes were downloaded from NCBI and used for manual positive selection analysis. Analysis included the VGCC (*Cacnα1D*) and several β-subunits, potassium-activated (BK) channel α subunit, and several Shaker (Kv) channels which have been implicated as major ion channels involved in electroreception in sharks and rays (Bellono, Leitch, & Julius, 2017). Transcripts for *P. pectinata* were identified with BLASTP using *L. erinacea* sequences from NCBI (acc. AJP74816.1, KY355736.1) as query and a protein database was constructed from the transcriptome by

Hmmer2GO (Altschul et al., 1990; Staton, 2014). BKβ and Shaker-type channels were identified using BLAST and confirmed using Interpro Scan(Jones et al., 2014). Multiple sequence alignments were performed using ClustalW with default parameters and 1000 bootstraps and visualized using the GGMsa package in R (Zhou et al., 2022). The phylogenetic tree was visualized with the GGTree package(Yu et al., 2017). GenBank accessions for the elasmobranch sequences most similar to the uncharacterized BK*β* subunit were: XP_041.47886.1 , XP_032892500.1, GCB66272.1, XP_038668782.1, and XP_043564107.1 (accessed 9/26/2022).

1.6 Data Availability Statement

Raw sequence data used to assemble the transcriptome can be found under NCBI accession PRJNA864825.

1.7 Author Contributions

NP, KF, and AF administered the project. GN contributed the high coverage reference genome assembly. AF, TJ, and NP designed and executed experiments. TJ and AF analyzed and interpreted data. CVE performed PCRs for *Ccdc103* and *HoxA5*. GP acquired and maintained the endangered species collection permit and collected the tissue samples from *Pristis pectinata*. TJ, AF, NP, GP, and GN wrote the manuscript and/or contributed to final edits.

1.8 Acknowledgments

Funding for this project was provided by NOAA Awards NA16NMF4720062 (field work) and NA18NMF0080237 (laboratory processing), start-up funds from the University of Florida to GN for the long read sequencing and assembly of the reference genome. The authors acknowledge HPC at The University of Southern Mississippi supported by the National Science Foundation under the Major Research Instrumentation (MRI) program via Grant #ACI 1626217. Thanks to David Morgan, Jeff Whitty, and the Western Australia Department of Fisheries for providing tissue samples of *P. clavata, P. pristis, and P. zijsron*.

CHAPTER II – Autosomal Markers In Pristis Pectinata

2.1 Introduction

Recovery efforts for the smalltooth sawfish have been underway since 2009 with the implementation of the Smalltooth Sawfish Recovery Plan, whose ultimate goal is the removal of the species from the List of Endangered and Threatened Wildlife (NMFS, 2009). While the U.S. population may have begun to stabilize in recent years, many criteria in the recovery plan for de-listing have not yet been met and warrant continued action (Wiley & Brame, 2018). One such criterium is obtaining genetic data for smalltooth sawfish, which would allow examination of relationships between sawfish in U.S. waters and those in nearby countries, as well as fine-scale population structure, sex-biased gene flow, and genetic diversity in the population.

Conservation management policy for endangered species relies on knowledge of population structure and shared evolutionary demography for delineation of evolutionarily significant management units. While the last remaining abundant population of smalltooth sawfish is currently found in southwest Florida, individuals have been observed in bordering countries such as the Bahamas, Cuba, and Mexico (Bonfil et al., 2018; Guttridge et al., 2015). Whether gene flow occurs between sawfish in these nearby countries and the U.S. population segment is currently unknown, but if so, would require reconsideration of management strategy in the U.S. population, as potential increases or decreases in one segment would have impacts on others.

Populations are structured by gene flow, or dispersal, which in many species of elasmobranchs can occur via philopatry, or the tendency of individuals to return to the same nursery area, breeding ground, or natal habitat. In *P. pectinata* females preferentially use nursery habitats or breeding grounds, whereas males are more transient and are the primary source of gene flow between regions (Feldheim et al., 2017a; Smith et al., 2021). Accordingly, understanding male gene flow is critical to managing the recovery of *Pristis pectinata*; however, to date no molecular tools have been developed to identify male markers in the absence of physical data for this species.

In addition to genetic information from current populations, past demography, including genetic diversity and effective population size, will be important for assessing the recovery of *P. pectinata.* Genetic diversity within and between populations is commonly measured by several types of markers, including microsatellites, allozymes, or single nucleotide polymorphisms (SNPs), with each tool having specificity to answer different questions. For example, microsatellites are cost-effective and can detect similar levels of genetic differentiation as single nucleotide polymorphism (SNP)-based markers, but are less common in the genome, experience variable rates of evolution, and are less able to detect very low levels of differentiation or fine population structure. SNP-based assays can be costly to develop for new species, and although individually less informative, are extremely common and offer the highest precision and statistical power to detect very low levels of genetic differentiation (Schaid et al., 2004). In addition, they can amplify shorter fragments of DNA than microsatellites, allowing genotyping of historical samples which can often be degraded, or environmental (eDNA). Baseline

estimates are critical to assess loss of diversity such as could occur in the event of a population bottleneck.

The U.S. population of smalltooth sawfish experienced a severe population decline during the mid-1900s when estimated abundance reached its lowest due to overfishing. Despite this, work using microsatellite markers found that *P. pectinata* retained high genetic diversity at all 8 loci tested and that there was no evidence of significant inbreeding (Chapman et al., 2011). However, the conclusions of this study are limited by the selectively neutral nature of microsatellites, which do not represent variation over coding regions and evolve at different rates than other noncoding regions. Furthermore, diversity in the contemporary population has not been compared to that of the pre-bottleneck population. The depth offered by high- throughput sequencing technologies allows the identification of large numbers of autosomal (present in males and females) and sex-specific variants over an entire genome, providing high statistical power even at low levels of genetic differentiation and a small number of loci (Morin et al., 2009). To address the lack of genetic data available for *P. pectinata*, we used high-throughput sequencing techniques to assemble a male smalltooth sawfish genome. This data will benefit the development of genetic assays which can be used in future population studies in *P. pectinata*. In addition, it will provide a foundation for future determination of male-specific variants in *Pristis pectinata*.

2.2 Methods

To obtain genomic DNA, blood samples were opportunistically collected from an adult male smalltooth sawfish (hereafter referred to as 881STL) (ESA Permit No. 21043).

DNA was extracted from the 881STL male using Gentra Puregene (Qiagen) Blood Kit with a MegaLong fragment separation kit using manufacturer protocols. DNA was suspended in a 50 μL elution buffer and quantified with a Qubit (ThermoFisher Waltham, MA). A Nanodrop (Thermofisher) was used to calculate 260/280 and 260/230 ratios to ensure sufficient DNA concentration. Eight DNA extracts were processed at the University of California Davis DNA Technologies Core and one sample was chosen for 10x sequencing based on fragment size following pulse-field electrophoresis. A library was constructed using the 10x Genomics Chromium Genome Reagent Kit (v2 Chemistry) and 10x Chromium system, and sequencing was performed on an Illumina NovaSeq 6000 machine at the University of California Davis DNA Technologies Core. Genome assembly was performed by SuperNova 10x Genomics software. Completeness of the assembly was assessed with Benchmarking Universal Single Copy Orthologs (BUSCO) software. Autosomal variants were called with GATK and phasing was performed with LongRanger software (10X) using the whole genome sequencing (--wgs) and --freebayes option for haplotype mode SNP-calling (McKenna et al., 2010).

2.3 Results and Discussion

The 881STL genome assembly had an N50 length of 23.71Kb and BUSCO analysis found 38.5% complete and 37.3% fragmented BUSCOs (Fig. 2.1B). Indels were called on 324 contigs and the average indel length was 21bp. In the longest 1,000 contigs (>100,000 bp) an average of 3,243 SNPs were called per contig. In the longest contig which was 356,205 bp in length, 9,616 SNPs were called (Fig. 2.1D).

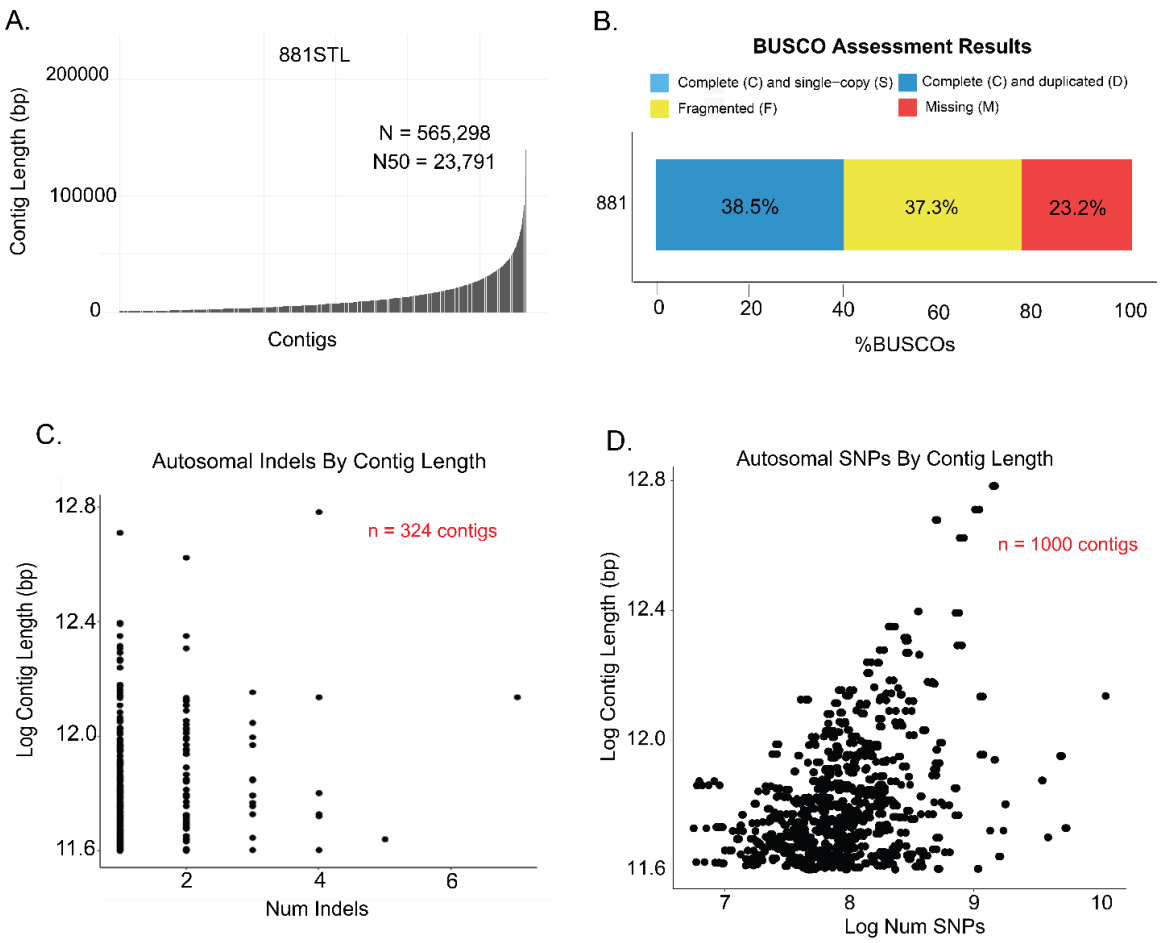


Figure 2.1 *Genome Statistics and Variant Caller Results*

A. Contig length distribution of 881STL genome. The genome contained 563,941 contigs and had an N50 value of 23,791 bp. **B.** Benchmarking Universal Single Copy Ortholog results showing the number of complete and single copy orthologs versus complete and duplicated, fragmented, and missing orthologs.

The data generated here as well as the large number of autosomal variants identified will be useful for many future applications in sawfish conservation, including assessment of fine-scale population structure, and genetic diversity not only in *Pristis pectinata* but also in other species of sawfish. Future studies could use the male genome

generated to identify male-specific variants and develop assays to sex sawfish when physical data is not available, such as in preserved rostra. Polymorphic genetic markers will be critical for assessing the capacity for adaptation to changing environmental conditions at a population level. In addition, genetic markers will be useful to determine whether low-level genetic differentiation exists in the current population or develops as *Pristis pectinata* recovers and returns to historically occupied habitats. Future work will be focused on assessing and refining SNP loci for suitability in genetic assays through further genome assembly refinement and identifying male specific markers. Once genetic tools such as autosomal and male marker assays are successfully developed for *Pristis pectinata*, they can be tailored for use in other sawfishes and other elasmobranchs.

CHAPTER III – Annotation of Small Noncoding RNAs in Non-Model Organisms

3.1 Background

3.1.1 RNA Interference and Small RNA Biogenesis

RNA interference (RNAi) is a collection of pathways that are found in most branches of the eukaryotic tree and which serve to regulate gene expression and suppress the activity of viral sequences and mobile elements, aiding in the maintenance of genome integrity and multicellular development (Carthew, 2006). The three primary pathways that participate in RNAi are micro-RNAs (miRNAs), short interfering (siRNAs) and piwi-interacting RNAs (piRNAs). Micro-RNAs and siRNAs are generated through cleavage of a double stranded RNA by a Dicer or Dicer-like ribonuclease enzyme, producing sequences which are 21-22nt and 21-24nt in length, respectively, and contain 5’ dinucleotide overhangs. These sequences bind complementary RNA transcripts, repressing translation of the gene, or in the case of siRNAs, viral or hairpin RNAs. In contrast, piRNAs are cleaved from a long single-stranded transcript into fragments by Zucchini, resulting in sequences of 26-29nt that either target complementary transcripts for degradation directly or enter the ping pong cycle. In the ping pong cycle, piRNA sequences brought by Ago3 or Piwi/Aub bind either target transposon sequences or other piRNAs, silencing the transposons and generating new piRNA sequences (le Thomas et al., 2014).

3.1.2 Annotation of Small RNAs

The advances of high-throughput sequencing have allowed unprecedented *ab initio* exploration of genomic data. Despite this, annotation of small RNAs, particularly in non-model or newly sequenced organisms, remains a non-trivial task. Due to the short length and sometimes repetitive nature of mature sRNAs, many map to the genome in multiple places, making determination of genomic origin and quantification of expression challenging. Other obstacles arise from the biological differences in the mechanisms of small RNA pathways between taxonomic groups. Prediction from genome sequence alone is computationally prohibitive and often insufficient to identify classes of small RNAs. Due to these challenges, a dizzying array of tools is available for the purpose of annotating, quantifying, or characterizing the many classes of small RNAs (Barturen et al., 2014; Friedländer et al., 2012; Gebert et al., 2017; Geles et al., 2021; Han et al., 2015; Huang et al., 2007; D. Li et al., 2016; Rosenkranz & Zischler, 2012; Wang et al., 2014; Wang et al., 2005; Wu et al., 2011). These tools use varying methods to predict or identify ncRNAs – some, such as miRNA annotation tools, utilize structure prediction or homology to known sequences, while some rely on machine learning algorithms (An et al., 2013; Mathelier & Carbone, 2010). The highest sensitivity and specificity are seen in those that consider multiple lines of evidence (Gomes et al., 2013).

3.2 MiSiPi: A Bioconductor For The Holistic Characterization of Endogenous Small RNA Pathways

Taiya M. Jarva[1], Alex S. Flynt[1]

[1]School of Biological, Environmental and Earth Sciences. The University of Southern Mississippi, 118 College Drive, Hattiesburg, MS, 39401

This manuscript has not yet been deposited in a pre-print server or submitted to a journal.

3.3 Abstract

RNA interference (RNAi) is mediated by small (20-30 nucleotide) RNAs that are produced by complex processing pathways. In animals, three main classes are recognized: including microRNAs (miRNAs), small-interfering RNAs (siRNAs) and piwi-interacting RNAs (piRNAs). Understanding of small RNA pathways has relied heavily on genetic models where key enzymatic events have been identified that lead to stereotypical positioning of small RNAs relative to precursor transcripts. Increasingly there is interest in using RNAi in non-model systems due to ease of generating synthetic small RNAs precursors or research and biotechnology. Unfortunately, small RNAs are often rapidly evolving, requiring investigation of a species endogenous small RNAs prior to deploying an RNAi approach. This can be accomplished through small non-coding RNA sequencing followed by applying various computational tools; however, the complexity and separately maintained packages lead to significant challenges for annotating small RNAs. To address this need, we developed a simple and efficient R package which can be used to characterize pre-selected loci and provide publication-ready plots and statistics, and aiding researchers in selecting the appropriate RNAi tactic

for their target species. Further, we provide new tools for assessing loci, bringing parity of tools for each small RNA class, and machine learning approaches to assist classification of loci.

3.4 Introduction

In the past two decades, RNA interference (RNAi) has gained popularity as a facile tool for genetic manipulation and for biotechnology such as in the control of pests as a replacement or supplement to traditional pesticides. Its specificity for repressing target gene expression has been demonstrated in several species and thus RNAi has emerged as a method for studying gene function in model species (Baum et al., 2007; Kamath, 2003; Mao et al., 2007; Zhu et al., 2011). Despite nearly twenty years of advancements in the applications of RNAi, identifying appropriate small RNA pathways and loci to target remains challenging (Joga et al., 2016; Willow et al., 2021). While there are a variety of tools available for small RNA annotation or quantification, many are limited to specific classes of RNAs, rely on sequence similarity to previously annotated noncoding RNAs, or require significant experience with bioinformatics systems and software. Furthermore, to our knowledge there are no existing tools for annotation of siRNAs or cis-NATs.

We designed the R package ‘MiSiPi’ to facilitate researchers with basic knowledge of computational tools and RNAi biology to easily and efficiently characterize various types of small RNA loci in hemipteran pest insects. The package is available to install via Bioconductor and requires simple input. Nearly all required

software components are contained in the package, apart from a precompiled RNAfold executable from the ViennaRNA package, which should be downloaded separately (Gruber et al., 2008). The required input is a binary alignment mapping (BAM) file of small RNA sequencing reads aligned to a reference genome and a 3-column BED file of regions of interest, e.g., regions of high small RNA expression. For the hairpin and miRNA modules, a reference genome is also required as a function argument. The package utilizes several Bioconductor genomics packages, including RSamtools, Biostrings, and GenomicRanges for fast manipulation of sequence data (Lawrence et al., 2013; Morgan et al., 2022; Pages et al., 2022).

3.5 Methods

3.5.1 Data Pre-processing

Small RNA sequencing data should be prepared for the MiSiPi package with standard computational methods (for example, adaptor trimming and quality checking), followed by aligning small RNA reads to a reference genome using Bowtie with appropriate arguments for small RNAs (Langmead & Salzberg, 2012). For example, supplying Bowtie with the options -a -m 200 -v0 --best --strata will return unique reads that contain no mismatches, allowing multi-mapping reads. Toggling the -m option will change the number of multi-mapping alignments allowed. The method of identifying regions of interest will be specific to the needs of the user and the depth of the sequencing data, however, the general processing should be done using SAMTools and BEDTools to identify regions of high small RNA expression using mapped sequence

reads and the reference genome (Fig. 3.1A, Supplementary Methods) (Li et al., 2009; Quinlan & Hall, 2010).

Supplementary Methods

Identifying small RNA regions of interest

miRNAs & siRNAs: get reads of only miRNA and siRNA length

```
awk 'BEGIN {OFS = "\n"} {header = $0; getline seq; getline qheader ; getline qseq ; if (length(seq) >= 19 && length(seq) <= 23) {print header, seq, qheader, qseq}}' \
< trimmed.fq > small.fastq
```

piRNAs: get reads of piRNA length

```
awk 'BEGIN {OFS = "\n"} {header = $0; getline seq; getline qheader ; getline qseq ; if(length(seq) >= 23 && length(seq) <= 30) {print header, seq, qheader, qseq}}' \
< trimmed.fq > large.fastq
```

Realign reads to genome

```
bowtie -p 10 -a -m100 --best --strata --no-unal genome.fna small.fastq -S | samtools view -@ 10 -q 10 -b |samtools sort -@ 10 -m 6G > small.bam
samtools index small.bam
```

```
bowtie -p 20 -a -m100 --no-unal genome.fna large.fastq -S | samtools view -@ 10 -q 10 -b | samtools sort -@ 10 -m 6G > large.bam
```

samtools index large.bam

Calculate coverage of all reads over genome

bedtools genomecov -bg -ibam aligned.bam | awk '$4 > 100' > HE.tmp.bedgraph

Get large regions of high expression

bedtools merge -d 500 -i HE.tmp.bedgraph > HE.tmp.merge.bed

awk '{n=$2; x=$3; print $1"\t"$2"\t"$3"\t"x-n}' < HE.tmp.merge.bed |

awk '$4 > 40' > HE.all.bed #all highly expressed

Get potential regions of high miRNA/siRNA RNA expresssion

bedtools multicov -bams small.bam aligned.bam -bed HE.all.bed |

awk '{n=$5; x=$6; print $1"\t"$2"\t"$3"\t"n/x}' |

awk '$4 > 0.5' > HE.small.bed # only high expressed 21-23 nt

Get potential regions of high piRNA expression

bedtools multicov -bams large.bam SRR12982892.bam -bed HE.all.bed |

awk '{n=$5; x=$6; print $1"\t"$2"\t"$3"\t"n/x}' | awk '$4 > 0.5' > HE.large.bed #only high expressed 23-30 nt

The package contains modules for each class of small RNAs which can be run separately or all at once for holistic locus characterization. With some exceptions, read data is initially processed similarly in all modules. The package RSamtools is used in accessory functions (OpenBamFile, makeBamDF, getChrMinus, and getChrPlus) to read the BAM file and create separate data frames containing reads from either the plus or minus strand for the region of interest, filtering out reads which are not of small RNA size (18-32nt), and retaining the minimal amount of information that is needed for downstream processing (Morgan et al., 2022). Reads are subsequently processed according to module- specific functions which leverage characteristics of the biogenesis of each type of small RNA (Fig. 3.1B).

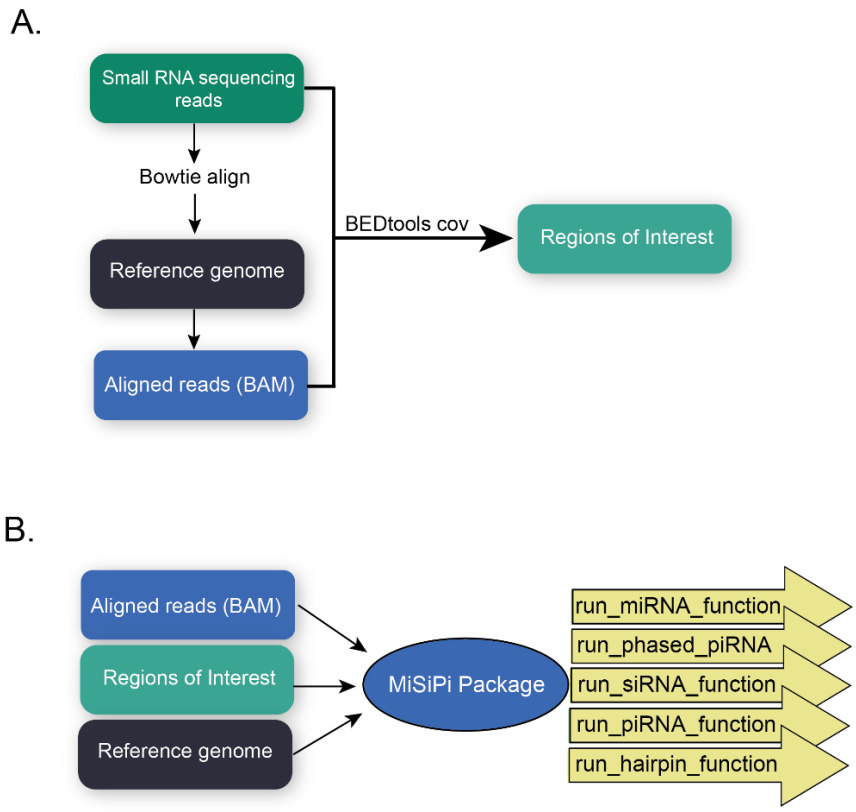


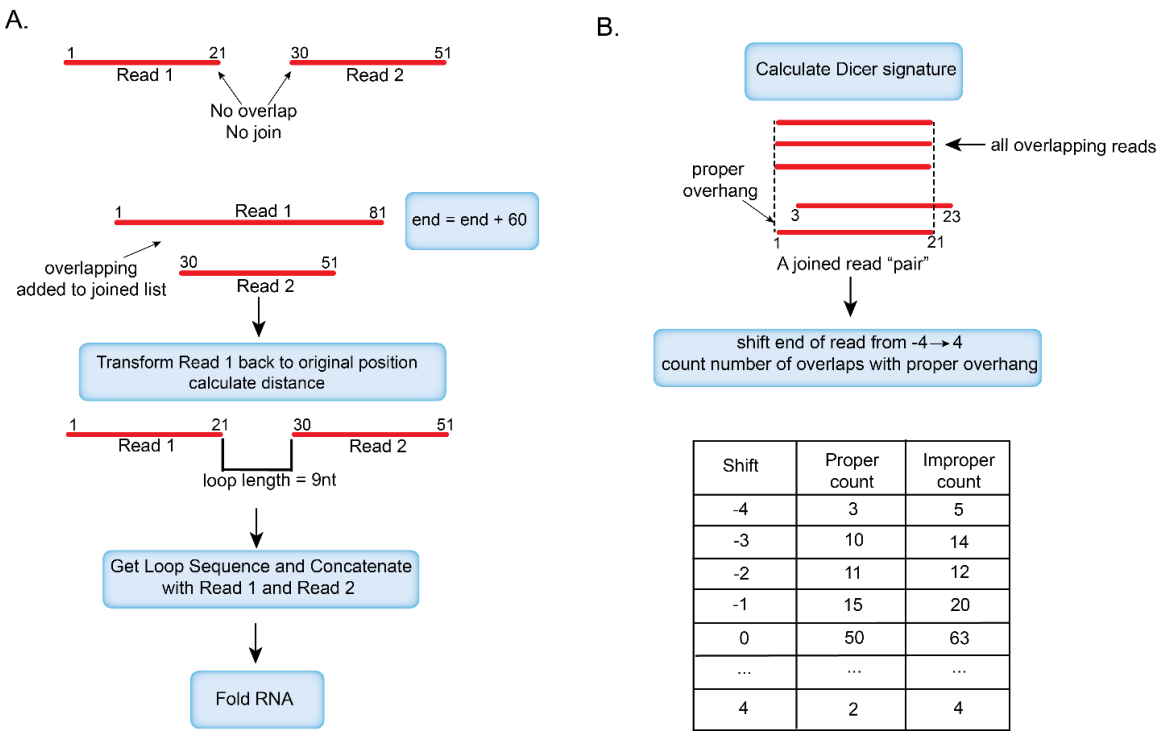
Figure 3.1 *Example MiSiPi Workflow.*

A. Raw sequence reads are aligned to a reference genome with Bowtie, giving a Binary Alignment Mapping file of aligned reads. The raw reads and the BAM file are used with BEDTools programs to identify regions of high small RNA expression. **B.** The BAM file, reference genome, and file containing regions of interest are provided as input to MiSiPi. The basic function commands for each module are shown.

3.5.2 Short Hairpin/microRNAs

miRNAs are ~22nt length sequences generated from longer, single-strand sequences containing regions of inverted repeats which causes the sequence to fold onto itself and create a hairpin (Lee, 2002). Before being exported from the nucleus, the loop sequence is cleaved by Drosha, resulting in a short segment of double-stranded RNA. One strand of the RNA becomes the ‘guide’ strand and goes on to participate in

inhibition of mRNA translation via the RISC complex, while the ‘passenger’ strand is preferentially degraded. This biogenesis pathway results in a characteristic expression pattern which can be identified in sequencing data. Because miRNAs are transcribed from a single strand, the miRNA module uses an algorithm to identify reads mapping within 60nt of each other on the same strand. During initial processing, a read data frame is made, along with a copy of the data frame in which the 3’ ends of the reads are transformed by adding 60 nucleotides. The findOverlaps() function from the GenomicRanges package is then used to find artificially “overlapping” reads. The ends are then transformed back to their original positions and the genomic sequence between read pairs (the loop sequence) is retrieved (Fig. 3.2A). Results which have greater than 3 reads aligning to the loop sequence are excluded from further analysis. Redundant results are merged using the IRanges function reduce(), and the remaining sequences are folded using RNAfold from the ViennaRNA package. The output plots include read density over the region, a secondary structure plot styled after miRBase hairpin plots, and a probability score of the number of overlaps which contain a Dicer signature. The probability is calculated by counting the number of reads that contain a proper 2-nt overhang, followed by shifting the read start position from a range of -4 to 4 and again counting the number of reads with proper overhangs. If most reads are indeed Dicer-generated, the highest z-score will be found at a shift of 0 (Fig. 3.2B). The output plots for the miRNA module include read depth over the locus (Fig. 3.3A), a secondary structure plot colored by approximate expression which can be used to identify the guide strand (Fig. 3.3B), read size distribution over the locus (Fig. 3.3C), and a Dicer probability plot Fig. 3.3D.



Shift	Proper count	Improper count
-4	3	5
-3	10	14
-2	11	12
-1	15	20
0	50	63
...	...	...
4	2	4

Figure 3.2 *Overview of miRNA module.*

A. Two identical data frames of reads from a single strand are made. The end positions of one data frame are incremented by 60nt to identify potential nearby read pairs by using GenomicRanges function findOverlaps. The end positions are transformed to the original position and the sequence between read pairs is extracted from the genome. The concatenated sequence is then folded by RNAfold. **B.** The dicer probability is calculated by counting the number of reads overlapping each individual read that have a proper 2-nt overhang. The position of the read is shifted downstream or upstream by a range of -4 to 4, and the number of overlapping reads with a 2-nt overhang are re-counted. The z-score is calculated from the ratios of "proper" to "improper" overhangs.

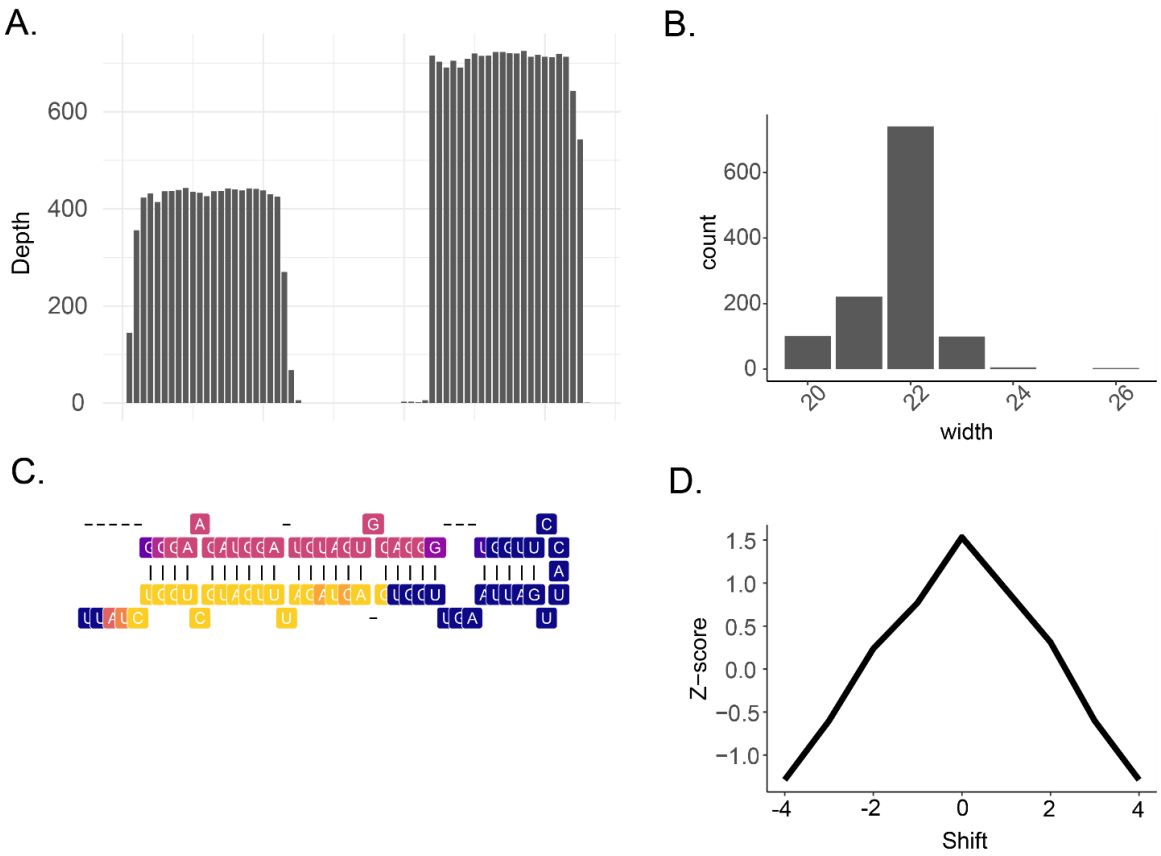


Figure 3.3 *Example miRNA output plots for Drosophila melanogaster mir-996.*

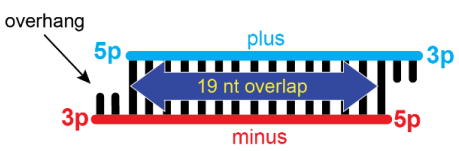
A. Read density over the locus. MiRNAs typically have high read depth on the guide strand and low depth on the passenger strand. **B.** Secondary structure plot colored by approximate expression of each nucleotide. **C.** Read size distribution plot over the locus. **D.** Dicer probability score. Properly overhanging reads are counted over a range of shifted positions to calculate a probability score.

3.5.3 Short Interfering RNAs Derived From Bi-directional Transcription

Short interfering RNAs (siRNAs) are short sequences (21-24nt) generated from double stranded RNA which bear characteristic overhangs of 2 nucleotides on their 3' ends due to processing by Dicer, and which generally overlap by 19 nucleotides (Tijsterman et al., 2002). The siRNA module finds overlapping reads from both strands and uses the functions get_si_overlaps() and proper_overlap() to search for "proper"

overlaps (pairs which have the 2nt overhang) over a range of overlap values (15-32), resulting in a matrix of counts by read size (Fig. 3.4A-C). The program also plots the read size distribution for the region of interest (Fig. 3.4D), and the probability that double stranded sequences were processed by Dicer (Fig. 3.4E).

A.



B.

Count reads with proper overlap

Query size	Target size	Overlap
15	15-32	13
16	15-32	14
17	15-32	15-16
18	15-32	15-17
...	...	...
...	...	...
...	...	...
...	...	...
32	15-32	15-32

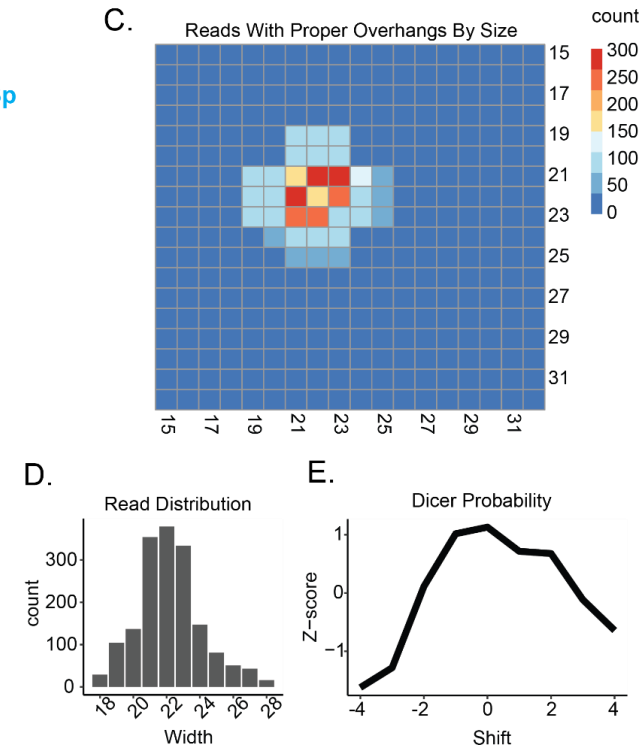


Figure 3.4 *siRNA module overview.*

Plot output from locus Chr3R: 27880965-27882867, a cis-NAT RNA located within Drosophila melanogaster CG34353. **A**. Example of siRNA overlap demonstrating 2nt 3' overhang and length of overlap between plus and minus strand reads. **B**. Table of query sizes, target sizes, and overlap sizes that are calculated by the siRNA module. **C**. Example heatmap plot from siRNA module showing counts of reads with proper overhang by read size. **D.** Read size distribution over locus. **E.** Dicer probability plot.

3.5.4 Short Interfering RNAs Derived From Long Hairpin Pre-cursors

Long hairpin RNAs are sequences containing inverted repeats, causing them to fold and have a double stranded secondary structure. Due to this double-stranded nature, they are processed by Dicer-2 and possess the 2-nt overhangs characteristic of siRNAs. Additionally, because they are cleaved in a processive manner approximately 21-22 nucleotides apart, they often present a phasing signature (Okamura et al., 2008). The hairpin module uses RNAfold to predict the secondary structure and calculates read depth over the region as well as the proportion of the reads which exhibit a Dicer signature. The secondary structure is predicted from the full genomic sequence by ViennaRNA and the arc diagram plotted with the R4RNA package (Lai et al., 2012). One of the outputs of R4RNA is a table of paired base positions where a nucleotide ‘i’ is paired with a nucleotide ‘j’. The intervals are collapsed into sequential ‘i’ ranges and ‘j’ ranges, and findOverlaps is used to create a set of reads which overlap with the ‘i’ ranges and a set of reads which overlap with the ‘j’ ranges. These two data sets are then compared with each other to find overlapping pairs. From the overlapping pairs, the dicer signature is calculated as shown in Fig. 3.2B. The hairpin module output plots include an arc diagram showing paired bases (Fig. 3.5A), read density. plotted by size (Fig. 3.5B), and the Dicer and phasing probability scores (Fig. 3.5C&D).

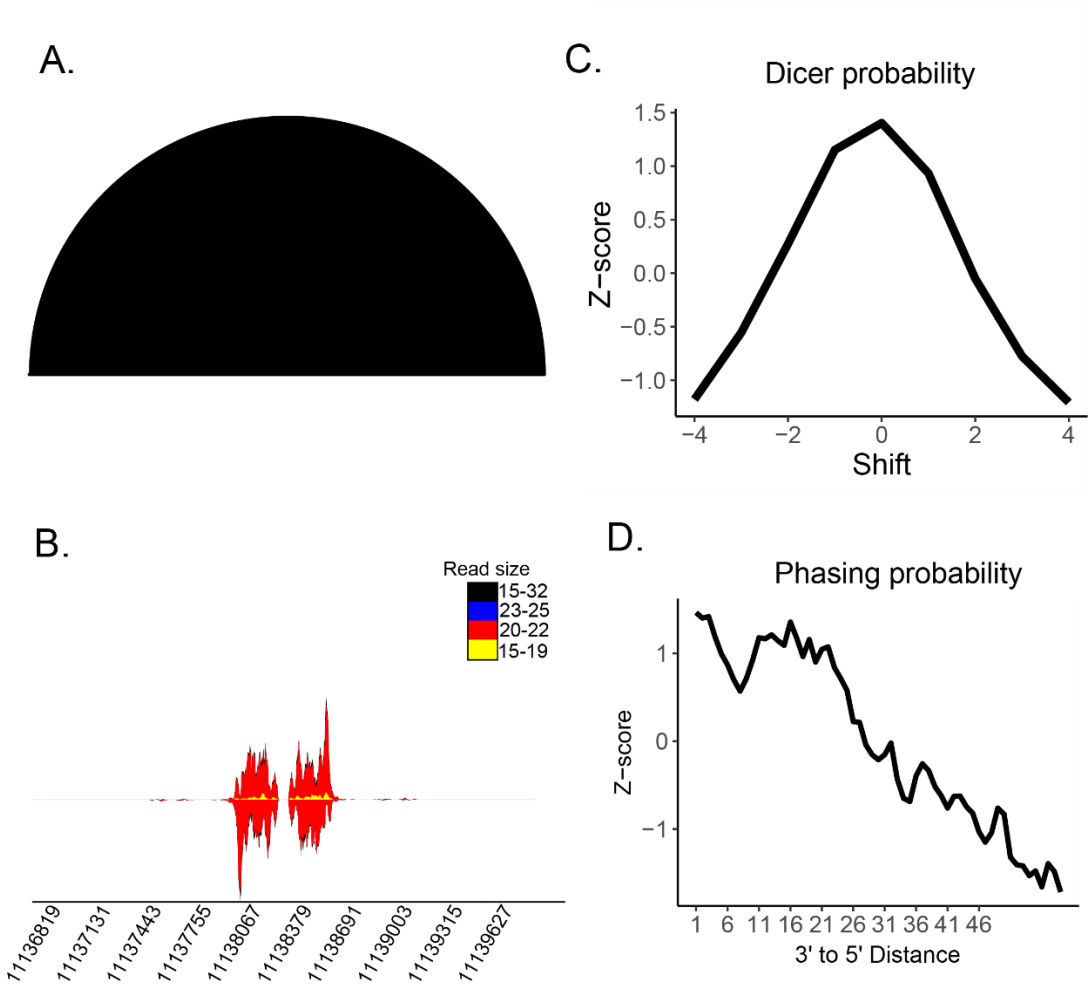


Figure 3.5 *Output plots from long hairpin module.*

A. Arc plot made with the R4RNA package showing paired bases over the Tmy locus in *Drosophila simulans*. **B.** Read density by size over the locus. **C.** Dicer probability plot. **D.** Phasing probability plot.

3.5.5 Piwi-Interacting RNAs

Biogenesis of piwi-interacting (piRNAs) comprises two pathways, ping- pong and phasing. In response to transcripts from transposable elements or during gametogenesis, pre-cursor piRNA sequences are transcribed and cleaved into short (26-29nt) sequences in a process known as phasing (Brennecke et al., 2007). These are bound by Aubergine

(Aub) or Argonaute (Ago) proteins by 10 nucleotides and guided to complementary antisense target transposons or piRNA sequences for cleavage, creating more piRNAs which go on to target further sequences (Gunawardane et al., 2007). Sequences generated via the ping- pong cycle overlap the 5' end of reads from the opposite strand by precisely 10 nucleotides (Fig. 3.6A). Thus, the module for piRNAs uses the GenomicRanges function findOverlaps to identify overlapping reads between the plus and minus strand and calculates the length of the overlap at a range of query sizes, target sizes, and overlap lengths (Fig. 3.6B). The overlap probability plot shows the probability of finding an overlap of 4-30nt and the heatmap shows the sizes of reads which overlap by 10 nucleotides (Fig. 3.6C&E).

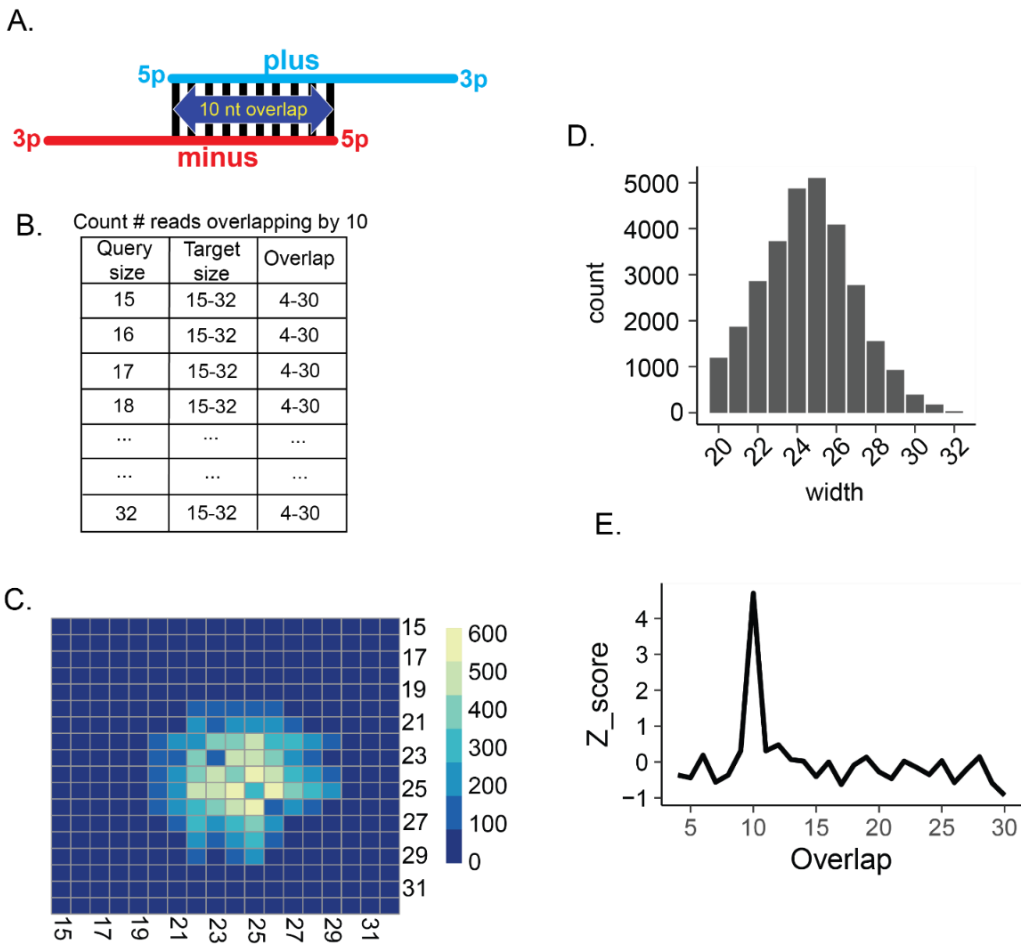


Figure 3.6 *Overview of piRNA module.*

A. Example overlapping piRNA pair. **B.** Range of read sizes and overlap lengths tested by piRNA module. **C.** Probability plot for each overlap size. **D.** Read size distribution of example piRNA locus. **E.** Heatmap of reads overlapping by 10 nucleotides by read size. Plots originate from *Drosophila melanogaster* 42AB locus using a public dataset (Table 3.1).

Table 3.1

Figure	Species	NCBI Accession	Tissue/Condition
Fig. 3	Drosophila melanogaster	DRR351386	Ovary/CG9925-KO
Fig. 4	Drosophila melanogaster	SRR1664731	Ovary/Ago1 IP sRNA
Fig. 5	Drosophila simulans	SRR7410589	Testis/w[XD1] & nmy[12-2-7] strains

Table 3.1 (continued)

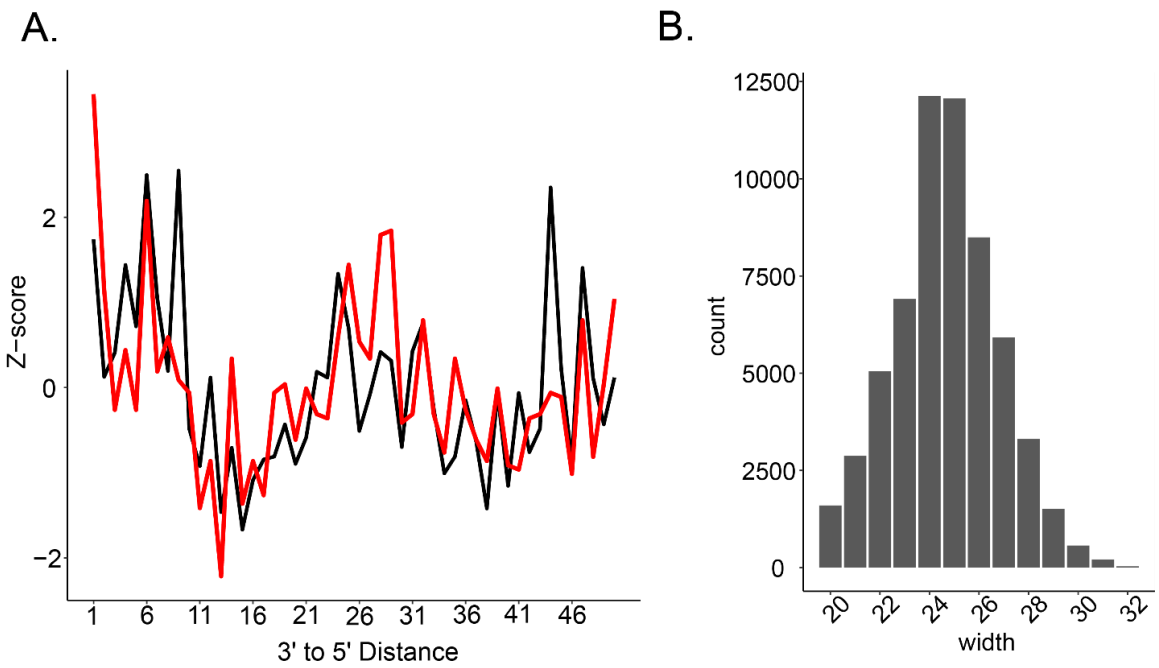
Fig. 6	Drosophila melanogaster	DRR351386	Ovary/control
Fig. 7	Drosophila melanogaster	DRR351386	Ovary/control

Table 3.1 *Summary of public datasets used.*

Species, NCBI accessions, tissue type, and conditions of datasets used to generate each figure.

3.5.6 Phased Pi-RNAs

Phased piRNAs are generated in a processive manner from a single transcript which is cleaved by Zucchini directly upstream of a uracil. This results in single-stranded reads beginning with a 5' uracil and which lie approximately 1-3 nt apart (Nishida et al., 2007; Vagin et al., 2006). The phased piRNA module uses a transformation method similar to the miRNA module to identify reads which are within 50 nucleotides, then calculates the distance between the original 3' position of one read and the 5' position of each read pair. Finally, it calculates the probability of finding a read that maps between 1 and 50 nucleotides away as a z-score. (Fig. 3.7A).



Example phased piRNA module plot outputs

A. Phasing probability in reads of all sizes (black) and reads larger than 26 nucleotides for 42AB region in *Drosophila melanogaster*. **B.** Read size distribution.

3.6 Results/Discussion

While many sRNA annotation tools are available, there is a clear lack of integrated software which can characterize all sRNA classes directly from sequencing data. The increasing interest in leveraging RNAi for therapeutics, bioengineering of crops, and biopesticides, especially in newly sequenced or un-annotated species, necessitates the development of computational tools which can facilitate the selection of appropriate RNAi strategies. Small RNA biology is well characterized in model species; however, recent studies highlight the diversity of biogenesis among and even within clades of plants, animals, and fungi, suggesting differences in susceptibility to RNAi and a potential to use atypical approaches to gene silencing (Flynt, 2021; Mondal et al.,

2021). Other studies have demonstrated the independent loss of certain small RNA pathways in some groups relative to closely- related members (Khanal et al., 2022).

MiSiPi is designed to be computationally efficient even over thousands of loci and on high-depth sequencing datasets and can be run on a laptop equipped with 8GB of RAM. The package achieves this by being conscious of R language's memory usage idiosyncrasies. For example, the package reads only one chromosome of data from the BAM file at once, extracting only reads from the region of interest, and retaining the minimal amount of data needed, e.g. the first letter of the nucleotide sequence and the start/stop positions instead of the whole sequence which reduces the amount of data that is kept in memory. In addition, it uses fast sequence manipulation packages such as GenomicRanges and Biostrings, and the most computationally complex functions are written in C++ through the Rcpp package (Eddelbuettel & François, 2011). The package modules have been validated using annotated loci in *Drosophila melanogaster* and *Drosophila simulans* and publicly available datasets (see Supplementary Table 1). Furthermore, MiSiPi is compatible with any species for which a reference genome is available. Here we have demonstrated the utility of the MiSiPi package in aiding RNAi research, particularly in validation of potential pathways or loci for potential biopesticide applications.

3.7 Acknowledgments

AF and TJ are supported by NSF 1845978, and Mississippi INBRE: funded by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P20GM103476. Alignment and processing of small RNA datasets was performed on Magnolia HPC at the University of Southern Mississippi, which is supported by the National Science Foundation under the Major Research Instrumentation (MRI) program via Grant # ACI 1626217.

REFERENCES

Abrantes, K., Barnett, A., Soetaert, M., Kyne, P., Laird, A., Squire, L., Seymour, J., Wueringer, B., Sleeman, J., & Huveneers, C. (2021). Potential of electric fields to reduce bycatch of highly threatened sawfishes. *Endangered Species Research*, *46*, 121–135. https://doi.org/10.3354/esr01146

Alexa A, & Rahnenfuhrer J. (2022). *topGO: Enrichment Analysis for Gene Ontology.* (2.48.0).

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

An, J., Lai, J., Lehman, M. L., & Nelson, C. C. (2013). miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. *Nucleic Acids Research*, *41*(2), 727–737. https://doi.org/10.1093/nar/gks1187

Antoku, S., & Mayer, B. J. (2009). Distinct roles for Crk adaptor isoforms in actin reorganization induced by extracellular signals. *Journal of Cell Science*, *122*(22), 4228–4238. https://doi.org/10.1242/jcs.054627

Babbin, B. A., Lee, W. Y., Parkos, C. A., Winfree, L. M., Akyildiz, A., Perretti, M., & Nusrat, A. (2006). Annexin I Regulates SKCO-15 Cell Invasion by Signaling through Formyl Peptide Receptors. *Journal of Biological Chemistry*, *281*(28), 19588–19599. https://doi.org/10.1074/jbc.M513025200

Balasubramanian, R., & Zhang, X. (2016). Mechanisms of FGF gradient formation during embryogenesis. *Seminars in Cell & Developmental Biology*, *53*, 94–100. https://doi.org/10.1016/j.semcdb.2015.10.004

Barturen, G., Rueda, A., Hamberg, M., Alganza, A., Lebron, R., Kotsyfakis, M., Shi, B.-J., Koppers-Lalic, D., & Hackenberg, M. (2014). sRNAbench: profiling of small RNAs and its sequence variants in single or multi-species high-throughput experiments. *Methods in Next Generation Sequencing*, *1*(1). https://doi.org/10.2478/mngs-2014-0001

Baum, J. A., Bogaert, T., Clinton, W., Heck, G. R., Feldmann, P., Ilagan, O., Johnson, S., Plaetinck, G., Munyikwa, T., Pleau, M., Vaughn, T., & Roberts, J. (2007). Control of coleopteran insect pests through RNA interference. *Nature Biotechnology*, *25*(11), 1322–1326. https://doi.org/10.1038/nbt1359

Bellono, Nicholas. Leitch, Duncan. Julius, David. (2017). Molecular Basis of Ancestral Vertebrate Electroreception. *Nature*, *543*(7645), 391–396. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5354974/

Bellono, Nicholas. Leitch, Duncan. Julius, David. (2018). Molecular tuning of electroreception in sharks and skates. *Nature*, *558*(7708), 122–126. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6101975/

Bennet, M. V. L., & Obara, S. (1986). Ionic mechanisms and pharmacology of electroreceptors. *Electroreception*, 157–181.

Bonfil, R., Ricaño-Soriano, M., Mendoza-Vargas, O., Méndez-Loeza, I., Pérez-Jiménez, J., Bolaño-Martínez, N., & Palacios-Barreto, P. (2018). Tapping into local ecological knowledge to assess the former importance and current status of sawfishes in Mexico. *Endangered Species Research*, *36*, 213–228. https://doi.org/10.3354/esr00899

Brame, A., Wiley, T., Carlson, J., Fordham, S., Grubbs, R., Osborne, J., Scharer, R., Bethea, D., & Poulakis, G. (2019). Biology, ecology, and status of the smalltooth sawfish Pristis pectinata in the USA. *Endangered Species Research*, *39*, 9–23. https://doi.org/10.3354/esr00952

Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., & Hannon, G. J. (2007). Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila. *Cell*, *128*(6), 1089–1103. https://doi.org/10.1016/j.cell.2007.01.043

Brenner, Robert. Jegla, Tim J. Wickenden, Alan. Liu, Yi. Aldritch, R. W. (2000). Cloning and Functional Characterization of Novel Large Conductance Calcium-activated Potassium Channel β Subunits, hKCNMB3 and hKCNMB4. *Journal of Biological Chemistry*, *275*, 6453–6461.

Carlson, J. K., Wiley, T., & Smith, K. (2013). *The IUCN Red List of Threatened Species*. http://www.iucnredlist.org/details/18175/0.

Carthew, R. W. (2006). Gene regulation by microRNAs. *Current Opinion in Genetics & Development*, *16*(2), 203–208. https://doi.org/10.1016/j.gde.2006.02.012

Chapman, D. D., Simpfendorfer, C. A., Wiley, T. R., Poulakis, G. R., Curtis, C., Tringali, M., Carlson, J. K., & Feldheim, K. A. (2011). Genetic Diversity Despite Population Collapse in a Critically Endangered Marine Fish: The Smalltooth Sawfish (Pristis pectinata). *Journal of Heredity*, *102*(6), 643–652. https://doi.org/10.1093/jhered/esr098

Cheng, S., Melkonian, M., Smith, S. A., Brockington, S., Archibald, J. M., Delaux, P.-M., Li, F.-W., Melkonian, B., Mavrodiev, E. v, Sun, W., Fu, Y., Yang, H., Soltis, D.

E., Graham, S. W., Soltis, P. S., Liu, X., Xu, X., & Wong, G. K.-S. (2018). 10KP: A phylodiverse genome sequencing plan. *GigaScience*, *7*(3). https://doi.org/10.1093/gigascience/giy013

Clusin, William T., Wu, Ting-Hsuan, Shi, Ling-Fang. Kao, P. N. (2019). Further Studies of Ion Channels in the Electroreceptor of the Skate Through Deep Sequencing, Cloning and Cross Species Comparisons. *Gene*, *718*. https://www.sciencedirect.com/science/article/abs/pii/S0378111919306481?via%3Dihub

Cocci, P., Mosconi, G., & Palermo, F. A. (2019). Gene expression profiles of putative biomarkers in juvenile loggerhead sea turtles (Caretta caretta) exposed to polycyclic aromatic hydrocarbons. *Environmental Pollution*, *246*, 99–106. https://doi.org/10.1016/j.envpol.2018.11.098

Cullen, J. A., Marshall, C. D., & Hala, D. (2019). Integration of multi-tissue PAH and PCB burdens with biomarker activity in three coastal shark species from the northwestern Gulf of Mexico. *Science of The Total Environment*, *650*, 1158–1172. https://doi.org/10.1016/j.scitotenv.2018.09.128

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dudgeon, C. L., Coulton, L., Bone, R., Ovenden, J. R., & Thomas, S. (2017). Switch from sexual to parthenogenetic reproduction in a zebra shark. *Scientific Reports*, *7*(1), 40537. https://doi.org/10.1038/srep40537

Dulvy, N. K., Davidson, L. N. K., Kyne, P. M., Simpfendorfer, C. A., Harrison, L. R., Carlson, J. K., & Fordham, S. v. (2016). Ghosts of the coast: global extinction risk and conservation of sawfishes. *Aquatic Conservation: Marine and Freshwater Ecosystems*, *26*(1), 134–153. https://doi.org/10.1002/aqc.2525

Dulvy, N. K., Pacoureau, N., Rigby, C. L., Pollom, R. A., Jabado, R. W., Ebert, D. A., Finucci, B., Pollock, C. M., Cheok, J., Derrick, D. H., Herman, K. B., Sherman, C. S., VanderWright, W. J., Lawson, J. M., Walls, R. H. L., Carlson, J. K., Charvet, P., Bineesh, K. K., Fernando, D., … Simpfendorfer, C. A. (2021). Overfishing drives over one-third of all sharks and rays toward a global extinction crisis. *Current Biology*, *31*(21), 4773-4787.e8. https://doi.org/10.1016/j.cub.2021.08.062

Eddelbuettel, D., & François, R. (2011). **Rcpp** : Seamless *R* and *C++* Integration. *Journal of Statistical Software*, *40*(8). https://doi.org/10.18637/jss.v040.i08

Emms, D. M., & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, *16*(1), 157. https://doi.org/10.1186/s13059-015-0721-2

Feldheim, K. A., Clews, A., Henningsen, A., Todorov, L., McDermott, C., Meyers, M., Bradley, J., Pulver, A., Anderson, E., & Marshall, A. (2017). Multiple births by a captive swellshark *Cephaloscyllium ventriosum via* facultative parthenogenesis. *Journal of Fish Biology*, *90*(3), 1047–1053. https://doi.org/10.1111/jfb.13202

Feldheim, K., Fields, A., Chapman, D., Scharer, R., & Poulakis, G. (2017a). Insights into reproduction and behavior of the smalltooth sawfish Pristis pectinata. *Endangered Species Research*, *34*, 463–471. https://doi.org/10.3354/esr00868

Feldheim, K., Fields, A., Chapman, D., Scharer, R., & Poulakis, G. (2017b). Insights into reproduction and behavior of the smalltooth sawfish Pristis pectinata. *Endangered Species Research*, *34*, 463–471. https://doi.org/10.3354/esr00868

Ferretti, F., Worm, B., Britten, G. L., Heithaus, M. R., & Lotze, H. K. (2010). Patterns and ecosystem consequences of shark declines in the ocean. *Ecology Letters*, *13*, 1055–1071. https://doi.org/10.1111/j.1461-0248.2010.01489.x

Fields, A. T., Feldheim, K. A., Poulakis, G. R., & Chapman, D. D. (2015). Facultative parthenogenesis in a critically endangered wild vertebrate. *Current Biology*, *25*(11), R446–R447. https://doi.org/10.1016/j.cub.2015.04.018

Flynt, A. S. (2021). Insecticidal <scp>RNA</scp> interference, thinking beyond long <scp>dsRNA</scp>. *Pest Management Science*, *77*(5), 2179–2187. https://doi.org/10.1002/ps.6147

Friedländer, M. R., Mackowiak, S. D., Li, N., Chen, W., & Rajewsky, N. (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Research*, *40*(1), 37–52. https://doi.org/10.1093/nar/gkr688

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., & Bairoch, A. (2005). Protein Identification and Analysis Tools on the ExPASy Server. In *The Proteomics Protocols Handbook* (pp. 571–607). Humana Press. https://doi.org/10.1385/1-59259-890-0:571

Gebert, D., Hewel, C., & Rosenkranz, D. (2017). unitas: the universal tool for annotation of small RNAs. *BMC Genomics*, *18*(1), 644. https://doi.org/10.1186/s12864-017-4031-9

Geles, K., Palumbo, D., Sellitto, A., Giurato, G., Cianflone, E., Marino, F., Torella, D., Mirici Cappa, V., Nassa, G., Tarallo, R., Weisz, A., & Rizzo, F. (2021). WIND (Workflow for pIRNAs aNd beyonD): a strategy for in-depth analysis of small RNA-seq data. *F1000Research*, *10*, 1. https://doi.org/10.12688/f1000research.27868.3

Gomes, C. P. C., Cho, J.-H., Hood, L., Franco, O. L., Pereira, R. W., & Wang, K. (2013). A Review of Computational Tools in microRNA Discovery. *Frontiers in Genetics*, *4*. https://doi.org/10.3389/fgene.2013.00081

Gonzalez-Perez, Vivian. Lingle, C. J. (2019). Regulation of BK Channels by Beta and Gamma Subunits. *Annual Review of Physiology*, *81*, 113–137. https://www.annualreviews.org/doi/10.1146/annurev-physiol-022516-034038?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub%3Dpubmed

Gotz, S., Garcia-Gomez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., Robles, M., Talon, M., Dopazo, J., & Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*, *36*(10), 3420–3435. https://doi.org/10.1093/nar/gkn176

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Graham, J., Kroetz, A. M., Poulakis, G. R., Scharer, R. M., Carlson, J. K., Lowerre-Barbieri, S. K., Morley, D., Reyier, E. A., & Grubbs, R. D. (2022). Commercial fishery bycatch risk for large juvenile and adult smalltooth sawfish (<scp> *Pristis pectinata* </scp>) in Florida waters. *Aquatic Conservation: Marine and Freshwater Ecosystems*, *32*(3), 401–416. https://doi.org/10.1002/aqc.3777

Graham, J., Kroetz, A., Poulakis, G., Scharer, R., Carlson, J., Lowerre-Barbieri, S., Morley, D., Reyier, E., & Grubbs, R. (2021). Large-scale space use of large juvenile and adult smalltooth sawfish Pristis pectinata: implications for management. *Endangered Species Research*, *44*, 45–59. https://doi.org/10.3354/esr01088

GREENFIELD, T. (2021). Corrections to the nomenclature of sawskates (Rajiformes, Sclerorhynchoidei). *Bionomina*, *22*(1). https://doi.org/10.11646/bionomina.22.1.3

Gruber, A. R., Lorenz, R., Bernhart, S. H., Neubock, R., & Hofacker, I. L. (2008). The Vienna RNA Websuite. *Nucleic Acids Research*, *36*(Web Server), W70–W74. https://doi.org/10.1093/nar/gkn188

Gunawardane, L. S., Saito, K., Nishida, K. M., Miyoshi, K., Kawamura, Y., Nagami, T., Siomi, H., & Siomi, M. C. (2007). A Slicer-Mediated Mechanism for Repeat-Associated siRNA 5’ End Formation in *Drosophila*. *Science*, *315*(5818), 1587–1590. https://doi.org/10.1126/science.1140494

Guris, D. L., Duester, G., Papaioannou, V. E., & Imamoto, A. (2006). Dose-Dependent Interaction of Tbx1 and Crkl and Locally Aberrant RA Signaling in a Model of del22q11 Syndrome. *Developmental Cell*, *10*(1), 81–92. https://doi.org/10.1016/j.devcel.2005.12.002

Guttridge, T. L., Gulak, S. J. B., Franks, B. R., Carlson, J. K., Gruber, S. H., Gledhill, K. S., Bond, M. E., Johnson, G., & Grubbs, R. D. (2015). Occurrence and habitat use of the critically endangered smalltooth sawfish *Pristis pectinata* in the Bahamas. *Journal of Fish Biology*, *87*(6), 1322–1341. https://doi.org/10.1111/jfb.12825

Hamers, T., Kamstra, J. H., Sonneveld, E., Murk, A. J., Kester, M. H. A., Andersson, P. L., Legler, J., & Brouwer, A. (2006). In Vitro Profiling of the Endocrine-Disrupting Potency of Brominated Flame Retardants. *Toxicological Sciences*, *92*(1), 157–173. https://doi.org/10.1093/toxsci/kfj187

Han, B. W., Wang, W., Zamore, P. D., & Weng, Z. (2015). piPipes: a set of pipelines for piRNA and transposon analysis via small RNA-seq, RNA-seq, degradome- and CAGE-seq, ChIP-seq and genomic DNA sequencing. *Bioinformatics*, *31*(4), 593–595. https://doi.org/10.1093/bioinformatics/btu647

Harmon, T. S., Kamerman, T. Y., Corwin, A. L., & Sellas, A. B. (2016). Consecutive parthenogenetic births in a spotted eagle ray *Aetobatus narinari*. *Journal of Fish Biology*, *88*(2), 741–745. https://doi.org/10.1111/jfb.12819

Hass, B. J. (n.d.). *TransDecoder* (5.5.0). Retrieved August 29, 2022, from https://github.com/TransDecoder/TransDecoder

Huang, T.-H., Fan, B., Rothschild, M. F., Hu, Z.-L., Li, K., & Zhao, S.-H. (2007). MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. *BMC Bioinformatics*, *8*(1), 341. https://doi.org/10.1186/1471-2105-8-341

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., & Bork, P. (2017). Fast Genome-Wide Functional Annotation through Orthology

Assignment by eggNOG-Mapper. *Molecular Biology and Evolution*, *34*(8), 2115–2122. https://doi.org/10.1093/molbev/msx148

Joga, M. R., Zotti, M. J., Smagghe, G., & Christiaens, O. (2016). RNAi Efficiency, Systemic Properties, and Novel Delivery Methods for Pest Insect Control: What We Know So Far. *Frontiers in Physiology*, *7*. https://doi.org/10.3389/fphys.2016.00553

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Kamath, R. (2003). Genome-wide RNAi screening in Caenorhabditis elegans. *Methods*, *30*(4), 313–321. https://doi.org/10.1016/S1046-2023(03)00050-1

Kassambra, A. (2020). *ggpubr: "ggplot2" Based Publication Ready Plots* (0.4.0). CRAN.

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Katoh, M., & Katoh, M. (2006). Cross-talk of WNT and FGF signaling pathways at GSK3ß to regulate ß-catenin and SNAIL signaling cascades. *Cancer Biology & Therapy*, *5*(9), 1059–1064. https://doi.org/10.4161/cbt.5.9.3151

Keilwagen, J., Wenk, M., Erickson, J. L., Schattat, M. H., Grau, J., & Hartung, F. (2016). Using intron position conservation for homology-based gene prediction. *Nucleic Acids Research*, *44*(9), e89–e89. https://doi.org/10.1093/nar/gkw092

Khanal, S., Zancanela, B. S., Peter, J. O., & Flynt, A. S. (2022). The Small RNA Universe of Capitella teleta. *Frontiers in Molecular Biosciences*, *9*. https://doi.org/10.3389/fmolb.2022.802814

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360. https://doi.org/10.1038/nmeth.3317

King, B. L., Gillis, J. A., Carlisle, H. R., & Dahn, R. D. (2011). A Natural Deletion of the *HoxC* Cluster in Elasmobranch Fishes. *Science*, *334*(6062), 1517–1517. https://doi.org/10.1126/science.1210912

King, Benjamin., Fang Shi, Ling., Kao, Peter., & Clusin, William. (2015). Calcium Activated K+ Channels in The Electroreceptor of the Skate Confirmed by Cloning. Details of Subunits and Splicing. *Gene*, *578*(1), 63–73. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4724458/#SD2

King, S. M., & Patel-King, R. S. (2015). The Oligomeric Outer Dynein Arm Assembly Factor CCDC103 Is Tightly Integrated within the Ciliary Axoneme and Exhibits Periodic Binding to Microtubules. *Journal of Biological Chemistry*, *290*(12), 7388–7401. https://doi.org/10.1074/jbc.M114.616425

Kyne, P., Oetinger, M., Grant, M., & Feutry, P. (2021). Life history of the Critically Endangered largetooth sawfish: a compilation of data for population assessment and demographic modelling. *Endangered Species Research*, *44*, 79–88. https://doi.org/10.3354/esr01090

Lai, D., Proctor, J. R., Zhu, J. Y. A., & Meyer, I. M. (2012). R- chie : a web server and R package for visualizing RNA secondary structures. *Nucleic Acids Research*, *40*(12), e95–e95. https://doi.org/10.1093/nar/gks241

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., & Carey, V. J. (2013). Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology*, *9*(8), e1003118. https://doi.org/10.1371/journal.pcbi.1003118

le Thomas, A., Tóth, K., & Aravin, A. A. (2014). To be or not to be a piRNA: genomic origin and processing of piRNAs. *Genome Biology*, *15*(1), 204. https://doi.org/10.1186/gb4154

Lee, Y. (2002). MicroRNA maturation: stepwise processing and subcellular localization. *The EMBO Journal*, *21*(17), 4663–4670. https://doi.org/10.1093/emboj/cdf476

Lehman, R. N., Poulakis, G. R., Scharer, R. M., Hendon, J. M., Court, A. G., Wooley, A. K., Williams, A. M., Ajemian, M. J., Hadden, J. P., Beal, J. L., McCallister, M. P., & Phillips, N. M. (2022). Environmental DNA evidence of the Critically Endangered smalltooth sawfish, Pristis pectinata, in historically occupied US waters. *Aquatic Conservation: Marine and Freshwater Ecosystems*, *32*(1), 42–54. https://doi.org/10.1002/aqc.3721

Li, D., Luo, L., Zhang, W., Liu, F., & Luo, F. (2016). A genetic algorithm-based weighted ensemble method for predicting transposon-derived piRNAs. *BMC Bioinformatics*, *17*(1), 329. https://doi.org/10.1186/s12859-016-1206-3

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, Q., & Yan, J. (2016). *Modulation of BK Channel Function by Auxiliary Beta and Gamma Subunits* (pp. 51–90). https://doi.org/10.1016/bs.irn.2016.03.015

Mao, Y.-B., Cai, W.-J., Wang, J.-W., Hong, G.-J., Tao, X.-Y., Wang, L.-J., Huang, Y.-P., & Chen, X.-Y. (2007). Silencing a cotton bollworm P450 monooxygenase gene by plant-mediated RNAi impairs larval tolerance of gossypol. *Nature Biotechnology*, *25*(11), 1307–1313. https://doi.org/10.1038/nbt1352

Martins, M. F., Costa, P. G., & Bianchini, A. (2021). Maternal transfer of polycyclic aromatic hydrocarbons in an endangered elasmobranch, the Brazilian guitarfish. *Chemosphere*, *263*, 128275. https://doi.org/10.1016/j.chemosphere.2020.128275

Mathelier, A., & Carbone, A. (2010). MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. *Bioinformatics*, *26*(18), 2226–2234. https://doi.org/10.1093/bioinformatics/btq329

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Modrell, Melinda S., Lyne, Mike. Carr, Adrian R. Zakon, Harold H. Buckley, David. Campbell, Alexander S. Davis, Marcus C. Micklem, Gos. Baker, C. VH. (2017).

Insights Into Electrosensory Organ Development, Physiology and Evolution From A Lateral Line-Enriched Transcriptome. *ELife*, *6*. https://elifesciences.org/articles/24197

Mondal, M., Peter, J., Scarbrough, O., & Flynt, A. (2021). Environmental RNAi pathways in the two-spotted spider mite. *BMC Genomics*, *22*(1), 42. https://doi.org/10.1186/s12864-020-07322-2

Morgan, M., Pages, H., Obenchain, V., & Hayden, N. (2022). *Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF) and tabix file import.* (2.14). Bioconductor.

MORIN, P. A., MARTIEN, K. K., & TAYLOR, B. L. (2009). Assessing statistical power of SNPs for population structure and conservation studies. *Molecular Ecology Resources*, *9*(1), 66–73. https://doi.org/10.1111/j.1755-0998.2008.02392.x

Newbern, J., Zhong, J., Wickramasinghe, R. S., Li, X., Wu, Y., Samuels, I., Cherosky, N., Karlo, J. C., O'Loughlin, B., Wikenheiser, J., Gargesha, M., Doughman, Y. Q., Charron, J., Ginty, D. D., Watanabe, M., Saitta, S. C., Snider, W. D., & Landreth, G. E. (2008). Mouse and human phenotypes indicate a critical conserved role for ERK2 signaling in neural crest development. *Proceedings of the National Academy of Sciences*, *105*(44), 17115–17120. https://doi.org/10.1073/pnas.0805239105

Nishida, K. M., Saito, K., Mori, T., Kawamura, Y., Nagami-Okada, T., Inagaki, S., Siomi, H., & Siomi, M. C. (2007). Gene silencing mechanisms mediated by Aubergine–piRNA complexes in *Drosophila* male gonad. *RNA*, *13*(11), 1911–1922. https://doi.org/10.1261/rna.744307

NMFS. (2009). *Smalltooth Sawfish Recovery Plan*. https://repository.library.noaa.gov/view/noaa/15983

Okamura, K., Chung, W.-J., Ruby, J. G., Guo, H., Bartel, D. P., & Lai, E. C. (2008). The Drosophila hairpin RNA pathway generates endogenous short interfering RNAs. *Nature*, *453*(7196), 803–806. https://doi.org/10.1038/nature07015

Ooms, J. (2014). *The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects*.

Pages, H., Aboyoun, P., Gentleman, R., & DebRoy, S. (2022). *Biostrings: Efficient manipulation of biological strings.* (2.66.0). Bioconductor.

Poulakis, G. R., Stevens, P. W., Timmers, A. A., Wiley, T. R., & Simpfendorfer, C. A. (2011). Abiotic affinities and spatiotemporal distribution of the endangered smalltooth sawfish, Pristis pectinata, in a south-western Florida nursery. *Marine and Freshwater Research*, *62*(10), 1165. https://doi.org/10.1071/MF11008

Poulakis, G., Urakawa, H., Stevens, P., DeAngelo, J., Timmers, A., Grubbs, R., Fisk, A., & Olin, J. (2017). Sympatric elasmobranchs and fecal samples provide insight into the trophic ecology of the smalltooth sawfish. *Endangered Species Research*, *32*, 491–506. https://doi.org/10.3354/esr00824

Price, M. N., Dehal, P. S., & Arkin, A. P. (2009). FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution*, *26*(7), 1641–1650. https://doi.org/10.1093/molbev/msp077

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Ritchie, E. G., Elmhagen, B., Glen, A. S., Letnic, M., Ludwig, G., & McDonald, R. A. (2012). Ecosystem restoration with teeth: what role for predators? *Trends in Ecology & Evolution*, *27*(5), 265–271. https://doi.org/10.1016/j.tree.2012.01.001

Rosenkranz, D., & Zischler, H. (2012). proTRAC - a software for probabilistic piRNA cluster detection, visualization and analysis. *BMC Bioinformatics*, *13*(1), 5. https://doi.org/10.1186/1471-2105-13-5

Sallan, L. C., & Coates, M. I. (2014). The long-rostrumed elasmobranch *Bandringa* Zangerl, 1969, and taphonomy within a Carboniferous shark nursery. *Journal of Vertebrate Paleontology*, *34*(1), 22–33. https://doi.org/10.1080/02724634.2013.782875

Salzberg, S. L. (2019). Next-generation genome annotation: we still struggle to get it right. *Genome Biology*, *20*(1), 92. https://doi.org/10.1186/s13059-019-1715-2

Schaid, D. J., Guenther, J. C., Christensen, G. B., Hebbring, S., Rosenow, C., Hilker, C. A., McDonnell, S. K., Cunningham, J. M., Slager, S. L., Blute, M. L., & Thibodeau, S. N. (2004). Comparison of Microsatellites Versus Single-Nucleotide Polymorphisms in a Genome Linkage Screen for Prostate Cancer–Susceptibility Loci. *The American Journal of Human Genetics*, *75*(6), 948–965. https://doi.org/10.1086/425870

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., & Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*(1), 539. https://doi.org/10.1038/msb.2011.75

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. v., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Smith, K., Feldheim, K., Carlson, J., Wiley, T., & Taylor, S. (2021). Female philopatry in smalltooth sawfish Pristis pectinata: conservation and management implications. *Endangered Species Research*, *45*, 85–98. https://doi.org/10.3354/esr01122

Smith, M. D., Wertheim, J. O., Weaver, S., Murrell, B., Scheffler, K., & Kosakovsky Pond, S. L. (2015). Less Is More: An Adaptive Branch-Site Random Effects Model for Efficient Detection of Episodic Diversifying Selection. *Molecular Biology and Evolution*, *32*(5), 1342–1353. https://doi.org/10.1093/molbev/msv022

Stanke, M., & Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, *19*(suppl_2), ii215–ii225. https://doi.org/10.1093/bioinformatics/btg1080

Staton, E. (2014). *Hmmer2GO* (0.17.8).

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, *34*(Web Server), W609–W612. https://doi.org/10.1093/nar/gkl315

Thauvin-Robinet, C., Duplomb-Jego, L., Limoge, F., Picot, D., Masurel, A., Terriat, B., Champilou, C., Minot, D., St-Onge, J., Kuentz, P., Duffourd, Y., Thevenon, J., Rivière, J.-B., & Faivre, L. (2016). Homozygous FIBP nonsense variant responsible of syndromic overgrowth, with overgrowth, macrocephaly, retinal coloboma and

learning disabilities. *Clinical Genetics*, *89*(5), e1–e4. https://doi.org/10.1111/cge.12704

Thorson, T. B. (1973). Sexual Dimorphism in Number of Rostral Teeth of the Sawfish, Pristis perotteti Müller and Henle, 1841. *Transactions of the American Fisheries Society*, *102*(3), 612–614. https://doi.org/10.1577/1548-8659(1973)102<612:SDINOR>2.0.CO;2

Tijsterman, M., Ketting, R. F., & Plasterk, R. H. A. (2002). The Genetics of RNA Silencing. *Annual Review of Genetics*, *36*(1), 489–519. https://doi.org/10.1146/annurev.genet.36.043002.091619

Tiktak, G. P., Butcher, D., Lawrence, P. J., Norrey, J., Bradley, L., Shaw, K., Preziosi, R., & Megson, D. (2020). Are concentrations of pollutants in sharks, rays and skates (Elasmobranchii) a cause for concern? A systematic review. *Marine Pollution Bulletin*, *160*, 111701. https://doi.org/10.1016/j.marpolbul.2020.111701

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*(5), 511–515. https://doi.org/10.1038/nbt.1621

Vagin, V. v., Sigova, A., Li, C., Seitz, H., Gvozdev, V., & Zamore, P. D. (2006). A Distinct Small RNA Pathway Silences Selfish Genetic Elements in the Germline. *Science*, *313*(5785), 320–324. https://doi.org/10.1126/science.1129333

Wang, K., Liang, C., Liu, J., Xiao, H., Huang, S., Xu, J., & Li, F. (2014). Prediction of piRNAs using transposon interaction and a support vector machine. *BMC Bioinformatics*, *15*(1), 419. https://doi.org/10.1186/s12859-014-0419-6

Wang, X., Zhang, J., Li, F., Gu, J., He, T., Zhang, X., & Li, Y. (2005). MicroRNA identification based on sequence and structure alignment. *Bioinformatics*, *21*(18), 3610–3614. https://doi.org/10.1093/bioinformatics/bti562

Welten, M., Smith, M. M., Underwood, C., & Johanson, Z. (2015). Evolutionary origins and development of saw-teeth on the sawfish and sawshark rostrum (Elasmobranchii; Chondrichthyes). *Royal Society Open Science*, *2*(9), 150189. https://doi.org/10.1098/rsos.150189

Wiley, T., & Brame, A. (2018). *5-Year Review: Summary and Evaluation of United States Distinct Population Segment of Smalltooth Sawfish.* https://repository.library.noaa.gov/view/noaa/19253

WILEY, T. R., SIMPFENDORFER, C. A., FARIA, V. v., & MCDAVITT, M. T. (2008). Range, sexual dimorphism and bilateral asymmetry of rostral tooth counts in the smalltooth sawfish Pristis pectinata Latham (Chondrichthyes: Pristidae) of the southeastern United States. *Zootaxa*, *1810*(1), 51. https://doi.org/10.11646/zootaxa.1810.1.3

Wiley, T., & Simpfendorfer, C. (2010). Using public encounter data to direct recovery efforts for the endangered smalltooth sawfish Pristis pectinata . *Endangered Species Research*, *12*(3), 179–191. https://doi.org/10.3354/esr00303

Willow, J., Taning, C. N. T., Cook, S. M., Sulg, S., Silva, A. I., Smagghe, G., & Veromann, E. (2021). RNAi Targets in Agricultural Pest Insects: Advancements,

Knowledge Gaps, and IPM. *Frontiers in Agronomy*, *3*. https://doi.org/10.3389/fagro.2021.794312

Wu, Y., Wei, B., Liu, H., Li, T., & Rayner, S. (2011). MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences. *BMC Bioinformatics*, *12*(1), 107. https://doi.org/10.1186/1471-2105-12-107

Wueringer, B. E. (2012). Electroreception in Elasmobranchs: Sawfish as a Case Study. *Brain, Behavior and Evolution*, *80*(2), 97–107. https://doi.org/10.1159/000339873

Wueringer BE, Jnr LS, Kajiura SM, Tibbetts IR, Hart NS, C. S. (2012). Electric Field Detection in Sawfish and Shovelnose Rays. *PLoS ONE*, *7*(7).

Wyffels, J. T., Adams, L. M., Bulman, F., Fustukjian, A., Hyatt, M. W., Feldheim, K. A., & Penfold, L. M. (2021). Artificial insemination and parthenogenesis in the whitespotted bamboo shark Chiloscyllium plagiosum. *Scientific Reports*, *11*(1), 9966. https://doi.org/10.1038/s41598-021-88568-y

Xia, X. (2018). DAMBE7: New and Improved Tools for Data Analysis in Molecular Biology and Evolution. *Molecular Biology and Evolution*, *35*(6), 1550–1552. https://doi.org/10.1093/molbev/msy073

Yan, H. F., Kyne, P. M., Jabado, R. W., Leeney, R. H., Davidson, L. N. K., Derrick, D. H., Finucci, B., Freckleton, R. P., Fordham, S. v., & Dulvy, N. K. (2021). Overfishing and habitat loss drive range contraction of iconic marine fishes to near extinction. *Science Advances*, *7*(7). https://doi.org/10.1126/sciadv.abb6026

Ye, J., Zhang, Y., Cui, H., Liu, J., Wu, Y., Cheng, Y., Xu, H., Huang, X., Li, S., Zhou, A., Zhang, X., Bolund, L., Chen, Q., Wang, J., Yang, H., Fang, L., & Shi, C. (2018).

WEGO 2.0: a web tool for analyzing and plotting GO annotations, 2018 update. *Nucleic Acids Research*, *46*(W1), W71–W75. https://doi.org/10.1093/nar/gky400

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T. (2017). ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*(1), 28–36. https://doi.org/10.1111/2041-210X.12628

Yutzey, K. E. (2010). DiGeorge Syndrome, Tbx1, and Retinoic Acid Signaling Come Full Circle. *Circulation Research*, *106*(4), 630–632. https://doi.org/10.1161/CIRCRESAHA.109.215319

Zhou, L., Feng, T., Xu, S., Gao, F., Lam, T. T., Wang, Q., Wu, T., Huang, H., Zhan, L., Li, L., Guan, Y., Dai, Z., & Yu, G. (2022). ggmsa: a visual exploration tool for multiple sequence alignment and associated data. *Briefings in Bioinformatics*, *23*(4). https://doi.org/10.1093/bib/bbac222

Zhu, F., Xu, J., Palli, R., Ferguson, J., & Palli, S. R. (2011). Ingested RNA interference for managing the populations of the Colorado potato beetle, *Leptinotarsa decemlineata*. *Pest Management Science*, *67*(2), 175–182. https://doi.org/10.1002/ps.2048