Summer 8-2-2022

A Genomic Investigation of Divergence Between Tuna Species

Pavel V. Dimens
University of Southern Mississippi

A GENOMIC INVESTIGATION OF DIVERGENCE BETWEEN TUNA SPECIES

by

Pavel V. Dimens

A Dissertation
Submitted to the Graduate School,
the College of Arts and Sciences
and the School of Ocean Science and Engineering
at The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

Approved by:

Dr. Eric Saillant, Committee Chair
Dr. Frank Hernandez
Dr. Leila Hamdan
Mr. James Franks
Dr. Kenneth Jones

May 2022

Published by the Graduate School

THE UNIVERSITY OF
SOUTHERN
MISSISSIPPI®

ABSTRACT

Effective management and conservation of marine pelagic fishes is heavily dependent on a robust understanding of their population structure, their evolutionary history, and the delineation of appropriate management units. The Yellowfin tuna (*Thunnus albacares*) and the Blackfin tuna (*Thunnus atlanticus*) are two exploited epipelagic marine species with overlapping ranges in the tropical and sub-tropical Atlantic Ocean. This work analyzed genome-wide genetic variation of both species in the Atlantic basin to investigate the occurrence of population subdivision and adaptive variation. A *de novo* assembly of the Blackfin tuna genome was generated using Illumina paired-end sequencing data and applied as a reference for population genomic analysis of specimens from 9 localities spanning most of the Blackfin tuna range. Analysis suggested the presence of four weakly differentiated units corresponding to the northwestern Atlantic Ocean, Gulf of Mexico, Caribbean Sea, and southwestern Atlantic Ocean, respectively. Significant spatial autocorrelation of genotypes was observed for specimens collected within 800 km of each other. A high-quality genome assembly generated for the Yellowfin tuna using PacBio and Illumina sequences was scaffolded by a linkage map developed through analysis of the segregation of genome wide Single Nucleotide Polymorphisms in 164 larvae offspring from a single pair produced by controlled breeding. The genome assembly was used as a reference for population genomic analysis of juvenile specimens from the 4 main nursery areas hypothesized in the Atlantic Ocean basin. Analyses corroborated previously reported population subdivision between the east and west Atlantic Ocean, but also suggested subdivision associated with individual nursery areas within the east and west regions. Draft reference assemblies were generated for Albacore, Bigeye and Longtail tunas and used in combination with the Yellowfin and Blackfin tuna genomes obtained in this work and existing assemblies for bluefin tunas in preliminary analyses of genome wide variation between species of the *Thunnus* genus. Whole-genome derived SNP-based phylogenetic analysis of the *Thunnus* genus suggests phylogenetic relationships may be more complex than suggested in earlier work based on Restriction-site Associated DNA sequencing or muscle transcriptome sequencing and prompt for further analysis of the genus using a more comprehensive sampling of taxa in each oceanic basin.

ACKNOWLEDGMENTS

The timely completion of this dissertation would not have been possible without Brooke McPeak providing so much laboratory and administrative assistance. I would also like to acknowledge Tami Hildahl, Noor Salam, and Heather King for their assistance and Becky O’Rourke walking me through my first genome assembly. I must acknowledge Maria Akopyan, my brilliant and compassionate partner who provided the wind in my sails to get to the finish line. Although she can’t read this, I want to acknowledge my dog Luna for being herself because life is just better with her around. Lastly, I will acknowledge my advisor Dr. Saillant and my committee for the environment that led to my ultimately completing this work.

There is a long list of collaborators that have supplied the samples for these projects, without whose generous assistance there would be no findings. These collaborators include Steven Davies and Sascha Cushner from the NOAA Pelagic Observer Program, Dr. Fabio Hazin and Pollyana Roque from Universidade Federal Rural De Perna Mbuco, Dr. Daniel Marguiles, Dr. Vernon Scholey, and Susana Cusatta from IATTC, Katrina Diener and Becky O’Rourke from UC Denver-Anschutz, Matthew Green from MS DMR, Drs. Cedric Pau and Lionel Reynal from Ifremer, Dr. Freddy Arocha from Universidad de Oriente, Dr. Fa Mbaye Ngom-Sow from ISRA, Dr. Paul Bannerman from MOFAD-Ghana, Brett Falternam and Ciji Marshall from LDWF, Amy Duke from SC DMR, Jay Rooker from TAMU-Galveston, Kelly Sullivan from FL FWC, Chris Manley, and Captains R. Whitaker, M. Hanke, Mike Gutierrez, and Daniel T. Powell, and Dr. Constance Diaha N'Guessan from Centre de Recherches Océanologiques.

DEDICATION

This is dedicated to Maria Slousch for making sure I did my math homework and Samuel Nalibotsky for reminding me of my strength. This is also dedicated to my family, who have no familiarity with post-graduate experience and whose patience and support since 2007 cannot be overstated. They always wanted a doctor in the family, too bad this isn't the kind they wanted specifically. This is also dedicated to Dr. Elizabeth Condon for taking a chance on me all those years ago, and quite literally everyone who has raised me up in any fashion, from telling a joke when I needed to hear one or imparting words of wisdom when I felt lost. I am forever a sum of your parts, and this work cannot exist without any of you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

LIST OF ABBREVIATIONS

SNP	Single Nucleotide Polymorphism
LD	Linkage Disequilibrium
UMI	Universal Molecular Identifiers
cM	centiMorgan
bp	base pairs
kb	kilobase pairs
Gb	gigabase pairs
Mb	megabase pairs
ddRAD	double-digest Restriction Associated DNA
US	United States of America
ATL	Atlantic Ocean
wATL	Western Atlantic Ocean
GOM	Gulf of Mexico
PNS	Pensacola, Florida
PR	Puerto Rico
TX	Texas
LA	Louisiana
VZ	Venezuela
MRT	Martinique
BRZ	Brazil
SEN	Senegal
IVC	Cote de Ivoire (Ivory Coast)

mtDNA	Mitochondrial DNA
PC	Principal Component
DAPC	Discriminate Analysis of Principal Components
AMOVA	Analysis of Molecular Variance
PAC	Pacific Ocean
COS	Cosmopolitan
SBF	Southern Bluefin Tuna

CHAPTER I – INTRODUCTION

1.1 Divergence within species

Understanding the processes that lead to divergence and the formation of new taxa is a central topic of evolutionary biology (Darwin 1859). These processes contribute to the diversification of phenotypes and are often associated with adaptations driven by natural selection occurring in new geographic areas or ecological niches colonized by a species (Schluter and Conte 2009). The initial step toward reproductive isolation is the formation of demographically independent populations that develop adaptations due to selection in their respective environment and/or geographic regions, and ultimately become reproductively isolated (Schluter and Conte 2009). Species rarely form a single panmictic population at any point in time, and are often subdivided in su bpopulations (demes) showing various levels of isolation and residual gene flow (Weir and Goudet 2017). Understanding the dynamics of these networks of connected populations or metapopulations is critical to assess the processes and drivers of divergence and speciation.

The initial definition of a metapopulation described a system of multiple semi-isolated demes over a given space experiencing episodic extinction and recolonization by other demes (Weir and Goudet 2017). In marine systems, a more appropriate model proposed by Kritzer and Sale (2004) considers demes or habitat patches that usually do not experience extinction and remain partially connected by migration. These demes maintain demographic independence such that there are important dynamics occurring on a su bpopulation level contributing to the dynamics of the metapopulation they form (Kritzer and Sale 2004; Sale et al. 2006). Oceans have historically been considered large circumglobal habitats where it was reasonable to expect seamless connectivity in vagile marine species (Avise 1998), but recent studies have revealed increasing evidence for population subdivision, sometimes at a small geographic scale (Hauser and Carvalho 2008) even if the forces responsible for fragmentation were not always clearly identified (Carlsson et al. 2006; Avise 1992; Reeb and Avise 1990).

Understanding metapopulation structures has direct applications for marine fisheries management, where it is necessary to identify stocks (discrete demographic units) expected to respond homogeneously to management measures within administrative jurisdictions. The dynamics of these fisheries stocks are affected by local processes such as birth and mortality rates, and they are expected to respond independently to exploitation and environmental variation (Grimes 1987; Carvalho and Hauser 1994; Begg et al. 1999). Therefore, failing to identify units comprising a subdivided stock may result in over-exploitation of some subunits and the loss of unique genetic characteristics carried by those local stocks that impart adaptions contributing to their regional sustainability (Smith et al. 1991; Begg et al. 1999; Hilborn et al. 2003).

1.2 Approaches to population studies in marine fishes

Accurate direct (visual-based) assessments of population size and movement between demes are usually impossible because natural habitats used by species are generally inaccessible to researchers (Palumbi 1994; Shaklee and Bentzen 1998). Similarly, tracking or observing specific reproductive events or the fitness of early life stages prior to recruitment is extremely challenging in the marine environment. Accordingly, indirect approaches are needed to assess these metapopulation parameters and test hypotheses regarding the factors driving the maintenance of the demographic trends we observe (Waples 1998). Despite these obstacles, several methods have been successfully implemented to study the structure and migration dynamics of marine fish populations. Passive and active tagging methods have been widely used to document fish movement and delineate stocks (Chapman et al. 2015; Metcalfe and Arnold 1997; Pollock 1991), but these approaches are hindered by low recapture rates for traditional physical tagging (Kohler and Turner 2001) and high costs limiting sample sizes for studies employing satellite and archival tags. Otolith chemistry methods (Secor et al. 1995) also document fish movements but are often limited in the pelagic environment by the lack of clear signature of geographic areas and shared chemical signatures between areas (Gibb et al. 2017). These approaches share the limitation that only physical movement of an individual within a portion of its lifespan is recorded rather than its actual contribution to the gene pool of recipient populations when movement to another deme is observed (Carlsson et al. 2006; Dimens et al. 2019). The latter is a high priority when assessing the validity of stock delineation for management, which relies on information on local spawning stock and recruitment. In contrast, genetic methods assessing the divergence of populations based on the distributions of genetic variants provide insight on the genetic contribution of individuals to local breeding stocks and gene pools. These methods also provide a breadth of information to understand the evolution of metapopulations including the adaptation of geographic populations, the contemporary and historical effective population size and connectivity of demes, and aspects of demographic history such as bottlenecks, range expansion, and isolation.

Recent developments of technologies for high throughput sequencing and genotyping enable the characterization of individuals within populations using a very large number of genetic markers such that divergence among populations is assessed in every region of the genome. Single nucleotide polymorphisms (hereafter SNPs) are presently the most widely used genetic markers in such genome scans (Puritz et al. 2014). SNPs are single base-pair substitutions at specific loci distributed throughout the entire genome in densities as high as 1 SNP every 64 base pairs as seen in rainbow trout *Oncorhynchus mykiss* (Gao et al., 2018). SNPs originate as the result of genetic mutations and their abundance, location, and allele frequencies are driven by evolutionary forces, namely mutations, natural selection, genetic drift, and migration (Castle 2011). High-density genome scans recover thousands to millions of polymorphic SNP loci and examining allele frequencies within and between populations at so many loci enables identifying genomic regions affected by natural selection in outlier analyses (Foll and Gaggiotti 2008; Lotterhos and Whitlock 2015). Applying high-density genome scans to species and populations also enables estimating parameters characterizing neutral processes affecting populations such as effective population size, population growth rate or migration rates among demes.

1.3 Challenges assessing population structure in marine species

In marine systems, apparent genetic homogeneity has been observed across broad areas in many taxa (Beatty et al. 2020; Kitada et al. 2017), yet in most cases the data collected could not rule out scenarios where populations would be partially isolated but failed to display genetic differences due to past connectivity or recent expansion from a common gene pool. However, despite of the high connectivity predicted to occur across large sections of open habitats for many marine species that feature tremendous dispersal potential, many instances of population subdivision have been reported, driven by environmental gradients such as salinity (André et al. 2011), temperature (Bradbury et al. 2010), limited movement of organisms or natal philopatry (Ferreira et al. 2015), patchiness of habitats (Selwyn et al. 2016), or by the direction and velocity of oceanic currents dispersing larvae restricting their transport to certain areas (Richardson et al. 2010). The degree of differentiation expected at equilibrium between partially isolated populations depends primarily on the number of migrants exchanged per generation, where only a few migrants (≥ 1 per generation) are usually sufficient to homogenize gene frequencies in connected populations and limit their divergence to very low levels (Mills and Allendorf 1996). However, the population sizes of marine species are often large such that divergence of demes towards equilibrium values under the effect of genetic drift is very slow, even when demes exchange few or no migrants (Waples 1998) leaving gene frequencies in su bpopulations highly homogeneous for extended periods of time. Demographic isolation may also occur temporarily, but periodic gene flow could be sufficient to maintain homogeneity among demes, as proposed for the reef-associated red snapper (Pruett et al. 2005). Under such conditions, the slow effects of genetic drift in large populations would prevent reaching a migration-drift equilibrium and cause the demes to remain genetically similar.

Highly migratory pelagic species are free swimming and can often move large distances compatible with the maintenance of connectivity across entire ocean basins (Sang et al. 1994; Schaefer et al. 2011). For these species, the physical characteristics of habitat patches are not as clearly defined as those identified for reef fishes and can be related to features that are not fixed in space such as floating structures (Druon et al. 2015) or frontal zones (Teo et al. 2007). However, many species have shown fidelity to specific geographic areas used for feeding, spawning or as nurseries (Luckhurst et al. 2001; Wells et al. 2012). Considering the dispersal potential of highly migratory marine fishes (Thorrold et al. 2001; Gibb et al. 2017), there would be an increased likelihood that these population patches are connected when they are present. Yet, many highly migratory fish species demonstrate geographic (Pecoraro et al. 2018; Portnoy et al. 2015) or sympatric (Daly-Engel et al. 2012; Tessier and Bernatchez 1999) population structure. This has been observed in teleosts with dispersal spawning (Carlsson et al. 2006; Bradman et al. 2011) as well as elasmobranchs with internal fertilization and live birth mating strategies (Jorgensen et al. 2010; Karl et al. 2011; Bernard et al. 2016). In pelagic systems, population structure may also occur among groups with overlapping geographic ranges, reflecting different patterns of habitat use or different migratory behaviors between stocks (Carlsson et al. 2006).

These adaptive traits may eventually lead to population divergence where gene flow between groups is restricted or eliminated over time, and ultimately to reproductive isolation and speciation. It is often difficult to understand the effects of all the evolutionary forces (mutations, genetic drift, selection, and migration) in empirical studies due to the *a priori* unknown and often complex demographic history of metapopulations, the multiplicity of scenarios one needs to consider, and the large number of candidate factors potentially acting concomitantly. It is possible to study the metapopulations formed by the same species across multiple oceanic basins (Pecoraro et al. 2018; Ward et al. 1997) used as replicates of evolutionary processes, but habitat constraints between metapopulations can vary, which complicates the conclusions drawn from such studies. Alternatively, one can investigate conserved genomic regions in congeners occupying similar or overlapping habitats to identify common evolutionary factors involved in divergence and speciation.

1.4 Divergence of sympatric congeners

Closely related species with similar life histories offer a good model to study patterns of genomic divergence associated with speciation. Speciation is the result of populations diverged from a common ancestor ceasing to produce viable offspring with one another (Palumbi 1994). Recently diverged taxa may be experiencing similar habitat-driven evolutionary constraints and may therefore have conserved genes impacting fitness in their shared habitats, while diverging at other genes during speciation. Allopatric speciation describes a population splitting into two or more geographically isolated populations with restricted gene flow between the isolated groups. Evolution of allopatric groups is driven by different selective pressures in the geographic habitats isolated during the split and genetic drift. Once enough time has passed and the isolated populations have diverged to a sufficient degree, they become reproductively incompatible and unable to exchange genes even if they were to come into contact again (Palumbi 1994). In contrast, sympatric speciation describes divergence occurring within a single geographical region where the range of one species overlaps the range of the other without physical barriers throughout the entire speciation process (Berlocher and Feder 2002). However, from a genetic perspective, the definition of sympatric speciation is debated and not necessarily related to spatial considerations. The varied genetic definitions of sympatric speciation include that an individual's birthplace should not affect its probability of dispersal (Berlocher and Feder 2002), that the probability of individuals mating relies solely on their genotypes and not on spatial or behavioral components (Kondrashov and Mina 1986; Howard and Berlocher 1998), that speciation occurs under panmixia or initial high gene flow (Coyne and Orr 2004), or with high migration (Coyne and Orr 2004). While these definitions differ from the geospatial definition of sympatric speciation as well as each other, the emphasis is that space is not driving divergence in any way.

Pelagic fishes can be considered ideal candidates for investigating sympatric speciation in contrast to sessile or structure/substrate associated fishes that are more prone to geographic isolation. Free swimming pelagic fishes experience few barriers to movement and have expansive ranges, usually exceeding thousands of miles (larger tunas, sharks, etc.), with some species, such as Yellowfin tuna (*Thunnus albacares*), having circumglobal distributions. The widespread sympatric distribution of congeners in these groups suggests that sympatric speciation may be the dominant process leading to the formation of new species. This seems especially true for the various recently diverged sympatric representatives of the genus *Thunnus* (family *Scombridae*), many of which are occurring across such an expansive range (Chow and Kishino 1995; Chow et al. 2006). The Atlantic Blackfin tuna (*Thunnus atlanticus*), one of the two species of focus in this work, is another highly migratory pelagic species, but whose range is restricted to the western Atlantic Ocean, where it overlaps with the Yellowfin tuna both spatially and temporally (Collette et al. 2010; Collette et al. 2011). Molecular studies suggest Yellowfin tuna may be more ancestral within the group, whereas Blackfin tuna would be a more recently derived species (Díaz-Arce et al. 2016). However, the relationships between these two species and among all tropical tunas may yet be unresolved because of the variability of inferences obtained when different sets of individuals and loci were used to characterize individual species (Chow and Kishino 1995; Chow et al. 2006), even when large numbers of genetic markers were deployed (Díaz-Arce et al. 2016; Guo et al. 2016).

Genomic technologies, particularly high-density genome scans, provide a breadth of information on the process of speciation. In the case of speciation with gene flow, genomic regions inferring adaptation or reproductive isolation and resisting homogenization by gene flow ("islands of divergence") can be identified by genome comparisons. Genetic sequencing technology continues to drop in price, increase throughput, and improve in sequence quality, lowering the barriers to generating quality genome assemblies on a limited budget and allowing one to realistically compare entire genomes between species. More complete genome assemblies can facilitate identifying islands of speciation or conserved regions by allowing larger syntenic alignments between species, or identifying critical genomic regions such as the centromeres (Ichikawa et al. 2017; Ferree and Barbash 2009) whose low recombination rates have been associated with speciation irrespective of postzygotic incompatibility (Noor and Bennett 2009). Such a comparative genomics approach can thus shed light on evolutionary trends between congeners and reveal patterns overlooked by genomic subsampling.

1.5 Species chosen for study

Tunas (family *Scombridae*) are highly specialized fast-swimming pelagic predators known to migrate large distances annually (Mariani et al. 2016; Reglero et al. 2017). They are therefore expected to form metapopulations connected over broad distances, up to the scale of entire oceanic basins. Genetic structure has been described between basins (Pecoraro et al. 2018), relating to the closure of the Isthmus of Panama, as well as within-basin for some of the Thunnus species indicating that despite the high potential for connectivity, this group is evolving in response to reproductive isolation and ecological factors with the formation of differentiated populations and new species. This study first focuses on the closely related congeners Yellowfin tuna and Blackfin tuna (Díaz-Arce et al. 2016). These two species are sympatric in the tropical Atlantic Ocean with similar life histories (Freire et al. 2005; Schaefer et al. 2007). Their estimated divergence time is estimated to be less than 5 MY (Chow and Kishino 1995; Chow et al. 2006; Ciezarek et al. 2019), there is an opportunity to understand genetic factors shared by the two species in response to common (recent) constraints in contrast to those involved in speciation in an overlapping habitat with limited noise from genetic drift. Applying genomic approaches can provide information on genomic regions, and ultimately genes, associated with differentiated groups, which can yield insights into factors driving divergence and speciation. This source of information can be valuable in these species because direct assessment of the phenotype of divergent groups and of environmental factors leading to isolation is particularly challenging considering the pelagic lifestyle of tunas.

1.6 Research objectives and hypotheses

This research first aims to elucidate the structure of the Atlantic Ocean metapopulations formed by Yellowfin and Blackfin tunas, by assessing whether there are demographically independent assemblages within the Atlantic Ocean basin. When distinct assemblages exist, this work attempts to identify the patterns of structure involved (e.g., occurrence of spatial or temporal barriers, isolation by distance), describe current patterns of gene flow between units, and assess evidence for selection and local adaptation of geographic populations. The mobility and passive larval dispersal of adult Yellowfin and Blackfin tuna suggests there will be minimal spatial structure, with large demes comprising highly connected metapopulations. Investigation of demographic assemblages will include pairwise relatedness, which provides information on cohort cohesiveness and indications of early stages of structuring. Overlaying patterns of structure, gene flow, and selection among the two congeners in the same environmental context and implementing a comparative genomics approach will provide information on processes and drivers of speciation and will also identify conserved or convergent genetic characteristics responding to shared habitat constraints contributing to the continued success of these taxa to their local environments**.** While investigating population genetics, this work also establishes data on stock structure beneficial to sound management of the two exploited taxa. These questions, hypotheses, and the methods to investigate them are addressed in the following chapters. The second chapter reports the development of genomic resources for Yellowfin tunas to enable the interpretation of genome scans in this species and comparative genomic analysis. The third chapter reports the analysis of spatial and temporal variation in Yellowfin tuna across the Atlantic basin. The fourth chapter reports the analysis of spatial and temporal variation in the Blackfin tuna across its West Atlantic range. The fifth chapter utilizes a comparative genomics approach to analyze divergence between and reconstruct phylogenies for all *Thunnus* species.

CHAPTER II – GENOMIC RESOURCES FOR THE YELLOWFIN TUNA *THUNNUS ALBACARES*

2.1 Introduction

Yellowfin tuna (*Thunnus albacares*) is a large epipelagic scombrid identified by its elongated yellow anal and second dorsal fins. It can grow to 2.2 m and weight up to 200 kg, making it the second largest extant tuna species. The species is globally distributed in the tropical and sub-tropical waters of all oceans (Collette and Nauen 1983) where it is targeted by major fisheries. Yellowfin tuna is the second most harvested tuna species worldwide (Pew Charitable Trusts 2020) with a global dock value of $4.4 billion (37.44% of the entire tuna evaluation) in 2018. The continued exploitation of this species necessitates well-informed management to maintain sustainable harvest. The analysis of genome wide variation using genome scans provides information on aspects of the biology and ecology of species. Studies of the subdivision of populations and delineation of appropriate stock units are critical to management (Carvalho and Hauser 1994). Genomic data also provide further insights on the structure of metapopulations by allowing estimating rates and patterns of gene flow among units, detecting barriers to gene flow within the range, estimating the size of demes, their demographic history, and the effects of natural selection and local adaptation on divergence of loci for examples.

DNA polymorphism has long been used to make inferences in population genetics (Allendorf et al. 2012), but assay methods compatible with population genetic surveys were limiting studies to low marker density in most cases until recently. The development of next-generation sequencing methods (NGS), yielding sequence throughput of several-fold coverage of the entire genome of a species, revolutionized the genotyping of genetic markers. Among the DNA polymorphisms revealed by sequencing, Single Nucleotide Polymorphisms (SNPs), which are single nucleotide substitutions occurring at specific positions in the genome or locus, have become the most popular markers for population genetic inference. SNPs are co-dominant, inherited in a Mendelian fashion, and are distributed throughout the entire genome at high density, e.g., 1 SNP every 64 bp (base pairs) in rainbow trout (Gao et al. 2018). Ascertaining genotypes at SNP positions using NGS methods reduced the costs associated with multilocus SNP assays by orders of magnitude (down to <$1 per SNP) and facilitated the expanded use of SNPs and other forms of structural variation for molecular population genetic studies by enabling genotyping thousands of loci for hundreds of samples at a time. Genotyping-by-sequencing approaches (e.g., Restriction Associated DNA sequencing) rely on sequencing SNP variants and the flanking DNA, allowing locus homology to be confirmed and reliable scoring. Since NGS studies typically reveal thousands of SNP loci, these methods provide for higher statistical power, along with higher genetic resolution than microsatellites and other earlier methods (Kwok 2001; Glaubitz et al. 2003; Koskinen et al. 2004; Hauser et al. 2011).

These methods discover and genotype SNP loci by mapping sequence reads onto a reference genome for the species of interest. Reference genomes are digital nucleotide sequence databases representative of the genome of a species under investigation. These databases are composed of consensus DNA segments (contigs) obtained by assembling large numbers of overlapping sequencing reads of the same genomic region (Schatz et al. 2010). Alignment of genotyping-by-sequencing reads on these reference genomes ensures the homology of the mapped sequences and reveals the occurrence of alternative sequence variants at the same locus (Figure 2.1). While the reference can be produced *de novo* during genotyping by assembling RAD sequencing reads, these sequencing reads are short and restrict the genomic contigs obtained to the immediate region surrounding restriction sites, making it difficult to control the risk of incorrectly merging or splitting loci (Alkan et al. 2011) and providing no information regarding locus physical proximity (linkage). Producing an independent reference assembly covering a high fraction of the genome with high contiguity improves the reliability of locus identification. Such references can be produced at moderate costs for non-model species by applying assembly algorithms to short-read (300 bp-500 bp) shotgun sequencing. Genome assemblies produced using only short-read sequences often fail to resolve structural elements such as genomic repeat regions (Schmid et al., 2018) and results in highly fragmented assemblies even when sequencing is conducted with high coverage (Salzberg et al. 2012). Third generation single molecule sequencing using the Pacific Biosciences or Oxford Nanopore platforms potentially address this problem by generating long-read sequences (>10 kilobases, kb). These platforms have higher error rates than Illumina short-read sequences resulting in higher sequencing coverage required for *de novo* assembly. These methods also yield less throughput, increasing the cost to achieve the coverage necessary for an assembly with minor fragmentation. To combine the low cost, high throughput, and accuracy of short reads with the ease of assembly of long reads, “hybrid” assembly methods combining both sequencing approaches have been developed to circumvent the prohibitive cost associated with chromosome-scale non-model species *de novo* genome assembly (Ye et al. 2016; Ma et al. 2019).

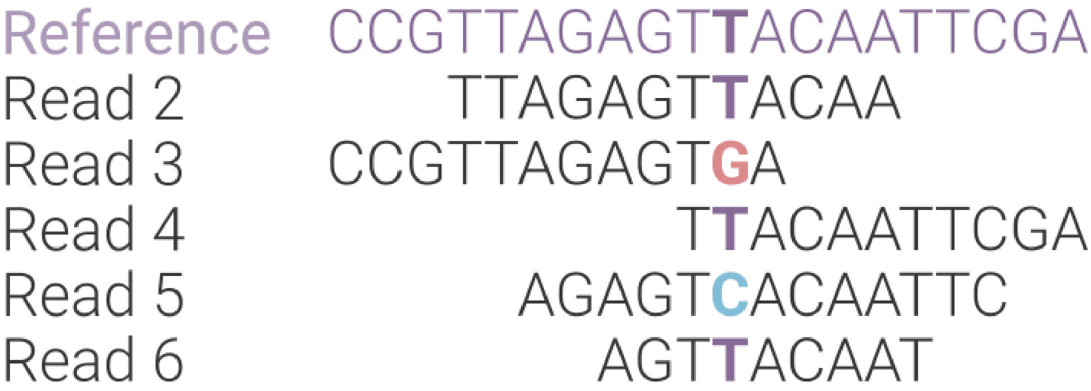

Figure 2.1 *Identifying sequence variants*
A generic diagram of SNP discovery. A series of sequences are aligned to a reference genome and regions where sequences are considered to overlap are investigated to find single base pair variants.

Despite the robust and varied algorithmic approaches to assemble long and/or short read sequences, the obtained reference genome sequences in new non-model species often remain fragmented in a few hundred to a few thousand contigs and scaffolds even after relatively large sequencing efforts. The mapping process of restriction associated DNA (RAD) sequencing reads allows positioning RAD loci onto these contigs, but their position in the overall genome will be unknown. Missing genomic arrangement information can be obtained through linkage mapping of genetic loci. Linkage is the study of the segregation of loci that occurs through recombination during meiosis, relying on the principle that loci that are closer together along a chromosome are more likely to migrate together during recombination. Thus, linkage maps provide information on the proximity of loci based on the frequency of recombination occurring between them (Sturtevant 1915). Once loci are positioned on the linkage map, they can be used to anchor contigs on which they were discovered so the genomic position of genome contigs is known. Furthermore, the location of any new SNP is determined once the sequencing reads revealing it are mapped on genome contigs (Tang et al. 2015). In sexually reproducing diploids, a linkage map can be created by comparing the genotypes of a mating pair (dam × sire) to their F1 progeny. While F2 progeny and backcross designs are possible and commonly implemented in other organisms like plants, it is not possible to implement this kind of experimental design in captivity for many marine fish species, particularly tunas, which cannot be cultured beyond late larval stages. Another approach is to produce haploids, but this design involves strip spawning, which is also currently not possible in many species, like Yellowfin tuna. However, the development of linkage maps in highly fecund fish is possible using single outbred crosses because of the extremely high fecundity (thousands to millions of eggs) that allow examining hundreds to thousands of offspring from the same cross. The Inter American Tropical Tuna Commission (IATTC) Achotines Laboratory has been rearing and captively spawning Yellowfin tuna since 1992 and provided F1 offspring samples and parental tissue necessary to produce a linkage map of the species in this project. Offspring from single pair crosses cannot be isolated in species like the Yellowfin tuna that spawn in groups in mass spawning tanks. Instead, molecular pedigrees can be used to identify a posteriori offspring from the same siblings during mass spawning events potentially involving multiple parents at the same time. Once a cross with progeny is obtained, genetic markers are genotyped in parents and offspring enabling estimating recombination rates between loci. Recombination rates estimates are then used to infer distances between markers and order these markers on chromosomes.

Mapping genetic loci on a reliable reference genome has several benefits. First, quality filtering and mapping of sequenced RAD loci are improved. Linkage information can also be used to make inferences regarding linkage disequilibrium (LD), a phenomenon where the frequency of association of different alleles deviates from the expected rate if the loci were independent and randomly associated (Hill and Robertson 1968; Slatkin 2008). LD information is useful in studying evolutionary processes such as the increase of linkage disequilibrium and reduced variation found in genomic regions surrounding loci bearing new mutations affecting fitness (selective sweep). With access to linkage information, sliding window analyses can also be employed to uncover genomic “islands” of divergent selection where proximal loci show congruent signals of divergence beyond neutral expectations. Recent and relatively minor changes in effective population size can be evaluated with the incorporation of linkage information (Hollenbeck et al. 2016; Waples and Do 2010) and used to elucidate the possible effects of anthropogenic forces (e.g., exploitation or management policies) or natural events over several generations. Also, the estimation of effective population size using the linkage disequilibrium (Waples 2006) relies on the assumption that SNP loci are not physically linked (Pritchard and Rosenberg 1999; Pritchard et al. 2000), but without linkage information, comparisons involving linked loci violating this assumption cannot be effectively removed.

2.2 Objective

The objective of this chapter was to develop the genomic resources needed to interpret genome scans generated during the population genetics study of Yellowfin (*Chapter III*) along with genome-level phylogenetic comparison with other tuna species (*Chapter V*). Progeny samples of Yellowfin tuna were made available by IATTC to generate a linkage map of the Yellowfin tuna genome, which was used to scaffold and improve the quality and contiguity of the de novo genome assembly to pseudochromosome level scaffolds. This reference will thus find many potential applications in future genomic studies of Yellowfin tuna.

2.3 Methods

2.3.1 Sample preparation and sequencing

To generate the reference genome, high-quality DNA from Yellowfin tuna samples collected in the western Atlantic Ocean was obtained by removing and storing heart and muscle tissue in 95% ethanol immediately after capture. The tissue was ground in liquid nitrogen using a mortar and pestle and directly processed with the Mag-Bind® Blood & Tissue DNA HDQ kit (Omega Bio-Tek, cat. M6399-01). Sequencing was executed with the intent of performing a “hybrid” assembly, described as a genome assembly created from multiple sequencing technologies. Longer reads (>1 kb), such as those produced by Pacific BioSciences (PacBio) and Oxford Nanopore Technologies (Nanopore) experience higher error rates (90-99.1%) and considerably lower throughput (by orders of magnitude) as compared to shorter reads (<800 bp) produced by Illumina technologies, whose read accuracy is greater than 99.99%. Combining the technologies is a cost-effective way of leveraging the strengths of both sequence types to achieve high quality assemblies (Haghshenas et al. 2020; Salzberg et al. 2012). Consequently, the extracted DNA was sequenced on the PacBio Sequel platform to obtain noisy long (>15 kb) reads and the Illumina Novaseq 6000 platform to obtain accurate short (300 bp) reads. The PacBio libraries were size selected to retain fragments larger than 10,000 bp in length for the first sample and 20,000 bp for the second sample and sequenced with a target throughput of 20 Gb (gigabase pairs) at the Duke Sequencing and Genomic

Technologies Shared Resource, yielding an expected average coverage over 20x considering the estimated 790 Mb (megabase pairs) genome size of Yellowfin tuna (McWilliam et al. 2016) The Illumina library was size selected to 300 bp fragments that were sequenced to generate a target of 600 million paired-end (2 x 150 bp) reads yielding an expected coverage over 160x. Illumina sequencing was performed at the University of Colorado – Denver Genomics Genomics Shared Resource facility.

2.3.2 Genome assembly

The short reads were first trimmed using fastp (Chen et al. 2018) to remove adaptor sequences, base calls with a quality score below 20, and sequencing read shorter than 50 bp, The size of the Yellowfin tuna genome was estimated using the k-mer frequency counting method using Jellyfish (Marçais and Kingsford 2011). The frequency distributions of k-mer depth were used to determine the mean coverage and the genome size was estimated as the total number of k-mers divided by the mean coverage (https://bioinformatics.uconn.edu/genome-size-estimation-tutorial/). Estimates were computed from Illumina sequence reads for K-mer sizes ranging from 17 to 25. Trimmed short reads were randomly subsampled down to ~70X using seqtk (https://github.com/lh3/seqtk) to reduce the computational load of certain steps in the assembly that would not run to completion on the full short read data. For polishing and scaffolding, the long reads were first corrected using the subsampled short reads. This was done by first correcting the subsampled short reads using the default parameters of Karect (Allam et al. 2015) to perform indel and substitution error correction from multiple sequence alignments. Then Brownie was used with default parameters to further correct the resulting graph. (https://github.com/biointec/brownie). Finally, the raw long reads were aligned to the corrected short read graph with a k of 75 to correct the long reads with Jabba (Miclotte et al. 2016).

Detailed assembly parameters are shown in Appendix I. Briefly, trimmed short reads were assembled using SparseAssembler (Ye et al. 2012) with a k of 51 and chimeric contig removal (ChimeraTh 2 ContigTh 2). The assembled short-read contigs and raw long reads were assembled using DBG2OLC (Ye et al. 2016) with a k of 17 and chimera removal. Detailed assembly parameters are available in Appendix I. The assembled short reads and raw long reads were concatenated to assess contig consensus using BLASR (Chaisson and Tesler 2012) and pbdagcon (https://github.com/PacificBiosciences/pbdagcon). This process consolidates all the assembled contigs into a final haploid assembly with the help of the original sequences, attempting to merge contigs when possible and remove duplicate ones. Polishing is a process by which genomic sequences are used to correct assembly regions such as unknown bases or gaps, minor assembly errors, small repeat regions, chimeric regions, or low-confidence bases. The corrected long reads were mapped to the consensus sequences using minimap2 (Li 2018), then used to polish the assembly using racon (Vaser et al. 2017). This process was repeated twice more for a total of 3 rounds of long-read polishing. The subsampled short reads were then mapped to the polished assembly using minimap2 and used to polish the assembly with the default parameters of Pilon (Walker et al. 2014).

The corrected long reads were then mapped to the polished genome using minimap2 to scaffold the assembly with LRScaf (Qin et al. 2019). Finally, the subsampled short reads were mapped to the scaffolded assembly using minimap2 and polished once more with Pilon. assembly quality and contiguity was assessed using QUAST (Gurevich et al. 2013). The reported assembly metrics include N50, the length of the shortest contig in a minimum set of contigs that includes 50% of the assembly (i.e., 50% of the nucleotides in the assembly are contained in contigs of length N50 or greater) and L50, the minimum number of contigs necessary to contain 50% of the assembly. Genome completeness was assessed using a core set of conserved domain-specific genes known as Benchmarking Universal Single Copy Orthologs (Simão et al. 2015). Identification of these genes was performed using the busco software, which leverages metaeuk (Levy Karin et al. 2020) to scan for the 3,354 conserved genes in the *vertebrata* database. The full assembly pipeline described in this section can be seen in Figure 2.2.

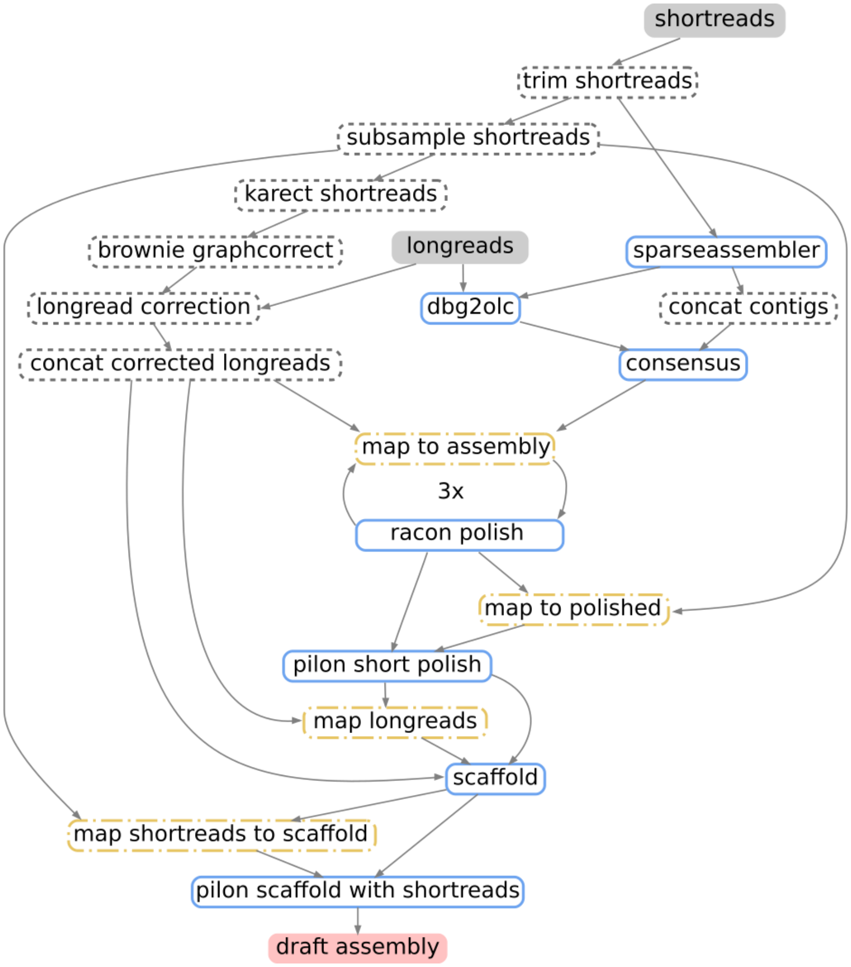

Figure 2.2 *Hybrid assembly of the Thunnus albacares genome*
A directed acyclic graph describing the genome assembly process. Solid grey nodes represent the raw sequences, nodes outlined in short dashes indicate sequence processing steps, nodes with solid blue outlines indicate assembly steps, and nodes with alternating short and long dashed outlines indicate sequence mapping steps. Lines indicate a direct dependency of the output of one step to the input of another, with arrow heads describing the direction of input to output.

2.3.3 High Density Linkage Map

2.3.3.1 Sample Acquisition and Sequencing

Larvae from a single outbred full sibling family cross of Pacific Yellowfin tuna bred in captivity were used to create ddRAD libraries to generate the high-density linkage map. Larvae were reared for 4 days post-hatch to increase the amount of DNA yielded from extraction. They were isolated without food for 3 hours to evacuate their digestive tracts and preserved in 95% ethanol for two days before transfer into 20% DMSO-EDTA. Breeding pairs of parental fish cannot be isolated for spawning; therefore, larvae were obtained from a spawning event in a mass spawning tank that contains 24 brooders (13 females and 11 males). To identify offspring from a single pair within the sampled offspring, genomic DNA was extracted using the Omega BioTek Mag-Bind Blood & Tissue DNA HDQ 96 Kit (cat. M6399-01) and 384 individual larvae were first genotyped at 16 previously developed 16 microsatellite markers (Antoni et al. 2014). Microsatellite primer sequences, specific annealing temperature, and fluorescent labeling for detection during electrophoresis are described in Antoni et al. (2014). Amplification products were electrophoresed on an ABI-377-96 automated sequencer (Applied Biosystems), electropherograms were processed using Genescan (v3.1.2), and alleles were called using Genotyper v2.5 to establish genotypes for each individual. Genotypes were used to assign larvae siblingship using the maximum likelihood approach implemented in COLONY (Jones and Wang 2010). When a full sibling family with at least 200 members was identified, it was selected to construct the linkage map. Family members were sequenced using the ddRAD sequencing protocol (Figure 2.3). DNA of individual larvae had to be amplified with Repli-G (Qiagen) following the manufacturer’s protocols to reach the minimum quantity needed for ddRAD library preparation (650 ng). The amplified DNA was digested using two restriction endonucleases (EcoRI and MspI) and adapter oligonucleotides with unique 6 bp barcodes were ligated onto the ends of the DNA fragments. Each sample was given a unique barcode so DNA fragments can be associated with each sample after sequencing. Additionally, these adapters contained randomized 8 bp Universal Molecular Identifiers (UMI) in fixed locations for downstream quality filtering to identify PCR duplicates. After adapter ligation, DNA fragments were PCR amplified and those ranging in size between 300 and 500 bp were size selected and retained for sequencing. Libraries were sequenced on an Illumina NovaSeq 6000 platform to generate paired end reads (2x150 bp) with a target sequencing depth of on average 6 million paired reads per sample.

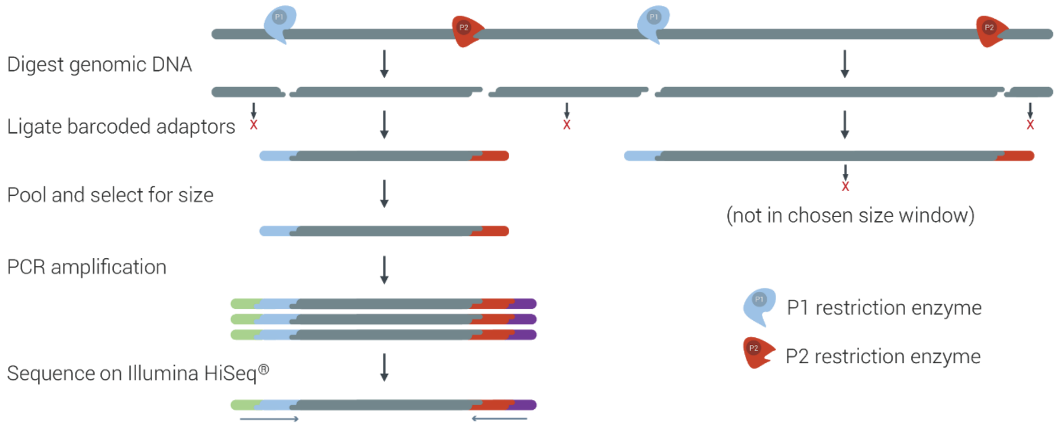

Figure 2.3 *Double Digest Restriction Associated DNA*
Genomic DNA is digested with two specific restriction enzymes and unique barcodes are ligated onto the fragmented ends. These modified DNA fragments are screened to be within a specific length range, then those passing screening are PCR amplified before sequencing.

2.3.3.2 Sequence Processing and SNP Discovery

Barcodes were used to demultiplex raw reads and assign them to individual samples. The reads were quality trimmed using fastp as above to retain only high-quality base pair calls in sequences greater than 50 bp. PCR duplicates were identified using UMIs and removed in fastp. The resulting filtered reads were then mapped to the draft genome assembly described above using BWA-MEM, and variants were called using the local haplotype variant caller Freebayes (Garrison and Marth 2012), as implemented in the dDocent pipeline (Puritz et al. 2014).

2.3.3.3 Data Filtering

The resulting VCF file was screened with VCFTools (Danecek et al., 2011) to remove markers with a depth below 10 and genotype quality below 20. The data were then filtered to remove markers with a minimum allele frequency <0.001 to screen out monomorphic loci that may have been generated by the previous filter. Subsequent filtration steps removed sites with >50% missing data, followed by individuals with >30% missing data, and sites with overall quality <20. Sites covered with a mean depth greater than twice the standard deviation of the mean site depth for the entire dataset that potentially represented repeated DNA (https://www.ddocent.com/filtering/) were removed. The next filtration step removed individuals with >60% missing data. One of the parents would have been screened out by the latter filter and was kept in the dataset. Complex haplotypes (multi-nucleotide polymorphisms) were deconstructed into SNPs using vcfalleleicprimatives from vcflib (https://github.com/vcflib/vcflib), and indels were removed restricting the data to only biallelic markers. The final biallelic SNP data was filtered to remove sites with >10% missing data.

2.3.3.4 Linkage Map Construction

Linkage analysis was performed using the software LepMap3 (Rastas et al. 2013; Rastas 2017). The ParentCall2 module was used to generate the input pedigree file from the filtered SNP data and remove non-informative markers. Loci were assigned to linkage groups in the module SeparateChromosomes2. Linkage group assignment was based on a critical LOD score (Morton 1955) computed from the distribution of LOD scores within the dataset. SeparateChromosomes2 was performed iteratively for varying critical LOD scores values ranging from 1 to 40. The final LOD score for assignment to linkage groups was chosen as the highest value partitioning of markers into 24 major linkage groups. Separate male, female, and sex averaged maps were generated by analyzing the segregation of markers that were informative in the male parent, the female parent, or both parents respectively.

Loci were ordered within linkage groups and phased using the OrderMarkers2 module and the algorithm described by McWilliam et al (2016) and the distances between markers were computed as Kosambi distances (Kosambi 1943). The ordering was replicated 100 times for each linkage group, and both ends of each linkage group were trimmed of poorly mapped marker clusters; tail end clusters that were more than 5 cM map distances apart from the next internal marker were removed (P. Rastas personal communication). The map with the highest likelihood was refined during 50 additional iterations of OrderMarkers2 using the additional parameters evaluateOrder to calculate map distances and improveOrder to improve existing map ordering.

2.3.4 assembly Orientation and Anchoring

The linkage maps generated from Lep-Map3 were used as input into Lep-Anchor (Rastas 2020), which uses linkage information to order and orient contigs of the genome assembly. Per the input requirements of Lep-Anchor, the draft assembly was repeat-masked using Red (Girgis 2015) and the chainfile was created per the recommendation of Lep-Anchor using LASTZ (Harris 2007) as implemented in HaploMerger2 (Huang et al. 2017). The input distance between markers and their physical positions was obtained from Lep-Map3. The corrected PacBio long reads described above were mapped onto the genome assembly using minimap2 to generate an alignment file to assist in contig anchoring. The PlaceAndOrientContigs procedure of Lep-Anchor was performed iteratively 3 times. Haplotigs (variant contigs) discovered by Lep-Anchor were removed prior to the initial iteration of PlaceAndOrientContigs and new haplotigs discovered during the initial iteration were removed prior to running the second iteration. A final iteration to improve ordering was completed without removing any additional contigs beforehand. Finally, the marker positions on the male, female and sex-averaged maps were converted to reflect the final anchored assembly (lifted) and the ends of each linkage group on each map were scanned for the presence of isolated clusters of markers at the end of each linkage group. When such clusters were found and they were separated from the immediate next markers in the linkage group by more than 5% of the total length of the linkage group in centiMorgans, they were removed as recommended by P. Rastas (personal communication). These analyses were performed using LepWrap (Dimens 2022), an executable Snakemake workflow (Köster and Rahmann 2012) developed during this work to facilitate the use of the various modules of both Lep-Map3 and Lep-Anchor.

2.3.5 Synteny

The obtained draft genome was compared against the medaka (*Oryzias latipes*) to assess chromosome level syntenic relationships. LAST (Kiełbasa et al. 2011) was used to align the 24 pseudochromosomes of the anchored assembly to the medaka genome (NCBI accession GCF_002234675.1). Alignments with estimated probabilities of mapping to a different part of the genome (“mismap”) greater than 10^{-5} were removed and the remaining alignments were converted to blasttab format. The alignments were then filtered to remove sequence overlaps below 200 bp and percent identity below 75%, and visualized using the circlize package (Gu et al. 2014) in the R statistical programming language (R Core Team 2013).

2.4 Results

2.4.1 Genome Size and assembly

Long read sequencing yielded 8,809,334,274 bp across 1,229,738 reads for the first sample and 19,172,364,391 bp across 4,398,082 reads for the second sample for a total of approximately 27.9 Gb across 5.6m reads (N50 = 9,928 bp). Hybrid long read error correction allowed correcting 1,388,026 reads, which were combined with the remaining uncorrected long reads for a final long read dataset covering 21,774,870,233 bp across 6,060,123 reads. Short read sequencing yielded 235,120,433,564 bp across 778,544,482 paired-end reads and 227,548,533,460 bp across 769,372,246 paired-end reads after trimming. Down sampling the short reads yielded 56,880,326,564 bp across 192,320,736 paired end reads. Estimates of genome size obtained varying K between 17 and 25 ranged from 757,518,950 bp to 775,865,895 bp, with a mean of 769,491,698 (σ = 5.93 Mb).

The assembly of short reads in SparseAssembler produced 4,386,682 contigs spanning a total length of 1,000,382,856 bp. The were 5,594 and 20,685, respectively. assembly of the short-read contigs obtained from SparseAssembler with the uncorrected long reads in DBG2OLC yielded 5,420 contigs (N50 = 342,887, L50 = 636) spanning 763,219,229 bp. The consensus assembly obtained from pbdagcon analysis of BLASR alignments reduced the assembly to a total length of 732,665,565 bp contained within 5,193 contigs (N50 = 339,167 bp, L50 = 622). After iterative mapping and polishing the assembly using corrected long reads and short reads, and scaffolding, the final draft assembly spanned 745,092,531 bp contained in 5,193 scaffolds (N50 = 351,587 bp, L50 = 608); the complete and partial BUSCO scores were 86.47% and 3.63%, respectively.

2.4.2 High Density Linkage Map

Most of the larvae genotyped (302 out of 384) were inferred to belong to one single full sibling. Individual larvae from the full sibling set with highest DNA quality (274) and the two parental DNA samples were processed through the ddRAD sequencing protocol to discover and genotype SNP markers. Sequence quality control, mapping and variant calling yielded 3,531,199 initial variant sites. Filtering of this initial dataset retained 19,469 biallelic SNP loci genotyped in 166 individuals (164 F1 offspring and the two parents).

Iterative runs of SeparateChromosomes2 increasing the values of the threshold LOD score for assignment of markers to linkage groups recovered 24 major groups when the threshold LOD score was 32. Applying this value clustered 16,244 informative markers into 24 linkage groups (Table 2.1, Figure 2.4). The mean number of recombinations per linkage group ranged from 1.77 to 2.73 ($\sigma = 0.23$), with a maximum of 7 recombinations within a linkage group, and a minimum recombination occurrence of 1. Prior to anchoring the genome, the female linkage map (2,707.76 cM, $\mu = 101.45$, $\sigma = 16.54$) was larger than the male linkage map (2,434.96 cM, $\mu = 112.82$, $\sigma = 22.53$), but after anchoring and lifting over coordinates, the male map (1,243.82 cM, $\mu = 51.82$, $\sigma = 7.18$) was slightly longer than the female one (1,222.9 cM, $\mu = 50.95$, $\sigma = 4.44$), (Table 2.2). The average interval between markers was 0.0902 cM ($\sigma = 0.0119$), 0.0912 cM ($\sigma =$

0.0127), and 0.0907 cM (σ = 0.0101) for the female, male, and sex averaged maps, respectively.

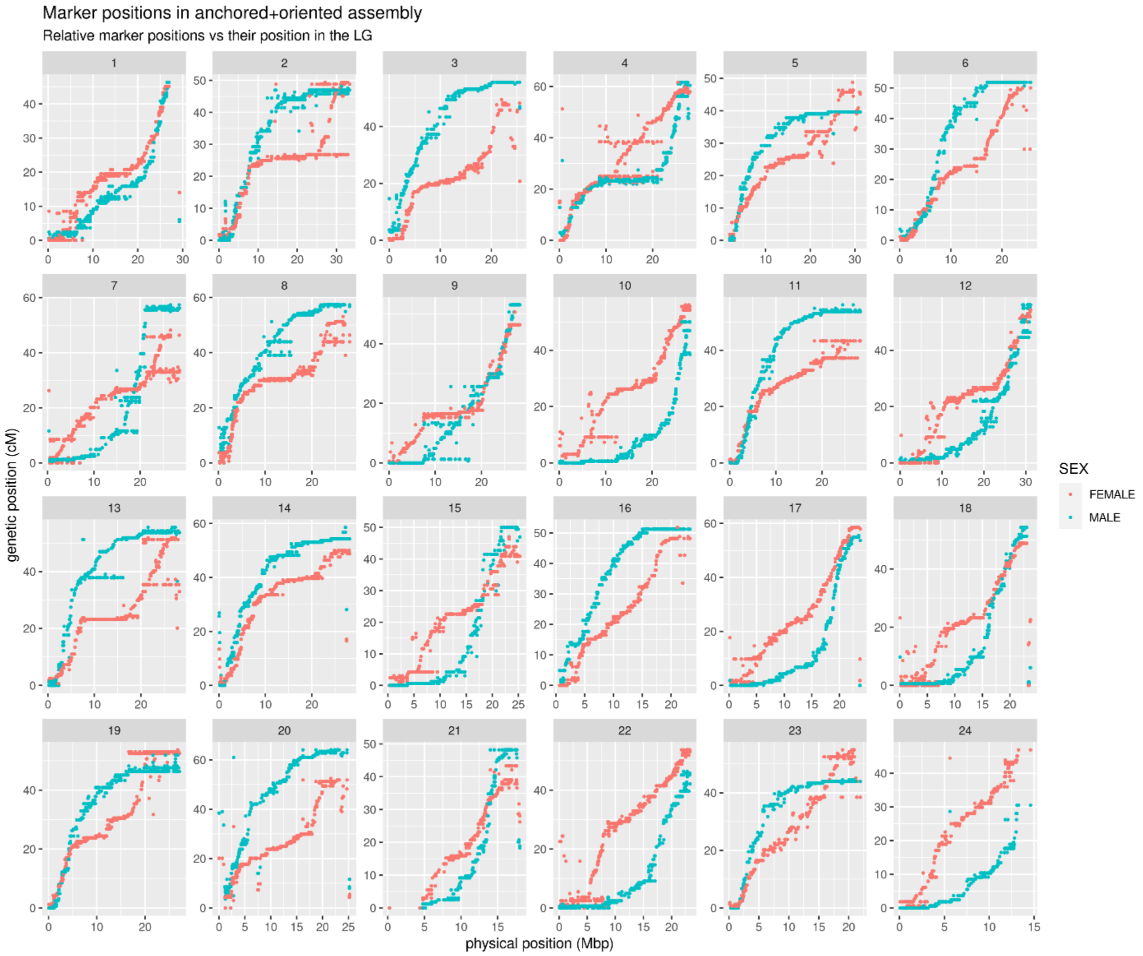

Figure 2.4 *Correspondence between marker positions on the physical map and position on the male (in blue) and female (in red) linkage maps*

Each plot represents a linkage group, identified by the number in the grey rectangle abutting the top of each plot. Every point represents a single ddRAD locus, with the position being its physical location in base pairs versus its genetic position in centiMorgans.

Table 2.1 *Distribution of SNPs clustered into linkage groups*

Linkage Group	SNP Count	Final SNP Count	Mean kb	SD kb
unclustered	1437	-	-	-
1	830	634	45.97	117.17

Table 2.1 (continued).

2	796	708	47.16	88.07
3	775	672	38.11	71.88
4	767	684	41.05	74.48
5	738	543	53.96	88.56
6	715	588	43.51	85.28
7	712	594	47.53	85.44
8	689	625	44.75	82.44
9	686	573	49.28	81.22
10	685	595	47.38	83.25
11	681	552	51.39	86.76
12	680	645	48.36	81.77
13	669	583	48.23	86.56
14	663	593	47.55	84.25
15	660	531	47.16	83.74
16	644	504	45.06	76.76
17	638	556	42.57	72.51
18	626	577	41.13	69.34
19	625	577	46.93	72.86
20	612	513	49.23	84.22
21	604	444	40.2	206.4
22	594	554	41.98	66.62
23	580	463	47.12	83.03
24	555	431	33.82	85.17
global average	706	572	45.39	91.35

SNP count refers to the number of markers clustered into linkage groups from Lep-Map3, whereas Final SNP Count refers to the final number of markers after removing spurious edge clusters when anchoring and orienting the assembly. Mean kb is the average kilobase distance between adjacent markers in a linkage group and SD kb is the standard deviation of the kilobase distance between adjacent markers in a linkage group.

Table 2.2 *Map lengths (Kosambi map function) of linkage groups in centimorgans (cM)*

	Draft assembly		Anchored and Oriented	
Linkage Group	Male	Female	Male	Female
1	105.239	80.209	46.356	45.136
2	78.361	121.157	48.182	49.407
3	84.391	107.261	55.505	49.412
4	106.968	150.924	61.62	61.602
5	74.364	130.538	39.642	48.793
6	106.761	121.733	51.849	50.013

Table 2.2 (continued).

7	103.792	139.466	57.38	48.181
8	98.627	108.912	57.336	53.062
9	87.299	120.526	53.063	50.628
10	90.11	122.008	50.021	56.11
11	133.071	102.924	56.113	43.301
12	103.451	94.535	56.127	54.295
13	112.012	116.116	55.506	51.843
14	112.573	103.513	58.566	50.02
15	83.026	103.252	50.028	46.963
16	105.354	133.205	51.235	51.842
17	87.528	150.584	55.512	58.548
18	87.33	124.95	54.29	48.792
19	113.06	80.344	51.847	53.669
20	115.618	111.621	64.05	52.453
21	119.53	71.375	48.191	43.302
22	108.504	84.804	46.363	53.669
23	136.42	142.408	44.535	54.895
24	81.574	85.393	30.499	46.96
Total	2434.96	2707.76	1243.82	1222.9

Map lengths are separated by sex to illustrate the differences between the male and female linkage maps. Map lengths are also presented before and after assembly anchoring, as anchoring genomes reduces map lengths.

Anchoring the draft assembly using the male and female linkage maps revealed 80 putative haplotigs which were removed from the assembly. The anchoring process oriented and combined 2,553 contigs (49.2% of the total number of contigs in the assembly) into 24 pseudochromosomes spanning 625,893,634 base pairs (79.2% of the assembly). Scaffolding the contigs using the linkage map led to a final N50 of 26.5 Mb (~75x greater than the N50 of the pre-anchored assembly), a L50 of 13 (~46x improvement), and a marginal increase of the BUSCO scores (Table 2.3). The average size of each linkage group was 50.95 cM (std = 4.44) and 51.82 cM (std = 7.18) for female and male linkage maps, respectively. The anchored genome assembly had a duplication rate of 76.15%, meaning that ¾ of the genome is composed of paralogous or repeat regions, which can arise in eukaryotes from genetic duplication, mutation, or DNA repair mechanisms. A summary of assembly metrics is shown in table 2.3.

Table 2.3 *Descriptive summary statistics of the Yellowfin tuna genome assembly*

	Draft assembly	Anchored assembly
expected genome length (bp)	790,000,000	790,000,000
unanchored contigs	5,193	2,640
% anchored contigs	0	49.2
linkage groups	-	24
bases in linkage groups	-	625,893,634
% genome anchored	0	79.2
Total contigs/scaffolds	5,193	2,664
Largest contig	1,896,512	33,236,316
contigs > 1000 bp	5,190	2,661
contigs >5000 bp	5,058	2,540
contigs >10000 bp	4,768	2,277
contigs >25000 bp	3,652	1,358
contigs >50000 bp	2,780	728
Total length	744,983,845	743,073,847
Total length >1000 bp	745,089,982	743,071,336
Total length >5000 bp	744,645,873	742,671,907
Total length >10000 bp	742,395,447	740,633,239
Total length >25000 bp	723,658,933	725,347,530
Total length >50000 bp	692,785,422	703,298,802
GC Content	40.10%	40.08%
N50	351,587	26,516,309
N75	164,428	23,232,453
L50	608	13
L75	1,374	21
N's per 100 kb	53.28	85.91
genome duplication	-	76.1%
complete BUSCO	83.8%	83.9%
partial BUSCO	8.4%	8.2%

This table compares the metrics of the Yellowfin tuna assembly before and after anchoring the genome using the linkage map. Contigs refer to contiguous assembled segments of DNA, lengths are described as base pairs (bp) or kilobase pairs (kb).

2.4.3 Synteny

Alignment of the Yellowfin tuna linkage groups to the Medaka reference genome yielded 272,277 syntenic alignments spanning a total of 76,169,144 bp after mismap filtering. Setting percent identity and alignment length thresholds retained 69,677 alignments spanning 28,649,042 bp. Alignment revealed a general 1-to-1 association between the assembled Yellowfin tuna linkage groups and the chromosomes of the medaka assembly, although inversions and translocations were evidenced affecting small portion of the chromosomes (Figure 2.5, Figure 2.6). The filtered alignments accounted for 4.57% of the Yellowfin tuna assembly forming syntenic regions with the Medaka chromosomes, with a mean of 4.51% (range = 3.01-6.03, $\sigma = 0.95$) of each Yellowfin tuna linkage group aligning to Medaka chromosomes.

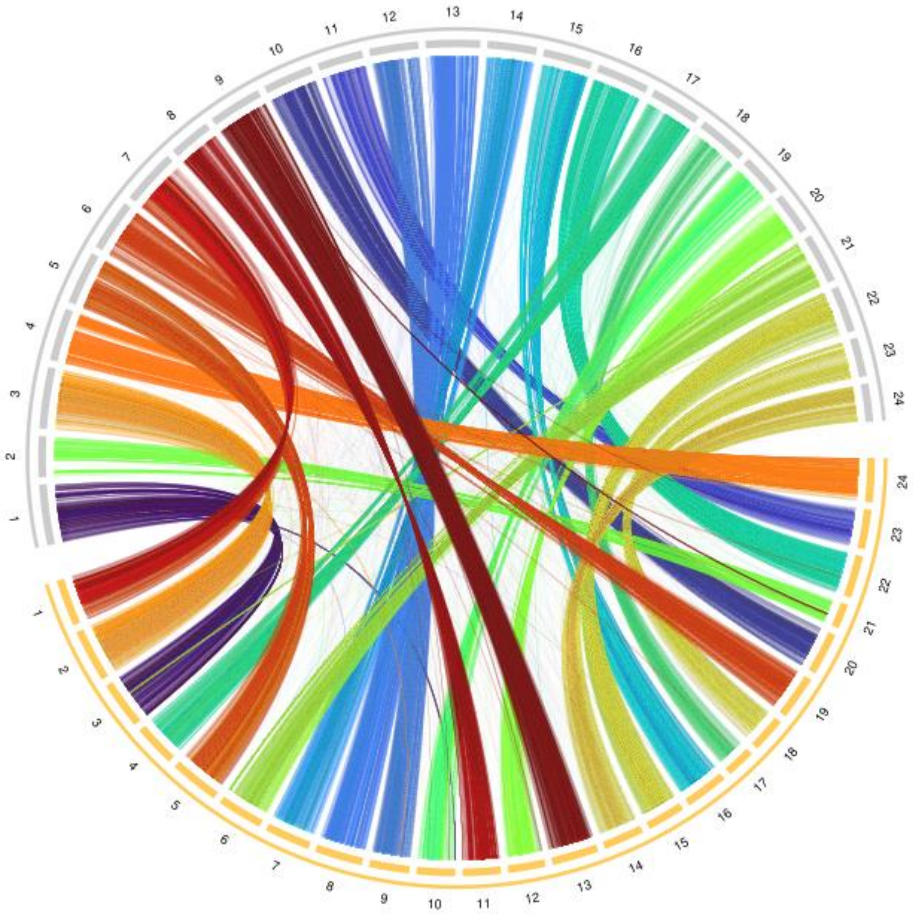

Figure 2.5 *Diagram representing syntenic blocks between the Yellowfin tuna and Medaka genomes.*

Each line represents a genomic sequence matched between the two genomes. Line width reflects the length of the alignment. Each chromosome is represented as a labelled block on the outer ring, with the bottom (yellow) corresponding to the Yellowfin tuna and the top (grey) corresponding to the medaka.

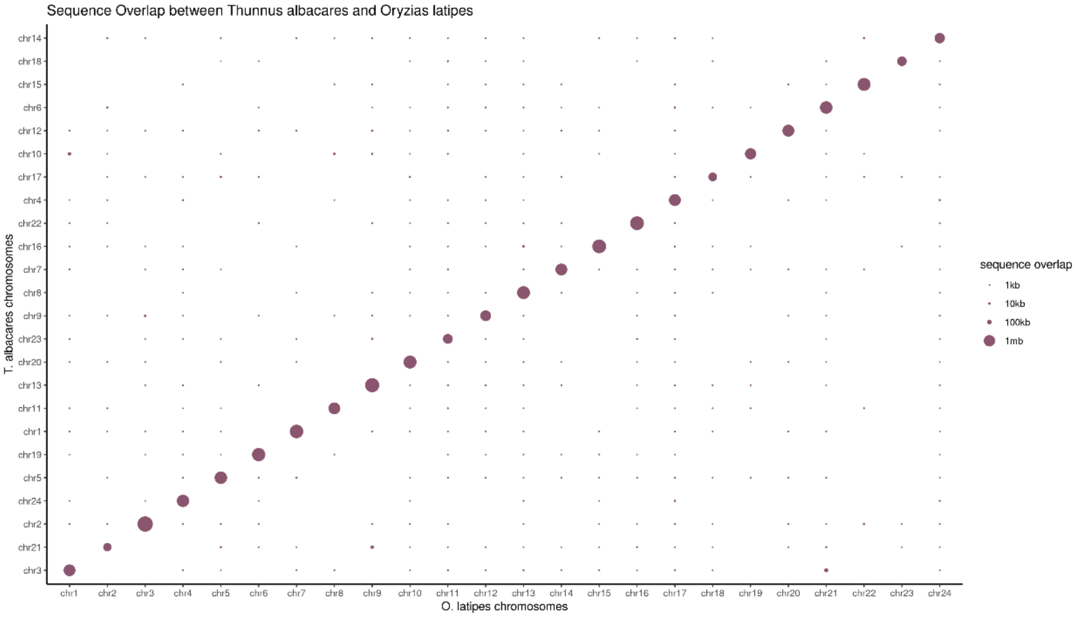

Figure 2.6 *Oxford grid dot plot depicting syntenic regions between Yellowfin tuna and Medaka*

Another representation of the syntenic blocks between Yellowfin tuna and medaka. Each point represents the number of base pairs (bp) of sequence overlap between the two species for those chromosomes. The x-axis is sorted in the order of chromosomes 1-24 in the Medaka genome, whereas the y-axis is sorted such that the greatest sequence overlaps correspond to the order of Medaka chromosomes on the x-axis to emphasize the visual analogues.

2.5 Discussion

The first objective of this work was to develop a reliable reference to map genotyping by sequencing reads during genome scans of Yellowfin tuna. Before anchoring with the linkage map, 93% of the total length was contained in 2,780 contigs with a N50 of 351,587 bp. Genomic windows of 200 kb or less are considered sufficient to capture the signal of selective sweeps in most cases (Catchen et al. 2017) and the assembled contigs from this assembly are therefore expected to allow mapping short sequencing reads from genotyping by sequencing surveys for SNP discovery across most of the genome in contigs sufficiently large to assess genomic regions affected by selection around them. These contigs were corrected for potential misassemblies through multiple rounds of polishing with short highly accurate Illumina sequencing reads as well as long reads. The obtained assembled contigs are therefore expected to provide a reliable reference to map sequencing reads during genomic studies of Yellowfin tuna.

The hybrid assembly was improved with the integration of the linkage map, resulting in approximately 79% of the genome residing in the 24 linkage groups corresponding to the expected 24 chromosomes. Anchoring led to significant improvement of the contiguity (92% of the genome contained in 728 scaffolds, N50 over 26 Mb and a final BUSCO score of 83.9%). The linkage map included 13,739 markers yielding an average interval between markers of 45.39 kb (σ = 91.35, Table 2.1). This interval is also compatible with detection of selective sweeps spanning tens of kb but could impact the detection of genomic regions affected by narrower sweeps or the separation of distinct but proximal ones, therefore caution needs to be exercised when assessing multiple genomic regions under selection within a chromosome.

The statistics of the assembly, prior to anchoring and scaffolding with the linkage map, were in the range of hybrid assemblies of fishes using similar sequencing strategies. An assembly of the European eel (*Anguilla anguilla*) using the TULIP hybrid sparse assembler with 18x Nanopore reads spanned 891.7 Mb in 2366 scaffolds (N50 = 1.23 Mb, Jansen et al. 2017). The assembly completeness, as reported with BUSCO assessment was 77.5% complete, 14.1% fragmented, and 8.4% missing. The sea lamprey (*Petromyzon marinus)* genome was assembled using 100x Illumina sequences, 300x Illumina 4 kb mate-pair reads, 600x Illumina 40 kb mate-pair reads and 17x Pacific BioSciences long reads (Smith et al. 2018). The assembly also incorporated 56x BioNano optical mapping and 325 million Hi-C reads. This approach heavily favored the use of mate-pair and linked reads to maximize long distance read information and produced an assembly of 34 super-scaffolds comprising 12,077 contigs (N50 = 12 Mb), with a BUSCO assessment of 90% completeness of vertebrate orthologs. The assembly presented in this work strikes a balance between these two examples, achieving near chromosome-level contiguity and completeness at a fraction of the cost and sequencing effort of the 34 super-scaffold assembly generated for the sea lamprey. Regarding the other tunas, the Southern Bluefin tuna (*Thunnus maccoyii*) was assembled into 54 scaffolds (N50 = 33.7mb, McWilliam et al. 2016), the Pacific bluefin tuna was assembled into 444 scaffolds (N50 = 7,922,002, accession PRJEB46021), and the Atlantic bluefin tuna (*T. Thynnus*) was assembled into 354,425 scaffolds (N50 = 3,045, Puncher et al. 2018), situating the (unanchored) Yellowfin tuna assembly as the third most contiguous tuna assembly to date.

The hybrid strategy employed in this study has been suggested to be a cost-effective solution to generate highly contiguous de novo genome assemblies in a variety of taxa including fish (Jaworski et al. 2019; Wiley and Miller 2020; Tan et al. 2018). In the hybrid assembly process, short reads first need to be self-assembled. The produced contigs are expected to contain few misassemblies thanks to the high accuracy of short reads and the high coverage usually achieved for most sequenced regions with Illumina sequencing. The short read assembly obtained in this work is usually highly fragmented (4.3 m contigs, N50 = 3,702, L50= 45,825), in part because of challenges assembling genomic regions featuring repeated elements (Tørresen et al. 2019). Self- assembly of the long reads was also attempted using Canu (Koren et al. 2017) prior to hybrid assembly but did not complete because there were too few long reads remaining after the self-correction process. The low yield of quality long reads after correction reflected the overall low success sequencing long reads in this study. Multiple tissue type and preservation and DNA extraction methods were attempted to improve the quality of the template DNA used for PacBio sequencing, but templates remained partially degraded with low yield of large fragments for SMRT sequencing after size selection. The library obtained for the first sample could only be size selected to retain fragments larger than 10,000 bp and still yielded an exceptionally low sequencing output which led to generating additional long sequencing reads from a second sample. The second sample featured a higher frequency of long fragments allowing size selection of the library to retain fragments larger than 20 kb but the yield in long molecule sequences remained low after correction.

The contiguity of the assembly could be improved by increasing the number of long reads. The Oxford Nanopore sequencing platform generates cost effectively ultra-long sequencing reads with error rates comparable to those of PacBio sequencing in its current implementation (Dumschott et al. 2020). Improving the DNA quality or at least quantity for this sequencing would also be valuable. Multiple extractions could be combined and size-selected to yield enough fragments for sequencing and assembly. Recent developments to achieve chromosome-level assemblies are also hybrid methods, which use Pacific BioSciences HiFi technology and scaffold using linked short reads (Hawkes et al. 2021; Lohse et al. 2021). These methods tend to produce chromosome-resolved assemblies through scaffolding. The HiFi approach is still cost prohibitive for larger eukaryotic genomes and the 10x Genomics link reads approach was unavailable at the time of sequencing for this project, but these approaches would be worth implementing to improve the scaffolding of the current assembly.

Another approach to improve the current reference is to perform a reference-guided assembly using the *Thunnus maccoyii* genome assembly (McWilliam et al. 2016) to potentially achieve higher initial contiguity. The reference guided approach was not attempted in this work because the Yellowfin tuna assembly was also intended for whole-genome phylogenetic reconstruction of the *Thunnus* genus, and it was imperative that the composition of the Yellowfin tuna genome assembly was not influenced by that of any congeners.

Linkage mapping recovered 24 major linkage groups consistent with the expected number of chromosomes in *T. albacares* (Lee et al. 2018). The male maps were slightly longer than the female maps, but most of the linkage groups have segments where markers in either the male or female maps could not get positioned confidently (e.g., LG13, female map, markers with coordinates >20 Mb on the physical map, Figure 2.4). This phenomenon was observed to some degree in almost half of the linkage groups (Figure 2.4), with the regions spanning as few as 10 cM/8 Mb (LG1) to as great as 25 cM/15 Mb (LG2). While the origin of these ambiguous marker placements is unclear, it may be related to missing data, particularly among the dam and sire, or to missassemblies in regions that are difficult to assemble, such as centromeres and telomeres. It is also possible that the whole-genome amplification led to large sections of the genome unrepresented in the amplicons leading to low or no SNPs available for mapping in these regions. A future improvement would be to optimize the ddRAD protocol to reduce the required amount of input DNA. Doing so would allow using larvae only a few days post-hatch without incurring the costs of additional rearing time and genome amplification reagents. Requiring less starting DNA for a ddRAD protocol would result in independence from whole genome amplification, reducing fragment representation bias that results from any DNA amplification (Becker et al. 2000; Aird et al. 2011).

Each linkage group in the Yellowfin tuna overwhelmingly aligned to a corresponding Medaka chromosome, often with only trace fractions of the alignments assigned to non-syntenic chromosomes. The last common medaka-Tetraodon-zebrafish ancestor (MTZ) occurred 336-404 million years ago (Kasahara et al. 2007), and had 24 chromosomes like the tuna species. The medaka is thought to have largely preserved the ancestral genomic arrangement for over 300 million years. Current evidence suggests that the last common ancestor between the medaka and the tunas occurred approximately 116.4 million years ago during the cretaceous period (Betancur-R et al. 2017) and the present findings suggest that tunas also largely retained the ancestral arrangement. Rearrangements were few but present and their size did not deviate from the average syntenic block size. However, caution needs to be exercised regarding inferences on the abundance and size of rearrangements because individual syntenic blocks tended to be short (only a few hundred base pairs), and there was an overall low percentage of confident alignment between the two species. Thus, this study suggests that small rearrangements such as translocations and/or inversions have occurred during evolution of the medaka and/or tunas from their common ancestor, but these are yet to be explored by dedicated methods and likely more suited for a study using a more complete genome assembly.

CHAPTER III – POPULATION STRUCTURE OF THE YELLOWFIN TUNA, THUNNUS ALBACARES, IN THE ATLANTIC OCEAN BASIN

3.1 Introduction

Effective management of marine fisheries is heavily dependent on a robust understanding of the structure of their metapopulations (Carvalho and Hauser 1994). The populations of many marine fishes are expected to display a high degree of connectivity across large geographic areas due to the open nature of marine habitats and the high dispersal potential of many species (Avise 1998; Waples 1998). These characteristics, combined with the large effective size of populations render studies of the genetic structure of marine metapopulations challenging due to the very slow and ultimately low levels of divergence among demes. These factors explain the lack of apparent structure, sometime over broad areas, reported in many marine species that recently observed (post-glacial) range expansion and isolation of populations but have not reached equilibrium due to insufficient time for genetic differences to accumulate between demes (Pruett et al. 2005; Domínguez-López et al. 2015; Cheng et al. 2018). These limitations are particularly important for highly migratory species such as marlins, tunas, or sharks (Chapman et al. 2015) that have high movement capability and, in some cases, large populations. While divergence among populations due to genetic drift is expected to be slow in many marine fishes as discussed above, natural selection has been proposed to be a major factor structuring these species and can lead to faster genetic change (Avise 1998; López et al. 2014), although in highly connected metapopulations such as those formed by large migratory fishes, gene flow may be sufficient to counterbalance effects of divergent selection and local adaptation (Lenormand 2002). Overall, a prediction that can be made, and has been verified in most studies of highly migratory pelagics, is that divergence among connected populations is weak and the signature of local adaptation may be restricted to genes experiencing very strong selection (Anderson et al. 2019; Graves 1998). Recent genetic studies did challenge the hypotheses that highly migratory pelagics formed panmictic populations across entire oceanic basins. For example Atlantic bluefin tuna Thunnus thynnus have been shown to be subdivided in an eastern and western stock (Carlsson et al. 2006) and cryptic units within the Eastern stock utilizing different spawning grounds and nursery grounds and/or utilizing habitats at separate times were also evidenced (Riccioni et al. 2010). Seasonal migration (e.g., movement between feeding and spawning grounds) has been hypothesized and observed in a range of marine fishes including tunas (Calvert et al. 2009; TinHan et al. 2018; Mariani et al. 2016), and would prevent detection of a pattern such as the one described by Riccioni et al (2010), which highlights the importance of sampling breeding sites to describe the breeding structure of the metapopulation. If spawning grounds are unknown or reproductively active individuals are difficult to sample, then nursery areas may provide information on metapopulation structure under the assumption that juveniles of the species have more limited movement capacity than the adults and remain proximal to spawning grounds. When these issues prevent sampling candidate demes, Bayesian clustering and relatedness analyses can be used to reveal the occurrence of individuals from demographically independent assemblages from genetic data (Carlsson et al. 2006; Pritchard et al. 2000).

The Yellowfin tuna (*Thunnus albacares*) is a large (~200 cm) epipelagic scombrid distributed in tropical and subtropical waters of all oceans. They are typically found above the thermocline in the top 100 m of surface water (Weng et al. 2009; Hoolihan et al. 2014; Brill et al. 1999) and have been observed in proximity of sargassum mats, sometimes in nearshore waters. Yellowfin tuna are an important species in the Atlantic Ocean for their role as a high-level predator (Buonaccorsi et al. 1999; Graham et al. 2006) with a diet consisting of crustaceans, squid, and fish (Rudershausen et al. 2010; Collette et al. 2011).

Spawning occurs during a protracted summer season (ICCAT 2019) and larvae are documented to metamorphose into juveniles at 30 days old (Kaji et al. 1999) where they show limited movement restricting them to regional nursery areas (Wells et al. 2012). Four nursery areas have been identified in the Atlantic basin, two each of the east and west of the basin. The East Atlantic nurseries include the Gulf of Guinea where spawning occurs from December to April (ICCAT 2019), and the West African coast in the area of Cabo Verde where spawning occurs from April to June (Diaha et al. 2016). In the West Atlantic, spawning occurs in the Gulf of Mexico from May to August (Lang et al. 1994; Franks et al. 2015) and in the Southern Caribbean from July to November (Arocha et al. 2001). Age and growth and reproductive traits have been documented in the West Atlantic (Lang et al. 1994; Fonteneau and Chassot 2013; Brown-Peterson et al. 2013) and East Atlantic (Pacicco et al. 2021; Diaha et al. 2016), although comparison of these traits between the two regions is challenging due to different exploitation rates affecting size distributions and potentially reproductive parameters.

Yellowfin tuna are exploited by major fisheries using a variety of gear including purse seine, longline, handline, and bait boat (ICCAT 2016). Landings for the United States (US) Atlantic Yellowfin tuna fisheries in 2019 were an estimated 730 mt for the combined recreational and commercial fisheries (https://www.st.nmfs.noaa.gov/stocksmart?stockname=Yellowfin%20tuna%20-%20Atlantic&stockid=10166). Abroad, Yellowfin are commonly captured in Angola, Cape Verde, Ivory Coast, the Republic of Guinea, Mexico, and along the South American Atlantic coast (Collette et al. 2011). Landings in 2018 for Senegal and Ivory Coast, both presumed nursery regions, were and 5029 mt (3988 mt in 2017) 116mt (952 mt in 2017) respectively (ICCAT 2019).

Atlantic Yellowfin tuna are managed under the dual authority of the Magnuson-Stevens Fishery Conservation and Management Act (Magnuson-Stevens Act) and the Atlantic Tunas Convention Act (ATCA). The species is currently assessed as a single stock for the Atlantic (ICCAT 2011; ICCAT 2016), despite occurrence of up to four distinct spawning areas associated with different spawning periods and substantial heterogeneity in the distribution of Yellowfin tuna within the basin (ICCAT 2019). Results of earlier stock assessments were ambiguous in that the Atlantic stock was considered overfished under international thresholds (ICCAT 2011). The most recent assessment concluded that the stock was not overfished and overfishing was not on-going (ICCAT 2019), although a gradual decline was noted since the 2004 and 2006 assessments. Accordingly, the species was reclassified as Least Concern with a Decreasing population trend in the most recent IUCN assessment (Collette et al. 2021).

The unique stock strategy currently applied in stock assessment is based on the continuous distribution of the species throughout the entire tropical Atlantic Ocean and on the observation that tags are recovered on a regular base from West to East (ICCAT 2019). Yellowfin tuna connectivity among geographic populations remains poorly understood and better understanding of stock structure is essential to determine if the single stock hypothesis for the Atlantic used in management is appropriate or if multiple stocks need to be considered. Movement across the Atlantic basin at the adult stage have been documented using tag-recapture methods (ICCAT 2019) and reflect the swimming capabilities of the species, which can cover at least 77km in a day (Schaefer et al. 2007), but these data are insufficient to quantify migration rates between regions. Recent investigation indicates that the movement of adult Yellowfin tuna tends to be spatially restricted (Schaefer et al. 2011) and transatlantic movement reported in tagging studies may be anecdotal. Dispersal could also occur at the larval stage through passive transport by surface oceanic currents, but the period of reduced motility is expected to be restricted to the few weeks of larval development, likely limiting transport to a few hundred km at best. Studies using natural tags indicated that yellowfin tuna show limited movements as juveniles (Wells et al. 2012) and remain in nursery areas colonized by post larvae resulting in locally recruiting stocks.

Most of the previous genetic studies of Yellowfin tuna were restricted to investigating variation between oceanic basins (Ward et al. 1997; Ely et al. 2005) or surveyed variation within the Pacific or Indian oceans (Appleyard et al. 2001; Dammannagoda et al. 2008; Aguila et al. 2015; Guo et al. 2016). Three of the studies documenting genetic structure of Yellowfin tuna in the Atlantic Ocean thus far (Talley-Farnham et al. 2004; Ely et al. 2005; Pecoraro 2016) suffer from low sample size per locality, sampling inadequate to describe the breeding structure, and low marker densities. A recent genetic study conducted in our laboratory used 16 microsatellites and larger sample sizes but did not reveal any clear pattern of divergence within the Atlantic basin either (Franks et al. 2015). However, the observation of significant spatial and temporal autocorrelation of genotypes suggested the occurrence of demographic assemblages with different habitat usage patterns. The study did not assess the role of natural selection on population structure. A recent study targeting juveniles by Pecoraro et al (2018) employed SNP markers to characterize Atlantic Ocean populations and revealed possible structure between the eastern and western Atlantic nursery areas. The structure was primarily supported by the divergence of the two groups at genetic markers putatively under divergent selection (outlier loci). However, the study was based on a relatively low density of markers (less than 1,000) and a highly fragmented short-read assembly (Malmstrøm et al. 2016) for read mapping and SNP calling, which reduced the power of the dataset to detect structure. The study also did not sample the putative nursery offshore West Africa and Cabo Verde and did not repeat sampling temporally. Temporal repetition of sampling seems important in this species considering the spatial and temporal autocorrelation noted by (Franks et al. 2015) and reported in other tunas (Carlsson et al. 2006).

In summary, the most recent work suggested breeding structure related to presumed spawning grounds or associated nursery areas is occurring and the four main nursery areas hypothesized in the basin (east: Gulf of Guinea and Cabo Verde, west: Gulf of Mexico and southern Caribbean; ICCAT 2016; ICCAT 2019) should be characterized to comprehensively describe the Atlantic metapopulation. The findings of Pecoraro et al (2018) suggest that adaptive variation may be a primary driver of population structure and should be also characterized. However, the findings of Franks et al (2015) suggest that cryptic structure may also be occurring and could be uncovered by a high-resolution genome scan.

3.2 Objective

The objective of this chapter is to describe the genetic stock structure of Yellowfin tuna in the Atlantic Ocean, namely identify sub-units of the stock if they exist and describe patterns of gene flow among them. Sampling was repeated for two consecutive years to assess temporal stability of the observed patterns. Based on previous genetic work (Pecoraro et al. 2018) and the presence of multiple potential spawning and nursery areas along the east and west Atlantic Ocean, I hypothesize that discrete demes occur corresponding to the four breeding regions yet showing some degree of connectivity, forming a metapopulation in the Atlantic Ocean.

3.3 Methods

3.3.1 Sample Acquisition

Because of practical challenges effectively sampling adults on spawning areas in the Atlantic through the protracted spawning season of Yellowfin tuna, candidate breeding stocks were characterized as juveniles collected in the nearby nursery areas. Samples from the four nursery areas hypothesized for Atlantic Yellowfin tuna (off the West African coast, the Gulf of Guinea, the Gulf of Mexico, and the Southern Caribbean, ICCAT 2016), were collected during fisheries dependent surveys conducted in these regions and provided to USM for genetic characterization. Sampling targeted juveniles (target young of the year or age 1, less than 50 cm) in Senegal (West Africa), Venezuela (South Caribbean), and the Gulf of Mexico (Table 3.1, Figure 3.1). Because only 8 juveniles were obtained from the Gulf of Mexico, this nursery area was also characterized using larvae collected between 2010 and 2015 and provided by Dr. J. Rooker from Texas A&M University. Samples from the Gulf of Guinea nursery (Ivory coast) were only provided late in this project (2020). Therefore, samples from two size-groups were analyzed (< 50 cm, 50-75 cm) to recover two cohorts of juveniles from the area.

Table 3.1 *Sample counts per location*

	Sampling Location	Abbreviation	Sequenced	Passed QC
🟨	Western Atlantic Ocean	ATL	114	77
🟨	Gulf of Mexico	GOM	113	64
🟨	Louisiana	LA	29	17
🟨	Mississippi / Alabama	MSAL	13	7
🟨	Texas	TX	115	31
🟨	Venezuela	VZ	88	82
🟦	Senegal	SEN	104	68
🟦	Côte d'Ivoire	IVC	94	72

Sampling location colors indicate whether those sites are in the western (yellow) or eastern (blue) Atlantic Ocean.

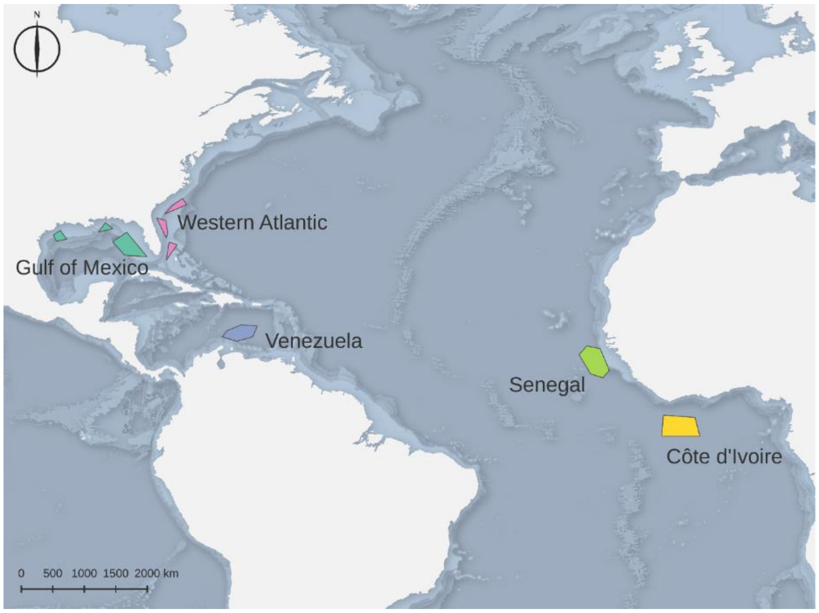

Figure 3.1 *Sample acquisition locations of Yellowfin tuna*
Location of capture was not available for every individual; therefore, polygons reflect general fishing areas and not the specific extents to which each locality was sampled.

Adults captured in the Gulf of Mexico and along the US east coast by US fishers were provided by the National Oceanic and Atmospheric Association pelagic observer program and analyzed to characterize the composition of US fishery and potential cryptic patterns within adults.

Tissue collections from all fisheries dependent sampling occurred between 2015 and 2020, with a target of up to 50 individuals per location for two years for each nursery area, and up to 200 samples from U.S fisheries in the Gulf of Mexico and along the East coast (target 50 per year per region repeated over two years). Actual numbers of samples received and analyzed are given in Table 3.1. Fin clips and muscle plugs from captured fish were collected postmortem and stored in 20% DMSO-EDTA, or 95% molecular-grade ethanol.

3.3.2 Sequencing

Samples were processed according to a modified double-digest restriction associated DNA protocol (ddRAD, Peterson, Weber, Kay, Fisher, & Hoekstra, 2012) as described in Norrell et al (2020). The genomic DNA from each sample was extracted using the Mag-Bind® Blood & Tissue DNA HDQ kit. Extracted DNA was digested using the EcoRI-HF and MspI restriction enzymes (New England Biolabs) and the digestion products were ligated with adapters featuring unique custom 6 bp barcodes that allow retrieving sequence reads from individual samples during multiplex sequencing on a single Illumina sequencing flow cell. Adapters also feature an 8 bp Unique Molecular Identifier (UMI, Schweyen et al. 2014) that allows identification and removal of PCR duplicates after sequencing. Samples were double-barcoded to mitigate the occurrence of sequence demultiplexing misidentification that may arise from the bridge-PCR method employed by the Illumina NovaSeq6000 sequencing platform (van der Valk et al. 2017). Library pools included equal numbers of samples from each location and year for multiplex sequencing to minimize bias that could result from variation among sequencing flow cells. Ligated fragments were amplified using PCR and size selected to 300-500 bp using a Pippin Prep (Sage Science). The final library pools were submitted for paired-end sequencing (2x150 bp) on an Illumina NovaSeq6000 platform at the University of Colorado Denver. Sequencing depth aimed to achieve an average coverage of sequenced regions per individual greater than 30x.

3.3.3 Sequence Processing and Filtering

Raw sequence reads were demultiplexed and duplicates were removed as described in *Chapter II*. Demultiplexed reads were trimmed, and initial quality filtering performed using the dDocent pipeline (Puritz et al. 2014). Trimmed reads were mapped using BWA-MEM (Li and Durbin 2009) onto the anchored Yellowfin tuna genome developed in *Chapter II*. The Gulf of Mexico larval samples tended to be preferentially removed from the dataset during filtering. Since these samples were needed to characterize the Gulf of Mexico nursery, the following strategy was implemented to increase the representation of these samples in the final dataset. SNPs were first identified in the larval samples using FreeBayes (Garrison and Marth 2012). The resulting variants were filtered using VCFtools and BCFtools (Danecek et al. 2011; Danecek et al. 2021), and vcflib (from the Freebayes package) to identify higher quality SNP positions with high call rates in the larval samples (Table 3.2). The obtained SNPs were called in all remaining (non-larvae) individuals in the dataset and included in the final filtering pipeline described below.

Table 3.2 *Filtering protocol applied to the larval specimens SNP dataset*

	Before		After		
Step	Samples	Markers	Samples	Markers	Filtering Parameters
1	97	7995261	97	341488	minDP=5 minGQ=20 max-missing=0.5
2	97	341488	97	285098	maf=0.001
3	97	285098	72	285098	individual missingness > 0.7
4	72	285098	72	125806	overall quality > 20
5	72	125806	72	92667	biallelic, no-indel

Filtering parameters reflect those employed in VCFtools (steps 1,2,5), vcflib (step 4), or a colloquial representation of which criteria were filtered (steps 3,4,5). Grey text indicates that the filtering step did not affect this component of the dataset.

Variants identified in the entire dataset were then subject to a more rigorous filtering pipeline described in Table 3.3 to maximize the yield of high-quality SNPs for population genetic analysis. Filtering steps included removal of sites with extreme allelic balance, improperly paired reads, and SNPs called from overlapping forward and reverse reads as described in dDocent_filters (provided by dDocent). Monomorphic sites, sites with extremely high coverage (>95% depth quantile) and individuals with >40% missing data were then removed. MNPs were decomposed into multiple SNPs and the data were restricted to biallelic SNPs with a minor allele frequency of 0.05 or greater. The data were then tested for conformance to Hardy-Weinberg Equilibrium expectation separately for each population, removing sites that departed significantly in two or more populations. Sites were then restricted to those with <10% missing data. Finally, GWAStools (Gogarten et al. 2012) was used to remove loci in close genomic proximity to each other (linkage disequilibrium coefficient $r^2 > 0.1$).

Table 3.3 *Filtering protocol applied to the entire SNP dataset*

	Before		After		
Step	Samples	Markers	Samples	Markers	Filtering Parameters
1	670	33625043	670	33625043	50% missing
2	670	33625043	670	30132923	depth < 10 and quality <20
3	670	30132923	670	867467	remove monomorphic (maf < 0.001)

Table 3.3 continued.

4	670	867467	546	867467	Individuals with <70% missing
5	546	867467	546	430370	Loci with overall quality > 20
6	546	430370	546	235872	Extreme allelic balance
7	546	235872	546	235221	Improperly paired reads
8	546	72599	546	72599	SNPs called from overlapping F+R reads
9	546	72599	546	68969	Extremely high coverage sites
10	546	68969	546	32721	Sites with >75% missing
11	546	32721	441	32721	Individuals with >40% missing
12	441	32721	441	25998	Sites with >15% missing in 2+ populations
13	441	27952	441	27952	Decompose MNPs into SNPs
14	441	27952	441	26231	Biallelic and no indels
15	441	26231	441	8432	Minor allele frequency 0.05
16	441	8432	441	8337	HWE outliers per pop
17	441	8337	441	7910	90% missing
18	441	7910	441	5771	Linkage disequilibrium filtering
19	441	5771	418	5771	Remove putative kin

Filtering parameters reflect a colloquial representation of which criteria were filtered. Grey text indicates that the filtering step did not affect this component of the dataset.

3.3.4 Population Genetic Analyses

3.3.4.1 Relatedness

Kinship was estimated for all pairs of samples using PC-Relate (Conomos et al. 2016), which builds on the KING method (Manichaikul et al. 2010) and is robust against the presence of population structure. The dataset was pruned to retain markers with linkage disequilibrium coefficients $r^2 < 0.1$ before performing the KING-robust estimation of relationship coefficients between all pairs of individuals. The resulting matrix was then partitioned into related and unrelated individuals using PC-Air (Conomos et al. 2015) with default parameters and PC-Relate was performed on the unrelated set of individuals. The eigenvalues obtained in PC-Relate were projected onto the subset of related individuals to obtain estimates of relatedness coefficients for each sample pair. Following Conomos et al (2015), pairs with coefficients ≥0.3535 were considered full siblings, those between 0.1767 and 0.3535 half siblings, those between 0.0883 and 0.1767 first cousin pairs, and those ≤0.0441 unrelated pairs. For each pair of samples identified as kin, one of the two samples involved in the pair (the sample with higher percent missing data), was removed from the dataset for subsequent analyses. To determine thresholds for assignment of pairs to kinship categories, 500 pairs each of full siblings, half siblings, and unrelated individuals were simulated using PopGenSims.jl (Dimens 2022) based on the whole dataset allele frequencies. The KING, PC-Air, and PC-Relate analyses were completed on the simulated data and the success rate assigning simulated sibship groups to the correct (simulated) category was calculated.

Summary statistics for the kin-removed empirical dataset were generated using PopGen.jl (Dimens and Selwyn 2022).

3.3.4.2 Outlier Loci Detection

Loci potentially impacted by natural selection were identified using an outlier analysis as implemented in outFLANK (Lotterhos and Whitlock 2015). A second outlier detection approach was implemented using the Bayesian resampling framework of Bayescan v2.1 (Foll and Gaggiotti 2008). The Bayesian approach was performed using 100,000 burn-in iterations, pilot runs of 15,000 iterations, final run with 15,000 resampling iterations, and prior odds of 100 for the neutral model. The resulting loci were sorted into putatively neutral and selected datasets using PopGen.jl and VCFtools and separately used to further study population structure (Danecek et al. 2011; Dimens and Selwyn 2022).

3.3.4.3 Population Structure

Cryptic subdivision was inferred using K-means clustering and Discriminant Analysis of Principal Components (DAPC, Jombart et al. 2010) to infer population clusters in the dataset and individual membership in the inferred clusters. The optimal number of principal components to retain was determined using the cross-validation method with 500 iterations accounting for varying numbers of PCs from 1 to the number needed to explain 85% of the variance in the data. K-means clustering was iterated 50 times for each value of the number of clusters (K) from 1 to 7 and the BIC for each value of K was computed to assess an optimal value (DAPC, Jombart et al., 2010). Cryptic structure was also tested with Bayesian clustering as implemented in fastStructure (Raj et al. 2014).

Divergence between sample groups (four nursery areas in the East and West Atlantic and adult fisheries samples from the Gulf of Mexico and East Atlantic) was estimated using pairwise F_{ST} indices (Hudson et al. 1992). Significance of estimates was tested using 10,000 permutations of individuals between groups. The False Discovery Rate correction (Benjamini and Hochberg 1995) was applied to account for multiple testing. Pairwise F_{ST} were estimated separately for the neutral and outlier markers using PopGen.jl.

A hierarchical analysis of molecular variance (Excoffier et al. 1992), implemented in Arlequin (Excoffier and Lischer 2010) was used to test the significance of variation among geographic region and between capture year within region. Significance of molecular variance components was assessed based on 10,000 permutations of haplotypes.

Because single-locus outlier analyses are impeded by artifacts and risks of false positives (Lotterhos and Whitlock 2015), estimates of F_{ST} were plotted as a function of mapping position on the reference genome to identify genomic regions with clusters of loci showing significantly high levels of divergence between populations. Significance of individual regions was assessed in a sliding window analysis (Hohenlohe et al. 2010; Bourret et al. 2013) of global and pairwise F_{ST} in VCFTools, where the average F_{ST} in windows of sizes of 1 cM (~640 kb) sliding across the genome with an offset of 0.5 cM (320 kb) was computed. Outlier windows were identified as those departing from the average (window) F_{ST} by more than two standard deviations of the genome wide distribution of F_{ST} window-averages.

Effective population size was estimated on the neutral dataset using the linkage disequilibrium method implemented in NeEstimator (Do et al. 2014) considering critical allele frequencies of 0.01, 0.02, and 0.05 and restricting the analysis to only compare loci across linkage groups (Sved et al. 2013) to avoid bias resulting from physical linkage. Non-parametric 95% confidence intervals were calculated over 1,000 bootstrap iterations.

3.4 Results

3.4.1 Variant Calling

Initial variant calling on the larval samples identified 7,995,261 variants. Filtration reduced the dataset to 92,667 candidate SNPs (Table 3.2). Variant calling of the entire dataset identified 33,625,043 putative SNP sites across all 670 samples. Filtering and removing closely linked loci reduced the dataset to 5,746 high quality biallelic SNPs across 441 samples (Table 3.3, Table 3.4). Details of each filtering step for the entire dataset are documented in Table 3.3.

Table 3.4 *Genomic distribution of SNPs surveyed in Yellowfin tuna*

Linkage Group	SNP count
unplaced	1989
1	197
2	156
3	194
4	140
5	167
6	163
7	194
8	163
9	185
10	183
11	174
12	143
13	152
14	99
15	177
16	152
17	107
18	179
19	164
20	147
21	112
22	123
23	173
24	135
10	183
11	174

3.4.2 Population Genetic Analysis

3.4.2.1 Relatedness

Relatedness coefficients estimated from simulated full sibling, half sibling and unrelated pairs were all within the published ranges for these classifications in PC-Relate (r > 0.352, 0.176 < r < 0.352, and r <0.044, respectively). Pairs in the empirical dataset were therefore classified as unrelated, half sibling or full sibling based on these thresholds (Figure 3.2a). Kinship analysis identified 51 pairs of full siblings, 29 pairs of half siblings, and the remainder were classified as unrelated (Figure 3.2b). Two large clusters of closely related individuals (full or half siblings) were identified in the data. The first one comprised 11 larval samples and 1 adult sample from the Gulf of Mexico, predominantly related as full siblings. The second cluster included 7 SEN samples predominantly related as half siblings. There was also an inferred half sibling pair of from VZ and a full sibling pair from IVC. Given the density and complexity of the interrelated kin clusters, all individuals in the TX and SEN kin clusters were removed, along with a single individual (the one with the most missing data) in each of the remaining kin pairs. The final sample count can be seen in Table 3.5. Summary statistics are shown in Table 3.7.

Table 3.5 *Final sample counts per location per year*

Location		Year							
		2010	2013	2015	2016	2017	2018	2019	2020
ATL	🟨	0	0	0	0	0	18	59	0
GOM	🟨	0	0	0	0	7	12	45	0
LA	🟨	0	0	0	0	0	17	0	0
MSAL	🟨	0	0	0	7	0	0	0	0
TX	🟨	10*	12*	9*	0	0	0	0	0
VZ	🟨	0	0	0	0	51	0	31	0
SEN	🟦	0	0	0	0	34	33	1	0
IVC	🟦	0	0	0	0	0	0	0	72+

* Larvae, + Juveniles from two size classes expected to represent primarily 1-2 year old fish. Sampling location colors indicate whether those sites are in the western (yellow) or eastern (blue) Atlantic Ocean.

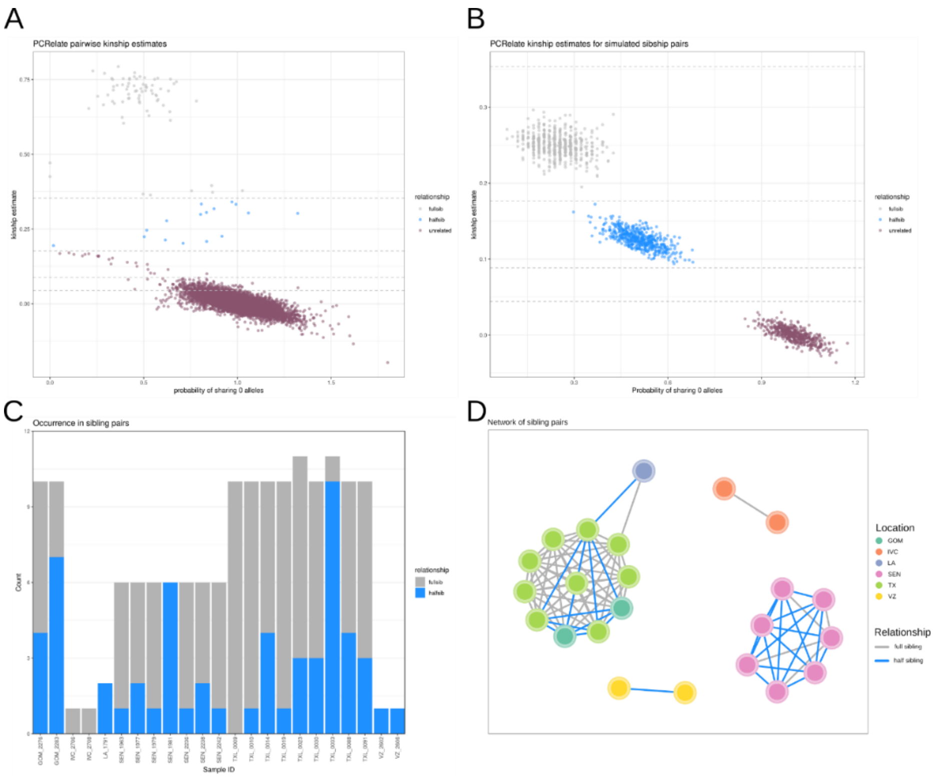

Figure 3.2 *Pairwise relatedness analysis of yellowfin tuna sampled in different regions of the Atlantic basin.*

Distribution of relatedness coefficients in the empirical dataset (A) and in simulated full sibs (in blue), half sibs (in grey), or unrelated (maroon) individuals (B). Dotted lines represent the thresholds used to assign pairs as full siblings, half siblings or unrelated. Number of half sib and full sib related to each of the 23 sampled individuals involved in full siblings or half sibling dyads (C). Network representing clusters of close kin identified within the dataset (D).

Table 3.6 *Summary statistics for the kin-removed dataset.*

Abbreviation: Description	Global	ATL	GOM	IVC	SEN	VZ
H_O : Observed Heterozygosity	0.0947	0.0994	0.0967	0.0892	0.0931	0.095
H_S : Within-pop gene diversity	0.0981	0.1023	0.0994	0.0926	0.0978	0.0985
H_T : Overall gene diversity	0.0982	0.1023	0.0994	0.0926	0.0978	0.0985
D_{ST} : Gene diversity among samples	0.0001					
$H_{T'}$: Overall gene diversity	0.0982					
$D_{ST'}$: Sample size adjusted	0.0001					
F_{ST} : su bpopulation vs total variance	0.0008					
$F_{ST'}$: Heterozygosity-adjusted F_{ST}	0.001					
F_{IS} Locus vs su bpopulation variance	0.035	0.0275	0.0275	0.0372	0.0473	0.0359
D_{EST} Population differentiation	0.0001					

3.4.2.2 Outlier Detection

Outlier analysis in outflank and Bayescan detected 5 and 4 putative outlier loci, respectively. All 4 loci detected in Bayesan were among the 5 outlier loci detected by outFLANK (Figure 3.3). Consequently, the 4 loci jointly detected by the two approaches were isolated from the dataset to generate separate putatively neutral and outlier datasets, which will be explicitly identified in the following analyses.

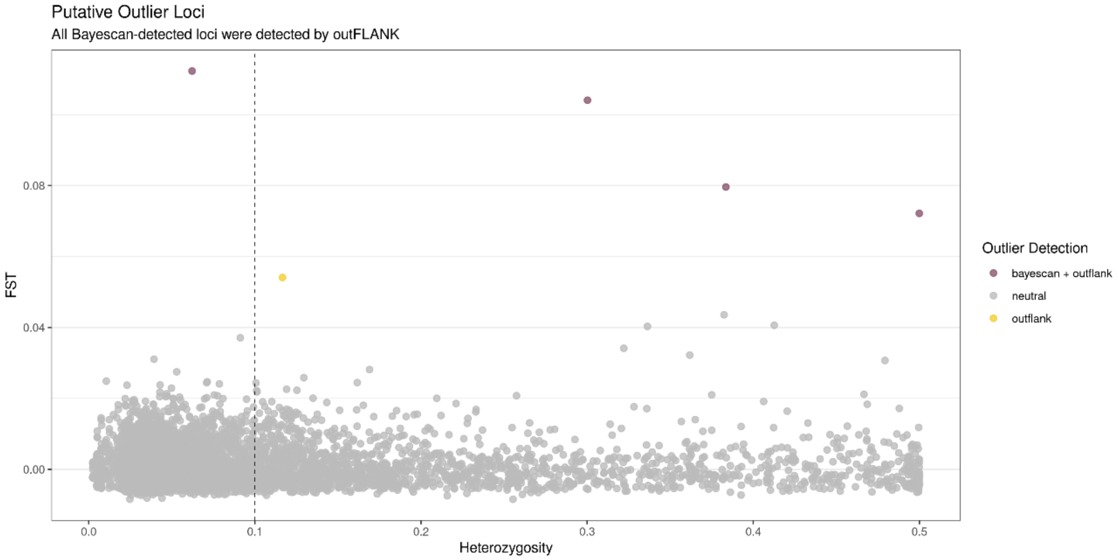

Figure 3.3 *Distribution of F_{ST} as a function of heterozygosity in neutral loci (grey), outliers detected by outflank and Bayescan (maroon) and in outflank only (yellow). The dashed vertical line is heterozygosity = 0.1.*

3.4.2.3 Population Structure

DAPC clustering was conducted on both the neutral and the outlier datasets. Cross-validation suggested retaining the first 180 principal components, capturing 61.6% of the variance in the data, for DAPC of the neutral dataset. The lowest BIC values were obtained for K=2 clusters during K-means clustering.

The DAPC clustering revealed a more complex pattern where the East Atlantic nursery samples (IVC and SEN) were distinct from the Western locations (Gulf of

Mexico, South Caribbean, and US Atlantic coast) and from each other (Figure 3.4A and B). The three Western samples showed some degree of separation between the South Caribbean nursery (VZ samples), the Gulf of Mexico and the US east coast (ATL) samples. The larval and adult samples from the Gulf of Mexico appeared closely related (Figure 3.4C) and were aggregated as an overall Gulf of Mexico cluster for the remaining analyses (Figure 3.4). Some overlap was observed between the ATL and IVC samples in the first 3 dimensions, and ATL was intermediate between the Western and Eastern nurseries (Figure 3.4A and B).

Cross validation retained 1 Principal component (41.3% variance) for DAPC of the outlier dataset and there was no clear clustering visible in the data (Figure 3.5). FastStructure results suggested K=1 was the most likely configuration for the neutral dataset, and K=2 for the outlier dataset (data not shown).

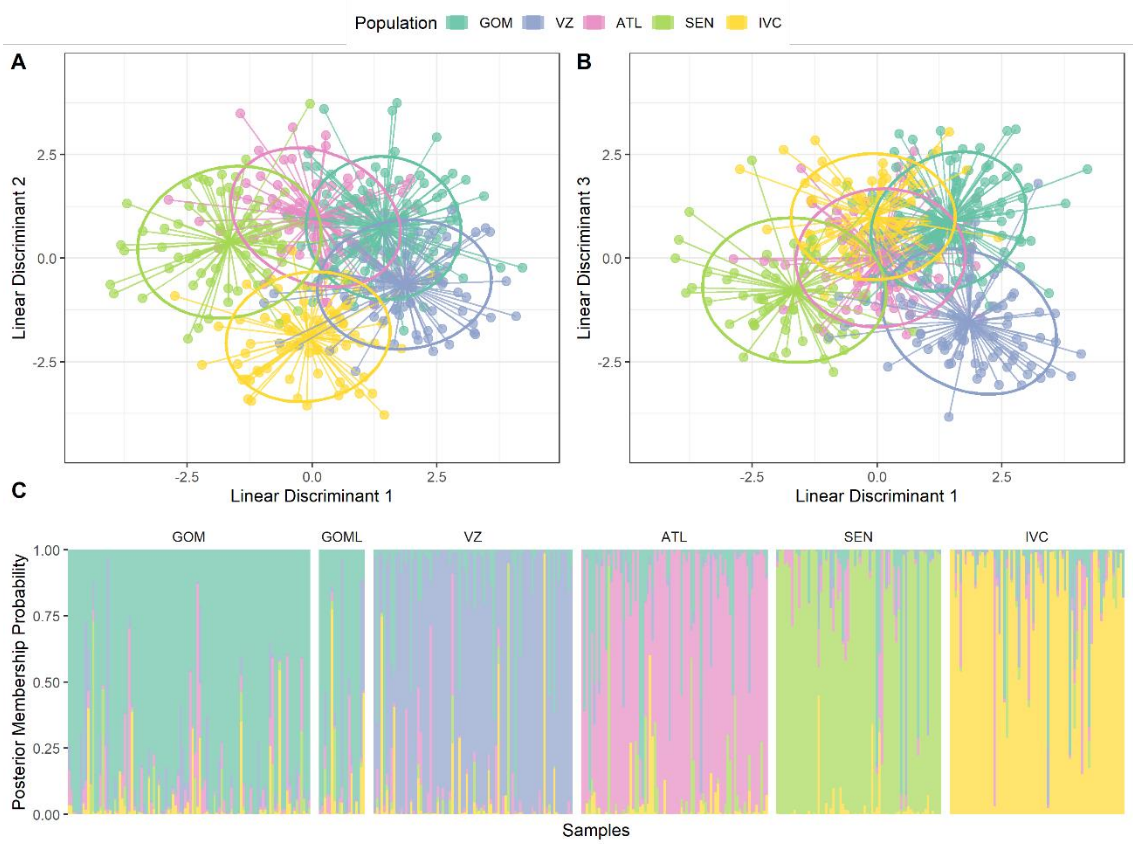

Figure 3.4 *Discriminant Analysis of Principal Components of the neutral dataset*
(A) K-means clustering using 180 (61% variance explained) principal components, (B) the first two linear discriminant functions separating samples by population, colored by location at capture, and (C) the posterior membership probability of each sample, where samples (each vertical bar) are sorted by location at capture and colors represent the percent membership identity to that genetic population, where "GOML" refers to the larvae sampled in the Gulf of Mexico.

F_{ST} estimates in the neutral data did not exceed 0.0104 (range = 0.0036 - 0.0104) but differed significantly from zero for all pairs of populations except for the comparison of Venezuela (VZ) and the western Atlantic Ocean (ATL) groups (Table 3.6). Pairwise F_{ST} in the outlier dataset followed the same pattern, although the F_{ST} estimates were greater (range = 0.003 - 0.0418).

Table 3.7 *Pairwise F_{ST} estimates between sampling locations.*

Neutral Loci

	ATL	GOM	VZ	SEN	IVC
ATL	-	0.0007	0.1432	0.0034	0.0003
GOM	0.0066	-	0.0014	0.0004	0.0003
VZ	0.0007	0.0054	-	0.0034	0.0003
SEN	0.0046	0.0076	0.0036	-	0.0003
IVC	0.0094	0.0104	0.0057	0.0082	-

Putative Outlier Loci

	ATL	GOM	VZ	SEN	IVC
ATL	-	0.0007	0.0192	0.0003	0.0030
GOM	0.0508	-	0.0005	0.0031	0.0003
VZ	0.0030	0.0485	-	0.0003	0.0003
SEN	0.0325	0.0428	0.0418	-	0.0004
IVC	0.0052	0.0717	0.0156	0.0322	-

F_{ST} (Hudson, 1992) values are shown below the diagonal and the corresponding FDR-adjusted P values generated after 10,000 permutations of genotypes over populations (above diagonal). P-values in shaded cells indicate significance at $\alpha = 0.01$. Sampling location colors indicate whether those sites are in the western (yellow) or eastern (blue) Atlantic Ocean.

Sliding window analyses of overall F_{ST} identified 17 1-cM outlier windows (Figure 3.6A). Sliding window analysis of pairwise F_{ST} between populations identified an average 3.4 ($\sigma = 1.83$) outlier windows per linkage group per population pair, amounting to a mean of 42.8 ($\sigma = 10.31$) outlier windows per population pair. While many of the significant windows only departed weakly from the mean window F_{ST} and did not show a clear geographic association, the highest genome wide window F_{ST} (linkage groups 1, 4, 17 and 22) did have a geographic pattern. The LG-1 outlier region appeared to differentiate IVC from other regions, the one on LG-17 and LG-22 were primarily associated with Gulf of Mexico divergence, and the region on LG-4 was associated with divergence between Eastern and Western Atlantic groups. (Figure 3.6B).

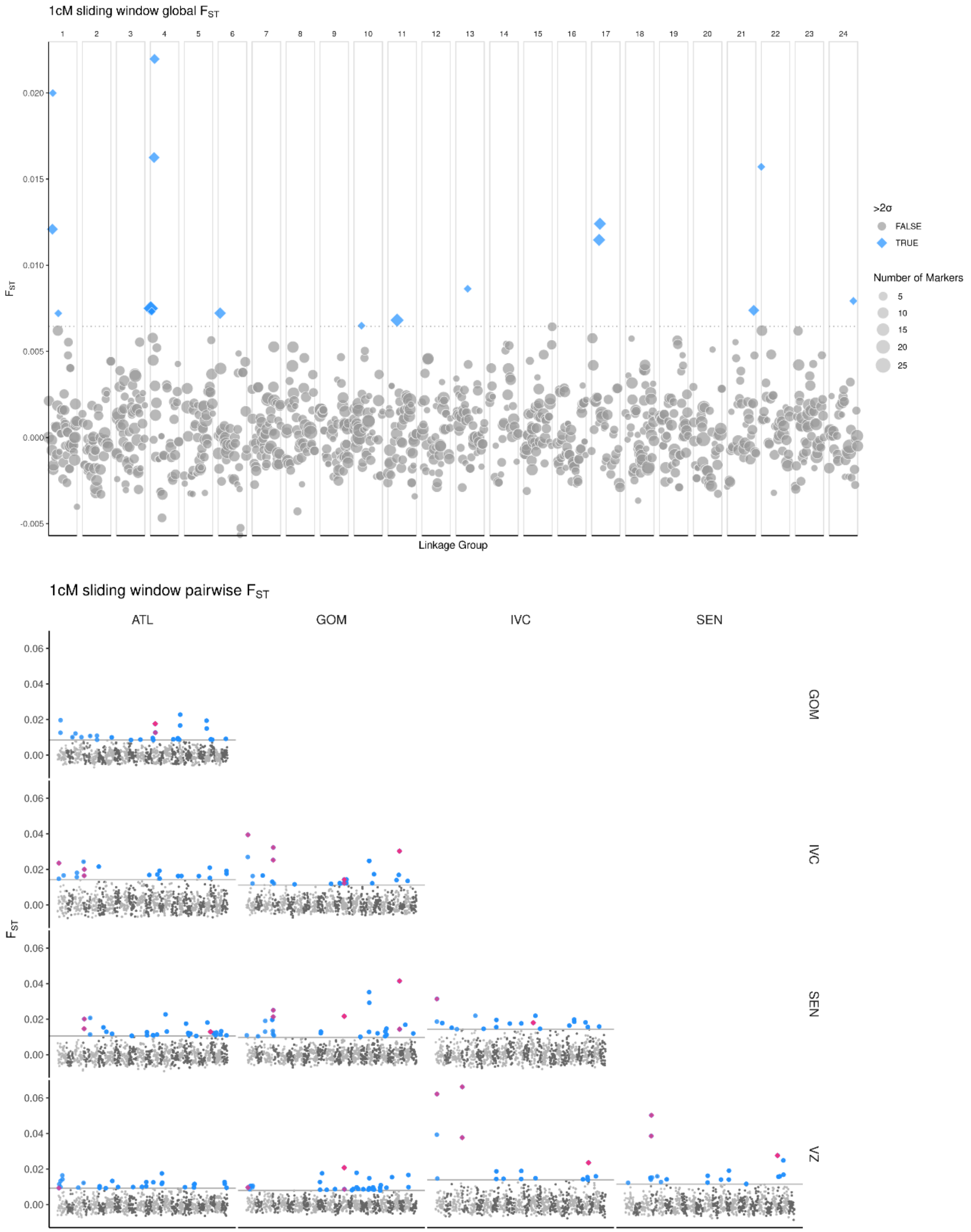

Figure 3.5 *Manhattan plots of overall (top) and pairwise (bottom) F_{ST} in 1 cM sliding windows relative to genomic position.*

Linkage groups are labeled using alternating shades of grey. Blue denotes genomic intervals with at least 3 SNPs and a mean F_{ST} greater than twice the standard deviation (2σ, dotted line). In the bottom panel, pink denotes significant windows appearing in at least 4 pairwise comparisons.

Hierarchical analysis of molecular variance in the neutral dataset and accounting for 5 regions (ATL, GOM, VZ, IVC, SEN) revealed significant variations among regions and no significant variation among years within region (Table 3.9). The analysis restricted to the 4 nursery areas and larvae or juvenile samples also yielded a positive among-nursery component, but it was not significant, likely due to limited sample sizes. The component among years within nursery region was not significant in both analyses. Similar results were obtained with the outlier loci dataset.

Table 3.8 *Analysis of Molecular Variance accounting for variation among regions and sampling year within region.*

Source of Variation	DF	Var. Comp	P	Dataset
Among regions	4	0.18	0.0004	All Data[1]
Among year within region	9	-0.14	0.9961	
Among regions	3	0.09	0.2182	Nursery Only[2]
Among year within region	4	-0.1	0.9508	

[1] Data from larvae, juveniles, and adults US Gulf of Mexico, US Atlantic Coast, South Caribbean (Venezuela), Gulf of Guinea (Ivory Coast), and West Africa (Senegal)

[2] Data from larvae and juveniles US Gulf of Mexico, South Caribbean (Venezuela), Gulf of Guinea (Ivory Coast), and West Africa (Senegal)

Estimates of effective population size using the linkage disequilibrium method varied widely between groups and were not reliable considering the small sample sizes available within cohort per location (below 30 in most cases) and are deemed negligible in comparison to the relatively large size of yellowfin tuna breeding stocks (Waples and Do 2010). Table 3.10 presents estimates for regions×year where sample sizes greater than 30 were available. Estimates for the East Atlantic (IVC and SEN) tended to be larger, particularly in the IVC sample.

Table 3.9 *Effective population size estimates by the linkage disequilibrium method*

Population	Year	Sample Size	N_e estimate	Jacknife CI
ATL	2019	59	53.8	35.9-91.6
GOM	2019	62	162.1	110.1-287.9
VZ	2017	51	167.2	114.9-292.2
VZ	2019	31	153.8	75.4-2286.0
SEN	2017	34	105.8	54.0-621.3
SEN	2018	34	208.8	64.2-Infinite
IVC	2020	59	26887.2	514-Infinite

Estimates only shown for year x region samples where more than 30 samples are available.

3.5 Discussion

The study intended to describe variation among the four main nursery areas hypothesized within the Atlantic, namely the Gulf of Mexico and the South Caribbean off Venezuela for the Western Atlantic and the Gulf of Guinea and the West African coast for the East Atlantic (ICCAT 2016). Nursery areas were characterized by sampling larvae and juveniles because such young specimens have limited dispersal potential and are therefore expected to remain proximal to the spawning grounds used by the breeding populations they derive from. The four presumptive nursery regions appeared to form genetically differentiated groups revealed by the DAPC, suggesting distinct breeding stocks are occurring. A fifth group was suggested comprising adult samples by fisheries along the East US coast. The among-region component of molecular variance was significant and greater than the temporal component of variance when analyzing the whole dataset indicating that the structure pattern was robust to temporal variations. When the AMOVA was conducted using the nursery dataset (only larvae and juveniles), the among-nursery region component of variance was still positive but not significant. The lack of significance was likely due to the smaller sample sizes in the dataset reducing inference power. Divergence observed in the East and West Atlantic nurseries corroborate recent findings by Pecoraro et al (2018), who sampled the South Caribbean nursery but did not detect differences between this nursery and the Gulf of Mexico. This study employed a larger number of loci which may have led to a higher power of inference. Temporal divergence could have affected the comparison between the Venezuela samples (collected in 2017 and 2019) and Gulf samples (collected between 2010 and 2015), but this seems unlikely because the overall hierarchical AMOVA showed that temporal variations are reduced compared to the among-region variation as discussed above. Additionally, the DAPC showed that the very large majority of individuals from the cohorts sampled in each locality had a very strong membership probability in the nursery region where they were collected, further supporting the hypothesis that overall migration between nursery areas is limited.

This study revealed divergence of the West African sample (Senegal) and the Gulf of Guinea (Ivory Coast). Pecoraro et al (2018) did not find heterogeneity between the two samples they examined in the East Atlantic, but these two samples were in the Gulf of Guinea and further South. Consequently, their study could not have detected the second eastern Atlantic cluster (the West African nursery), which is located north of their sampling range. Collectively, these results suggest that the East Atlantic yellowfin tuna may comprise at least two breeding stocks. One with offspring utilizing the Gulf of Guinea as nursery and the second one utilizing waters offshore West Africa. Spawning has been reported around Cabo Verde (ICCAT 2019) and may produce juveniles sampled offshore Senegal, a hypothesis that could be tested by genotyping adult samples from Cabo Verde. The Gulf of Guinea nursery may extend south of the Gulf along the coasts of Namibia and Angola as Pecoraro et al (2018) did not detect heterogeneity between yellowfin tunas sampled from the Gulf of Guinea and those collected offshore Angola. The West Atlantic would feature at least two distinct breeding stocks with offspring utilizing the Gulf of Mexico and the Southern Caribbean, respectively.

The analysis of the entire dataset suggested occurrence of a fifth group that was primarily represented in the samples from the West Atlantic area off the US East Coast (ATL group) where it was dominant. This group may correspond to a 5th breeding stock, producing offspring recruiting primarily along the US East coast as adults. However, the ATL group appeared intermediate between the East and West Atlantic stocks and an alternative hypothesis could be that yellowfin tuna in the region represent a mixture of the two East Atlantic stocks and possibly some of the Western stocks. Sampling of juveniles offshore the US southeast coast and northern Caribbean, if they occur, would be useful to further explore the hypothesis of a distinct breeding stock contributing to fisheries along the East US coast.

One last area that was not characterized in this study and that of Pecoraro et al (2018) is the Southwest Atlantic offshore the Brazilian coast. Specimens found in the northeast of Brazil appear to be primarily juveniles (da Silva et al. 2019) but Costa et al (2005) reported both juveniles and adults in the South of Brazil. Adults landed in Southeast Brazil were hypothesized to move to the South Caribbean Sea for spawning when temperatures decrease during the southern hemisphere winter. Sampling in Brazil was not possible during this project, but these hypotheses would deserve to be tested with juvenile and adult samples from South and North Brazil to determine whether additional breeding stocks occur or recruitment in that region is seeded by the Caribbean spawning population as hypothesized by Costa et al (2005). The divergence observed among regional samples in this study was notably low, e.g., the F_{ST} value between Venezuela and the other stocks in the West Atlantic was very low, suggesting that divergence of this group may be recent.

A second objective of this study was to assess the contribution of putative breeding stocks to adults sampled by US fisheries in the Gulf of Mexico and along the East Coast and the possible occurrence of mixed stock fisheries in these regions. This could not be achieved with confidence because of the very low level of divergence among groups identified during the study. The membership probability estimates from the DAPC provide initial information on this topic. The Gulf of Mexico seems to be relatively largely self-recruiting with most adults showing high membership in the cluster shared with the larvae although some migration, particularly from the US East Coast, were suggested by the high membership in that cluster of a few individuals caught in the Gulf of Mexico. Some moderate exchange between the Gulf of Mexico and South Caribbean stocks and the Gulf of Guinea were also suggested. The greatest amount of admixture was between the West African groups and the US East coast with numerous individuals with high ancestry in the West African stock sampled among the adults caught offshore the US East coast and some individuals with high ancestry in the US East Coast cluster present in the IVC and SEN samples. This suggests transatlantic movement of significance as already suggested by tagging studies (Zagaglia et al. 2004; Fonteneau and Hallier 2015; ICCAT 2019). The occurrence of differentiated stocks in the West Atlantic and West Africa, however, suggests that the reproductive success of migrants may be more limited than suggested by the rate of possible F0 migrants inferred from DAPC.

This study also documented adaptive variation through analysis of outlier loci. Single locus outlier analyses detected only a few loci under putative selection. Divergence at the outlier loci was on average 6 times greater than that at neutral loci. Clustering of the outlier dataset using DAPC did not recover the clear pattern identified with the neutral dataset. This finding contrasts with those of Pecoraro et al (2018) who found that divergence between East and West Atlantic stocks was primarily supported by outlier loci. This study found a much lower prevalence of outlier loci (only 4 outlier loci out of 5,772 loci versus 33 outliers out of 972 loci in the study of Pecoraro et al.). The levels of divergence at loci also differed substantially between the two studies with pairwise F_{ST} between regions, averaging 0.0062 at neutral and 0.0344 at outlier loci in this study versus 0.01 and 0.16 in Pecoraro et al. Discrepancies in the magnitude of divergence could be due to in part to differences in the genotyping approach in the two studies (2b-RAD in Pecoraro et al. versus ddRAD in this study), which employed different restriction enzymes to sample the genome and likely revealed different loci. According to this hypothesis, the study of Pecoraro *et al.* would have captured more divergent outlier loci by chance. This scenario is possible, at least for outlier loci, because the marker densities of both studies (one locus every 136 kb on average for this study and one locus every 813 kb for Percoraro et al.) were too low to capture most selective sweep windows, which are thought to span from less than a kb to 200 kb in most cases (Catchen et al. 2017). Differences in sampling, processing samples and filtration algorithms may also account for the unequal divergence magnitude inferred in the two studies. Sequencing bias could occur if individuals from different populations were sequenced on different sequencing runs, leading to confounding effects of sequencing batches with geographic origin. This potential bias was controlled in this study by mixing individuals from all localities in each sequencing pool. Because single locus outlier analyses are more prone to impacts of sampling or assay artifacts, a sliding window analysis was applied in this work to detect clusters of co-located loci showing significantly high divergence among regions, thus providing a potentially more robust assessment of genomic regions affected by divergent selection. Most outlier windows departed only mildly from the neutral distribution of F_{ST} and did not reveal a clear geographic pattern, but the few windows showing highest F_{ST} did show some geographic patterns differentiating the eastern and western Atlantic groups and specific nursery areas such as the Gulf of Guinea (IVC sample) or the Gulf of Mexico. The corresponding genomic regions may be affected by divergent selection and local adaptation of these stocks and could be investigated further by developing genomic scans targeting them in follow-up studies.

Finally, this study revealed multiple pairs of closely related individuals including two large groups clusters, one in the Gulf of Mexico composed of 12 sampled individuals (9 larvae and 3 adults), and a second one that was composed of 9 SEN juvenile samples. The remaining siblings identified were a pair of half siblings sampled in VZ and a pair of full siblings sampled in IVC. Co-location of closely related individuals was already reported in a study of Pacific yellowfin tuna (Anderson et al. 2019) and in the study of the closely related blackfin tuna *Thunnus altanticus* (*Chapter IV* of this dissertation). The co-location of larvae/juveniles (but also adults) in the Gulf of Mexico is consistent with the philopatric behavior observed in some tagging studies (Schaefer et al. 2011; Wells et al. 2012) and the detection of population structuring in this study. However, it is likely the full siblings identified across sampling years (e.g., full siblings involving larvae sampled in 2010 and 2015 in the Gulf of Mexico) were incorrectly classified. Consequently, close kin pairs were removed prior to analyses of population structure. The occurrence of close kin may also reflect sweepstake recruitment in the affected regions where small numbers of brooders contribute large fractions of sampled cohorts (Hedgecock and Pudovkin 2011). The small sample sizes available per cohort in this study prevented reliable estimation of the effective size of cohorts and evaluation of potential sweepstakes effect, but the observation of co-located close kin in adult samples suggests close kin aggregation may be occurring and persist to the adult stage in some cases.

The structure patterns obtained in this study are very weak and must be taken cautiously, particularly the DAPC analysis, which is prone to overfitting. Future efforts should focus on increasing the density of genetic markers to uncover more effectively genomic regions affected by divergent selection. Outlier loci exhibit a disproportionately higher degree of divergence among populations that may assist in better defining population structure and delineating units (Vaux et al. 2021). The equator has been suggested to be a cryptic barrier to gene flow in the sympatric congeners *Thunnus alalunga* (Vaux et al. 2021) and *Thunnus atlanticus* (*Chapter IV*, this study), and future effort should include samples below the equator to test if the equator plays a similar role in structuring Yellowfin tuna populations.

CHAPTER IV – POPULATION STRUCTURE AND DEMOGRAPHY OF BLACKFIN TUNA, *THUNNUS ATLANTICUS*

4.1 Introduction

Tunas (family *Scombridae*) are highly specialized fast-swimming pelagic predators known to migrate large distances annually (Mariani et al. 2016; Reglero et al. 2017). They are therefore expected to form metapopulations connected over broad distances, possibly at the scale of entire oceanic basins. The Blackfin Tuna (*Thunnus atlanticus*) is a small tuna growing to approximately 100 cm and weighing up to 21 kg, making it the smallest of the Thunnus genus. The species occupies the narrowest geographic range of all Atlantic true tuna species. It is restricted to the western Atlantic basin where it has been reported from Massachusetts to as far south as Brazil, although it is mostly found in tropical and sub-tropical waters where the temperature is likely to exceed 20° C. In the United States, Blackfin tunas are abundant throughout the Gulf of Mexico and South Atlantic Bight regions (Collette et al. 2010). They can be found at depths from 20 m to 700 m but are most common at 40 m to 50 m (Maghan and Rivas 1971). Their distribution has been linked to several factors such as water clarity, steepness of the continental shelf, and plankton concentrations correlated with terrestrial runoff and upwelling zones (De Sylva et al. 1987). Their diet consists of surface and deep-sea fishes, squid, and arthropods including amphipods, shrimps, and crabs (Collette et al. 2010; Frimodt and Dore 1995). Spawning occurs from late spring to early fall when water temperatures are at or above 27° C, with a peak of activity in the early summer months (Idyll and De Sylva 1963; Juárez 1978; Bezerra et al. 2013; Richardson et al. 2010).

Blackfin tunas are harvested by commercial and recreational fisheries across their range. Historically, they were not popular for recreational fishing in the US, but they have been increasingly targeted in recent years by recreational fishers along the US east coast, off the Florida Keys and around Puerto Rico. The species is seldom targeted by commercial boats in the US, although it can be captured as bycatch of other tuna fisheries. It is harvested commercially using longlines and purse seines in the Caribbean and South America with highest landings recorded in Cuba, the Dominican Republic, the Lesser Antilles, Venezuela, and Brazil (Mathieu et al. 2013). Blackfin tunas are managed at the basin level under the international jurisdiction of the International Commission for the Conservation of Atlantic Tunas (ICCAT) for international waters, relayed by domestic management entities such as the Highly Migratory Species (HMS) division of the National Oceanic and Atmospheric Association (NOAA) in the US for captures within the Exclusive Economic Zone. Considering the rising popularity of Blackfin tunas in the US and other countries exploiting them in western Atlantic waters, stock structure needs to be documented to design appropriate units for management.

Based on available records of sexually mature individuals, eggs or larvae, Mathieu et al (2013) suggest that Blackfin tunas reproduce over most of their distribution range, thus possibly forming a metapopulation composed of many demes. Mark-recapture studies by Luckhurst et al (2001) in Bermuda and (Singh-Renton and Renton 2007) in St Vincent and the Grenadines revealed some instances of site fidelity where some Blackfin tunas were recaptured in the tagging area, sometimes after very long periods (4 years). However, long-distance movement was also suggested by Luckhurst et al (2001) for individuals tagged in the Bermuda Islands where recaptures only occurred during the summer months while Blackfin tunas were absent during cold months and hypothesized to move South during those periods. These results suggest that gene flow across geographic populations of Blackfin tuna may be partially restricted by some degree of phylopatry.

Information on genetic stock structure is limited to a study by Saxton (2009) comparing the Gulf of Mexico and the US East coast using 6 microsatellites and sequences of the control region of mitochondrial DNA and a more comprehensive study by Saillant et al. (in review) using 13 microsatellite markers surveyed in 9 geographic population from Brazil to North Carolina. Saxton (2009) reported significant divergence between the US East Coast and the Gulf of Mexico. Saillant et al. reported very weak divergence across the sampling surface with a possible isolation of the Brazilian population from the rest of the range and a weak pattern of isolation by distance. Both studies were limited by the small numbers of genetic loci used, which prevented assessing occurrence of population structure related to divergent selection, and by the lack of or incomplete temporal replication of sampling. The advent of next generation sequencing and the development of genotyping by sequencing methods have enabled cost-effective generation of high-density genome scans including thousands of genetic loci (Peterson et al. 2012). The Restriction Site Associated DNA (RAD) sequencing methods have become the most popular genotyping option in molecular ecology studies due to their immediate applicability to non-model species (O'Leary et al. 2018). The reliability of genotyping and the number of polymorphic loci that can be recovered are dependent on rigorous data filtering and are improved when a reference genome is available and used to map RAD sequencing reads (Shafer et al. 2017).

This study addresses the limitations of previous population genetic studies of Blackfin tuna by employing high-density genome scans to describe genetic variation in geographic populations across the species' geographic range and multiple sampling years. A draft reference genome was developed and used to map RAD sequences obtained from population samples and both neutral and non-neutral patterns of structure were investigated to assess comprehensively genetic structure accounting for local adaptation of populations.

4.2 Methods

4.2.1 Draft Genome Development

Fin tissue from a single representative individual captured in the north central Gulf of Mexico was processed using the MagBind Blood and Tissue kit (Omega Bio-Tek, cat. M6399-01) to isolate high quality genomic DNA. The sample was sequenced on the Illumina NovaSeq6000 platform to obtain 150 bp paired end reads. The raw Illumina reads were trimmed using fastp (v0.20.0 Chen et al. 2018) as described in the previous chapter. The filtered reads were then used to estimate the size of the Blackfin tuna genome using the K-mer frequency counting method (https://bioinformatics.uconn.edu/genome-size-estimation-tutorial/). The K-mer frequency distribution in Illumina reads was calculated using the program jellyfish (Marçais and Kingsford 2011), for K-mer sizes varying from 17 to 25. Detailed assembly methods are shown in Appendix B. Briefly, trimmed short reads were assembled using SparseAssembler (Ye et al. 2012), with a K-mer size of 90. DB2OLC (Ye et al. 2016) was then used on the short-read contigs to produce a consensus with a K-mer size of 31. The trimmed raw reads were mapped onto the assembly using BWA-MEM (Li and

Durbin 2009) and then used to polish the consensus with Pilon (Walker et al. 2014). The trimmed short reads were then mapped onto the assembly again and used to identify and remove haplotigs using purge_haplotigs (Roach et al. 2018). Finally, trimmed short reads were mapped to the obtained assembly for a final round of polishing with Pilon using default parameters. Assemblies were assessed using metrics produced by Quast (Gurevich et al. 2013). Genome completeness was assessed using the *Eukaryota* database of Benchmark Universal Single Copy Orthologs (Simão et al. 2015).

4.2.2 Sample Acquisition

A total of 650 adult Blackfin tuna samples from 9 geographic localities were analyzed during the study (Figure 4.1, Table 4.1). Localities surveyed were offshore the US east coast (South Carolina), the Florida Keys, the north central Gulf of Mexico around Pensacola, the Western Gulf of Mexico around Corpus Christi, the US Caribbean (offshore Puerto Rico), the French Antilles (La Martinique), Venezuela, and Brazil (offshore Baía de Formosa, and St Peter and St Paul Archipelago) providing samples from across the species range. Samples were taken post-mortem from fish carcasses through fishery dependent sampling. Localities were sampled across two different years to allow testing the temporal stability of spatial patterns of structure, with a target of 50 specimens per locality per year. The two northern Gulf of Mexico localities, La Martinique Island, and St Peter and St Paul archipelago could only be sampled once. Exact coordinates were not known for samples collected during port sampling from fishermen and capture location was assumed within 150 km of the landing port in those cases. Tissue samples, a 1 cm^2 fin clip or 0.5 cm^3 muscle sample, were taken from each fish and stored in either 20% DMSO-EDTA, 95% ethanol, or Sarkosyl urea lysis buffer

(8 M urea, 1% sarkosyl, 20 mM sodium phosphate, 1 mM EDTA) until DNA isolation. Sampling targeted fish of reproductive size (FL > 50 cm).

Table 4.1 *Number of samples obtained for each locality and sampling year*

Locality	ID	2015	2016	2017	2018/19
South Carolina	SCA	-	50 (23)	49 (29)	-
Florida Keys	KEY	-	48 (32)	-	41 (24)
Pensacola	PNS	-	46 (30)	-	-
Texas	TX	-	-	51 (28)	-
Puerto Rico	PR	-	44 (23)	34 (16)	-
La Martinique	MRT	-	64 (39)	-	-
Venezuela	VZ	50 (31)	50 (14)	-	-
Brazil Baia Formosa	BRZ	-	46 (6)	-	49 (17)
Brazil St. Peter/Paul	BRZ-SP	-	-	-	28 (14)

Values in parenthesis reflect the number of samples remaining after sequence quality filtering

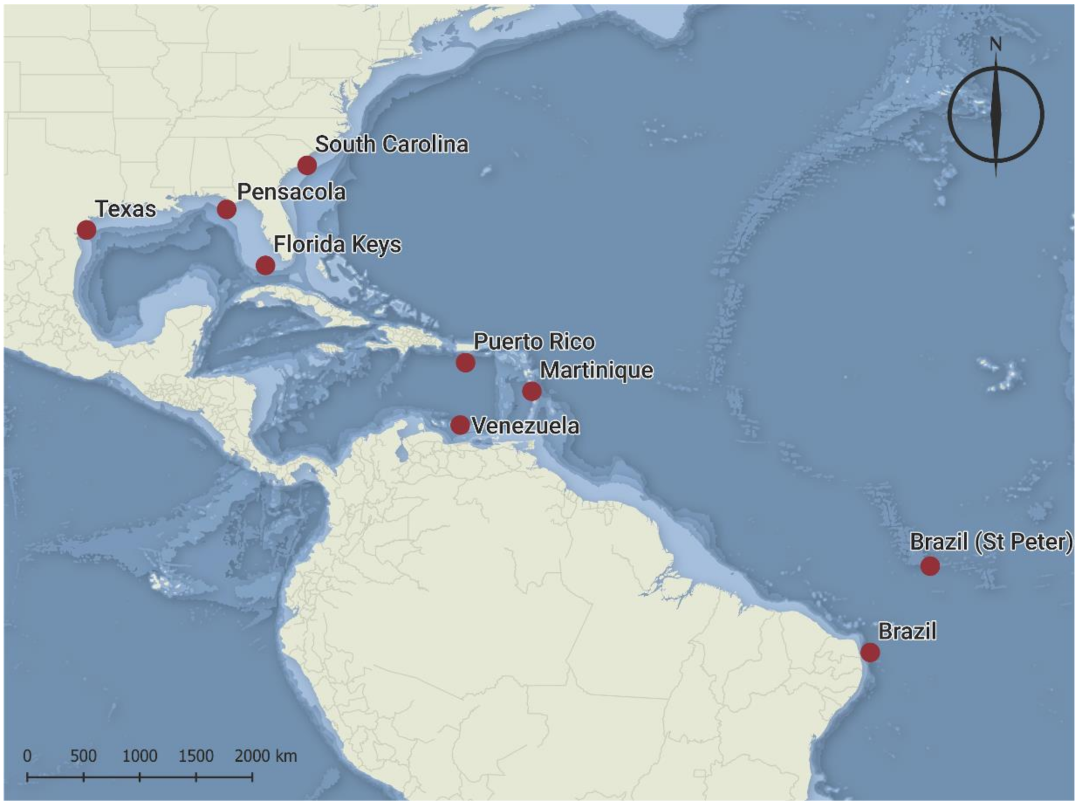

Figure 4.1 *Sampling localities for Blackfin tuna*

4.2.3 Sequencing

DNA was isolated using either Blood & Tissue DNA HDQ 96 Kit or EZ-96 tissue kit (Omega Bio-Tek, cat. D1196-01). After DNA isolation, samples were partitioned into 9 sequencing libraries, each receiving an equal number of samples from each sampling location and year to minimize the impacts of sequencing bias that could occur if samples from individual localities and year were sequenced on separate sequencing runs. Samples were prepared for sequencing using a modified version of the Double Digest Restriction Associated DNA protocol (Peterson et al. 2012). The modifications to the protocol include the use of EcoRI and MspI restriction endonucleases (New England Biolabs), along with custom adapters fitted with 6 bp unique barcodes allowing multiplexing up to 100 unique individuals in the same sequencing run and an 8 bp Universal Molecular Identifier (UMI, EuroFins) to isolate PCR duplicates in downstream analyses. The barcode was included in both the P1 (‘forward’) and P2 (‘reverse’) adapters to ensure proper demultiplexing of reverse sequencing reads and prevent errors due to “barcode hopping” (van der Valk et al. 2017). Samples were pooled and size selected using a 300-500 bp window on a Pippin Prep (Sage Science), and the DNA concentration and fragment size distribution of the pool were assessed on a NanoDrop 2000 and an Agilent 2100 BioAnalyzer DNA chip system, respectively. The obtained libraries were sequenced at the University of Colorado Genomics and Microarray Core facility to generate on average 6 million paired end reads (150 bp x 2) per individual using the Illumina NovaSeq6000 platform.

4.2.4 Data Filtering

The raw sequence data were demultiplexed at the sequencing facility and processed using the dDocent pipeline (Puritz et al. 2014). Briefly, raw sequences were trimmed and filtered to remove low quality bases using fastp (Chen et al. 2018). Reads were then mapped on the draft reference genome using BWA (Li and Durbin 2009) and SNPs were called using FreeBayes (Garrison and Marth 2012). The settings used for dDocent are described in the previous chapter.

The resulting raw SNP dataset was initially filtered using VCFtools (Danecek et al. 2011) to retain loci with less than 50% missing data, a minimum quality of 30, a minimum depth of 10, and a minimum allele frequency of 10^{-6}, and to remove individuals with more than 20% missing data. Data were then filtered for site depth, quality versus depth, strand representation, allelic balance at heterozygous individuals, and paired read representation. The vcfallelicprimatives transformation from vcflib library (https://github.com/vcflib/vcflib) was used to deconstruct multi-nucleotide polymorphisms into SNPs. Once markers were decomposed into SNPs, indels and multi-allelic SNPs were removed in VCFtools and a maximum missing data tolerance of 20% was applied. Individuals with significantly high or low heterozygosity were identified and removed using VCFtools. Loci departing significantly from Hardy-Weinberg equilibrium within populations were identified in VCFtools and removed. The Benjamini-Hochberg false discovery rate correction was applied to determine significance of within-population exact tests of Hardy Weinberg Equilibrium with an alpha (-h) of 0.0055 to account for the 9 tests (9 locality-samples) performed simultaneously for each locus. This method was applied as it calculates heterozygosity on a per-locality basis to minimize the influence of the Wahlund effect on allele frequencies if structure is present. The details of filtration follow those presented in *Chapter III*. Finally, we retained only loci with a minimum minor allele frequency of 0.01. A combination of the R package Radiator (Gosselin et al. 2019) and software PGDSpider2 (Lischer and Excoffier 2012) were used to convert the datasets between file formats for subsequent analyses.

4.2.5 Population Genetic Analysis

4.2.5.1 Relatedness

Pairwise relatedness was estimated using PC-Relate (Conomos et al. 2016), which builds on the KING method (Manichaikul et al. 2010) and is robust against the presence of population structure (Conomos et al. 2016). The pairwise relatedness matrix was partitioned into related and unrelated individuals using PC-Air (Conomos et al. 2015) with default parameters and PC-Relate was performed on the unrelated set of individuals. The resulting eigenvalues were projected onto the subset of related individuals to obtain the relatedness coefficients for each sample pair. The results were validated by simulating 1,000 pairs of full siblings, half siblings, and unrelated individuals using PopGenSims.jl (Dimens 2022) and performing the full analyses on those simulated data.

The probability that two members of a close kin dyad inferred in PC-Relate was collected in the same locality or the same locality x year was compared to a random distribution of kins with respect to geographic locality or geographic locality by year using a resampling approach implemented in Poptools v. 3.2.5 (Hood 2010). The two members of inferred dyads were assigned to geographic localities independently during resampling accounting for the sample size in each locality during 10,000 Monte Carlo simulations and the probability of finding co-located dyad members at the observed frequency or higher by random chance was estimated based on percentile distributions in the simulated datasets.

4.2.5.2 Outlier Loci

Loci potentially impacted by natural selection were identified using an outlier analysis implemented in outFLANK (Lotterhos and Whitlock 2015) using a *q* threshold of 0.05. We also performed a second outlier analysis using Bayescan (Foll and Gaggiotti 2008), accounting for a prior odds of 100.

4.2.5.3 Population Structure

The mean number of alleles and expected heterozygosity in each locality sample were computed using Arlequin. Pairwise F_{ST} estimates (Hudson et al. 1992; Bhatia et al. 2013) were calculated using PopGen.jl (Dimens and Selwyn 2022), with significance tested using 10,000 permutation of individual genotypes across populations. P-values were adjusted for multiple testing using FDR correction and a false discovery rate of 0.05 (Benjamini and Hochberg 1995). Hierarchical analyses of molecular variance (Excoffier et al. 1992) were conducted accounting for geographic localities and sampling year within locality. A Discriminant Analysis of Principal Components (DAPC) was implemented using the R language package adegenet (Jombart 2008; Jombart et al. 2010; R Core Team 2013). The optimal number of principal components to retain was determined using the cross-validation method with 500 iterations accounting for 1 to 200 components. The optimal number of genetic clusters present in the data *a priori* was inferred by applying k-means clustering. Cryptic structure within the sampled range was also examined using model-based Bayesian clustering in fastStructure (Raj et al. 2014) accounting for a range of 1 to 9 clusters.

Structuring according to an isolation by distance model was first assessed by testing the correlation between genetic and geographic distance using a Mantel test in Genalex 6.5.1 (Smouse et al. 1986; Peakall and Smouse 2012). Geographic distance between samples was estimated based on geographic coordinates in Genalex. Genetic distance was estimated using the multilocus distance of Smouse and Peakall (1999). The logarithm of geographic distance was used in the computations to account for dispersal in a 2-dimensional habitat. Occurrence of spatial structuring was also examined using spatial autocorrelation analysis in Genalex. This analysis allows detecting patterns of spatial structure (through analysis of correlation of genotypes) even if variation does not follow the strict isolation by distance model across the entire distance range sampled as assumed in Mantel tests. The multilocus spatial autocorrelation coefficient *r* was computed based on geographic distance and the multilocus genetic distance described by Smouse and Peakall (1999). When spatial autocorrelation is occurring, the estimated value of *r* among proximal samples differs significantly from zero and decreases with increasing geographic distance. Because the estimation of spatial autocorrelation is influenced by the size of the distance class (Peakall et al. 2003), *r* was computed based on a series of increasing distances between sampling locations. The distance at which *r* no longer differs significantly from zero provides an approximation of the distance at which genetic divergence (population structure) can be inferred (Peakall et al. 2003). Significance of *r* was determined via 1,000 random permutations of genotypes among distance classes; significance of spatial autocorrelation coefficients was inferred when the observed estimate of *r* lied beyond the upper 95% limit of the distribution of *r* values obtained during the 1,000 permutations (Peakall and Smouse 2012).

4.3 Results

4.3.1 Genome sequencing and assembly

Illumina sequencing produced 730,100,108 raw reads for a total of 110,245,116,308 bp. Sequence filtering retained 710,740,790 reads for a total of 107,321,859,290 bp. The total length of the assembly was 514,764,407 bp in 203,667 contigs with a GC content of 39.71%. A large fraction of the assembly was in small contigs with only 365,013,046 bp in contigs over 5 kb. The N50 and L50 were 10,873 bp and 11,820 bp respectively and the largest contig was 158,390 bp. The estimate of the size of the Blackfin tuna genome using the K-mer frequency spectrum counting method obtained with varying K-mer sizes were all between 773 Mb and 791 Mb (μ = 785,121,418 bp, σ = 5,416,115). According to this estimate, the assembly included approximately 65% of the Blackfin tuna genome and sequencing covered the genome at a depth of 139X with filtered reads. BUSCO assessment of the completeness of the assembly indicated that it contained 44.55% and 18.81% full and partial orthologs, respectively.

4.3.2 Population genetics analysis

The filtering process reduced the data to 2,139 biallelic SNPs across 334 samples. The numbers of individuals retained per geographic population averaged 36 and ranged between 14 and 57 (Table 4.1). Grouping the two Brazilian localities brought the average sample size to 40.75 (range 28-56).

The KING and PCRelate analysis were performed using 2 principal components and identified 4 full sibling pairs, and 8 half sibling pairs (Figure 4.2, Table 4.2). Among these observed kin pairs, two pairs of putative full siblings had >98% identical genotypes, suggesting the samples could be duplicates or a contamination may have occurred during processing. Representatives occurring in multiple kinship pairs were removed. For the remaining sibling pairs, the member of each kin pair with the most missing data was also removed. The distribution of close kin (full siblings or half siblings) members of a dyad was not random with respect to sampling locality (P < 0.0001). In all 4 full sibling dyads and in 4 out of the 8 half sibling dyads, the two dyad members (putative full siblings or half siblings) were collected in the same locality (Figure 4.2, bottom). These findings were similar to the distribution of siblingship inferred by TrioML (Wang, 2007).

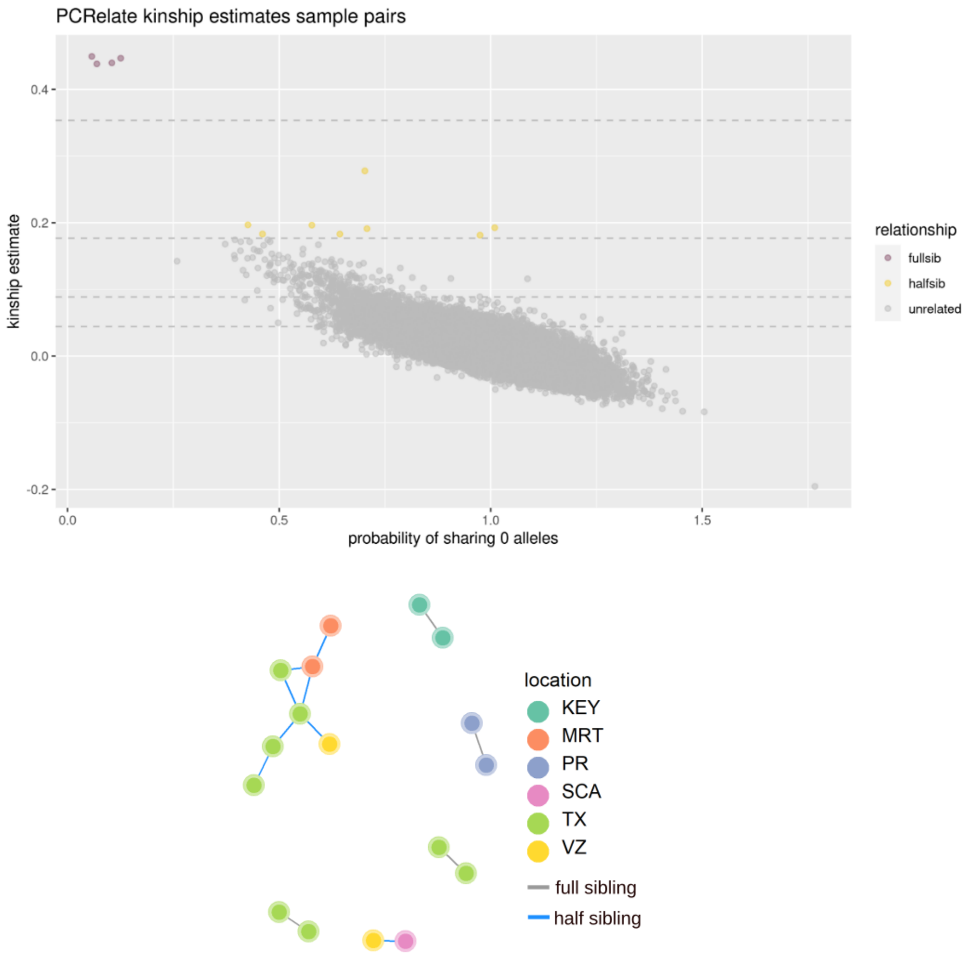

Figure 4.2 *Pairwise relatedness estimates in blackfin tuna samples from 9 localities*

Figure 4.2 continued. *Pairwise relatedness (r) estimates between 334 Blackfin tunas collected from 9 localities in the western Atlantic Ocean. The distribution of estimates of r (x-axis) is represented as a function of the probability of sharing no alleles (x-axis) on the top panel and a network diagram illustrates the relationships between individuals involved in putative kin pairs (bottom panel).*

Table 4.2 *Number of sibling pairs identified using PCRelate*

Relationship	Total Pairs	From Same Locality
Full Sib	4	4
Half Sib	8	4

OutFLANK detected 44 outlier SNPs although only 5 of these had expected heterozygosity values above 0.1 and were the most robust candidate outliers (Figure 4.3). Bayescan identified 1 SNP outlier, which was among the 5 putative outliers identified by outFLANK discussed above. Because outFLANK is expected to reduce the rate of false positives in limited datasets (Whitlock and Lotterhos 2015), all 44 outliers identified by outFLANK were conservatively removed to generate a neutral dataset (non-outlier) and were retained for the outlier dataset. Further analyses proceeded separately for putatively outlier and neutral loci.

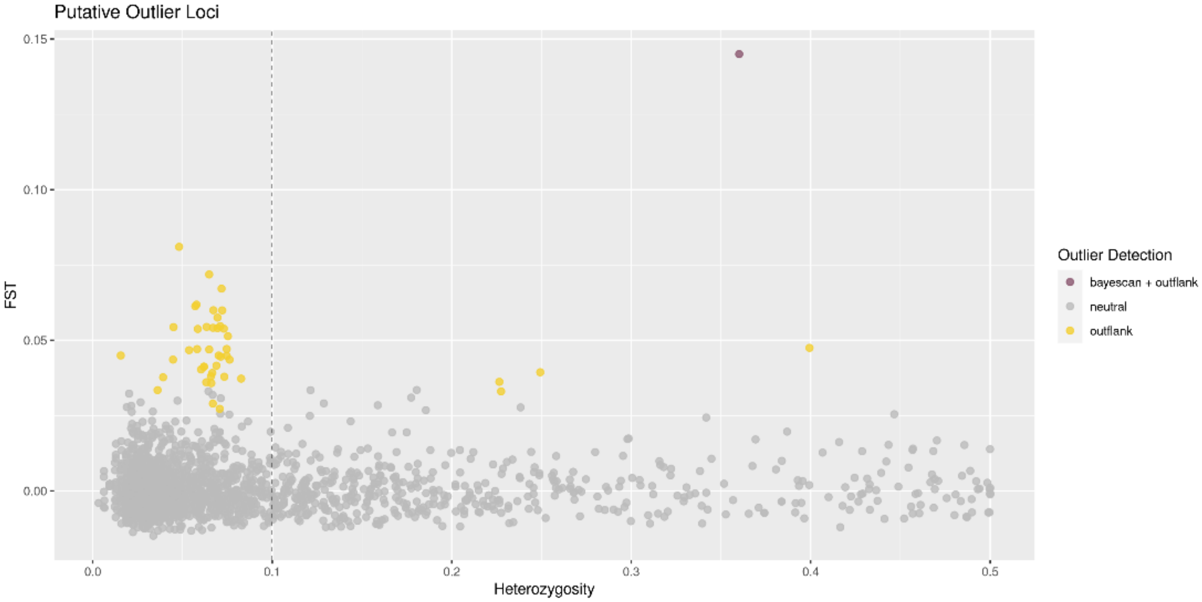

Figure 4.3 *Distribution of F_{ST} as a function of heterozygosity in neutral loci (grey), outliers detected by outFLANK and Bayescan (maroon) and in outFLANK only (yellow). The dashed vertical line is heterozygosity = 0.1.*

Cross validation led to retaining 98 principal components and 3 linear discriminant functions for the neutral dataset and 15 principal components and 3 linear discriminant functions for the outlier data. Samples were grouped as follows to facilitate visualization of the results: BRZ and BRZ_SP as a single “Brazil" group, SCA as an “Atlantic” group (ATL); MRT, VZ, PR as a “Caribbean” group; and TX, PR, PNS, KEY as a “Gulf of Mexico” group. The first two linear discriminant components of the DAPC (Figure 4.4, Figure 4.5) separate three groups with distinct but overlapping centroids: the Brazilian samples, the Atlantic samples, and the Gulf of Mexico and Caribbean samples as a third group. The third linear discriminant component differentiates the Gulf of Mexico and Caribbean samples although the divergence of the two groups is weaker with some overlap. DAPC of the outlier dataset also showed divergence of the Brazilian samples but there was no clear sub-structuring within the rest of the samples (Gulf, Caribbean, and US East coast).

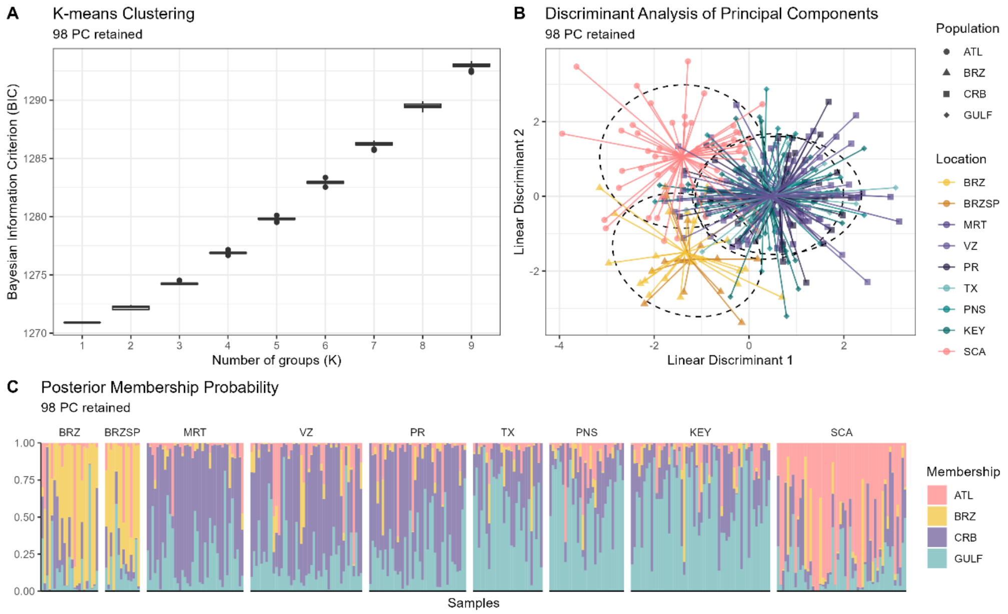

Figure 4.4 *Discriminant analysis of principal components of neutral markers*

Discriminant analysis of principal components of neutral markers (DAPC) for 326 Blackfin tuna from 9 localities in the Western Atlantic Ocean based on 2,096 putatively neutral SNPs (98 Principal Components, 3 Discriminant Analysis functions). (A) Kmeans clustering performed iteratively 50 times for each K. (B) Representation of individuals from each locality on the linear discriminant functions 1 and 2, and (C) Posterior membership probability of a sample to a population (populations aggregated in 4 groups, see text for description of the groups employed).

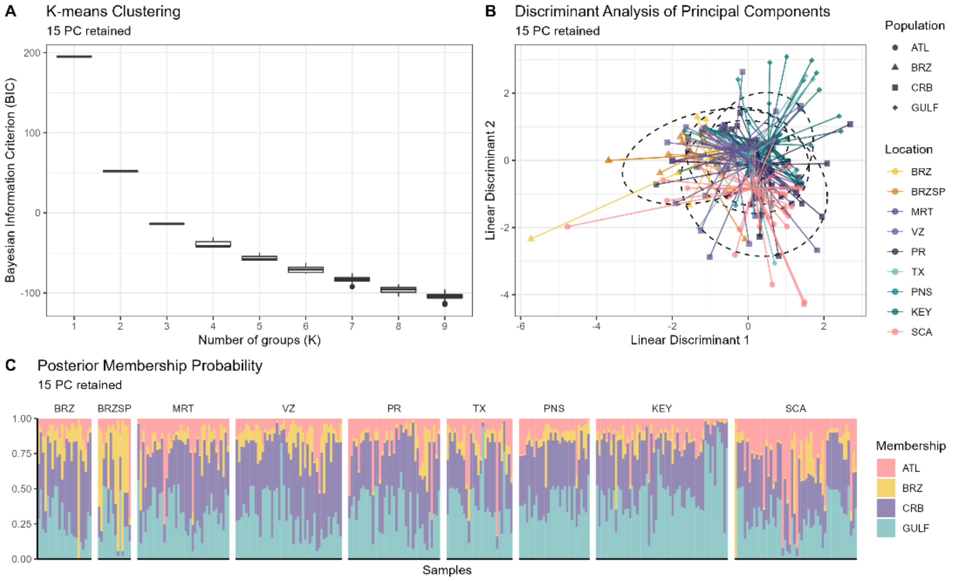

Figure 4.5 *Discriminant analysis of principal components of outlier markers*

Discriminant analysis of principal components (DAPC) for 326 Blackfin tuna from 9 localities in the Western Atlantic Ocean based on 44 putatively outlier SNPs (15 Principal Components, 3 Discriminant Analysis functions). (A) Kmeans clustering performed iteratively 50 times for each K. (B) Representation of individuals from each locality on the linear discriminant functions 1 and 2, and

(C) Posterior membership probability of a sample to a population (populations aggregated in 4 groups, see text for description of the groups employed).

The most supported configuration in fastStructure in the neutral data was one single cluster, whereas k=2 had the highest likelihood in the outlier dataset. In the latter dataset, the second cluster was represented in a variety of samples from all locations except Puerto Rico.

Pairwise F_{ST} estimates in the neutral dataset (Table 4.3) ranged from 0.0002 (MRT-PNS) to 0.0025 (KEY-TX) and after FDR correction, two population pairs had significant pairwise F_{ST} ($\alpha = 0.05$) in the neutral dataset (BRZ-KEY and TX-KEY). Pairwise F_{ST} estimates in the outlier dataset (Table 4.4) ranged from -0.0003 (SCA-KEY) to 0.1174 (TX-PR) and after FDR correction, 16 populations pairs had significant pairwise F_{ST}.

Table 4.3 *Pairwise F_{ST} estimates of neutral SNPs*

	Brazil	Gulf of Mexico			Caribbean			Atlantic
	BRZ	TX	PNS	KEY	MRT	PR	VZ	SCA
BRZ		0.0131	0.1505	0.0017*	0.0500	0.5010	0.2216	0.2661
TX	0.0019		0.0575	0.0001*	0.0994	0.0964	0.1295	0.1193
PNS	0.0012	0.0016		0.0259	0.8459	0.3205	0.5627	0.2810
KEY	0.0016	0.0025	0.0016		0.1322	0.0299	0.0097	0.0007*
MRT	0.0013	0.0010	0.0002	0.0009		0.21081	0.5420	0.0280
PR	0.0008	0.0015	0.0011	0.0013	0.0010		0.6274	0.0974
VZ	0.0009	0.0013	0.0007	0.0012	0.0005	0.0006		0.2251
SCA	0.0009	0.0015	0.0011	0.0014	0.0012	0.0011	0.0008	

Pairwise F_{ST} estimates of neutral SNP data (below diagonal) and associated P-values (above diagonal) comparing samples of Blackfin tuna geographic populations. Results are rounded to 4 decimal places. An asterisk (*) denotes a P-value significant at $\alpha = 0.05$ after FDR correction.

Table 4.4 *Pairwise F_{ST} estimates of outlier SNPs*

	Brazil	Gulf of Mexico			Caribbean			Atlantic
	BRZ	TX	PNS	KEY	MRT	PR	VZ	SCA
BRZ		0.0001*	0.5014	0.048	0.0016*	0.0001*	0.3478	0.1534
TX	0.104		0.0002*	0.0007*	0.0443	0.0001*	0.0002*	0.0011*
PNS	0.0098	0.0675		0.2066	0.0206	0.0001*	0.3958	0.2409
KEY	0.0002	0.0878	0.0049		0.0396*	0.0017*	0.2347	0.3664
MRT	0.0457	0.0243	0.0148	0.029		0.0001*	0.0104*	0.0258*
PR	0.0273	0.1174	0.0166	0.0245	0.0491		0.0001*	0.0001*
VZ	0.0014	0.083	0.0028	0.0019	0.0317	0.0211		0.1663
SCA	0.0049	0.0763	-0.0003	0.0034	0.0213	0.024	0.0049	

Pairwise F_{ST} estimates (below diagonal) and associated P-values (above diagonal) comparing samples of Blackfin tuna geographic populations. Results are rounded to 4 decimal places. An asterisk (*) denotes a P-value significant at α = 0.05 after FDR correction.

Analyses of molecular variance revealed no significant temporal or spatial component of molecular variance (P = 0.33 for year of capture and P = 0.99 for location of capture, data not shown), which was consistent with the very low F_{ST} estimates reported above. Similarly, the temporal and spatial components of molecular variance were not significant during the AMOVA conducted on the outlier dataset (P = 0.92 for year and P = 0.99 for location).

The Mantel test yielded a non-significant correlation between genetic distance and the logarithm of geographic distance (r = 0.014, P = 0.161). Spatial autocorrelation analysis runs using distance classes in increments of 100 km revealed that the highest correlation of genotypes was observed when samples were aggregated within a 500 km distance (Figure 4.6). Spatial autocorrelation remained significant for distances up to 800km.

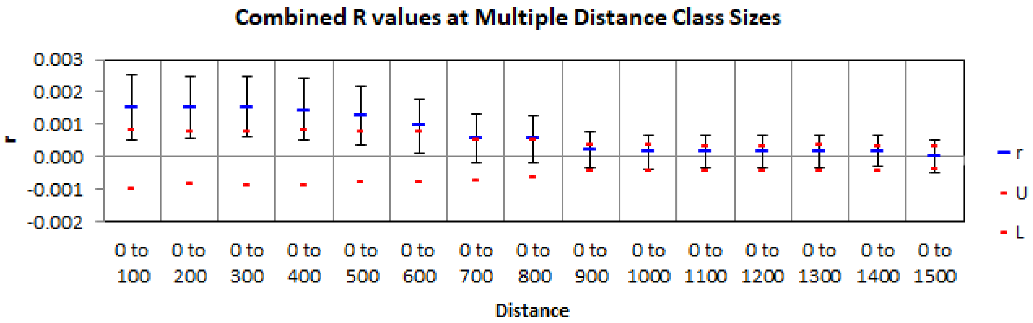

Figure 4.6 *Spatial autocorrelation*
Correlograms illustrating the influence of geographic distance on spatial autocorrelation. Correlation (*r*) of genotypes sampled in proximal locations (at distances less or equal to the first distance class) is estimated when the first distance class is increased in increments of 100 km. x-axis: distance class (km); y-axis: spatial autocorrelation (*r*). 95% confidence error bars for *r* were estimated by bootstrapping over pairs of samples; Red dash sy Mbols represent upper and lower bounds of a 95% CI for *r* generated under the null hypothesis of a random geographic distribution of Blackfin tuna.

4.4 Discussion

In this work, samples from localities spanning most of the Blackfin tuna's distribution range were characterized using 2,139 SNP loci, providing substantially improved inference power compared to previous studies of genetic variation in this species. The dataset also provided a first assessment of loci putatively under divergent selection.

All pairwise F_{ST} values were very low (<0.005) indicating divergence among geographic populations was very weak across the sampled range, a finding consistent with past surveys of Blackfin tuna populations using microsatellite markers alone (Saillant et al., in review) or in combination with mtDNA sequence variation (Saxton 2009). Weak divergence across large geographic surfaces is common in tunas (Barth et al. 2017; Pecoraro et al. 2018; Anderson et al. 2019) and other large pelagics and likely reflects high gene flow facilitated by their ability to travel long distances, combined with reduced effects of genetic drift. Blackfin tuna display high levels of genetic diversity (Antoni et al. 2014), which suggests they harbor large population sizes and their differentiation under genetic drift is therefore expected to be slow. Blackfin tunas, like other species in the region, are presumed to have expanded their ranges following the last glacial maxima (Pruett et al. 2005; Ely et al. 2005). Accordingly, some geographic populations may be currently isolated, but not have accumulated enough genetic difference to be detectable with present methods, especially if populations experience periodic residual gene flow (Pruett et al. 2005). This scenario is plausible for Blackfin tuna due to the species' high mobility during early life stages (passive dispersal) and adult life stages (active migration).

Population structure was detected during DAPC and spatial autocorrelation analyses. DAPC suggested the occurrence of up to 4 units. A first group supported by both neutral and outlier SNPs included the two Brazilian locations. The divergence of Brazilian populations from those located further north was also reported during a recent analysis of population structure using microsatellites (Saillant et al., in review). A first possible factor that could contribute to isolation between Brazilian and northern populations is the Amazon-Orinoco plume and the Mid Atlantic barrier which have been proposed to explain isolation of reef fishes dependent on pelagic larval dispersal (Luiz et al. 2012). If Blackfin tunas demonstrate regional fidelity as adults, as suggested by tagging studies, and connectivity between geographic stocks is mediated by passive larval dispersal, these barriers could be effective at maintaining Brazilian populations partially isolated. A second factor potentially involved in divergence is the differences in spawning seasons between northern and Southern hemispheres which could lead to unfavorable conditions encountered by larvae migrating from the southern to northern hemisphere and/or low reproductive success of adult migrants that would need to adjust to a 6-month shift in reproductive period (Saillant et al., in review). Sampling Blackfin tuna populations between north Brazil and the Southern Caribbean Sea would be useful to confirm occurrence of a discontinuity and assess patterns and rates of gene flow between the two groups.

The neutral SNP dataset provided further insights on subdivision within Blackfin tunas stocks located north of Brazil. The first two discriminant components separated a northern Atlantic group consisting of samples collected offshore South Carolina while the third discriminant component suggested some divergence between the Gulf of Mexico samples (including the Keys) and the Caribbean samples (Venezuela and La Martinique) although with some overlap between these two groups. Divergence between the US East coast and the Gulf of Mexico was suggested by an earlier study using mitochondrial DNA and 6 heterologous microsatellites (Saxton 2009) but not confirmed in the study of Saillant et al. (in review) with 13 homologous microsatellites, who only reported a weak isolation by distance pattern and no subdivision within Blackfin tunas sampled north of Brazil. These inconsistencies likely reflect, in part, that previous datasets had insufficient power to detect the very fine divergence between the three “northern” groups. Blackfin tunas from the three regions could be isolated because of reduced adult movements and/or limited larval transport. Tagging studies to date were conducted in Bermuda (Luckhurst et al. 2001) and the Southern Caribbean (Singh-Renton and Renton 2007) and were uninformative on movement between Gulf of Mexico, Caribbean, and northwest

Atlantic, although both studies indicated site fidelity of tagged fish (Luckhurst 2014), tentatively suggesting that site fidelity of adults could contribute to the isolation of the three groups. Connectivity between the Gulf of Mexico, Caribbean and East US coast could occur at the larval stage through passive transport; *Thunnus* larvae tend to be widely distributed in the continental shelf and the continental slope in the north central Gulf of Mexico (Cornic et al. 2017). The loop current which becomes the Florida current and then the Gulf Stream (http://oceancurrents.rsmas.miami.edu/atlantic/atlantic.html) was discussed to promote favorable conditions for *Thunnus* larvae when it extends farthest north (Cornic et al. 2017). Larvae distribution overlapped with the current itself and those interacting with the current (e.g., located East of the Mississippi river) could therefore be transported to the Keys or the United States east coast if conditions are favourable to their survival. Mesoscale structures, such as eddies and fronts have been hypothesized to create suitable foraging habitat for early life stages of tunas (Lang et al. 1994). Eddies spinning off the loop current promote opportunities for movement as they propagate, typically westward (i.e., from Central Gulf to the western Gulf, Damien et al. 2021) but larvae caught within the main Loop Current and transported towards the East coast would be expected to be outside of the favorable conditions promoted by eddies and may have low survival. Accordingly, recruitment would be promoted in the northern Gulf (yet with mixing within the Gulf), isolating this group from the East coast. A similar mechanism may contribute to prevent effective transport of larvae from the Caribbean to the Gulf and northwestern Atlantic through the Caribbean current or from the Caribbean Islands to the northwestern Atlantic via the Antilles current. In addition, direct transport from Caribbean islands to the Gulf of Mexico or northwestern Atlantic would be unlikely according to larval dispersal envelopes estimated by Roberts (1997) although these envelopes were determined for reef fish and would need to be generated for blackfin tunas that spawn farther offshore.

Evidence for spatial structuring was also provided by isolation by distance analysis. In this study, the slope of the isolation by distance model was not significantly different from zero, but significant spatial autocorrelation of samples collected within 800 km was observed. Isolation by distance was also inferred from the study of variation at microsatellites in a previous study (Saillant et al., in review) and is consistent with the site fidelity of adults and/or restrictions to effective dispersal at the larval stage discussed above. The 800 km distance at which spatial structure is detected is close to the maximum distance separating individual localities within each of the three “northern” groups discussed above. Accordingly, it is possible that the clustering obtained in DAPC is driven in part by the isolation by distance pattern, as clustering methods are prone to infer false barriers in populations evolving under isolation by distance (Blair et al. 2012). Disentangling the role of isolation by distance and that of possible discontinuities within the range is difficult with this dataset because of the distribution of the sampled localities. Characterizing additional localities within the range to increase sampling density would be helpful to formally determine whether discontinuities occur between the Caribbean Islands, the Gulf of Mexico, and northwestern Atlantic groups or if the genetic structure is primarily explained by an isolation by distance model. While isolation by distance may also be involved in the divergence between the Brazilian samples and the northern groups, the occurrence of a discontinuity between these two groups seems more likely. Divergence of the Brazilian group was already suggested by the analysis of microsatellite data (Saillant et al., in review) as discussed above and, in the present analysis, it was supported by DAPCs conducted on both the neutral and the outlier datasets. Divergence of the outlier dataset tentatively would indicate occurrence of adaptive variation contributing to divergence in South America, another argument for true isolation of this group. In contrast, the outlier dataset did not provide a clear pattern of divergence between the three northern groups, a finding consistent with the hypothesis that residual gene flow is occurring.

Information on the geographic location of capture was limited for some of the localities where samples were obtained from fishing boats at landing who did not communicate the exact coordinates of captures (captures were assumed to have occurred within 150 km of the landing port in those cases). Therefore, comparisons of genotypes collected at small distances were lacking from the dataset and may have prevented detection of isolation by distance in Mantel tests. In non-equilibrium situations, isolation by distance establishes first at short distance scales (Robledo-Arnuncio and Rousset 2010) and reaches a plateau when geographic distance between samples exceeds $0.56\sigma/\sqrt{2\mu}$, where σ is the standard deviation of parental position relative to offspring position and μ is the mutation rate (Rousset 2008). Future studies incorporating a larger number of proximal localities with accurate capture coordinates would be valuable to refine the isolation by distance model and estimate dispersal distance parameters.
The low F_{ST} between localities was a major challenge in this study and likely contributed to the lack of significance in most of the spatial analyses. Low F_{ST} has been shown to create clustering inaccuracies (Miller et al. 2020) and F_{ST} values in the range of those obtained here are incompatible with detection of subdivision in Structure (Chen et al.

2007). This issue can be overcome in future studies by increasing the sampling density as discussed above as the power to detect isolation by distance patterns is improved when samples separated by short distances are included during estimation (Leblois et al. 2003). Increasing the density of the genome scan would also improve the likelihood of detecting structure related to local adaptation when it occurs. Population structure of other tunas was indeed revealed by markers under selection, even when groups were homogeneous at neutral markers (e.g., Pecoraro et al. 2018).

In this study, 44 putatively outlier loci were identified in outFLANK. However, only 5 of these loci had heterozygosity greater than 0.1 and can be considered robust candidate loci experiencing selection (Whitlock and Lotterhos 2015). Similarly, analyses in Bayescan only identified one outlier. The small number of candidate outlier loci found in this work suggests that, if they exist, the genomic regions affected by divergent selection and local adaptation may be very limited. However, considering the number of loci surveyed in this genome scan (2,139) and estimates of the size of Blackfin tuna genome (774 Mb), the average interval between markers was 362 kb such that a selected locus would be expected to be within 181 kb of one of the markers surveyed in this study. Genomic regions affected by selection may have remained undetected considering that the average size of linkage blocks in studies of other fish is only a few kilobases (Lowry et al. 2017). We note that 3 pairs of the 44 candidate outliers identified by outFLANK were SNPs defined on the same genomic contigs, which strengthens the inference of selection at these loci. Further monitoring of these strong candidate outliers is warranted to confirm the pattern detected in this study and determine if the signal is stable over time. Increasing the density of the genome scan is also warranted to capture a greater fraction of adaptive variation in the species. The estimates of pairwise F_{ST} identified in this work are on average 27.5 times higher in the outlier dataset than the neutral dataset and sampling a greater proportion of the genome not only will provide more information on local adaptation but will also improve the power to detect population subdivision. Greater genomic sampling can be achieved with a more complete genome assembly and a larger set of loci across the entire genome such as those derived from low coverage whole-genome sequencing (Therkildsen and Palumbi 2017). Information on the genomic proximity of genetic loci would also allow performing a sliding window analysis where F_{ST} is assessed in groups of markers located in the same genomic regions. This approach is expected to reduce the occurrence of false positive outliers by observing the lack of signal in neighboring loci (Hohenlohe et al. 2010; Bourret et al. 2013). The draft reference assembly generated in this study was incomplete and highly fragmented due to the type of sequencing data used to generate the assembly. In the absence of a linkage map, information on the genomic proximity of candidate outlier loci was limited to only those sharing the same contig. A more complete genome assembly, scaffolded to a chromosome level draft genome, would therefore be valuable by enabling sliding window analyses of Blackfin tuna to identify putative genomic regions of selection, if they exist.

The marginal evidence for divergent selection and local adaptation may also be related to the high levels of gene flow in Blackfin tuna. High gene flow is expected to counterbalance the differentiation caused by divergent selection and local adaptation, effectively preventing local adaptation from occurring, or limiting it to loci affected by strong selective pressures (Lenormand 2002; Conover et al. 2005; Cheviron and Brumfield 2009). Genomic studies of other marine species revealed the occurrence of outliers in metapopulations that were also exhibiting structure at neutral markers (Nielsen et al. 2009; Bradbury et al. 2010; Limborg et al. 2012; Laconcha et al. 2015). However, outlier loci were also discovered in metapopulations where no significant spatial structure was observed at neutral loci (Lamichhaney et al. 2012; Grewe et al. 2015; Pujolar et al. 2014). It is possible that Blackfin tunas utilize their high capacity for movement to select habitats with favorable characteristics across their range leading to little or no local selection, although the wide range utilized by the species suggests that regional populations would differ by some environmental characteristics such as the differences in reproductive season timing discussed above for the South American group.

The observed unimodal distribution of pairwise relatedness coefficients across the samples indicated the absence of large groups of related individuals. However, based on simulated distributions of unrelated, half siblings, and full siblings, a few sample pairs were inferred as kin, and 8 out of 12 of inferred kin had members sampled in the same locality. The observation of close kins within localities is inconsistent with the large population sizes hypothesized for Blackfin tuna. A hypothesized mechanism potentially contributing to the unexpectedly high incidence of co-located kins involves some behavioral cohesion (“close kin co-dispersal”), where larvae spawned by the same parents remain together through early life stages and, in some cases, may stay together through sexual maturity (Anderson et al. 2019). Another potential explanation is sweepstake recruitment where cohorts of a regional population include a disproportionate contribution of a few siblings (Hedgecock and Pudovkin 2011). The present study cannot distinguish between these hypotheses, but further sampling to investigate patterns of relatedness across various life stages is warranted to better evaluate the close kin co-dispersal hypothesis. Investigation of patterns of relatedness in larvae at close spatial scales would also be interesting to assess variance in reproductive success and further evaluate the sweepstake hypothesis.

CHAPTER V – EVOLUTIONARY HISTORY AND DIVERGENCE OF TUNA SPECIES USING WHOLE GENOME COMPARISONS

5.1 Introduction

Investigations of the evolutionary processes of divergence and speciation have documented numerous examples of allopatric speciation, where geographic isolation due to barriers to gene flow leads to the accumulation of genetic differences resulting from genetic drift and mutations (Feder et al. 2012). The marine environment challenges this notion of speciation in that strict barriers to gene flow are much less frequent, leading to speciation processes in situations where residual gene flow exists, and new species arise following isolation and adaptation along environmental or depth gradients (Ingram and Mahler 2011) or result from factors such as differences in resource use (Miglietta et al. 2011). These examples of sympatric speciation are consistent with the increasing evidence that marine populations are genetically structured and display extensive adaptive differentiation (Hauser and Carvalho 2008). Significant structure was observed in benthopelagic species such as Atlantic cod (Berg et al. 2017), small pelagics such as herring (Ruzzante et al. 2006), and even the highly mobile migratory bluefin tuna within the Atlantic basin (Carlsson et al. 2006). Structure in many of these cases reflected different environment preferences or habitat use within the same geographic range and may be leading to sympatric speciation in these taxa following divergent selection followed by divergence hitchhiking and genomic hitchhiking (Feder et al. 2012).

High density genome scans provide an opportunity to study the process of divergence and speciation in detail. Genetic divergence is expected to be heterogeneous across the genome (Nosil et al. 2009; Michel et al. 2010), where neutral regions will diverge under the effects of genetic drift at a rate relative to the effective size of the two species, while regions experiencing divergent selection involved in speciation are expected to show increased divergence as islands of speciation (Bradbury et al. 2013). Conversely, genomic regions under the same evolutionary constraints in the diverging species would retain a higher degree of similarity. Regions neutral to speciation would evolve under the stochastic effects of genetic drift and residual migration with new mutations taking a greater role as gene flow decreases in the newly formed species. Genome-wide species comparison enables detecting outlier regions (islands of speciation), where divergence is more pronounced than is expected for neutral regions (Hofer et al. 2012). Theoretically, regions conserved between taxa due to shared selective constraints may be detected as outliers showing reduced divergence as compared to the rest of the genome during genomic scans (Feng et al. 2015; Dalongeville et al. 2018). Detecting these regions can be challenging due to the potential confusion with neutral regions, some also expected to show reduced divergence by chance, but high-density genome scans may circumvent this issue if the marker density is sufficient to yield several markers in linkage disequilibrium with the affected loci. Congeners occupying overlapping ranges (sympatry) are a unique system to study factors involved in sympatric speciation, where the identification of sister species is essential to effectively analyze genomic signatures of speciation as it allows assuming a simple model of isolation with residual migration.

The genus *Thunnus* features several groups of sympatric species (Figure 5.1) with high dispersal, thus providing opportunities to study the sympatric speciation process. Tunas are characterized by large populations sizes (Qiu et al. 2013; Laconcha et al. 2015;

Waples et al. 2018; Qiu and Miyamoto 2011), which limits the effects of genetic drift on neutral regions and slow the swamping of islands of divergence over time (Quilodrán et al. 2020). Another advantage of this group is that the ranges of some of these species overlap which provides the opportunity to study shared evolutionary constraints such as geographic barriers, environmental conditions or other factors impacting fitness.

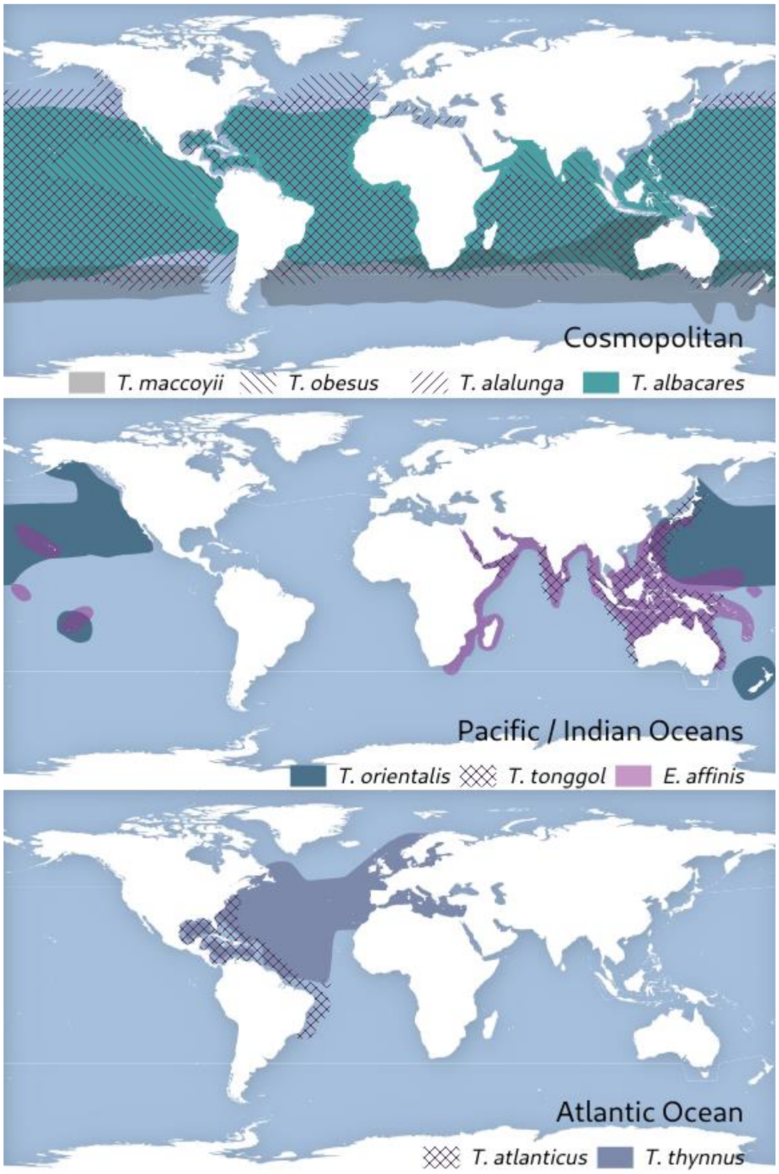

Figure 5.1 *Ranges of scombroid species under investigation (source: IUCN Red List)*

The tuna group itself is taxonomically challenging to define, as they are members of the Scombridae family (sub-family Scombrinae) but defined under the tribe sub-classification *Thunnini*. Within *Thunnini* there is a further distinction between the genus *Thunnus* ("true tunas" e.g., Yellowfin tuna and Albacore tuna) and several genera for "lesser" tunas (e.g., mackerel tuna, frigate tuna). The *Thunnus* genus consists of 8 species within two subgenera, of which 5 species occur in the Atlantic Ocean: Albacore (*T. alalunga*), Bigeye tuna (*T. obesus*), Atlantic bluefin (*T. thynnus*), Yellowfin tuna (*T. albacares*), and Blackfin tuna (T. atlanticus, de Sylva 1955). Initial molecular phylogeny using sequences of the ITS1 nuclear gene and the mitochondrial genes ATPase 6 and Cytochrome oxydase III suggest a close relationship (monophyly) between *T. albacares*, *T. atlanticus*, *T. tonggol, and T. obesus* and a close relationship between *T. thynnus*, *T. orientalis, and T. alalunga*. (Chow et al. 2006). Later, a single nucleotide polymorphism (SNP) based phylogeny derived from restriction site associated DNA methods (Peterson et al. 2012) resolved a phylogeny using up to ~70,000 genome wide markers putatively from both coding and non-coding regions of the genome (Díaz-Arce et al. 2016). This new phylogeny suggests *T. albacares* and *T. obesus* are sister taxa (recently diverged), as are *T. altlanticus* and *T. tonggol,* along with *T. thynnus* and *T. orientalis*. Introgression between some tuna species has been suggested, rendering mtDNA-based phylogenies inaccurate (Viñas and Tudela 2009; Díaz-Arce et al. 2016). A recent analysis of the transcriptome of species in the genus *Thunnus* corroborated the SNP-derived phylogeny and suggested selection was occurring in the tropical tuna clade (*T. albacares, obesus, atlanticus* and *tonggol*) related to growth and endothermy (Ciezarek et al. 2019). However, the authors noted that many more genes related to reproduction or other phenotypic traits were likely not captured in the muscle transcriptome analyzed in the study. A genome wide assessment of divergence, not limited to a single tissue transcriptome or subset of genomic positions, may thus shed more insights on the nature of the adaptations involved in speciation events and their location in the genome. A genomewide analysis would also clarify evolutionary relationships among global tunas. The two recent phylogenies challenge the two original molecular phylogenies presented by Chow et al (2006) and Viñas and Tudela (2009), and each acknowledges the limitations of marker density, representativity of genomic regions, or missing data, potentially impacting inferences on phylogenetic relationships.

A formal evaluation of evolutionary relationships between tunas requires robust molecular phylogenies accurately reflecting genome-wide variation and not limited to reduced representation or tissue-specific transcripts. The state of the art for molecular sequencing can generate assemblies spanning most of the genome of eukaryotic species with medium genome size such as tunas for less than $1,000. Sequencing and assembly of the genomes of the 3 remaining representatives of the genus *Thunnus* (of 8 species total) that do not already have genome assemblies available can thus be achieved with moderate effort and would be sufficient to complete a thorough evaluation of genome wide variation between all eight species by constructing a cross-species SNP-based phylogeny and applying it to study patterns of divergence between congeners. The obtained resources can be used to study genomic regions involved in the divergence of specific pairs of taxa or groups of taxa (e.g., Atlantic-exclusive tunas versus Pacific-exclusive tunas or temperate versus tropical tunas). This genome-wide approach can also be applied to study the divergence process within species occupying overlapping ranges.

As an example, both *T. atlanticus* and *T. albacares* species are sympatric in the Western Atlantic Ocean and often share habitats, so they may be experiencing similar selective pressures. The outlier SNPs identified in Atlantic *T. albacares* populations in *Chapter III* and *T. atlanticus* populations in *Chapter IV* may provide insight into whether selective pressures may be affecting both species in similar genomic regions, and if otherselective pressures occur in regions associated with areas of highest divergence between the species, acting as drivers of speciation.

5.2 Objective

The objective of this chapter was to develop genomic resources for use in comparative studies of genome wide variation both within and between the *Thunnus* species and conduct initial investigations of 1) the phylogenetic relationships between the tropical and temperate tunas, 2) the genome wide patterns of divergence between species and groups of species and possible evidence for genomic islands of speciation, 3) divergence between species or groups of species related to specific functions, and 4) adaptive divergence affecting sympatric congeners targeting the same genomic regions. Directions for further investigation are discussed.

5.3 Methods

5.3.1 Genome assembly

For each of *Thunnus. Obesus* (Atlantic Ocean), *T. alalunga* (Atlantic Ocean), and *T. tonggol* (Pacific Ocean), DNA was extracted from a representative individual using the Blood & Tissue DNA HDQ 96 Kit and sequenced on an Illumina NovaSeq6000. Sequencing was performed at 2x150 bp paired end reads for a target of 100x genomic coverage. The raw sequences were trimmed using fastp as described in *Chapter II* and

Chapter IV. assembly details are shown in Appendix C. Briefly, trimmed reads for each species were then assembled using ABySS (Simpson et al. 2009; Jackman et al. 2017). The resulting assemblies were then processed in redundans (Pryszcz and Gabaldón 2016) to remove haplotigs. A separate assembly for each species was then performed using SparseAssembler (Ye et al. 2012) with a Kmer size of 90 (EdgeCovTh 3) and removing chimeras. The resulting primary contigs were consolidated into consensus sequences using DBG2OLC (Ye et al. 2016) with a Kmer size of 31. Trimmed reads were then aligned to the consensus assembly using minimap2 (Li 2018) and the alignments were filtered using SAMTools (Li et al. 2009; Danecek et al. 2021) to remove empty and low-quality alignments, then sorted and indexed. These curated alignments were then used to polish the assembly using HyPo (Kundu et al. 2019). The polished assembly was screened for haplotigs using purge_haplotigs (Roach et al. 2018). assembly contiguity has been shown to improve by merging multiple assemblies generated by different assemblers (Chakraborty et al. 2016). Accordingly, the separate assemblies (ABySS/redunans, SparseAssembler/DBG2OLC) were then merged using quickmerge (Chakraborty et al. 2016), where the length cutoff was determined by the N50 of the less-contiguous assembly in each pair. In each species, the query assembly was the one with fewer contigs and the reference was the assembly with greater contigs. The merged assembly was screened for haplotigs, scaffolded, and gap filled using redundans. Reads were mapped to the scaffolded assembly and polished using HyPo.

Given the improvement in short read assembly contiguity achieved with this pipeline, the draft assembly of *T. atlanticus* was reassembled using these methods and the resulting assembly was used in genomic comparisons.

5.3.2 Phylogenetic analysis

The evolutionary relationships of the Atlantic *Thunnus* species were investigated using molecular phylogenies constructed based on whole-genome comparisons. The genome for *T. albacares* was made available by the work in *Chapter II* and the genome for *T. atlanticus* was reassembled as described above. Using the NCBI SRA, genomes for the other species for evolutionary comparison were retrieved: the Atlantic Bluefin tuna (T. thynnus, accession GCA_003231725.1, Puncher et al. 2018)), Pacific Bluefin tuna (T. orientalis, accession GCA_009176245.1 Suda et al. 2019), the Southern Bluefin tuna (T. maccoyii, accession GCA_910596095.1, McWilliam et al. 2016), and the Mackerel Tuna (Euthynnus affinis, , accession GCA_019973915.1, Havelka et al. 2021)

5.3.3 Identifying variants

The genomes of the 9 species had extremely varied contiguities (58 to >180 k scaffolds), therefore a SNP-based approach was used to mitigate structural variant bias due to assembly quality. A manual approach to the process outlined in the REALPHY phylogenetic software (Bertels et al. 2014) was used to be more efficient with runtime and memory requirements, relying on well tested existing third-party software instead of custom Java implementations of routine variant calling practices. First, the genomes are converted from FASTA to FASTQ format with seqtk (https://github.com/lh3/seqtk), then the assemblies are "fragmented" using seqkit (Shen et al. 2016) by creating a 150 bp sliding window of the sequences that advances by 1 bp. The southern bluefin tuna genome was the most contiguous (58 scaffolds) and was therefore chosen as the reference to align the other genomes to. The fragmented assemblies, including the fragmented version of the reference assembly, were then mapped onto the Southern bluefin tuna genome using bwa (Li 2013; Li and Durbin 2009) with a K-mer size of 22 and outputting all alignments, maintaining congruence with the read mapping done using bowtie2 (Langmead and Salzberg 2012) within REALPHY (Bertels et al. 2014). Alignments were filtered to remove unaligned regions, converted to binary BAM format, sorted, and indexed using samtools (Danecek et al. 2021; Li et al. 2009). The alignments were then used to identify variants using freebayes (Garrison and Marth 2012), which was performed assuming a haploid model because input genomes were haploid assemblies, i.e., a single consensus allele could be represented in the sequences for each base. Freebayes was run by parallelizing over 5 kb genomic windows (https://github.com/freebayes/), with the additional parameters of requiring a minimum of 1 alternate observation to call a SNP, a minimum coverage of 5, the –standard-filters option, and treating each species as a separate population.

5.3.4 Filtering Variants

The raw variants were filtered using bcftools (Danecek et al. 2021). First, variants with low genotype quality (<30), low mapping quality (<40), and depth (<10) were removed. Next, sites with depth greater than twice the mean depth in the dataset were removed as possible paralogs. Monomorphic sites and sites with any missing data were also removed. Sites containing multi-nucleotide polymorphisms were decomposed into separate adjacent SNPs and indels using vcfallelicprimatives. It was assumed that any site for which the reference genome sample did not have the reference allele was a false positive SNP generated from Freebayes and such sites were removed. Finally, data size and redundancy were reduced by thinning the data to retain 10 sites for every 20,000 bp using the maximum allele frequency model in bcftools +prune plugin (Danecek et al. 2021).

5.3.5 Phylogenetic reconstruction

Filtered variant data were converted into FASTA file format using vcf2phylip (Ortiz 2019). MAFFT (Katoh and Standley 2013), was used to align the sequences with the automatic model-choosing parameter –auto and a maximum of 1,000 iterations. The alignment generated by MAFFT was used to build phylogenetic trees in RaxML-NG (Kozlov et al. 2019) which was run with the GTR+G mutation model (Rodríguez et al. 1990; Abadi et al. 2019; Miura 1986), 100 bootstrapped trees, scaled branch lengths, 25 initial parsimonious trees, 25 initial random trees, and the *E. affinis* specified as the outgroup. The resulting tree topology was then evaluated with RaxML-NG to optimize the model parameters and the trees were reconstructed with the same number of bootstrapped and initial trees, but with the optimized mutation model parameters. The resulting phylogenetic tree was visualized with Dendroscope (Huson and Scornavacca 2012).

5.3.6 Species divergence

To estimate genomic divergence between the different scombroid species under investigation, the assemblies were aligned with an all vs. all strategy in a pairwise manner. Alignments were performed using nucmer from mummer (Marçais et al. 2018) configured to align with a maximum gap length of 2,000 bp and minimum cluster length of 1000 bp. In this analysis, alignments <500 bp were removed with delta-filter (provided by mummer) to ensure homology by capturing larger syntenic regions and to span large fractions of genes in coding regions. Divergence was then estimated as the global mean percent sequence difference (100 - percent sequence identity) across every alignment and visualized using the R language package corrplot (Wei and Simko 2021). To calculate mean divergence across a set of conserved alignments, the nucmer alignments of each species to *T. maccoyii* were consolidated and merged using bedtools (Quinlan and Hall 2010) to create a superset of *T. maccoyii* intervals that species aligned to. This was done by combining overlapping alignments, merging intervals that were within 1 kb, and retaining only those intervals with representation in all the species. This alignment superset was then used to identify alignments in the remaining species that correspond to the *T. maccoyii* interval superset, and the intervals from each species were merged if overlapping or 1 kb apart, creating a superset of shared intervals for each species. The species-specific interval supersets were then used to restrict the all vs. all nucmer alignment data to only pangenomic alignments (shared by all species), followed by calculating mean sequence divergence between each pair of species from these regions.

5.3.7 Candidate genes associated with divergence between species

To identify genes putatively responsible for divergence between sympatric tunas, the species groups were classified based on distribution patterns: Atlantic-only (ATL: *T. atlanticus, T. thynnus),* Pacific-only (PAC: *T. orientalis, T.tonggol*), cosmopolitan (COS: *T. albacares*, *T. alalunga*, *T. obesus*), and Southern Bluefin Tuna (SBF: *T. maccoyii*), which was classified separately due to its exclusively cosmopolitan-temperate distribution in the southern hemisphere. Alignments with a divergence greater than or equal to 20% (>99th percentile for all true-tuna pairs, Table 5.3) between ATL-PAC, ATL-COS, PAC-COS, COS-COS, and COS-SBF were isolated and the alignment positions were extracted with an additional buffer of ±1000 bp in an attempt to capture entire genes that may not be fully represented by the sequence alignment intervals and to conservatively capture surrounding regions likely affected by divergence hitchhiking (Feder and Nosil 2010). Overlapping alignment intervals and alignments within 10 kb had their intervals merged using bedtools, also under the assumption of divergence hitchhiking (Via 2012; Feder and Nosil 2010), and these merged alignments were then used to extract the sequences from one of the pair of genomes associated with the alignment. These highly divergent sequences were used for gene prediction using the default parameters of AUGUSTUS (Stanke et al. 2006) with the model trained on the zebrafish *Danio rerio*. The putative genes were then compared against the Non-Redundant Protein Sequences database using blastp (Camacho et al. 2009). Protein names were queried using the EMBL-EBI QuickGO API (https://www.ebi.ac.uk/QuickGO/) to identify the top 10 gene ontology ("GO") terms to assess general gene function.

5.3.8 Comparison of genomic regions under selection within species

The location of the putative outlier loci identified in the population studies of *Chapter III* and *Chapter IV* for *T. albacares* and *T. atlanticus*, respectively, were compared to assess if outlier loci within one species were occurring in the same genomic regions as outlier loci found in the second, (i.e., possibly responding to the same selective forces). Outliers found in the blackfin tuna study were mapped against the *T. albacares* genome using nucmer as described above. The SNP positions for each species were used to compare if the SNPs occurred in syntenic regions between the two species.

5.4 Results

5.4.1 Draft reference genomes for *T.* alalunga, *T. obesus* and *T. tonggol*

The draft genome assembly for *T. alalunga* spanned 744,485,078 bp in 179,788 scaffolds with a GC content of 39.78% (Table 5.1). The largest contig was 272,877 bp, and the N50 and L50 were 9,436 bp and of 19,363 contigs, respectively. The average gap size from scaffolding was 0.06 N’s per 100 kb. Genome completeness scores assessed based on the *vertebrata* database were 53.14% complete BUSCO score and 33.33% partial BUSCO.

Table 5.1 *Genome assembly metrics*

	T. tonggol	*T. obesus*	*T. alalunga*	*T. atlanticus*
Common name	Longtail Tuna	Bigeye Tuna	Albacore Tuna	Blackfin Tuna
Read Coverage	100X	100X	100X	200X
Total contigs	135,500	102,435	179,788	96,974
contigs > 1 kb	102,048	78,547	136,948	79,144
contigs > 5 kb	41,730	34,497	40,413	37,001
contigs > 10 kb	21,805	20,932	18,008	21,231
contigs > 25 kb	5,103	7,716	3,455	7,333
contigs > 50 kb	734	1,817	394	1,857
Total length	753,306,505	748,118,547	744,485,078	759,896,579
length > 1 kb	733,203,651	733,495,221	716,330,768	748,740,756
length > 5 kb	594,159,279	634,265,403	501,768,818	650,761,017
length > 10 kb	449,828,572	538,112,887	345,017,414	539,149,055
length > 25 kb	192,952,682	327,183,517	125,796,818	319,116,997
length > 50 kb	47,712,188	125,265,747	25,227,543	132,518,985
Largest contig	193,111	243,792	272,877	234,409
GC (%)	39.72	39.73	39.78	39.72
N50	13,615	21,706	9,436	20,494
N75	6,640	9,290	4,005	8,833
L50	14,670	9,417	19,363	9,777
L75	34,115	22,180	48,360	23,614
# N's per 100 k bp	0.73	0.07	0.06	0.93
Complete BUSCO (%)	59.9	67.6	53.14	68.9
Partial BUSCO (%)	27.0	20.8	33.33	21.5

The draft genome assembly for *T. obesus* spanned 748,118,547 bp in 102,435 scaffolds comprising 39.73% GC (Table 5.1). The largest contig was 243,792 bp, and the N50 and L50 were 21,706 bp and 9,417 contigs respectively. The average gap size from scaffolding was 0.07 N's per 100 kb. Genome completeness scores were 67.6% complete BUSCO score and 20.8% partial BUSCO.

The draft genome assembly for *T. tonggol* yielded 135,500 scaffolds comprising 753,306,505 bp and 39.72% GC (Table 5.1). The largest contig was 193,111 bp, and the N50 and L50 were 13,615 bp and 14,670, respectively. The average gap size from scaffolding was 0.73 N's per 100 kb. Genome completeness scores were 59.9% complete BUSCO and 27% partial BUSCO.

The reassembled *T. atlanticus* genome spanned 759,896,579 bp in 96,974 scaffolds comprising 39.72% GC (Table 5.1). The largest contig was 243,409 bp, and the N50 and L50 were 20,494 bp and 9,777. The average gap size from scaffolding was 0.93 N's per 100 kb. Genome completeness scores were 68.9% complete BUSCO score and 21.5% partial BUSCO.

5.4.2 Construction of a genome-wide phylogeny of true tunas

Variant calling and genotyping on the genomes aligned against *T. maccoyii* yielded 43,845,749 raw SNPs. Quality filters and thinning reduced the data to 93,451 biallelic SNP markers. The phylogenetic trees built from RaxML-NG converged to a single topology with 100% bootstrap value support for all bifurcations across varying numbers of initial trees and mutation model parameters (data not shown). Rebuilding phylogenetic trees after model optimization also yielded this same single topology, which is reported on Figure 5.2. The inferred phylogeny suggests *T. atlanticus* and *T. obesus* are sister taxa, as are *T. orientalis* and *T. thynnus*. The topology also suggests that the tropical tunas (*T. atlanticus*, *T. obesus*, *T. tonggol*, and *T. albacares*) are monophyletic and more derived compared to the other true tunas, with *T. albacares* being polyphyletic to the sister taxa pair *T. atlanticus* and *T. obesus*.

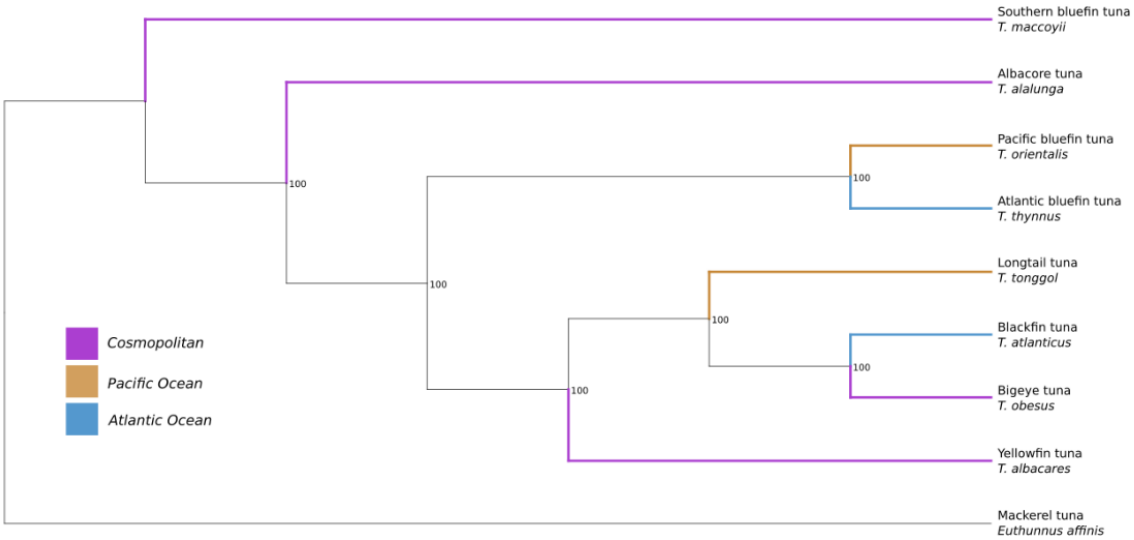

Figure 5.2 *Thunnid phylogeny inferred from genome-derived SNPs using a Maximum Likelihood algorithm.*

Sequence divergence between all alignments of all species had a mean of 5.56% (σ = 2.75) with an average genome alignment of 74.91% (σ = 16.25) between species pairs (Table 5.2, Figure 5.3). The mean sequence divergence between only representatives from the *Thunnus* genus was 3.54% (σ = 2.504) with and average genome alignment of 81.93% (σ = 9.67) between species pairs. There were 509 intervals identified in *T. maccoyii* that were putatively syntenic in all 9 species, and 1,024 intervals putatively syntenic among the 8 *Thunnus* species.

The mean sequence divergence in the syntenic intervals between all species was 5.45% (σ = 2.66) with an average genome alignment of 68.23% (σ = 13.89). The mean sequence divergence between the 8 *Thunnus* species for alignments present in all 9 species was 3.41% (σ = 2.39) with an average genomic alignment of 75.02% (σ = 9.39). Among the true tuna pairs, *T. obesus* and *T. atlanticus* had the least divergence (1.67% divergence, σ = 1.66, 91.69% alignment), and *T. orientalis* and *T. albacares* had the most divergence (5.57%, σ = 3.84, 65.67% genomic alignment). True tunas diverged from *E. affinis* (Figure 5.3, Table 5.3) by an average percent sequence divergence of 12.63% (σ = 3.61) with an average genomic alignment of 50.7% (σ = 7.17). *E. affinis* divergence from true tunas ranged from 11.2% (σ = 3.69, 45.55% genomic alignment, *T. thynnus*) to 13.2% (σ = 3.59, 53.58% genomic alignment, *T. orientalis*).

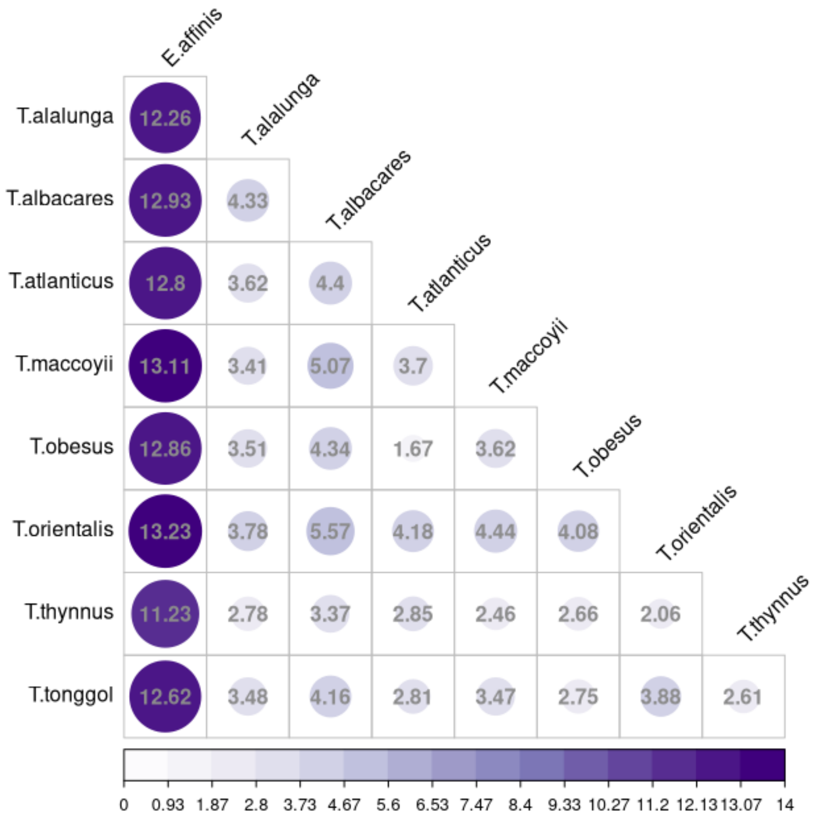

Figure 5.3 *Mean sequence divergence between Scombroid species across syntenic regions*

Table 5.2 *Genomic divergence between species*

Species 1	Species 2	% Divergence	Std Dev	20% Percentile score
E. affinis	*T. obesus*	12.86 (12.86)	3.6 (3.6)	0.969
E. affinis	*T. tonggol*	12.62 (12.62)	3.59 (3.59)	0.973
T. alalunga	*E. affinis*	12.26 (12.2)	3.54 (3.52)	0.978
T. alalunga	*T. atlanticus*	3.62 (3.37)	2.31 (2.06)	0.999
T. alalunga	*T. maccoyii*	3.41 (3.23)	2.18 (2.02)	0.999
T. alalunga	*T. obesus*	3.51 (3.3)	2.26 (2)	0.999
T. alalunga	*T. orientalis*	3.78 (3.51)	2.31 (2.05)	0.999
T. alalunga	*T. thynnus*	2.78 (2.65)	2.01 (1.86)	0.999
T. alalunga	*T. tonggol*	3.48 (3.26)	2.25 (2.02)	0.999
T. albacares	*E. affinis*	12.93 (12.9)	3.74 (3.73)	0.963
T. albacares	*T. alalunga*	4.33 (4.27)	3 (2.96)	0.998
T. albacares	*T. atlanticus*	4.4 (4.34)	3.24 (3.22)	0.998
T. albacares	*T. maccoyii*	5.07 (4.87)	3.89 (3.79)	0.996
T. albacares	*T. obesus*	4.34 (4.28)	3.22 (3.19)	0.998
T. albacares	*T. orientalis*	5.57 (5.42)	3.84 (3.78)	0.995
T. albacares	*T. thynnus*	3.37 (3.34)	2.73 (2.7)	0.999
T. albacares	*T. tonggol*	4.16 (4.09)	3.15 (3.11)	0.998
T. atlanticus	*E. affinis*	12.8 (12.73)	3.57 (3.55)	0.971
T. atlanticus	*T. maccoyii*	3.7 (3.34)	2.59 (2.24)	0.999
T. atlanticus	*T. obesus*	1.67 (1.52)	1.66 (1.49)	0.999
T. atlanticus	*T. orientalis*	4.18 (3.95)	2.62 (2.44)	0.999
T. atlanticus	*T. thynnus*	2.85 (2.75)	2.09 (1.99)	0.999
T. atlanticus	*T. tonggol*	2.81 (2.63)	2.05 (1.86)	0.999
T. maccoyii	*E. affinis*	13.11 (13.11)	3.61 (3.61)	0.964
T. maccoyii	*T. obesus*	3.62 (3.63)	2.51 (2.51)	0.999
T. maccoyii	*T. tonggol*	3.47 (3.47)	2.47 (2.47)	0.999
T. orientalis	*E. affinis*	13.23 (13.23)	3.59 (3.59)	0.962
T. orientalis	*T. maccoyii*	4.44 (4.34)	3.2 (3.13)	0.997
T. orientalis	*T. obesus*	4.08 (4.07)	2.56 (2.55)	0.999
T. orientalis	*T. thynnus*	2.06 (2.06)	1.77 (1.77)	0.999
T. orientalis	*T. tonggol*	3.88 (3.87)	2.46 (2.46)	0.999
T. thynnus	*E. affinis*	11.23 (11.18)	3.69 (3.66)	0.985
T. thynnus	*T. maccoyii*	2.46 (2.32)	1.95 (1.77)	0.999
T. thynnus	*T. obesus*	2.66 (2.54)	1.9 (1.79)	0.999
T. thynnus	*T. tonggol*	2.61 (2.51)	1.89 (1.79)	0.999
T. tonggol	*T. obesus*	2.75 (2.56)	2 (1.80)	0.999

Species 1 and Species 2 refer to the species pair being compared, % Divergence is the mean sequence divergence across all alignments, Std Dev is the standard deviation of the mean sequence divergence, \and 20% Percentile score is the percentile of the percent divergence distribution between the two species that has a divergence score below 20%. Values in parenthesis are the same calculations for putatively syntenic regions between all species.

5.4.3 Genome wide patterns of divergence between species and within species (*T. Albacares vs. T. atlanticus*)

The genome wide pattern of divergence between *T. atlanticus* and *T. albacares* is shown in Figure 5.4. Formal tests of the occurrence of islands of speciation were not conducted but most of the linkage groups show an increased level of divergence at the two ends of the linkage group. A similar pattern was observed during pairwise comparisons of other tuna species to *T. maccoyii* (Figure 5.5). Figure 5.4 also shows the location of outliers identified during comparisons of geographic populations of *T. atlanticus* and *T. albacares*. None of the single locus outliers identified in outFLANK are co-located in the two species. The outlier genomic regions mostly supported during sliding window analysis of yellowfin tuna in *Chapter III* (LGs1, 4, 17 and 22) are not co-located with blackfin tuna outliers either.

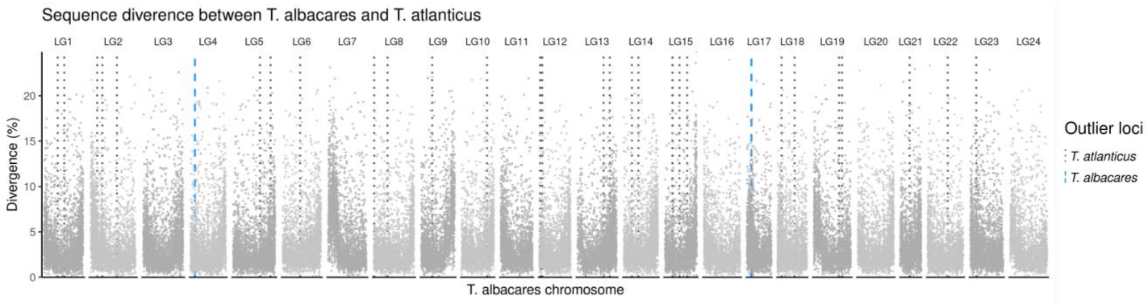

Figure 5.4 *Genome-wide sequence divergence between T. atlanticus and T. albacares.* Points represent the percent sequence divergence between *T. atlanticus* sequences that mapped to *T. albacares* chromosomes. Vertical lines denote outlier loci detected in the population genetic studies for each species that were present in alignments between the two species.

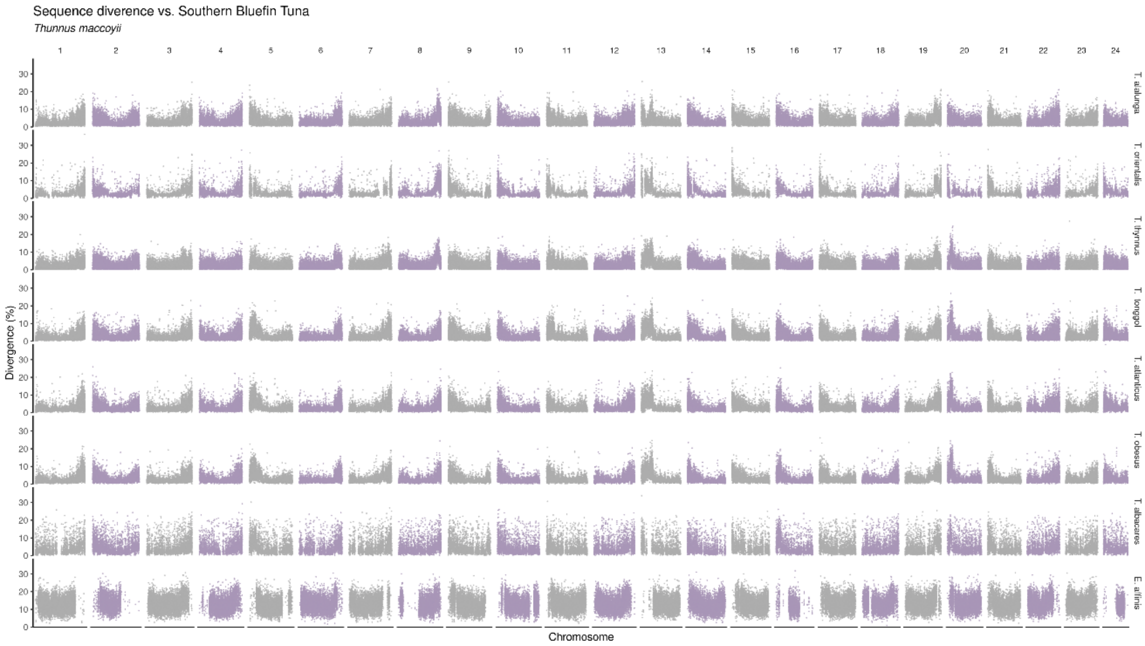

Figure 5.5 *Genome-wide sequence divergence of various scombroids to the Southern Bluefin Tuna, Thunnus maccoyii.*

Points represent the sequence divergence of sequence alignments between each species and *T. maccoyii* relative to the genomic position of those alignments in the *T. maccoyii* genome.

5.4.4 Initial annotation of genes found in genomic regions showing highest divergence between Thunnus species

There were 1,699 sequence alignments with >20% divergence between the true tuna species (globally); Among those, 81 alignments were between ATL-PAC, 312 between COS-ATL, 522 between COS-PAC, 368 between COS-COS, and 278 between COS-SBF. These genomic intervals are associated with 223 putatively divergent genes identified in Augustus, of which 22 genes (7 exclusive) were between ATL-PAC, 59 genes (30 exclusive) between COS-ATL, 85 genes (49 exclusive) between COS-PAC, 69 genes (40 exclusive) between COS-COS, and 63 (41 exclusive) between COS-SBF

(Table 5.3). The full list of predicted genes and their association to oceanic basins is shown in Appendix E.

Table 5.3 *Summary of genes putatively associated with species divergence*

Genes	ATL-PAC	COS-ATL	COS-PAC	COS-COS	COS-SBF
Total	22	59	85	69	63
Exclusive	7	30	40	49	41

Genes associated with highly divergent sequences between species categories ATL (Atlantic-only), PAC (Pacific-only), COS (cosmopolitan), and SBF (Southern Bluefin Tuna). The raw "Exclusive" features the number of genes associated only with the comparison of the two classes in a column and not associated with divergence in any other comparison pair.

5.5 Discussion

The first objective of this work was to develop reference assemblies for the three species of the *Thunnus* tribe for which no reference genome was available (*T. alalunga, T. obesus*, and *T. tonggol)*. Short read assemblies were generated for each of these species and the assembly of *T. atlanticus* developed in *Chapter IV* was improved. While these assemblies are still fragmented, the short read sequencing approach cost effectively generated contigs spanning high fractions of the studied genomes (>91% of the genome in contigs longer than 1,000 bp or more, >63% in contigs longer than 5,000 bp). The use of accurate short reads with low error rates sequenced with high coverage yielded reliable draft reference genomes that can be used to map low depth re-sequencing reads to discover and genotype SNPs and other variants and study genetic variation within each species.

Initial investigation of genome-wide variation between tuna species were conducted using the new assemblies developed in this chapter, the draft reference genome obtained in *Chapter II* for Yellowfin tuna, and those published for Bluefin tunas.

Studies of genomic variation in relation to speciation require knowledge of phylogenetic relationships within the studied group (*Thunnus* genus) so that sister species can be identified and compared under simple assumptions of isolation with migration in future studies. Whole genome comparisons using SNPs yielded a phylogeny similar to the two most recently published phylogenies of the group based on RAD sequencing (Díaz-Arce et al. 2016) and transcriptome sequencing (Ciezarek et al. 2019) datasets in that the endothermic species utilizing temperate waters (*T. thynnus*, *T. orientalis*, *T. maccoyii*, and *T. alalunga*) occupy basal positions in the tree and the tropical tunas form a derived monophyletic clade. The first difference between the current phylogeny and the recently published ones involves the placement of *T. alalunga*, which was inferred to be the most basal taxon by Ciezarek et al. and Diaz-Arce et al. while the SNP approach employed here situates *T. alalunga* between the Southern Bluefin tuna and the two northern Bluefin tunas, a finding also inconsistent with the phenotypic similarity of the three bluefin tunas. The second main difference concerns the relationships within the tropical tunas; Ciezarek et al (2019) and Díaz-Arce et al (2016) found *T. atlanticus* and *T. tonggol* were sister taxa, a finding consistent with their similar body sizes, coastal distribution, and restricted latitudinal ranges. The phylogeny obtained in this work suggests that *T. obesus* and *T. atlanticus* are sister taxa. *T. obesus* features adaptations to cold temperature shared with the temperate tuna species (*T. Thynnus*, *T. orientalis*, *T. maccoyii*, and *T. alalunga*). Accordingly, this species was discussed to be intermediate between the temperate group and the tropical tunas (*T. Albacares*, *T. atlanticus* and *T. tonggol)* (Ciezarek et al. 2019; Gibbs and Collette 1967), a hypothesis potentially inconsistent with the placement of *T. obesus* within the tropical tuna clade although Díaz-Arce et al (2016) further discussed that *T. obesus* may have acquired thermoregulation adaptations independently from the temperature tunas, which would provide a potential explanation for its placement in the tropical tuna clade. This work employed a sufficiently high marker density (~1 SNP per 2kb, Wortley et al. 2005) to provide a reliable assessment of phylogenetic relationships and capture both neutral and most of the adaptive variation. The topology obtained was concordant with the estimates of sequence divergence between pairs of genomes (Figure 5.3, Table 5.3, where *T. atlanticus* and *T. obesus* have a sequence divergence of 1.67%, the lowest of any species pair, with >90% total genomic alignment). This work also provided complete datasets for the SNP markers employed while the RAD-sequencing approach used by (Díaz-Arce et al. 2016) allowed for varying levels of missing data. Additionally, the whole-genome approach to identify and thin SNPs promoted more genomic homogeneity in the distribution of SNPs used for investigation, limiting the bias of signal concentration that may result from RAD-based approaches, where SNPs are only identified near non-uniformly distributed restriction sites. The transcriptome approach of Ciezarek et al (2019) may have been biased by restricting the transcriptome to only one tissue type (muscle). The use of muscle tissue transcripts primarily focused on endothermy as a driver of evolution and was not necessarily reflective of other drivers of divergence and speciation.

However, the current dataset has some important limitations that potentially affect the inferred phylogeny. Only one consensus sequence was used to characterize each species. Accordingly, comparisons between species confounded variation among individuals and populations within species and true variation between species. The impact of such errors on inferences could be substantial, inflating differences between species at some loci or completely erasing differences at others. This limitation can be addressed by sampling additional specimens of each species, performing low depth sequencing, and aligning sequencing reads on the reference assemblies developed in this work for whole genome SNP calling, which will allow accounting for within species variation in the obtained dataset. Individuals should be sampled in various parts of the range, including from populations currently isolated in different oceanic basins (Atlantic versus Pacific). A second potential limitation of the dataset was related to ascertaining homology of sequences from different species. Genomes were split into 150 kb windows for mapping to the southern bluefin genome (McWilliam et al. 2016) and these alignments were used for variant calling. The congruence of the topology obtained from the SNP dataset and that from the long sequence alignments used to infer sequence divergence (Appendix D) suggests that alignments were robust. The mixing of coding and non-coding regions could impact the selection of an appropriate mutation model for use in Maximum likelihood inference on phylogenies. A strategy to improve this aspect could be to restrict SNPs to coding regions as identified through annotation.

The geographic distribution of tuna species (Figure 5.1) will be a key factor to future efforts to clarify phylogenetic relationships within this group. One consideration is the formation of the Isthmus of Panama ~3.5mya (Coates et al. 1992; O’Dea et al. 2016), which restricted gene flow between demes in a larger multi-basin metapopulation of ancestral tunas. The timing of the formation of the Isthmus coincides with the molecular clock presented by Ciezarek et al (2019), which proposes that the true tunas began diverging approximately 5-7 mya. This suggests gene flow between these basins may have been attenuating for many generations before the completion of the Isthmus of

Panama and accordingly that some of the speciation events within the group may have occurred within basins. The hypothesis of a pre-closure isolation is supported by the analysis of vicariant events inferred in sea catfish from genome wide data. Isolation was related to the emergence of the isthmus and dated in the late Myocene period, millions of years before the closure (Stange et al. 2018). This scenario is supported by the basin-scale population subdivision in globally distributed tunas (Gonzalez et al. 2008; Pecoraro et al. 2018; Montes et al. 2012) and the differentiation of bluefin tuna in 3 species. Speciation post isolation of the Atlantic and Pacific could explain a more distant relationship between *T. tonggol* and *T. atlanticus*. Acquiring specimens from the globally distributed species in the different basins where they occur will be a priority to achieve a thorough understanding of the relationship within this group.

A second objective of this work was to examine patterns of divergence between species across the genome. Due to high assembly fragmentation in most of the species compared, full chromosome-scale genomic alignments between all pairs of species were not possible. However, performing these alignments against the *T. maccoyii* genome, the most contiguous of the assemblies in this work, identified chromosomal regions with disproportionately high divergence (putative “islands of divergence”) consistent across species (Figure 5.5). In many of these cases, sequence divergence rose to upwards of 30%, and these regions typically occurred towards the ends of the chromosomes. It has been previously suggested that there is higher genetic diversity in subtelomeric regions which may be responsible for promoting speciation (Zhang et al. 2015). The pattern of higher sequence divergence in telomeric/subtelomeric regions was observed in other eukaryotic organisms (Shao et al. 2018) where it has been suggested that speciation is driven by chromosomal extension within a common ancestor, leading to reproductive isolation and ultimately speciation (Shao et al. 2018). The divergence of genomic alignments between *T. atlanticus* and *T. albacares*, the two species of primary focus in this dissertation, followed the pattern discussed above where higher divergence occurred towards chromosomal ends (Figure 5.4). Comparison of these two species, the smaller-bodied *T. atlanticus* with a restricted coastal range in the western Atlantic Ocean, and the larger-bodied *T. albacares*, with a global offshore range (Figure 5.1), revealed notable genomic segments with disproportionately elevated divergence (e.g., linkage groups 7, 9, and 19), where the mean sequence divergence was significantly higher than the rest of the genome.

Recent literature challenges the historical implications of "islands of divergence" (Bay and Ruegg 2017; Renaut et al. 2013; Shao et al. 2018) suggesting these islands may be associated with introgression or regions of reduced recombination rather than harboring genes involved in divergent selection and reproductive isolation. Further research on patterns of recombination and linkage disequilibrium may provide insights into their potential contributions to reproductive isolation between the two sympatric species.

The population structure study conducted in *T. atlanticus* and *T. albacares* in the Atlantic basin also provided the opportunity to assess whether outlier loci identified within the two species reflected shared selective constraints occurring in their overlapping ranges. Within species outliers detected in this work were found in different genomic positions in the two species, usually on different chromosomes, suggesting different genes were affected by divergent selection. The SNP loci analyzed in this work were generated using a relatively low density reduced-representation genome scan, likely missing a large portion of adaptive variation occurring in each species. More dense genome scans will be necessary to assess potential shared genomic targets of natural selection for these two species if they occur.

This work also conducted a preliminary investigation of genes annotated in regions showing highest divergence between species. Only genes in genomic regions of extreme divergence (>99.9th percentile) were examined, with the implication they have an increased likelihood to be associated with observed speciation. As an example, seven genes were exclusively associated between highly divergent genomic regions of Atlantic-exclusive and Pacific-exclusive species, in addition to the other 15 genes that are associated with other pairwise comparisons involving these species. These 7 exclusively ATL-POS diverging genes are associated with protein kinases, calcium transport, histone and metal ion binding, respiration, and spermatocyte progression. The 5 genes presented in (Table 1 in Ciezarek et al. 2019) putatively associated with endothermy-driven evolution in tunas did not appear among the 223 putatively divergent genes presented in this study. This may be due to the >99.9th percentile threshold that was applied for alignments to be considered “highly divergent” and excluded most sequence alignments two species shared, only retaining extreme cases. However, the absence of these putatively endothermy-associated genes among the extremely divergent alignments suggests that either different endothermy-associated genes are associated with the regions of greatest divergence, or that endothermy may not be the primary driver of evolution between the tuna species. The distribution of true tunas is strongly influenced by temperature but has also been linked to a myriad of characteristics, such as salinity, dissolved oxygen, and oceanographic features like currents and eddy types (Arrizabalaga et al. 2015; Hsu et al. 2015).

This work involved generating *de novo* draft reference genome assemblies for three tuna species (*T. alalunga, T. obesus, T. tonggol*) which are now publicly available for future research. Genomes were assembled using short reads only to reduce costs while achieving reliable contigs owing to the accuracy and high coverage obtained during Illunina sequencing. While short-read sequencing is not compatible with achieving chromosome-scale vertebrate genomes assemblies (Kuhl et al. 2020; Lischer and Shimizu 2017), the method of merging two analogous assembly processes (ABySS + redundans and Sparseassembler + purge_haplotigs) applied in this work resulted in much improved contiguity as compared to all the other *de novo* assembly strategies attempted (not shown). Incorporation of the haplotig reduction step (regardless of tool) was key to achieving better contiguity and genome length in both assembly methods, as both the ABySS and Sparseassembler assemblers resulted in initial assemblies with over 1 million contigs and nearly double the expected length. A single pass with either of redundans (easier and automated) or purge_haplotigs (manual and slow but more aggressive) reduced the initial assemblies to within a few dozen megabases of their expected haploid lengths and reduced the number of contigs at least threefold. The merging of these haplotig-filtered assemblies further improved the resulting assemblies. Reference-guided assembly software (Lischer and Shimizu 2017; Alonge et al. 2019; Bao et al. 2014) may result in higher assembly contiguity and could employ either the *T. albacares* genome developed in *Chapter II* or the chromosome-scale *T. maccoyii* genome assembly developed by McWilliam et al (2016). The assemblies produced in this work were intended to be used for phylogenomic comparisons and were therefore generated *de novo* to avoid introducing structural biases that may result from reference-guided assembly or scaffolding methods, but the obtained assemblies could be improved by the reference guided approach in future efforts to increase contiguity. An exclusively long read or hybrid technology approach may also become fiscally viable to generate new draft reference genomes soon considering the increasing accuracy of long read sequencing (e.g., PacBio HiFi, Nanopore R10.x chemistry) and the decreasing cost of these long-read sequencing platforms.

Many whole-genome phylogenetic analyses rely on annotated genomes or chromosome-scale genome assemblies and the generation of either may not be feasible for every study. This study was unable to successfully implement existing SNP-based alternatives, such as REALPHY (Bertels et al. 2014) which consistently required more than 256 gb of RAM available to our computational servers, or phame (Shakya et al. 2020), whose errors causing premature termination could not be determined. Notably, these methods failed to produce viable results due to their implementation rather than the incompatibility of the datasets. The pipeline described in this work approximates the method described in Bertels et al (2014) using ubiquitous and thoroughly tested 3rd party bioinformatics tools to generate variant information from genome assemblies. RAM usage peaked at 7 gb, which is less than the capacity of common mid-range laptops. These tools have been combined into a reusable and configurable Snakemake workflow (Köster and Rahmann 2012) called gust (https://github.com/pdimens/gust), which is freely available in a public repository with versioned software dependencies provided in a file compatible with conda-derived virtual environment frameworks (e.g., conda, mamba).

The holistic genomic approach to phylogeny applied in this work and others (e.g., based on transcriptome sequencing , Ciezarek et al. 2019), allows studying the specific biological domains affected by the putatively divergent genes, thereby providing further understanding of the mechanisms of the speciation process. While this work compared only the most diverged sequences between species pairs, there may be a wealth of information hidden within the regions that did not align between species, namely, what (presumably adaptive) genes are present in one species but not another, along with what added or redundant purposes the unaligned DNA serve to the species. With the availability of reference genomes, modern methods, and software to study homology and functional genomics, continued exploitation of the genomic tools developed in this work will provide valuable information on the mechanisms responsible for the sympatric speciation of the true tuna species.

CHAPTER VI– SUMMARY AND CONCLUSIONS

Analysis of genome-wide SNPs genotyped using the ddRAD sequencing method revealed weak but present structuring of Yellowfin and blackfin tuna in the Atlantic Ocean. The levels of divergence between geographic populations as measured by F_{ST} estimates are very low, a common situation in marine species (Waples 1998). This weak differentiation was incompatible with detection of structure by several current methods such as Bayesian Clustering (Pritchard et al. 2000). Accordingly, the findings must be taken cautiously and continued monitoring of temporal stability of the patterns described in this work is warranted. However, the results suggest that Yellowfin tuna form a large connected metapopulation within the Atlantic Ocean basin. Documented transatlantic movement of adult Yellowfin tuna individuals (ICCAT 2019) is supported in this study by the observation of individuals with high membership probability in the east clusters sampled in the western locations and vice versa. The population clusters inferred for each nursery area were not sufficiently differentiated to distinguish F0 migrants from offspring of migrants such as F1 and document effective gene flow (reproduction of migrants in the recipient area). However, the low level of divergence observed in the study suggests that movement between regions are associated with gene flow. The study also suggested structuring between nursery areas within the west and particularly the east Atlantic regions, where individuals from Ivory Coast were more similar to each other than to juveniles from the adjacent nursery in Senegal, suggesting philopatry is occurring towards the proposed spawning area in the Gulf of Guinea and West Africa. Current management of yellowfin tuna assumes a single stock for the entire Atlantic Basin, whereas this work and a recent study by Pecoraro et al (2018) indicate that the distinction of east and west Atlantic stocks may be warranted. This works shows that additional subdivision within the west and the east side of the basin according to nursery areas may also be needed.

The presence of many co-located close kin Yellowfin tuna larvae suggests co-dispersal of larval cohorts is occurring. Further work sampling more adults in close geographic proximity is warranted to determine if co-location is maintained across ontogeny as suggested by Anderson et al (2019). Kinship in co-located adult yellowfin tuna could not be formally studied because information on capture location was limited to large fishing areas for fisheries samples, but a few adult Blackfin tuna close kin pairs were identified, and members of these pairs were co-located in most cases. Blackfin tuna formed a metapopulation with four possible weakly differentiated demes in the Gulf of Mexico, US Atlantic, the South Caribbean, and Brazil. This species is currently not regulated and future management plans would need to consider units in Brazil and possibly multiple units in the northern part of the range corresponding to the different groups suggested above. Delineation of the Southern (Brazil) and northern stocks with targeted sampling at the transition between South America and the Caribbean Sea also is warranted to determine if a barrier is occurring or if isolation by distance if the main driver of isolation. Analyses of outlier loci yielded little evidence for local adaptation in either species. However, the sliding window analysis conducted in Yellowfin tuna using the draft reference genome identified candidate outlier genomic regions that will warrant further targeted study using higher density genome scans.

Whole genome derived SNP-based phylogeny and genome alignment data challenged partly previous phylogenetic reconstruction for *Thunnus* species. Genome comparisons between species revealed chromosomal intervals with clear patterns of inflated divergence towards chromosome ends. Gene prediction of regions of extreme divergence (>99.9th percentile) identified gene sets that were unique to species-group comparisons based on global distribution patterns. Genomic comparisons were limited by assembly contiguity, and a more thorough investigation is warranted using chromosome-scale assemblies, which would reveal larger patterns of structural variants between species (e.g., inversions, translocations). A continuation of this work should include diploid genotypes and several individuals representing each species and their basin-specific populations to account for genetic variation within and across demes. The genes associated with the divergence of distribution-based species groups warrant reevaluation with a less conservative threshold, along with a more exhaustive gene ontology analysis to identify generalized domains for these genes (e.g., thermoregulation, reproduction, salinity tolerance).

This work highlights improvements in inferential power in population genetic analyses when 1) using a genome assembly of the target species, 2) sampling across both space and time, and 3) rigorous sample preparation and data filtering to remove sources of bias. It also highlights the difficulty of population inferences when neutral divergence between groups is extremely low, suggesting a focus on identifying adaptive loci and using them as the basis for inference. Additionally, the whole-genome approach to phylogeny underpins the use of holistic data for phylogenetic inference rather than the previous approaches of phylogenetic reconstruction using data that do not represent comprehensively variation across the genome (RADseq, muscle transcripts). Sequencing technology continues to advance rapidly and decrease in cost, enabling more analyses relying on genome assemblies for a host of non-model species of interest to conservation. The work presented here laid the groundwork for future genetic investigation of both the Yellowfin and Blackfin tuna species and will enable further understanding of isolation, gene flow and speciation of these important species.

APPENDIX A – Yellowfin Tuna Genome assembly

Trimmed short reads were assembled using SparseAssembler, which uses memory-saving sparse kmer graphs to effectively subsample the kmer space to skip intermediate overlaps between pairs of reads (Ye et al. 2012). The assembly was performed with a k of 51, node coverage threshold (NodeCovTh) of 2, edge coverage threshold (EdgeCovTh) of 1, intermediate k-mer skipping (g) of 15, and chimeric contig removal (ChimeraTh 2 ContigTh 2). The assembled short-read contigs and raw long reads (as recommended by the software) were assembled using DBG2OLC (Ye et al. 2016) with a k of 17, kmer coverage threshold (KmerCovTh) of 2, an adaptive theta (AdaptiveTh) of 0.001, minimum kmer overlap (MinOverlap) of 15, and chimera removal (RemoveChimera 1). The assembled short reads and raw long reads were concatenated to assess contig consensus using BLASR (Chaisson and Tesler 2012) and pbdagcon (https://github.com/PacificBiosciences/pbdagcon). This process consolidates all the assembled contigs into a final haploid assembly with the help of the original sequences, attempting to merge contigs when possible and remove duplicate ones. The corrected long reads were mapped to the consensus sequences using minimap2 (Li, 2018), then used to polish the assembly using racon (Vaser et al. 2017). This process was repeated two more times for a total of 3 rounds of long-read polishing. The subsampled short reads were then mapped to the polished assembly using minimap2 and used to polish the assembly with the default parameters of Pilon (Walker et al. 2014). The corrected long reads were then mapped to the polished genome using minimap2 to scaffold the assembly with LRScaf (Qin et al. 2019). Finally, the subsampled short reads were mapped to the scaffolded assembly using minimap2 and polished once more with Pilon.

APPENDIX B – Blackfin Tuna Genome assembly

The assembly was performed using SparseAssembler (Ye et al. 2012) with a kmer size of 90, node coverage threshold of 5, edge coverage threshold of 3, skipping 15 intermediate kmers, chimera removal enabled with a threshold of 2 with a contig threshold of 2, and a genome size imputed from the Jellyfish k-mer counting method. A consensus was performed using DBG2OLC (Ye et al. 2016), with a minimum overlap of 115 bp, a path coverage threshold of 3, Kmer coverage threshold of 0, and a k of 31. Sequence mapping for genome assembly was performed with default BWA parameters. The low, medium, and high thresholds for purge_haplotigs were set to 20, 75, and 185 respectively.

APPENDIX C – Phylogenetic topography based on sequence divergence

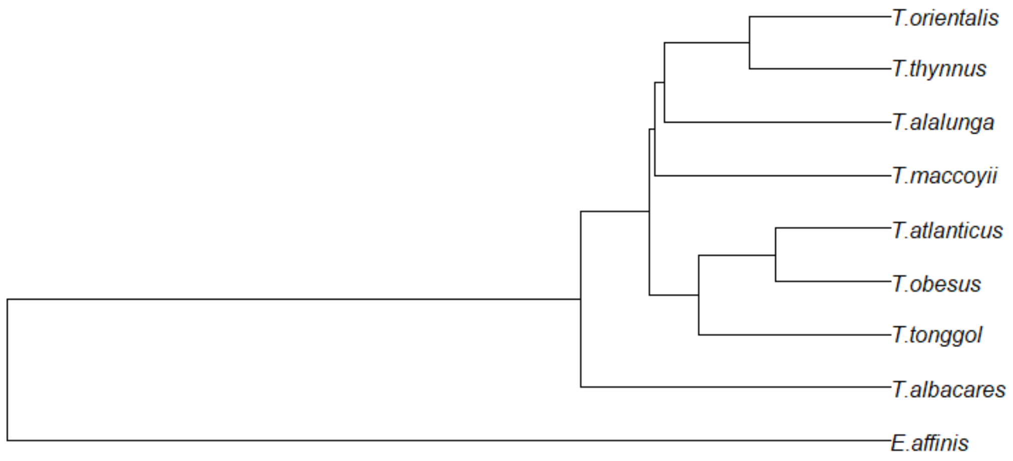

Figure C.1 The tree represents the phylogenetic relationships between species based on the sequence divergence between species pairs as shown in Table 5.3.

APPENDIX D – Tuna Short-Read Genome Assemblies

Trimmed reads for each species were assembled using ABySS (Simpson et al. 2009; Jackman et al. 2017) with 3 Bloom filter hash functions (H=3), a K-mer size of 81 (k=81) and a minimum K-mer count threshold of 3 (kc=3). The resulting assemblies were then processed with the default parameters of redundans (Pryszcz and Gabaldón 2016), with the exception of using BWA (Li, 2013) for mapping, to remove haplotigs (no scaffolding or gap filling). A separate assembly for each species was then performed using SparseAssembler (Ye et al., 2012) with a K of 90, node coverage threshold of 5 (NodeCovTh 5), edge coverage threshold of 3 (EdgeCovTh 3), skipping 15 intermediate K-mers (g 15), and removing chimeras with a threshold of 2 (RemoveChimera 1 ChimeraTh 2 ContigTh2). The resulting primary contigs were consolidated into consensus sequences using DBG2OLC (Ye et al. 2016) with a K of 31, minimum overlap of 115 (MinOverlap 15), K-mer coverage threshold of 0 (K-merCovTh 0), and path coverage threshold of 3 (PathCovTh). Trimmed reads were then aligned to the consensus assembly using minimap2 (Li, 2018) using the short-read presets (-ax sr) along with the additional mapping parameters of ignoring secondary alignments (--secondary=no), outputting the MD tag (--MD), and output only SAM hits (--sam-hit-only). The alignments were filtered using SAMTools (Li et al. 2009; Danecek et al. 2021) to remove empty alignments.(Danecek et al., 2021; Li et al., 2009) to remove empty alignments, alignments with a mapping quality less than 5, then sorted and indexed. These curated alignments were then used to polish the assembly using HyPo (Kundu et al. 2019) with default parameters. Using new alignments to the polished assembly generated with the same mapping parameters, the polished assembly was screened for haplotigs using purge_haplotigs (Roach et al. 2018). The resulting K-mer coverage histograms had to be assessed visually for bimodal peaks and valleys, resulting in estimated low (low=80), mid (mid=176), and high (high=368) cutoffs for each of *T. obesus*, *T. alalunga*, and *T. tonggol*.

The separate assemblies (ABySS/redunans, SparseAssembler/DBG2OLC) were then merged using quickmerge (Chakraborty et al. 2016), where the length cutoff was determined by the N50 of the less-contiguous assembly in each pair. In each species, the query assembly was the one with fewer contigs and the reference was the assembly with greater contigs. The merged assembly was screened for haplotigs, scaffolded, and gap filled using redundans. Reads were mapped to the scaffolded assembly and polished using HyPo using the same parameters described above. The quality of the assembled genomes was assessed using Quast (Gurevich et al. 2013) and completeness was evaluated using Benchmarking Universal Single Copy Orthologs, as implemented in busco (Simão et al. 2015) and metaeuk (Levy Karin et al. 2020) to search the genomes for the 3,354 genes in the *vertebrata* database.

APPENDIX E – Genes putatively associated with species divergence

Table E.1 *Genes putatively associated with species divergence*

Accession	**Gene Name**	**ATL-PAC**	**COS-ATL**	**COS-PAC**	**COS-COS**	**COS-SBF**
XP_042260979.1	1-phosphatidylinositol 4,5-bisphosphate phosphodiesterase		+	+	+	+
XP_040009476.1	6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase isoform X				+	
XP_044229836.1	A-kinase anchor protein 6 isoform X2		+		+	
XP_044215408.1	acidic leucine-rich nuclear phosphoprotein 32 family member				+	
XP_039893151.1	activating molecule in BECN1-regulated autophagy protein 1b					+
KAF6722036.1	Adenomatous polyposis coli protein 2				+	
XP_044221816.1	AF4/FMR2 family member 3			+	+	
XP_042252718.1	ankyrin repeat and SOCS box protein 3 isoform X3		+			
XP_042259301.1	antho-RFamide neuropeptides-like			+		
XP_042292307.1	antigen peptide transporter 2a			+		
TKS87287.1	Band 4.1-like protein 3 4.1B					+
XP_042252634.1	Bardet-Biedl syndrome 1 protein			+		
XP_044216819.1	basigin isoform X2			+		
XP_042365660.1	beta-adrenergic receptor kinase 2-like				+	
XP_044219707.1	biorientation of chromosomes in cell division protein 1-like 1				+	
XP_042257680.1	butyrophilin subfamily 1 member A1-like isoform X5		+			
XP_042245767.1	CAD protein isoform X1		+			
XP_044189749.1	calcium-binding mitochondrial carrier protein SCaMC-2 isoform					+
XP_042256689.1	calmodulin-binding transcription activator 2				+	+
XP_042264187.1	carbohydrate sulfotransferase 11-like isoform X1			+	+	
XP_042283052.1	caveolae-associated protein 2a					+

Table E.1 (continued)

XP_044189124.1	centrosomal protein of 120 kDa	+	+			
XP_044190720.1	clumping factor A-like				+	
XP_034566810.1	collagen alpha-1(I) chain-like isoform X1		+	+	+	+
XP_044209712.1	collagen alpha-4(IV) chain				+	
XP_042270390.1	complement C1q-like protein 4		+	+	+	+
XP_045072135.1	CUB and sushi domain-containing protein 2-like					+
XP_042273117.1	cyclic AMP-dependent transcription factor ATF-6 alpha	+				
XP_044231415.1	D-glutamate cyclase, mitochondrial isoform X1			+		
XP_044214888.1	DENN domain-containing protein 4B-like isoform X1				+	
XP_044213199.1	diacylglycerol kinase zeta isoform X1					+
XP_042283827.1	disco-interacting protein 2 homolog C isoform X1			+		
XP_042245738.1	DNA-binding protein inhibitor ID-2b		+			
XP_042278045.1	docking protein 2 isoform X2				+	
XP_029282623.1	double C2-like domain-containing protein beta			+		
XP_026170866.1	dual specificity mitogen-activated protein kinase kinase 6				+	
KAA0725373.1	Dynein heavy chain 9, axonemal				+	+
XP_044197525.1	dysferlin isoform X9		+		+	
XP_044201158.1	dystrotelin				+	
XP_044225577.1	E3 ubiquitin-protein ligase RNF26-like		+		+	
XP_044228054.1	echinoderm microtubule-associated protein-like 6 isoform X1		+			
XP_044227205.1	enteropeptidase				+	
XP_042249882.1	envoplakin		+	+	+	+
XP_042243762.1	estrogen receptor 2b isoform X1	+		+		
XP_042243765.1	estrogen receptor 2b isoform X2	+		+		

Table E.1 (continued)

XP_044230097.1	estrogen receptor 2b isoform X3	+			+	
XP_031430915.1	eukaryotic initiation factor 4A-II isoform X2					+
XP_034721565.1	eukaryotic initiation factor 4A-III-A-like		+			
XP_042252479.1	exocyst complex component 7 isoform X4			+		
XP_044215379.1	exopolyphosphatase PRUNE1					+
XP_029495459.1	extensin-like			+		
XP_042276653.1	FHF complex subunit HOOK interacting protein 2B			+		+
XP_044189512.1	filaggrin-2-like				+	
XP_022061671.1	folliculin-interacting protein 1-like isoform X2		+	+	+	
XP_044222129.1	frizzled-5			+		
XP_044229365.1	galectin-related protein B-like		+			
XP_044209791.1	glucose 1,6-bisphosphate synthase				+	
XP_029630028.1	glutamine-rich protein 2-like			+		
XP_026225578.1	golgin subfamily A member 6-like protein 1					+
XP_044224583.1	GREB1-like protein isoform X4		+			
XP_042257457.1	HAUS augmin-like complex subunit 6 isoform X1		+			+
XP_044192410.1	heat shock factor protein 1			+		
XP_023252975.1	histone acetyltransferase p300-like			+	+	+
KAG5276892.1	*hyp:* AALO_G00110960		+		+	
KAG7269102.1	*hyp:* CRUP_004371					+
KAG7253440.1	*hyp:* CRUP_037317	+	+			
KTF88672.1	*hyp:* cypCar_00041010		+			
KTG42443.1	*hyp:* cypCar_00044540				+	

Table E.1 (continued)

KAE8283796.1	*hyp:* D5F01_LYC17121					+
TNN59813.1	*hyp:* EYF80_029998		+			
TNN27718.1	*hyp:* EYF80_062135			+		
KAF0035558.1	*hyp:* F2P81_013316				+	
KAF0028640.1	*hyp:* F2P81_019727					+
KAF0025904.1	*hyp:* F2P81_022785				+	
KAF3841610.1	*hyp:* F7725_023561			+		
KAF3834529.1	*hyp:* F7725_027087			+		
TNM98120.1	*hyp:* fugu_014366					+
KAG7231563.1	*hyp:* INR49_011555	+				
KAG7215238.1	*hyp:* INR49_022677				+	
KAG7240845.1	*hyp:* INR49_023419		+		+	
KAG7239647.1	*hyp:* INR49_028583		+			
KAG7239156.1	*hyp:* INR49_029907			+		+
KAG7238877.1	*hyp:* INR49_030424			+		
KAG7461491.1	*hyp:* JOB18_049994			+		
KAG9348175.1	*hyp:* JZ751_001910				+	
KAF7643850.1	*hyp:* LDENG_00231980			+		
KAF7659711.1	*hyp:* LDENG_00294050					+
KAF1383200.1	*hyp:* PFLUV_G00128850	+		+		
XP_044210162.1	hypoxia inducible factor 1 subunit alpha, like		+			
XP_042278506.1	inactive phospholipid phosphatase 7-like					+
XP_042247784.1	interferon a3-like			+		+
KAF3707941.1	Intersectin-1 EH and SH3 domains protein 1					+

Table E.1 (continued)

XP_020484142.1	involucrin-like					+
XP_042265117.1	IQ motif and SEC7 domain-containing protein 2-like				+	
XP_044190288.1	kazal-type serine peptidase inhibitor domain 3			+		+
XP_044191192.1	kelch-like protein 31			+		
XP_005169283.1	keratin-associated protein 4-8-like				+	
XP_044208471.1	ketohexokinase isoform X1					+
XP_042276129.1	kinesin-like protein KIFC3 isoform X2			+		
XP_044196178.1	laminin subunit alpha-1			+		
XP_042255542.1	late secretory pathway protein AVL9 homolog				+	
KAF6720348.1	Leucine-rich repeat and coiled-coil domain-containing protein		+			
XP_044206393.1	leucine-rich repeat neuronal protein 1 isoform X2		+	+		
XP_042255446.1	leucine-rich repeat-containing protein 30-like			+		
XP_042244807.1	LOW QUALITY PROTEIN: MAX dimerization protein MGA a				+	
XP_042257406.1	lysophospholipid acyltransferase LPCAT4			+		
XP_042288471.1	male-enhanced antigen 1				+	
XP_044212753.1	malonyl-CoA decarboxylase, mitochondrial			+		
XP_044219112.1	mastermind-like protein 1 isoform X2			+		
XP_042286506.1	meiosis regulator and mRNA stability factor 1 isoform X5		+			
TWW56070.1	Melanoma-associated antigen D2 11B6		+			
XP_042250165.1	methyl-CpG-binding domain protein 2					+
XP_042273217.1	microtubule-associated serine/threonine-protein kinase 2		+			
XP_044195317.1	mitochondrial enolase superfamily member 1 isoform X2					+
XP_044224047.1	mitofusin-1					+
XP_030208297.1	mucin-2-like					+

Table E.1 (continued)

XP_042250799.1	multidrug resistance-associated protein 1-like isoform X5					+
XP_042289274.1	myoferlin isoform X1	+		+		
XP_031697665.1	myosin light chain kinase, smooth muscle-like	+	+			
XP_042244622.1	nesprin-2			+		
XP_038859211.1	neurturin-like				+	
XP_044189667.1	NF-kappa-B inhibitor-like protein 1		+	+		+
XP_042278600.1	opioid receptor, delta 1b			+		
XP_044197842.1	OTU domain-containing protein 4	+	+			
XP_034055202.1	oxoglutarate (alpha-ketoglutarate) dehydrogenase a (lipoamide)				+	
XP_044219098.1	phospholipid-transporting ATPase ABCA1-like isoform X2				+	
XP_044216872.1	polypyrimidine tract-binding protein 1b isoform X3			+		
XP_018554262.1	*pr:* centrosomal protein of 55 kDa-like isoform X5			+		+
XP_010783701.1	*pr:* choline-phosphate cytidylyltransferase A-like		+		+	+
XP_013889121.1	*pr:* E3 ubiquitin-protein ligase RNF123-like			+	+	
XP_018544223.1	*pr:* janus kinase and microtubule-interacting protein			+		
XP_014037514.1	*pr:* leucine-rich repeat extensin-like protein 1	+		+		
XP_016347478.1	*pr:* myosin regulatory light polypeptide 9-like			+	+	
XP_016426160.1	*pr:* protein argonaute-4-like					+
XP_018549751.1	*pr:* regulator of nonsense transcripts 3B-like					+
XP_019968233.1	*pr:* sushi, von Willebrand factor type A, EGF and			+		
XP_010764314.1	*pr:* thyrotropin-releasing hormone-degrading				+	
XP_015259182.1	*pr:* transme Mbrane protein 198-B-like				+	
XP_014064096.1	*pr:* uncharacterized PE-PGRS family protein PE_PGRS36-li			+		
XP_014041969.1	*pr: unchar: LOC106595118*	+		+		

Table E.1 (continued)

XP_042277028.1	probable ATP-dependent RNA helicase DDX31					+
XP_042281424.1	proline-rich protein 36-like	+		+		
XP_044201517.1	proteasome adapter and scaffold protein ECM29 isoform X2			+	+	
XP_044200349.1	protein CREG1		+			
XP_042254363.1	protein HID1b		+		+	
XP_042286733.1	protein KIBRA				+	
XP_006810892.2	protein KIBRA-like			+	+	
XP_042259839.1	protein mono-ADP-ribosyltransferase PARP12b					+
XP_020779526.1	protein patched homolog 2		+	+	+	+
XP_042288404.1	puromycin-sensitive aminopeptidase			+		
XP_044211462.1	putative ATP-dependent RNA helicase TDRD12					+
XP_044230504.1	putative leucine-rich repeat-containing protein DDB_G0290503				+	
XP_020513591.2	receptor-type tyrosine-protein phosphatase N2-like		+			
TNN52118.1	Receptor-type tyrosine-protein phosphatase zeta		+		+	
XP_042260384.1	receptor-type tyrosine-protein phosphatase-like N isoform X3		+			
XP_028272829.1	regulator of nonsense transcripts 3B		+		+	
XP_044209277.1	remodeling and spacing factor 1 isoform X3	+				
XP_042267885.1	retinoblastoma-like protein 2 isoform X2	+		+		
XP_042290510.1	rho family-interacting cell polarization regulator 2 isoform X		+			
XP_037342016.1	rho GTPase-activating protein 5					+
XP_044224318.1	ribonuclease P protein subunit p40 isoform X2			+		
XP_042292491.1	ribosome biogenesis protein bop1 isoform X2					+
XP_042281494.1	ribosome-binding protein 1b isoform X4		+		+	
XP_044204700.1	RING finger protein 207 isoform X3			+		

Table E.1 (continued)

XP_044206126.1	RNA-binding protein 39-like isoform X7			+		
XP_042257603.1	RPA-related protein RADX		+			
XP_041650245.1	secreted protein C-like			+		
XP_042252289.1	serine-rich coiled-coil domain-containing protein 2-like isofo					+
XP_042288345.1	serine/arginine repetitive matrix protein 2 isoform X4				+	
XP_042246237.1	serine/threonine-protein kinase MRCK beta isoform X2		+			
XP_044197827.1	serine/threonine-protein kinase/endoribonuclease IRE2-like		+			+
TMS00855.1	Serine/threonine-protein phosphatase 2B catalytic subunit				+	
XP_034721965.1	SH3 domain-containing YSC84-like protein 1	+				
XP_044221284.1	sodium/potassium/calcium exchanger 1-like	+		+		
XP_042286362.1	solute carrier family 23 member 1 isoform X3					+
XP_044218439.1	solute carrier organic anion transporter family member 2A1				+	
XP_035516581.1	son of sevenless homolog 1-like isoform X2					+
XP_042371562.1	spectrin beta chain, non-erythrocytic 1-like			+		
XP_042280249.1	sphingosine kinase 1-like isoform X1		+			
XP_027135729.1	spindlin-1 isoform X3		+			
XP_024001672.1	structural maintenance of chromosomes protein 3-like			+		
XP_044193548.1	SUN domain-containing protein 1-like isoform X6			+		
XP_044225010.1	supervillin-like isoform X11		+	+	+	+
XP_042249989.1	suppressor of cytokine signaling 7-like isoform X2					+
XP_023261120.1	surfeit locus protein 4-like			+		
XP_044193814.1	syntaxin-binding protein 4 isoform X3				+	
XP_044207193.1	tafazzin	+				
KAE8283764.1	Tankyrase-2				+	+

Table E.1 (continued)

XP_042285102.1	THO complex subunit 1				+	
XP_042263287.1	TIMELESS-interacting protein		+			
TMS01456.1	Transcription factor COE3				+	
XP_042247104.1	transcriptional repressor protein YY1b isoform X3		+			
XP_042274314.1	transforming growth factor-beta-induced protein ig-h3			+		
XP_030219775.1	transme Mbrane protein C17orf113 homolog					+
XP_044189756.1	TSC complex subunit 1a isoform X2				+	
XP_035517746.1	U3 small nucleolar ribonucleoprotein protein IMP4-like		+	+		
XP_044232212.1	ubiquitin carboxyl-terminal hydrolase 37-like			+		
XP_044232322.1	*unchar:* bub1ba isoform X2			+		
XP_023147949.1	*unchar:* LOC111583207		+			
XP_041651867.1	*unchar:* LOC121515259		+			
XP_042250304.1	*unchar:* LOC121885174 isoform X2	+				
XP_042261517.1	*unchar:* LOC121893587 isoform X1			+		+
XP_042264126.1	*unchar:* LOC121895221			+		+
XP_044186060.1	*unchar:* LOC122966129					+
XP_044206307.1	*unchar:* LOC122981697 isoform X1	+		+		
XP_044215416.1	*unchar:* LOC122987555			+		+
XP_044221307.1	*unchar:* LOC122991901 isoform X6		+			
XP_044221309.1	*unchar:* LOC122991901 isoform X8		+			
XP_044230523.1	*unchar:* si:dkey-33c12.4					+
CDQ88249.1	unnamed protein product		+		+	
CAG14839.1	unnamed protein product			+		
XP_042291316.1	upstream-binding protein 1 isoform X1			+		

Table E.1 (continued)

XP_042284405.1	vacuolar protein sorting-associated protein 13C isoform X3				+	
XP_033973361.1	vacuolar protein sorting-associated protein 13C-like					+
XP_031723088.1	vacuolar protein sorting-associated protein 37C-like			+		
XP_042336920.1	voltage-dependent L-type calcium channel subunit alpha-1D-like	+				
XP_040898821.1	voltage-dependent N-type calcium channel subunit alpha-1B-like			+		
XP_044221792.1	WD repeat, SAM and U-box domain-containing protein 1-like					+
XP_021328157.1	zinc finger CCCH domain-containing protein 13 isoform X2					+
XP_044194192.1	zinc finger CCCH domain-containing protein 7B isoform X1			+		
XP_044221194.1	zinc finger protein 142					+
XP_042258929.1	zinc finger protein 180-like isoform X1			+		
XP_042249632.1	zinc finger protein ZAT1				+	

Genes associated with highly divergent sequences between species categories ATL (Atlantic-only), PAC (Pacific-only), COS (cosmopolitan), and SBF (Southern Bluefin Tuna). A plus sign (+) indicates the gene is associated with divergence between thesee comparison groups. Gene/protein names use abbreviated terms hyp (hypothetical protein), unchar (uncharacterized), and pred (predicted).

BIBLIOGRAPHY

Abadi, S., Azouri, D., Pupko, T., & Mayrose, I. (2019). Model selection may not be a mandatory step for phylogeny reconstruction. *Nature Communications*, *10*(1), 934. doi: 10.1038/s41467-019-08822-w

Aguila, R. D., Perez, S. K. L., Catacutan, B. J. N., Lopez, G. V., Barut, N. C., & Santos, M. D. (2015). Distinct Yellowfin Tuna (Thunnus albacares) Stocks Detected in Western and Central Pacific Ocean (WCPO) Using DNA Microsatellites. *Plos One*, *10*(9), e0138292. doi: 10.1371/journal.pone.0138292

Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., … Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, *12*(2), R18. doi: 10.1186/gb-2011-12-2-r18

Alkan, C., Sajjadian, S., & Eichler, E. E. (2011). Limitations of next-generation genome sequence assembly. *Nature Methods*, *8*(1), 61–65. doi: 10.1038/nmeth.1527

Allam, A., Kalnis, P., & Solovyev, V. (2015). Karect: accurate correction of substitution, insertion and deletion errors for next-generation sequencing data. *Bioinformatics*, *31*(21), 3421–3428. doi: 10.1093/bioinformatics/btv415

Allendorf, F. W., Luikart, G. H., & Aitken, S. N. (2012). *Conservation and the Genetics of Populations* (2nd ed., p. 620). Wiley-Blackwell.

Alonge, M., Soyk, S., Ramakrishnan, S., Wang, X., Goodwin, S., Sedlazeck, F. J., … Schatz, M. C. (2019). Fast and accurate reference-guided scaffolding of draft genomes. *BioRxiv*. doi: 10.1101/519637

Anderson, G., Hampton, J., Smith, N., & Rico, C. (2019). Indications of strong adaptive population genetic structure in albacore tuna (Thunnus alalunga) in the southwest and central Pacific Ocean. *Ecology and Evolution*, *9*(18), 10354–10364. doi: 10.1002/ece3.5554

Anderson, G., Lal, M., Hampton, J., Smith, N., & Rico, C. (2019). Close Kin Proximity in Yellowfin Tuna (Thunnus albacares) as a Driver of Population Genetic Structure in the Tropical Western and Central Pacific Ocean. *Frontiers in Marine Science*, *6*. doi: 10.3389/fmars.2019.00341

André, C., Larsson, L. C., Laikre, L., Bekkevold, D., Brigham, J., Carvalho, G. R., … Ryman, N. (2011). Detecting population structure in a high gene-flow species, Atlantic herring (Clupea harengus): direct, simultaneous evaluation of neutral vs putatively selected loci. *Heredity*, *106*(2), 270–280. doi: 10.1038/hdy.2010.71

Antoni, L., Luque, P. L., Naghshpour, K., Reynal, L., & Saillant, E. A. (2014). Development and characterization of microsatellite markers for blackfin tuna (Thunnus atlanticus) with the use of Illumina paired-end sequencing. *First Break*, *112*(4), 322–325. doi: 10.7755/FB.112.4.8

Appleyard, S., Grewe, P., Innes, B., & Ward, R. (2001). Population structure of yellowfin tuna (*Thunnus albacares*) in the western Pacific Ocean, inferred from microsatellite loci. *Marine Biology*, *139*(2), 383–393. doi: 10.1007/s002270100578

Arocha, F., Lee, D. W., Marcano, L. A., & Marcano, J. S. (2001). Update information on the spawning of yellowfin tuna, Thunnus albacares, in the western central Atlantic. *Col. Vol. Sci. Pap. ICCAT*, *52*(1), 167–176.

Arrizabalaga, H., Dufour, F., Kell, L., Merino, G., Ibaibarriaga, L., Chust, G., … Bonhomeau, S. (2015). Global habitat preferences of commercially valuable tuna. *Deep Sea Research Part II: Topical Studies in Oceanography*, *113*, 102–112. doi: 10.1016/j.dsr2.2014.07.001

Avise, John C. (1992). Molecular Population Structure and the Biogeographic History of a Regional Fauna: A Case History with Lessons for Conservation Biology. *Oikos*, *63*(1), 62. doi: 10.2307/3545516

Avise, J C. (1998). Conservation genetics in the marine realm. *Journal of Heredity*, *89*(5), 377–382. doi: 10.1093/jhered/89.5.377

Bao, E., Jiang, T., & Girke, T. (2014). AlignGraph: algorithm for secondary de novo genome assembly guided by closely related references. *Bioinformatics*, *30*(12), i319–i328. doi: 10.1093/bioinformatics/btu291

Barth, J. M. I., Damerau, M., Matschiner, M., Jentoft, S., & Hanel, R. (2017). Genomic Differentiation and Demographic Histories of Atlantic and Indo-Pacific Yellowfin

Tuna (Thunnus albacares) Populations. *Genome Biology and Evolution*, *9*(4), 1084–1098. doi: 10.1093/gbe/evx067

Bay, R. A., & Ruegg, K. (2017). Genomic islands of divergence or opportunities for introgression? *Proceedings. Biological Sciences / the Royal Society*, *284*(1850). doi: 10.1098/rspb.2016.2414

Beatty, W. S., Lemons, P. R., Sethi, S. A., Everett, J. P., Lewis, C. J., Lynn, R. J., … Wenburg, J. K. (2020). Panmixia in a sea ice-associated marine mammal: evaluating genetic structure of the Pacific walrus (Odobenus rosmarus divergens) at multiple spatial scales. *Journal of Mammalogy*, *101*(3), 755–765. doi: 10.1093/jmammal/gyaa050

Becker, S., Böger, P., Oehlmann, R., & Ernst, A. (2000). PCR bias in ecological analysis: a case study for quantitative Taq nuclease assays in analyses of microbial communities. *Applied and Environmental Microbiology*, *66*(11), 4945–4953. doi: 10.1128/AEM.66.11.4945-4953.2000

Begg, G. A., Friedland, K. D., & Pearce, J. B. (1999). Stock identification and its role in stock assessment and fisheries management: an overview. *Fisheries Research*, *43*(1–3), 1–8. doi: 10.1016/S0165-7836(99)00062-4

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Berg, P. R., Star, B., Pampoulie, C., Bradbury, I. R., Bentzen, P., Hutchings, J. A., … Jakobsen, K. S. (2017). Trans-oceanic genomic divergence of Atlantic cod ecotypes is associated with large inversions. *Heredity*, *119*(6), 418–428. doi: 10.1038/hdy.2017.54

Berlocher, S. H., & Feder, J. L. (2002). Sympatric speciation in phytophagous insects: moving beyond controversy? *Annual Review of Entomology*, *47*, 773–815. doi: 10.1146/annurev.ento.47.091201.145312

Bernard, A. M., Feldheim, K. A., Heithaus, M. R., Wintner, S. P., Wetherbee, B. M., & Shivji, M. S. (2016). Global population genetic dynamics of a highly migratory, apex predator shark. *Molecular Ecology*, *25*(21), 5312–5329. doi: 10.1111/mec.13845

Bertels, F., Silander, O. K., Pachkov, M., Rainey, P. B., & van Nimwegen, E. (2014). Automated reconstruction of whole-genome phylogenies from short-sequence reads. *Molecular Biology and Evolution*, *31*(5), 1077–1088. doi: 10.1093/molbev/msu088

Betancur-R, R., Wiley, E. O., Arratia, G., Acero, A., Bailly, N., Miya, M., … Ortí, G. (2017). Phylogenetic classification of bony fishes. *BMC Evolutionary Biology*, *17*(1), 162. doi: 10.1186/s12862-017-0958-3

Bezerra, N. P. A., Fernandes, C. A. F., Albuquerque, F. V., Pedrosa, V., Hazin, F., & Travassos, P. (2013). Reproduction of Blackfin tuna Thunnus atlanticus (Perciformes: Scombridae) in Saint Peter and Saint Paul Archipelago, Equatorial Atlantic, Brazil. *Revista de Biologia Tropical*, *61*(3), 1327–1339.

Bhatia, G., Patterson, N., Sankararaman, S., & Price, A. L. (2013). Estimating and interpreting FST: the impact of rare variants. *Genome Research*, *23*(9), 1514–1521. doi: 10.1101/gr.154831.113

Blair, C., Weigel, D. E., Balazik, M., Keeley, A. T. H., Walker, F. M., Landguth, E., … Balkenhol, N. (2012). A simulation-based evaluation of methods for inferring linear barriers to gene flow. *Molecular Ecology Resources*, *12*(5), 822–833. doi: 10.1111/j.1755-0998.2012.03151.x

Bourret, V., Kent, M. P., Primmer, C. R., Vasemägi, A., Karlsson, S., Hindar, K., … Lien, S. (2013). SNP-array reveals genome-wide patterns of geographical and potential adaptive divergence across the natural range of Atlantic salmon (Salmo salar). *Molecular Ecology*, *22*(3), 532–551. doi: 10.1111/mec.12003

Bradbury, I. R., Hubert, S., Higgins, B., Borza, T., Bowman, S., Paterson, I. G., … Bentzen, P. (2010). Parallel adaptive evolution of Atlantic cod on both sides of the Atlantic Ocean in response to temperature. *Proceedings. Biological Sciences / the Royal Society*, *277*(1701), 3725–3734. doi: 10.1098/rspb.2010.0985

Bradbury, I. R., Hubert, S., Higgins, B., Bowman, S., Borza, T., Paterson, I. G., … Bentzen, P. (2013). Genomic islands of divergence and their consequences for the resolution of spatial structure in an exploited marine fish. *Evolutionary Applications*, *6*(3), 450–461. doi: 10.1111/eva.12026

Bradman, H., Grewe, P., & Appleton, B. (2011). Direct comparison of mitochondrial markers for the analysis of swordfish population structure. *Fisheries Research*, *109*(1), 95–99. doi: 10.1016/j.fishres.2011.01.022

Brill, R. W., Block, B. A., Boggs, C. H., Bigelow, K. A., Freund, E. V., & Marcinek, D. J. (1999). Horizontal movements and depth distribution of large adult yellowfin tuna (Thunnus albacares) near the Hawaiian Islands, recorded using ultrasonic telemetry: implications for the physiological ecology of pelagic fishes. *Marine Biology*, *133*(3), 395–408. doi: 10.1007/s002270050478

Brown-Peterson, N. J., Franks, J. S., Gibson, D. M., & Marshall, C. (2013). Aspects of the Reproductive Biology of Yellowfin Tuna, Thunnus albacares, in the Northern Gulf of Mexico. *66th Gulf and Caribbean Fisheries Institute*, *66*, 509–510. 66th Gulf and Caribbean Fisheries Institute.

Buonaccorsi, V. P., Reece, K. S., Morgan, L. W., & Graves, J. E. (1999). Geographic distribution of molecular variance within the blue marlin (makaira nigricans): a hierarchical analysis of allozyme, single-copy nuclear dna, and mitochondrial dna markers. *Evolution*, *53*(2), 568–579. doi: 10.1111/j.1558-5646.1999.tb03792.x

Calvert, A. M., Walde, S. J., & Taylor, P. D. (2009). Nonbreeding-Season Drivers of Population Dynamics in Seasonal Migrants: Conservation Parallels Across Taxa. *Avian Conservation and Ecology*, *4*(2). doi: 10.5751/ACE-00335-040205

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, *10*, 421. doi: 10.1186/1471-2105-10-421

Carlsson, J., McDowell, J. R., Carlsson, J. E. L., & Graves, J. E. (2006). Genetic Identity of YOY Bluefin Tuna from the Eastern and Western Atlantic Spawning Areas. *Journal of Heredity*, *98*(1), 23–28. doi: 10.1093/jhered/esl046

Carvalho, G. R., & Hauser, L. (1994). Molecular genetics and the stock concept in fisheries. *Reviews in Fish Biology and Fisheries*, *4*(3), 326–350. doi: 10.1007/BF00042908

Castle, J. C. (2011). SNPs occur in regions with less genomic sequence conservation. *Plos One*, *6*(6), e20660. doi: 10.1371/journal.pone.0020660

Catchen, J. M., Hohenlohe, P. A., Bernatchez, L., Funk, W. C., Andrews, K. R., & Allendorf, F. W. (2017). Unbroken: RADseq remains a powerful tool for understanding the genetics of adaptation in natural populations. *Molecular Ecology Resources*, *17*(3), 362–365. doi: 10.1111/1755-0998.12669

Chaisson, M. J., & Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, *13*, 238. doi: 10.1186/1471-2105-13-238

Chakraborty, M., Baldwin-Brown, J. G., Long, A. D., & Emerson, J. J. (2016). Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Research*, *44*(19), e147. doi: 10.1093/nar/gkw654

Chapman, D. D., Feldheim, K. A., Papastamatiou, Y. P., & Hueter, R. E. (2015). There and back again: a review of residency and return migrations in sharks, with implications for population structure and management. *Annual Review of Marine Science*, *7*, 547–570. doi: 10.1146/annurev-marine-010814-015730

Chen, C., Durand, E., Forbes, F., & François, O. (2007). Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Molecular Ecology Notes*, *7*(5), 747–756. doi: 10.1111/j.1471-8286.2007.01769.x

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. doi: 10.1093/bioinformatics/bty560

Cheng, J., Han, Z., Song, N., Gao, T., Yanagimoto, T., & Strüssmann, C. A. (2018). Effects of Pleistocene glaciation on the phylogeographic and demographic histories of chub mackerel Scomber japonicus in the north-western Pacific. *Marine and Freshwater Research*, *69*(4), 514. doi: 10.1071/MF17099

Cheviron, Z. A., & Brumfield, R. T. (2009). Migration-selection balance and local adaptation of mitochondrial haplotypes in rufous-collared sparrows (Zonotrichia capensis) along an elevational gradient. *Evolution*, *63*(6), 1593–1605. doi: 10.1111/j.1558-5646.2009.00644.x

Chow, S., & Kishino, H. (1995). Phylogenetic relationships between tuna species of the genus Thunnus (Scombridae: Teleostei): inconsistent implications from morphology, nuclear and mitochondrial genomes. *Journal of Molecular Evolution*, *41*(6), 741–748. doi: 10.1007/BF00173154

Chow, S., Nakagawa, T., Suzuki, N., Takeyama, H., & Matsunaga, T. (2006). Phylogenetic relationships among Thunnus species inferred from rDNA ITS1 sequence. *Journal of Fish Biology*, *68*(A), 24–35. doi: 10.1111/j.0022-1112.2006.00945.x

Ciezarek, A. G., Osborne, O. G., Shipley, O. N., Brooks, E. J., Tracey, S. R., McAllister, J. D., … Savolainen, V. (2019). Phylotranscriptomic Insights into the Diversification of Endothermic Thunnus Tunas. *Molecular Biology and Evolution*, *36*(1), 84–96. doi: 10.1093/molbev/msy198

Coates, A. G., Jackson, J. B. C., Collins, L. S., Cronin, T. M., Dowsett, H. J., Bybell, L. M., … Obando, J. A. (1992). Closure of the Isthmus of Panama: The near-shore marine record of Costa Rica and western Panama. *Geological Society of America Bulletin*, *104*(7), 814–828. doi: 10.1130/0016-7606(1992)104<0814:COTIOP>2.3.CO;2

Collette, B, Amorim A, F., Boustany, A., E, C. K., Dooley, J., de Oliviera Leite, N. Jr., … Pires Ferreira Travassos, P. E. (2010, September 15). Thunnus atlanticus. Retrieved

February 3, 2021, from The IUCN Red List of Threatened Species website: http://www.iucnredlist.org/details/155276/0

Collette, Bruce, BoustanY, A., Fox, W., Graves, J., Juan Jorda, M., & Restrepo, V. (2021). Thunnus albacares. Retrieved February 2, 2021, from The IUCN Red List of Threatened Species website: https://dx.doi.org/10.2305/IUCN.UK.2021-2.RLTS.T21857A46624561.en.

Collette, Bruce B., Acer, A., Amorim, A. F., Boustany, A., Canales-Ramirez, C., Cardenas, G., … Yanez, E. (2011). Thunnus albacares. Retrieved February 3, 2021, from The IUCN Red List of Threatened Species website: http://dx.doi.org/10.2305/IUCN.UK.2011-2.RLTS.T21857A9327139.en

Collette, B B, & Nauen, C. E. (1983). *FAO species catalogue. Scombrids of the world. An annotated and illustrated catalogue of tunas, mackerels, bonitos and related species known to date* (Vol. 2). Rome: Food and Agriculture Organization of the United Nations.

Conomos, M. P., Miller, M. B., & Thornton, T. A. (2015). Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genetic Epidemiology*, *39*(4), 276–293. doi: 10.1002/gepi.21896

Conomos, M. P., Reiner, A. P., Weir, B. S., & Thornton, T. A. (2016). Model-free Estimation of Recent Genetic Relatedness. *American Journal of Human Genetics*, *98*(1), 127–148. doi: 10.1016/j.ajhg.2015.11.022

Conover, D. O., Arnott, S. A., Walsh, M. R., & Munch, S. B. (2005). Darwinian fishery science: lessons from the Atlantic silverside (*Menidia menidia*). *Canadian Journal of Fisheries and Aquatic Sciences*, *62*(4), 730–737. doi: 10.1139/f05-069

Cornic, M., Smith, B. L., Kitchens, L. L., Alvarado Bremer, J. R., & Rooker, J. R. (2017). Abundance and habitat associations of tuna larvae in the surface water of the Gulf of Mexico. *Hydrobiologia*, *806*(1), 1–18. doi: 10.1007/s10750-017-3330-0

Costa, F. E. S., Braga, F. M. S., & Amorim, A. F. (2005). *Fishery Biology of the Yellowfin Tuna, Thunnus albacares in Southern Brazil* (No. 58; 1st ed., pp. 309–349). ICCAT.

Coyne, J. A., & Orr, H. A. (2004). *Speciation* (Vol. 37, p. 545). Sinauer Associates.

da Silva, G. B., Hazin[1], H. G., & Hazin, F. H. V. (2019). The tuna fisheries on associated schools in Brazil: description and trends. *Collective Volume of Scientific Papers*, *75*(7), 1924–1934.

Dalongeville, A., Benestan, L., Mouillot, D., Lobreaux, S., & Manel, S. (2018). Combining six genome scan methods to detect candidate genes to salinity in the Mediterranean striped red mullet (Mullus surmuletus). *BMC Genomics*, *19*(1), 217. doi: 10.1186/s12864-018-4579-z

Daly-Engel, T. S., Seraphin, K. D., Holland, K. N., Coffey, J. P., Nance, H. A., Toonen, R. J., & Bowen, B. W. (2012). Global phylogeography with mixed-marker analysis reveals male-mediated dispersal in the endangered scalloped hammerhead shark (Sphyrna lewini). *Plos One, 7*(1), e29986. doi: 10.1371/journal.pone.0029986

Damien, P., Sheinbaum, J., Pasqueron de Fommervault, O., Jouanno, J., Linacre, L., & Duteil, O. (2021). Do Loop Current eddies stimulate productivity in the Gulf of Mexico? *Biogeosciences, 18*(14), 4281–4303. doi: 10.5194/bg-18-4281-2021

Dammannagoda, S. T., Hurwood, D. A., & Mather, P. B. (2008). Evidence for fine geographical scale heterogeneity in gene frequencies in yellowfin tuna (Thunnus albacares) from the north Indian Ocean around Sri Lanka. *Fisheries Research, 90*(1–3), 147–157. doi: 10.1016/j.fishres.2007.10.006

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics, 27*(15), 2156–2158. doi: 10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., … Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience, 10*(2). doi: 10.1093/gigascience/giab008

Darwin, C. (1859). *The origin of the species by means of natural selection.* (p. 365). London. doi: 10.5962/bhl.title.24329

De Sylva, D. P., Rathjen, W. F., Higman, J. B., Caábro, S., José, A., & Ramirez-Flores, A. (1987). *Fisheries development for underutilized Atlantic tunas: Blackfin and little tunny*. (No. NOAA technical memorandum NMFS-SEFC ; 191). NOAA.

de Sylva, D. P. (1955). The osteology and phylogenetic relationships of the blackfin tuna, Thunnus atlanticus (Lesson). *Bulletin of Marine Science*, *5*(1), 1–41.

Diaha, N. C., Zudaire, I., Chassot, E., & Barrigah, B. D. (2016). *Annual monitoring of reproductive traits of female yellowfin tuna (Thunnus albacares) in the eastern Atlantic Ocean* (No. 72; pp. 534–548). ICCAT.

Díaz-Arce, N., Arrizabalaga, H., Murua, H., Irigoien, X., & Rodríguez-Ezpeleta, N. (2016). RAD-seq derived genome-wide nuclear markers resolve the phylogeny of tunas. *Molecular Phylogenetics and Evolution*, *102*, 202–207. doi: 10.1016/j.ympev.2016.06.002

Dimens, P. V., & Selwyn, J. (2022). BioJulia/PopGen.jl. *Zenodo*. Retrieved from https://doi.org/10.5281/zenodo.6077851

Dimens, P. V., Willis, S., Dean Grubbs, R., & Portnoy, D. S. (2019). A genomic assessment of movement and gene flow around the South Florida vicariance zone in the migratory coastal blacknose shark, Carcharhinus acronotus. *Marine Biology*, *166*(7), 86. doi: 10.1007/s00227-019-3533-1

Dimens, P. V. (2022a). pdimens/LepWrap. *Zenodo*. Retrieved from https://doi.org/10.5281/zenodo.6326228

Dimens, P. V. (2022b). pdimens/PopGenSims.jl: v0.3.1. *Zenodo*. doi: 10.5281/zenodo.6077864

Dimens, P. V. (2022c). pdimens/PopGenSims.jl. *Zenodo*. Retrieved from https://doi.org/10.5281/zenodo.6325983

Do, C., Waples, R. S., Peel, D., Macbeth, G. M., Tillett, B. J., & Ovenden, J. R. (2014). NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. *Molecular Ecology Resources*, *14*(1), 209–214. doi: 10.1111/1755-0998.12157

Domínguez-López, M., Díaz-Jaimes, P., Uribe-Alcocer, M., & Quiñonez-Velázquez, C. (2015). Post-glacial population expansion of the Monterey Spanish mackerel Scomberomorus concolor in the Gulf of California. *Journal of Fish Biology*, *86*(3), 1153–1162. doi: 10.1111/jfb.12580

Druon, J. N., Chassot, E., Floch, L., & Maufroy, A. (2015, October 28). *Preferred habitat of tropical tuna species in the Eastern Atlantic and Western Indian Oceans: a comparative analysis between FAD-associated and free-swimming schools.* 23. Montpellier, France: 17ème groupe de travail sur les thons tropicaux. Retrieved from https://www.documentation.ird.fr/hor/fdi:010065858

Dumschott, K., Schmidt, M. H.-W., Chawla, H. S., Snowdon, R., & Usadel, B. (2020). Oxford Nanopore sequencing: new opportunities for plant genomics? *Journal of Experimental Botany*, *71*(18), 5313–5322. doi: 10.1093/jxb/eraa263

Ely, B., Viñas, J., Alvarado Bremer, J. R., Black, D., Lucas, L., Covello, K., … Thelen, E. (2005). Consequences of the historical demography on the global population structure of two highly migratory cosmopolitan marine fishes: the yellowfin tuna (Thunnus albacares) and the skipjack tuna (Katsuwonus pelamis). *BMC Evolutionary Biology*, *5*, 19. doi: 10.1186/1471-2148-5-19

Excoffier, L, Smouse, P. E., & Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, *131*(2), 479–491. doi: 10.1093/genetics/131.2.479

Excoffier, Laurent, & Lischer, H. E. L. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, *10*(3), 564–567. doi: 10.1111/j.1755-0998.2010.02847.x

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

Feder, J. L., & Nosil, P. (2010). The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution*, *64*(6), 1729–1747. doi: 10.1111/j.1558-5646.2009.00943.x

Feng, X.-J., Jiang, G.-F., & Fan, Z. (2015). Identification of outliers in a genomic scan for selection along environmental gradients in the bamboo locust, Ceracris kiangsu. *Scientific Reports*, *5*, 13758. doi: 10.1038/srep13758

Ferree, P. M., & Barbash, D. A. (2009). Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in Drosophila. *PLoS Biology*, *7*(10), e1000234. doi: 10.1371/journal.pbio.1000234

Ferreira, D. G., Galindo, B. A., Frantine-Silva, W., Almeida, F. S., & Sofia, S. H. (2015). Genetic structure of a Neotropical sedentary fish revealed by AFLP, microsatellite and mtDNA markers: a case study. *Conservation Genetics (Print)*, *16*(1), 151–166. doi: 10.1007/s10592-014-0648-2

Foll, M., & Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, *180*(2), 977–993. doi: 10.1534/genetics.108.092221

Fonteneau, A, & Chassot, E. (2013). An overview of yellowfin tuna growth in the Atlantic Ocean: von Bertalanffy or multistanza growth?. *Collect. Vol. Sci. Pap. ICCAT*, *69*(5), 2059–2075.

Fonteneau, Alain, & Hallier, J.-P. (2015). Fifty years of dart tag recoveries for tropical tuna: A global comparison of results for the western Pacific, eastern Pacific, Atlantic, and Indian Oceans. *Fisheries Research*, *163*, 7–22. doi: 10.1016/j.fishres.2014.03.022

Franks, J. S., Saillant, E. A., & Brown-Peterson, N. (2015). *Studies of reproductive biology, feeding ecology and conservation genetics of Yellowfin Tuna (*thunnus albacares*) in the northern gulf of mexico* (No. CFMS#718119; Final Report, p. 80). Louisiana Department of Wildlife and Fisheries.

Freire, K. M. F., Lessa, R., & Lins-Oliveira, J. E. (2005). Fishery and Biology of Blackfin Tuna Thunnus atlanticus off Northeastern Brazil. *Gulf and Caribbean Research*, *17*, 15–24. doi: 10.18785/gcr.1701.02

Frimodt, C., & Dore, I. (1995). *Multilingual Illustrated Guide to the World's Commercial Coldwater Fish ("Fishing News" Books)* (1st ed., p. 264). Oxford: Wiley-Blackwell.

Gao, G., Nome, T., Pearse, D. E., Moen, T., Naish, K. A., Thorgaard, G. H., … Palti, Y. (2018). A new single nucleotide polymorphism database for rainbow trout generated through whole genome resequencing. *Frontiers in Genetics*, *9*, 147. doi: 10.3389/fgene.2018.00147

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *ArXiv Preprint ArXiv:1207.3907*.

Gibb, F. M., Régnier, T., Donald, K., & Wright, P. J. (2017). Connectivity in the early life history of sandeel inferred from otolith microchemistry. *Journal of Sea Research*, *119*, 8–16. doi: 10.1016/j.seares.2016.10.003

Gibbs, R. J., & Collette, B. B. (1967). Comparative anatomy and systematics of the tunas, genus Thunnus. *Fish. Bull., Fish Wildl. Serv*, *66*(1), 65–130.

Girgis, H. Z. (2015). Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. *BMC Bioinformatics*, *16*, 227. doi: 10.1186/s12859-015-0654-5

Glaubitz, J. C., Rhodes, O. E., & Dewoody, J. A. (2003). Prospects for inferring pairwise relationships with single nucleotide polymorphisms. *Molecular Ecology*, *12*(4), 1039–1047. doi: 10.1046/j.1365-294X.2003.01790.x

Gogarten, S. M., Bhangale, T., Conomos, M. P., Laurie, C. A., McHugh, C. P., Painter, I., … Laurie, C. C. (2012). GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics*, *28*(24), 3329–3331. doi: 10.1093/bioinformatics/bts610

Gonzalez, E. G., Beerli, P., & Zardoya, R. (2008). Genetic structuring and migration patterns of Atlantic bigeye tuna, Thunnus obesus (Lowe, 1839). *BMC Evolutionary Biology*, *8*, 252. doi: 10.1186/1471-2148-8-252

Gosselin, T., Lamothe, M., Devloo-Delva, F., & Grewe, P. (2019). thierrygosselin/radiator. *Zenodo*. Retrieved from https://doi.org/10.5281/zenodo.2595083

Graham, B. S., Grubbs, D., Holland, K., & Popp, B. N. (2006). A rapid ontogenetic shift in the diet of juvenile yellowfin tuna from Hawaii. *Marine Biology*, *150*(4), 647–658. doi: 10.1007/s00227-006-0360-y

Graves, J. E. (1998). Molecular insights into the population structures of cosmopolitan marine fishes. *Journal of Heredity*, *89*(5), 427–437. doi: 10.1093/jhered/89.5.427

Grewe, P. M., Feutry, P., Hill, P. L., Gunasekera, R. M., Schaefer, K. M., Itano, D. G., … Davies, C. R. (2015). Evidence of discrete yellowfin tuna (Thunnus albacares) populations demands rethink of management for this globally important resource. *Scientific Reports*, *5*, 16916. doi: 10.1038/srep16916

Grimes, C. B. (1987). Delineation of king mackerel (Scomberomorus cavalla) stocks along the US east coast and in the Gulf of Mexico. *Proceedings of Stock Identification Workshop 186-187*, 186.

Guo, L., Li, M., Zhang, H., Yang, S., Chen, X., Meng, Z., & Lin, H. (2016). Next-generation sequencing of the yellowfin tuna mitochondrial genome reveals novel phylogenetic relationships within the genus Thunnus. *Mitochondrial DNA. Part A, DNA Mapping, Sequencing, and Analysis*, *27*(3), 2089–2090. doi: 10.3109/19401736.2014.982570

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075. doi: 10.1093/bioinformatics/btt086

Gu, Z., Gu, L., Eils, R., Schlesner, M., & Brors, B. (2014). circlize Implements and enhances circular visualization in R. *Bioinformatics, 30*(19), 2811–2812. doi: 10.1093/bioinformatics/btu393

Haghshenas, E., Asghari, H., Stoye, J., Chauve, C., & Hach, F. (2020). HASLR: fast hybrid assembly of long reads. *IScience*, *23*(8), 101389. doi: 10.1016/j.isci.2020.101389

Harris, R. S. (2007). Improved Pairwise Alignment of Genomic DNA. *The University of Pennsylvania, Doctoral Dissertation*, 85. Retrieved from http://lynx.lib.usm.edu/dissertations-theses/improved-pairwise-alignment-genomic-dna/docview/304835295/se-2?accountid=13946

Hauser, L., Baird, M., Hilborn, R., Seeb, L. W., & Seeb, J. E. (2011). An empirical comparison of SNPs and microsatellites for parentage and kinship assignment in a wild sockeye salmon (Oncorhynchus nerka) population. *Molecular Ecology Resources*, *11 Suppl 1*, 150–161. doi: 10.1111/j.1755-0998.2010.02961.x

Hauser, L., & Carvalho, G. R. (2008). Paradigm shifts in marine fisheries genetics: ugly hypotheses slain by beautiful facts. *Fish and Fisheries*, *9*(4), 333–362. doi: 10.1111/j.1467-2979.2008.00299.x

Havelka, M., Sawayama, E., Saito, T., Yoshitake, K., Saka, D., Ineno, T., … Matsubara, T. (2021). Chromosome-Scale Genome Assembly and Transcriptome Assembly of Kawakawa Euthynnus affinis; A Tuna-Like Species. *Frontiers in Genetics, 12*, 739781. doi: 10.3389/fgene.2021.739781

Hawkes, W., Wotton, K., University of Oxford and Wytham Woods Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, … Darwin Tree of Life Consortium. (2021). The genome sequence of the tapered dronefly, Eristalis pertinax (Scopoli, 1763). *Wellcome Open Research, 6*, 292. doi: 10.12688/wellcomeopenres.17267.1

Hedgecock, D., & Pudovkin, A. I. (2011). Sweepstakes Reproductive Success in Highly Fecund Marine Fish and Shellfish: A Review and Commentary. *Bulletin of Marine Science*, *87*(4), 971–1002. doi: 10.5343/bms.2010.1051

Hilborn, R., Quinn, T. P., Schindler, D. E., & Rogers, D. E. (2003). Biocomplexity and fisheries sustainability. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(11), 6564–6568. doi: 10.1073/pnas.1037274100

Hill, W. G., & Robertson, A. (1968). Linkage disequilibrium in finite populations. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *38*(6), 226–231. doi: 10.1007/BF01245622

Hofer, T., Foll, M., & Excoffier, L. (2012). Evolutionary forces shaping genomic islands of population differentiation in humans. *BMC Genomics*, *13*, 107. doi: 10.1186/1471-2164-13-107

Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A., & Cresko, W. A. (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, *6*(2), e1000862. doi: 10.1371/journal.pgen.1000862

Hollenbeck, C. M., Portnoy, D. S., & Gold, J. R. (2016). A method for detecting recent changes in contemporary effective population size from linkage disequilibrium at linked and unlinked loci. *Heredity*, *117*(4), 207–216. doi: 10.1038/hdy.2016.30

Hood, G. M. (2010). PopTools. A module of the Biological ESTEEM Collection. *BioQUEST Curriculum Consortium, 3.2.5*. Retrieved from http://bioquest.org/esteem/esteem_details.php?product_id=248

Hoolihan, J. P., Wells, R. J. D., Luo, J., Falterman, B., Prince, E. D., & Rooker, J. R. (2014). Vertical and horizontal movements of yellowfin tuna in the gulf of mexico. *Marine and Coastal Fisheries*, *6*(1), 211–222. doi: 10.1080/19425120.2014.935900

Howard, D. J., & Berlocher, S. H. (1998). Theory and models of sympatricspeciation. In *Endless Forms: Species and Speciation* (Illustrated, pp. 79–89). New York: Oxford University Press.

Hsu, A. C., Boustany, A. M., Roberts, J. J., Chang, J.-H., & Halpin, P. N. (2015). Tuna and swordfish catch in the U.S. northwest Atlantic longline fishery in relation to mesoscale eddies. *Fisheries Oceanography*, *24*(6), 508–520. doi: 10.1111/fog.12125

Huang, S., Kang, M., & Xu, A. (2017). HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. *Bioinformatics*, *33*(16), 2577–2579. doi: 10.1093/bioinformatics/btx220

Hudson, R. R., Slatkin, M., & Maddison, W. P. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics*, *132*(2), 583–589. doi: 10.1093/genetics/132.2.583

Huson, D. H., & Scornavacca, C. (2012). Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Systematic Biology*, *61*(6), 1061–1067. doi: 10.1093/sysbio/sys062

ICCAT. (2011, September 5). *Report of the 2011 ICCAT Yellowfin Tuna Stock Assessment*. 113. ICCAT. Retrieved from https://www.iccat.int/Documents/Meetings/Docs/2011_YFT_ASSESS_REP.pdf

ICCAT. (2016). *Report of the 2016 ICCAT Yellowfin Tuna Stock Assessment meeting* (p. 103). San Sebastian, Spain: ICCAT. Retrieved from ICCAT website: https://www.iccat.int/Documents/SCRS/DetRep/YFT_SA_ENG.pdf

ICCAT. (2019a, October 4). *Reportof the standing committee on research and statistics (SCRS)* . 24–43. Madrid, Spain: ICCAT.

ICCAT. (2019b, July 16). *Report of the 2019 ICCAT Yellowfin tuna stock assessment meeting*. 117. Grand-Bassam, Cote d'Ivoire: ICCAT. Retrieved from https://www.iccat.int/Documents/SCRS/DetRep/YFT_SA_ENG.pdf

Ichikawa, K., Tomioka, S., Suzuki, Y., Nakamura, R., Doi, K., Yoshimura, J., … Morishita, S. (2017). Centromere evolution and CpG methylation during vertebrate speciation. *Nature Communications*, *8*(1), 1833. doi: 10.1038/s41467-017-01982-7

Idyll, C. P., & De Sylva, D. P. (1963). Synopsis of biological data on the blackfin tuna Thunnus atlanticus (Lesson) 1830 (Western Atlantic). *FAO Fisheries Biology Synopsis No, 68, Species synopsis No 25*, 11. Rome: Food and Agriculture Organization.

Ingram, T., & Mahler, D. L. (2011). Niche diversification follows key innovation in Antarctic fish radiation. *Molecular Ecology*, *20*(22), 4590–4591. doi: 10.1111/j.1365-294x.2011.05321.x

Jackman, S. D., Vandervalk, B. P., Mohamadi, H., Chu, J., Yeo, S., Hammond, S. A., … Birol, I. (2017). ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. *Genome Research*, *27*(5), 768–777. doi: 10.1101/gr.214346.116

Jansen, H. J., Liem, M., Jong-Raadsen, S. A., Dufour, S., Weltzien, F.-A., Swinkels, W., … Henkel, C. V. (2017). Rapid de novo assembly of the European eel genome from nanopore sequencing reads. *Scientific Reports*, *7*(1), 7213. doi: 10.1038/s41598-017-07650-6

Jaworski, C. C., Allan, C. W., & Matzkin, L. M. (2019). Chromosome-level hybrid *de novo* genome assemblies as an attainable option for non-model organisms. *BioRxiv*. doi: 10.1101/748228

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, *11*, 94. doi: 10.1186/1471-2156-11-94

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403–1405. doi: 10.1093/bioinformatics/btn129

Jones, O. R., & Wang, J. (2010). COLONY: a program for parentage and sibship inference from multilocus genotype data. *Molecular Ecology Resources*, *10*(3), 551–555. doi: 10.1111/j.1755-0998.2009.02787.x

Jorgensen, S. J., Reeb, C. A., Chapple, T. K., Anderson, S., Perle, C., Van Sommeran, S. R., … Block, B. A. (2010). Philopatry and migration of Pacific white sharks.

Proceedings. Biological Sciences / the Royal Society, *277*(1682), 679–688. doi: 10.1098/rspb.2009.1155

Juárez, M. (1978). Distribución de las larvas de la familia Scombridae en aguas adyacentes a las Bahamas. *Cuba. Rev. Invest*, *3*, 69–77.

Kaji, T., Tanaka, M., Oka, M., Takeuchi, H., Ohsumi, S., Teruya, K., & Hirokawa, J. (1999). Growth and Morphological Development of Laboratory-Reared Yellowfin Tuna *Thunnus albacares* Larvae and Early Juveniles, with Special Emphasis on the Digestive System. *Fisheries Science*, *65*(5), 700–707. doi: 10.2331/fishsci.65.700

Karl, S. A., Castro, A. L. F., Lopez, J. A., Charvet, P., & Burgess, G. H. (2011). Phylogeography and conservation of the bull shark (Carcharhinus leucas) inferred from mitochondrial and microsatellite DNA. *Conservation Genetics (Print)*, *12*(2), 371–382. doi: 10.1007/s10592-010-0145-1

Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., … Kohara, Y. (2007). The medaka draft genome and insights into vertebrate genome evolution. *Nature*, *447*(7145), 714–719. doi: 10.1038/nature05846

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. doi: 10.1093/molbev/mst010

Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., & Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Research*, *21*(3), 487–493. doi: 10.1101/gr.113985.110

Kitada, S., Nakajima, K., & Hamasaki, K. (2017). Population panmixia and demographic expansion of a highly piscivorous marine fish Scomberomorus niphonius. *Journal of Fish Biology*, *91*(5), 1435–1448. doi: 10.1111/jfb.13466

Kohler, N. E., & Turner, P. A. (2001). Shark tagging: A review of conventional methods and studies. In *The behavior and sensory biology of elasmobranch fishes: an anthology in memory of Donald Richard Nelson* (pp. 191–244).

Kondrashov, A. S., & Mina, M. V. (1986). Sympatric speciation: when is it possible? *Biological Journal of the Linnean Society*, *27*(3), 201–223. doi: 10.1111/j.1095-8312.1986.tb01734.x

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736. doi: 10.1101/gr.215087.116

Kosambi, D. D. (1943). The estimation of map distances from recombination values. *Annals of Eugenics*, *12*(1), 172–175. doi: 10.1111/j.1469-1809.1943.tb02321.x

Koskinen, M. T., Hirvonen, H., Landry, P.-A., & Primmer, C. R. (2004). The benefits of increasing the number of microsatellites utilized in genetic population studies: an empirical perspective. *Hereditas*, *141*(1), 61–67. doi: 10.1111/j.1601-5223.2004.01804.x

Köster, J., & Rahmann, S. (2012). Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. doi: 10.1093/bioinformatics/bts480

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, *35*(21), 4453–4455. doi: 10.1093/bioinformatics/btz305

Kritzer, J. P., & Sale, P. F. (2004). Metapopulation ecology in the sea: from Levins' model to marine ecology and fisheries science. *Fish and Fisheries*, *5*(2), 131–140. doi: 10.1111/j.1467-2979.2004.00131.x

Kuhl, H., Li, L., Wuertz, S., Stöck, M., Liang, X.-F., & Klopp, C. (2020). CSA: A high-throughput chromosome-scale assembly pipeline for vertebrate genomes. *GigaScience*, *9*(5). doi: 10.1093/gigascience/giaa034

Kundu, R., Casey, J., & Sung, W.-K. (2019). Hypo: super fast & accurate polisher for long read genome assemblies. *BioRxiv*. doi: 10.1101/2019.12.19.882506

Kwok, P. Y. (2001). Methods for genotyping single nucleotide polymorphisms. *Annual Review of Genomics and Human Genetics*, *2*, 235–258. doi: 10.1146/annurev.genom.2.1.235

Laconcha, U., Iriondo, M., Arrizabalaga, H., Manzano, C., Markaide, P., Montes, I., … Estonba, A. (2015). New Nuclear SNP Markers Unravel the Genetic Structure and Effective Population Size of Albacore Tuna (Thunnus alalunga). *Plos One*, *10*(6), e0128247. doi: 10.1371/journal.pone.0128247

Lamichhaney, S., Martinez Barrio, A., Rafati, N., Sundström, G., Rubin, C.-J., Gilbert, E. R., … Andersson, L. (2012). Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(47), 19345–19350. doi: 10.1073/pnas.1216128109

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

Lang, K. L., Grimes, C. B., & Shaw, R. F. (1994). Variations in the age and growth of yellowfin tuna larvae,Thunnus albacares, collected about the Mississippi River plume. *Environmental Biology of Fishes*, *39*(3), 259–270. doi: 10.1007/BF00005128

Leblois, R., Estoup, A., & Rousset, F. (2003). Influence of mutational and sampling factors on the estimation of demographic parameters in a “continuous” population under isolation by distance. *Molecular Biology and Evolution*, *20*(4), 491–502. doi: 10.1093/molbev/msg034

Lee, Y.-H., Yen, T.-B., Chen, C.-F., & Tseng, M.-C. (2018). Variation in the Karyotype, Cytochrome b Gene, and 5S rDNA of Four Thunnus (Perciformes, Scombridae) Tunas. *Zoological Studies (Taipei, Taiwan)*, *57*, e34. doi: 10.6620/ZS.2018.57-34

Lenormand, T. (2002). Gene flow and the limits to natural selection. *Trends in Ecology & Evolution*, *17*(4), 183–189. doi: 10.1016/S0169-5347(02)02497-7

Levy Karin, E., Mirdita, M., & Söding, J. (2020). MetaEuk-sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome*, *8*(1), 48. doi: 10.1186/s40168-020-00808-x

Limborg, M. T., Helyar, S. J., De Bruyn, M., Taylor, M. I., Nielsen, E. E., Ogden, R., … Bekkevold, D. (2012). Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish, the Atlantic herring (Clupea harengus). *Molecular Ecology*, *21*(15), 3686–3703. doi: 10.1111/j.1365-294X.2012.05639.x

Lischer, Heidi E L, & Shimizu, K. K. (2017). Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC Bioinformatics*, *18*(1), 474. doi: 10.1186/s12859-017-1911-6

Lischer, H E L, & Excoffier, L. (2012). PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, *28*(2), 298–299. doi: 10.1093/bioinformatics/btr642

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv*. doi: 10.48550/arxiv.1303.3997

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. doi: 10.1093/bioinformatics/bty191

Lohse, K., Taylor-Cox, E., Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, & Darwin Tree of Life Consortium. (2021). The genome sequence of the speckled wood butterfly, Pararge aegeria (Linnaeus, 1758). *Wellcome Open Research*, *6*, 287. doi: 10.12688/wellcomeopenres.17278.1

López, M. E., Neira, R., & Yáñez, J. M. (2014). Applications in the search for genomic selection signatures in fish. *Frontiers in Genetics*, *5*, 458. doi: 10.3389/fgene.2014.00458

Lotterhos, K. E., & Whitlock, M. C. (2015). The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*, *24*(5), 1031–1046. doi: 10.1111/mec.13100

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A. (2017). Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Molecular Ecology Resources*, *17*(2), 142–152. doi: 10.1111/1755-0998.12635

Luckhurst, B. E., Trott, T., & Manuel, S. (2001). *Landings, Seasonality, Catch per Unit Effort, and Tag-Recapture Results of Yellowfin Tuna and Blackfin Tuna at Bermuda.* 225–234. American Fisheries Society Symposium 25.

Luckhurst, B. E. (2014). Elements of the ecology and movement patterns of highly migratory fish species of interest to ICCAT in the Sargasso Sea. *Collect. Vol. Sci. Pap. ICCAT*, *70*(5), 2183–2206.

Luiz, O. J., Madin, J. S., Robertson, D. R., Rocha, L. A., Wirtz, P., & Floeter, S. R. (2012). Ecological traits influencing range expansion across large oceanic dispersal barriers: insights from tropical Atlantic reef fishes. *Proceedings. Biological Sciences / the Royal Society*, *279*(1730), 1033–1040. doi: 10.1098/rspb.2011.1525

Maghan, B. W., & Rivas, L. R. (1971). The blackfin tuna (Thunnus atlanticus) as an underutilized fishery resource in the tropical western Atlantic Ocean. *FAO Fish Rep, 71*(2), 163–172.

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*(22), 2867–2873. doi: 10.1093/bioinformatics/btq559

Ma, Z. (Sam), Li, L., Ye, C., Peng, M., & Zhang, Y.-P. (2019). Hybrid assembly of ultra-long Nanopore reads augmented with 10x-Genomics contigs: Demonstrated with a human genome. *Genomics*, *111*(6), 1896–1901. doi: 10.1016/j.ygeno.2018.12.013

Malmstrøm, M., Matschiner, M., Tørresen, O. K., Star, B., Snipen, L. G., Hansen, T. F., … Jentoft, S. (2016). Evolution of the immune system influences speciation rates in teleost fishes. *Nature Genetics*, *48*(10), 1204–1210. doi: 10.1038/ng.3645

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology*, *14*(1), e1005944. doi: 10.1371/journal.pcbi.1005944

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*(6), 764–770. doi: 10.1093/bioinformatics/btr011

Mariani, P., Křivan, V., MacKenzie, B. R., & Mullon, C. (2016). The migration game in habitat network: the case of tuna. *Theoretical Ecology*, *9*(2), 219–232. doi: 10.1007/s12080-015-0290-8

Mathieu, H., Pau, C., Reynal, L., & Theophille, D. (2013). Chapter 2.1.10.7 Thon a Nageoires Noires. In *ICCAT Publications*: *Vol. 65*. *ICCAT Manual*. International Commission for the Conservation of Atlantic Tuna.

McWilliam, S., Grewe, P. M., Bunch, R. J., & Barendse, W. (2016). A draft genome assembly of southern bluefin tuna Thunnus maccoyii. *ArXiv*. doi: 10.48550/arxiv.1607.03955

Metcalfe, J. D., & Arnold, G. P. (1997). Tracking fish with electronic tags. *Nature*, *387*(6634), 665–666. doi: 10.1038/42622

Michel, A. P., Sim, S., Powell, T. H. Q., Taylor, M. S., Nosil, P., & Feder, J. L. (2010). Widespread genomic divergence during sympatric speciation. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(21), 9724–9729. doi: 10.1073/pnas.1000939107

Miclotte, G., Heydari, M., Demeester, P., Rombauts, S., Van de Peer, Y., Audenaert, P., & Fostier, J. (2016). Jabba: hybrid error correction for long sequencing reads. *Algorithms for Molecular Biology*, *11*, 10. doi: 10.1186/s13015-016-0075-7

Miglietta, M. P., Faucci, A., & Santini, F. (2011). Speciation in the sea: overview of the symposium and discussion of future directions. *Integrative and Comparative Biology*, *51*(3), 449–455. doi: 10.1093/icb/icr024

Miller, J. M., Cullingham, C. I., & Peery, R. M. (2020). The influence of a priori grouping on inference of genetic clusters: simulation study and literature review of the DAPC method. *Heredity*, *125*(5), 269–280. doi: 10.1038/s41437-020-0348-2

Mills, L. S., & Allendorf, F. W. (1996). The One-Migrant-per-Generation Rule in Conservation and Management. *Conservation Biology*, *10*(6), 1509–1518. doi: 10.1046/j.1523-1739.1996.10061509.x

Miura, R. M. (1986). *Some Mathematical Questions in Biology: DNA Sequence Analysis* (p. 124). Providence, R.I: Amer Mathematical Society.

Montes, I., Iriondo, M., Manzano, C., Arrizabalaga, H., Jiménez, E., Pardo, M. Á., … Estonba, A. (2012). Worldwide genetic structure of albacore Thunnus alalunga revealed by microsatellite DNA markers. *Marine Ecology Progress Series*, *471*, 183–191. doi: 10.3354/meps09991

Morton, N. E. (1955). Sequential tests for the detection of linkage. *American Journal of Human Genetics*, *7*(3), 277–318.

Nielsen, E. E., Hemmer-Hansen, J., Poulsen, N. A., Loeschcke, V., Moen, T., Johansen, T., … Carvalho, G. R. (2009). Genomic signatures of local directional selection in a high gene flow marine organism; the Atlantic cod (Gadus morhua). *BMC Evolutionary Biology*, *9*, 276. doi: 10.1186/1471-2148-9-276

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

Norrell, A. E., Jones, K. L., & Saillant, E. A. (2020). Development and characterization of genomic resources for a non-model marine teleost, the red snapper (Lutjanus campechanus, Lutjanidae): Construction of a high-density linkage map, anchoring of genome contigs and comparative genomic analysis. *Plos One*, *15*(4), e0232402. doi: 10.1371/journal.pone.0232402

Nosil, P., Funk, D. J., & Ortiz-Barrientos, D. (2009). Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, *18*(3), 375–402. doi: 10.1111/j.1365-294X.2008.03946.x

O’Dea, A., Lessios, H. A., Coates, A. G., Eytan, R. I., Restrepo-Moreno, S. A., Cione, A. L., … Jackson, J. B. C. (2016). Formation of the isthmus of panama. *Science Advances*, *2*(8), e1600883. doi: 10.1126/sciadv.1600883

O’Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M., & Portnoy, D. S. (2018). These aren’t the loci you’e looking for: Principles of effective SNP filtering for molecular ecologists. *Molecular Ecology*, *27*(16), 3193–3206. doi: 10.1111/mec.14792

Ortiz, E. M. (2019). vcf2phylip. *Zenodo*. Retrieved from https://doi.org/10.5281/zenodo.2540861

Pacicco, A. E., Allman, R. J., Lang, E. T., Murie, D. J., Falterman, B. J., Ahrens, R., & Walter, J. F. (2021). Age and growth of yellowfin tuna in the U.S. gulf of mexico and western atlantic. *Marine and Coastal Fisheries : Dynamics, Management , and Ecosystem Science*, *13*(4), 345–361. doi: 10.1002/mcf2.10158

Palumbi, S. R. (1994). Genetic divergence, reproductive isolation, and marine speciation. *Annual Review of Ecology and Systematics*, *25*(1), 547–572. doi: 10.1146/annurev.es.25.110194.002555

Peakall, R., Ruibal, M., & Lindenmayer, D. B. (2003). Spatial autocorrelation analysis offers new insights into gene flow in the Australian bush rat, Rattus fuscipes. *Evolution*, *57*(5), 1182–1195. doi: 10.1111/j.0014-3820.2003.tb00327.x

Peakall, R., & Smouse, P. E. (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. *Bioinformatics*, *28*(19), 2537–2539. doi: 10.1093/bioinformatics/bts460

Pecoraro, C., Babbucci, M., Franch, R., Rico, C., Papetti, C., Chassot, E., … Tinti, F. (2018). The population genomics of yellowfin tuna (Thunnus albacares) at global geographic scale challenges current stock delineation. *Scientific Reports*, *8*(1), 13890. doi: 10.1038/s41598-018-32331-3

Pecoraro, C. (2016). Global Population Genomic Structure and Life History Trait Analysis of Yellowfin Tuna (Thunnus Albacares). *Doctoral Dissertation*, 200.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *Plos One*, *7*(5), e37135. doi: 10.1371/journal.pone.0037135

Pew Charitable Trusts. (2020, October). Netting Billions 2020: A Global Tuna Valuation. Retrieved November 3, 2021, from https://www.pewtrusts.org/en/research-and-analysis/reports/2020/10/netting-billions-2020-a-global-tuna-valuation

Pollock, K. H. (1991). Modeling Capture, Recapture, and Removal Statistics for Estimation of Demographic Parameters for Fish and Wildlife Populations: Past, Present, and Future. *Journal of the American Statistical Association*, *86*(413), 225. doi: 10.2307/2289733

Portnoy, D. S., Puritz, J. B., Hollenbeck, C. M., Gelsleichter, J., Chapman, D., & Gold, J. R. (2015). Selection and sex-biased dispersal in a coastal shark: the influence of philopatry on adaptive variation. *Molecular Ecology*, *24*(23), 5877–5885. doi: 10.1111/mec.13441

Pritchard, J. K., & Rosenberg, N. A. (1999). Use of unlinked genetic markers to detect population stratification in association studies. *American Journal of Human Genetics*, *65*(1), 220–228. doi: 10.1086/302449

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959. doi: 10.1093/genetics/155.2.945

Pruett, C. L., Saillant, E., & Gold, J. R. (2005). Historical population demography of red snapper (Lutjanus campechanus) from the northern Gulf of Mexico based on analysis of sequences of mitochondrial DNA. *Marine Biology*, *147*(3), 593–602. doi: 10.1007/s00227-005-1615-8

Pryszcz, L. P., & Gabaldón, T. (2016). Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Research*, *44*(12), e113. doi: 10.1093/nar/gkw294

Pujolar, J. M., Jacobsen, M. W., Als, T. D., Frydenberg, J., Munch, K., Jónsson, B., … Hansen, M. M. (2014). Genome-wide single-generation signatures of local selection in the panmictic European eel. *Molecular Ecology*, *23*(10), 2514–2528. doi: 10.1111/mec.12753

Puncher, G. N., Cariani, A., Maes, G. E., Van Houdt, J., Herten, K., Cannas, R., … Tinti, F. (2018). Spatial dynamics and mixing of bluefin tuna in the Atlantic Ocean and Mediterranean Sea revealed using next-generation sequencing. *Molecular Ecology Resources*, *18*(3), 620–638. doi: 10.1111/1755-0998.12764

Puritz, J. B., Hollenbeck, C. M., & Gold, J. R. (2014). dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ*, *2*, e431. doi: 10.7717/peerj.431

Puritz, J. B., Matz, M. V., Toonen, R. J., Weber, J. N., Bolnick, D. I., & Bird, C. E. (2014). Demystifying the RAD fad. *Molecular Ecology*, *23*(24), 5937–5942. doi: 10.1111/mec.12965

Qin, M., Wu, S., Li, A., Zhao, F., Feng, H., Ding, L., & Ruan, J. (2019). LRScaf: improving draft genomes using long noisy reads. *BMC Genomics*, *20*(1), 955. doi: 10.1186/s12864-019-6337-2

Qiu, F., Kitchen, A., Beerli, P., & Miyamoto, M. M. (2013). A possible explanation for the population size discrepancy in tuna (genus Thunnus) estimated from mitochondrial DNA and microsatellite data. *Molecular Phylogenetics and Evolution*, *66*(2), 463–468. doi: 10.1016/j.ympev.2012.05.002

Qiu, F., & Miyamoto, M. M. (2011). Use of Nuclear DNA Data to Estimate Genetic Diversity and Population Size in Pacific Bluefin and Yellowfin Tuna (Thunnus orientalis and T. albacares). *Copeia*, *2011*(2), 264–269. doi: 10.1643/CI-10-112

Quilodrán, C. S., Ruegg, K., Sendell-Price, A. T., Anderson, E. C., Coulson, T., & Clegg, S. M. (2020). The multiple population genetic and demographic routes to islands of genomic divergence. *Methods in Ecology and Evolution*, *11*(1), 6–21. doi: 10.1111/2041-210X.13324

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

Raj, A., Stephens, M., & Pritchard, J. K. (2014). fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, *197*(2), 573–589. doi: 10.1534/genetics.114.164350

Rastas, P., Paulin, L., Hanski, I., Lehtonen, R., & Auvinen, P. (2013). Lep-MAP: fast and accurate linkage map construction for large SNP datasets. *Bioinformatics*, *29*(24), 3128–3134. doi: 10.1093/bioinformatics/btt563

Rastas, P. (2017). Lep-MAP3: robust linkage mapping even for low-coverage whole genome sequencing data. *Bioinformatics*, *33*(23), 3726–3732. doi: 10.1093/bioinformatics/btx494

Rastas, P. (2020). Lep-Anchor: automated construction of linkage map anchored haploid genomes. *Bioinformatics*, *36*(8), 2359–2364. doi: 10.1093/bioinformatics/btz978

Reeb, C. A., & Avise, J. C. (1990). A genetic discontinuity in a continuously distributed species: mitochondrial DNA in the American oyster, Crassostrea virginica. *Genetics*, *124*(2), 397–406. doi: 10.1093/genetics/124.2.397

Reglero, P., Santos, M., Balbín, R., Laíz-Carrión, R., Alvarez-Berastegui, D., Ciannelli, L., … Alemany, F. (2017). Environmental and biological characteristics of Atlantic bluefin tuna and albacore spawning habitats based on their egg distributions. *Deep Sea*

Research Part II: Topical Studies in Oceanography, *140*, 105–116. doi: 10.1016/j.dsr2.2017.03.013

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

Riccioni, G., Landi, M., Ferrara, G., Milano, I., Cariani, A., Zane, L., … Tinti, F. (2010). Spatio-temporal population structuring and genetic diversity retention in depleted Atlantic bluefin tuna of the Mediterranean Sea. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(5), 2102–2107. doi: 10.1073/pnas.0908281107

Richardson, D. E., Llopiz, J. K., Guigand, C. M., & Cowen, R. K. (2010). Larval assemblages of large and medium-sized pelagic species in the Straits of Florida. *Progress in Oceanography*, *86*(1–2), 8–20. doi: 10.1016/j.pocean.2010.04.005

Roach, M. J., Schmidt, S. A., & Borneman, A. R. (2018). Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, *19*(1), 460. doi: 10.1186/s12859-018-2485-7

Roberts, C. M. (1997). Connectivity and management of caribbean coral reefs. *Science*, *278*(5342), 1454–1457. doi: 10.1126/science.278.5342.1454

Robledo-Arnuncio, J. J., & Rousset, F. (2010). Isolation by distance in a continuous population under stochastic demographic fluctuations. *Journal of Evolutionary Biology*, *23*(1), 53–71. doi: 10.1111/j.1420-9101.2009.01860.x

Rodríguez, F., Oliver, J. L., Marín, A., & Medina, J. R. (1990). The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology*, *142*(4), 485–501. doi: 10.1016/S0022-5193(05)80104-3

Rousset, F. (2008). Genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources*, *8*(1), 103–106. doi: 10.1111/j.1471-8286.2007.01931.x

Rudershausen, P. J., Buckel, J. A., Edwards, J., Gannon, D. P., Butler, C. M., & Averett, T. W. (2010). Feeding Ecology of Blue Marlins, Dolphinfish, Yellowfin Tuna, and Wahoos from the North Atlantic Ocean and Comparisons with other Oceans. *Transactions of the American Fisheries Society*, *139*(5), 1335–1359. doi: 10.1577/T09-105.1

R Core Team. (2013). R: A language and environment for statistical computing. (Version 3.6.3) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Ruzzante, D. E., Mariani, S., Bekkevold, D., André, C., Mosegaard, H., Clausen, L. A. W., … Carvalho, G. R. (2006). Biocomplexity in a highly migratory pelagic marine fish, Atlantic herring. *Proceedings. Biological Sciences / the Royal Society*, *273*(1593), 1459–1464. doi: 10.1098/rspb.2005.3463

Sale, P. F., Hanski, I., & Kritzer, J. (2006). The merging of metapopultion theory and marine ecology: establishing the historical context. *Marine Metapopulations*, (1967), 3–28.

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., … Yorke, J. A. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research*, *22*(3), 557–567. doi: 10.1101/gr.131383.111

Sang, T. K., Chang, H. Y., Chen, C. T., & Hui, C. F. (1994). Population structure of the Japanese eel, Anguilla japonica. *Molecular Biology and Evolution*, *11*(2), 250–260. doi: 10.1093/oxfordjournals.molbev.a040107

Saxton, B. L. (2009). Historical demography and genetic population structure of the Blackfin tuna (Thunnus atlanticus) from the Northwest Atlantic Ocean and the Gulf of Mexico. *Master Thesis*, 101.

Schaefer, K. M., Fuller, D. W., & Block, B. A. (2007). Movements, behavior, and habitat utilization of yellowfin tuna (Thunnus albacares) in the northeastern Pacific Ocean, ascertained through archival tag data. *Marine Biology*, *152*(3), 503–525. doi: 10.1007/s00227-007-0689-x

Schaefer, K. M., Fuller, D. W., & Block, B. A. (2011). Movements, behavior, and habitat utilization of yellowfin tuna (Thunnus albacares) in the Pacific Ocean off Baja California, Mexico, determined from archival tag data analyses, including unscented

Kalman filtering. *Fisheries Research*, *112*(1–2), 22–37. doi: 10.1016/j.fishres.2011.08.006

Schatz, M. C., Delcher, A. L., & Salzberg, S. L. (2010). Assembly of large genomes using second-generation sequencing. *Genome Research*, *20*(9), 1165–1173. doi: 10.1101/gr.101360.109

Schluter, D., & Conte, G. L. (2009). Genetics and ecological speciation. *Proceedings of the National Academy of Sciences of the United States of America*, *106 Suppl 1*, 9955–9962. doi: 10.1073/pnas.0901264106

Schmid, M., Frei, D., Patrignani, A., Schlapbach, R., Frey, J. E., Remus-Emsermann, M. N. P., & Ahrens, C. H. (2018). Pushing the limits of de novo genome assembly for complex prokaryotic genomes harboring very long, near identical repeats. *Nucleic Acids Research*, *46*(17), 8953–8965. doi: 10.1093/nar/gky726

Schweyen, H., Rozenberg, A., & Leese, F. (2014). Detection and removal of PCR duplicates in population genomic ddRAD studies by addition of a degenerate base region (DBR) in sequencing adapters. *The Biological Bulletin*, *227*(2), 146–160. doi: 10.1086/BBLv227n2p146

Secor, D. H., Henderson-Arzapalo, A., & Piccoli, P. M. (1995). Can otolith microchemistry chart patterns of migration and habitat utilization in anadromous fishes? *Journal of Experimental Marine Biology and Ecology*, *192*(1), 15–33. doi: 10.1016/0022-0981(95)00054-U

Selwyn, J. D., Hogan, J. D., Downey-Wall, A. M., Gurski, L. M., Portnoy, D. S., & Heath, D. D. (2016). Kin-Aggregations Explain Chaotic Genetic Patchiness, a Commonly Observed Genetic Pattern, in a Marine Fish. *Plos One*, *11*(4), e0153381. doi: 10.1371/journal.pone.0153381

Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., & Wolf, J. B. W. (2017). Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods in Ecology and Evolution*, *8*(8), 907–917. doi: 10.1111/2041-210X.12700

Shaklee, J. B., & Bentzen, P. (1998). Genetic Identification of Stocks of Marine Fish and Shellfish. *Bulletin of Marine Science*, *62*(2), 589–621.

Shakya, M., Ahmed, S. A., Davenport, K. W., Flynn, M. C., Lo, C.-C., & Chain, P. S. G. (2020). Standardized phylogenetic and molecular evolutionary analysis applied to species across the microbial tree of life. *Scientific Reports*, *10*(1), 1723. doi: 10.1038/s41598-020-58356-1

Shao, H., Zhou, C., Cao, M. D., & Coin, L. J. M. (2018). Evolutionary analysis of chromosome end extension. *PeerJ Preprints*, *6*(e26624v1), 12.

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *Plos One*, *11*(10), e0163962. doi: 10.1371/journal.pone.0163962

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. doi: 10.1093/bioinformatics/btv351

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., & Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Research*, *19*(6), 1117–1123. doi: 10.1101/gr.089532.108

Singh-Renton, S., & Renton, J. (2007). Cframp's large pelagic fish tagging program. *Gulf and Caribbean Research*, *19*(2), 99–102. doi: 10.18785/gcr.1902.12

Slatkin, M. (2008). Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nature Reviews. Genetics*, *9*(6), 477–485. doi: 10.1038/nrg2361

Smith, J. J., Timoshevskaya, N., Ye, C., Holt, C., Keinath, M. C., Parker, H. J., … Amemiya, C. T. (2018). The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution. *Nature Genetics*, *50*(2), 270–277. doi: 10.1038/s41588-017-0036-1

Smith, P. J., Francis, R. I. C. C., & McVeagh, M. (1991). Loss of genetic diversity due to fishing pressure. *Fisheries Research*, *10*(3–4), 309–316. doi: 10.1016/0165-7836(91)90082-Q

Smouse, Peter E., Long, J. C., & Sokal, R. R. (1986). Multiple regression and correlation extensions of the mantel test of matrix correspondence. *Systematic Zoology*, *35*(4), 627. doi: 10.2307/2413122

Smouse, P E, & Peakall, R. (1999). Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity*, *82 (Pt 5)*, 561–573. doi: 10.1038/sj.hdy.6885180

Stange, M., Sánchez-Villagra, M. R., Salzburger, W., & Matschiner, M. (2018). Bayesian Divergence-Time Estimation with Genome-Wide Single-Nucleotide Polymorphism Data of Sea Catfishes (Ariidae) Supports Miocene Closure of the Panamanian Isthmus. *Systematic Biology*, *67*(4), 681–699. doi: 10.1093/sysbio/syy006

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research*, *34*(Web Server issue), W435-9. doi: 10.1093/nar/gkl200

Sturtevant, A. H. (1915). The behavior of the chromosomes as studied through linkage. *Zeitschrift Für Induktive Abstammungs- Und Vererbungslehre*, *13*(1), 234–287. doi: 10.1007/BF01792906

Suda, A., Nishiki, I., Iwasaki, Y., Matsuura, A., Akita, T., Suzuki, N., & Fujiwara, A. (2019). Improvement of the Pacific bluefin tuna (Thunnus orientalis) reference genome and development of male-specific DNA markers. *Scientific Reports*, *9*(1), 14450. doi: 10.1038/s41598-019-50978-4

Sved, J. A., Cameron, E. C., & Gilchrist, A. S. (2013). Estimating effective population size from linkage disequilibrium between unlinked loci: theory and application to fruit fly outbreak populations. *Plos One*, *8*(7), e69078. doi: 10.1371/journal.pone.0069078

Talley-Farnham, T. C., Stéquert, B., & Alvarado-Bremer, J. R. (2004). Preliminary analysis of the comparison in levels of variation between juvenile and adult yellowfin tuna samples from the Atlantic Ocean using both mtDNA and microsatellite data. *Col. Vol. Sci. Pap., ICCAT*, *56*(2), 694–703.

Tang, H., Zhang, X., Miao, C., Zhang, J., Ming, R., Schnable, J. C., … Lu, J. (2015). ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biology*, *16*, 3. doi: 10.1186/s13059-014-0573-1

Tan, M. H., Austin, C. M., Hammer, M. P., Lee, Y. P., Croft, L. J., & Gan, H. M. (2018). Finding Nemo: hybrid assembly with Oxford Nanopore and Illumina reads greatly improves the clownfish (Amphiprion ocellaris) genome assembly. *GigaScience*, *7*(3), 1–6. doi: 10.1093/gigascience/gix137

Teo, S. L. H., Boustany, A. M., & Block, B. A. (2007). Oceanographic preferences of Atlantic bluefin tuna, Thunnus thynnus, on their Gulf of Mexico breeding grounds. *Marine Biology*, *152*(5), 1105–1119. doi: 10.1007/s00227-007-0758-1

Tessier, N., & Bernatchez, L. (1999). Stability of population structure and genetic diversity across generations assessed by microsatellites among sympatric populations of landlocked Atlantic salmon (*Salmo salar* L.). *Molecular Ecology*, *8*(2), 169–179. doi: 10.1046/j.1365-294X.1999.00547.x

Therkildsen, N. O., & Palumbi, S. R. (2017). Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. *Molecular Ecology Resources*, *17*(2), 194–208. doi: 10.1111/1755-0998.12593

Thorrold, S. R., Latkoczy, C., Swart, P. K., & Jones, C. M. (2001). Natal homing in a marine fish metapopulation. *Science*, *291*(5502), 297–299. doi: 10.1126/science.291.5502.297

TinHan, T. C., Mohan, J. A., Dumesnil, M., DeAngelis, B. M., & Wells, R. J. D. (2018). Linking Habitat Use and Trophic Ecology of Spotted Seatrout (Cynoscion nebulosus) on a Restored Oyster Reef in a Subtropical Estuary. *Estuaries and Coasts*, *41*(6), 1–13. doi: 10.1007/s12237-018-0391-x

Tørresen, O. K., Star, B., Mier, P., Andrade-Navarro, M. A., Bateman, A., Jarnot, P., … Linke, D. (2019). Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Research*, *47*(21), 10994–11006. doi: 10.1093/nar/gkz841

van der Valk, T., Vezzi, F., Ormestad, M., Dalen, L., & Guschanski, K. (2017). Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. *BioRxiv*. doi: 10.1101/179028

Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, *27*(5), 737–746. doi: 10.1101/gr.214270.116

Vaux, F., Bohn, S., Hyde, J. R., & O'Malley, K. G. (2021). Adaptive markers distinguish North and South Pacific Albacore amid low population differentiation. *Evolutionary Applications*, *14*(5), 1343–1364. doi: 10.1111/eva.13202

Via, S. (2012). Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1587), 451–460. doi: 10.1098/rstb.2011.0260

Viñas, J., & Tudela, S. (2009). A validated methodology for genetic identification of tuna species (genus Thunnus). *Plos One*, *4*(10), e7606. doi: 10.1371/journal.pone.0007606

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., … Earl, A. M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *Plos One*, *9*(11), e112963. doi: 10.1371/journal.pone.0112963

Wang, J. (2007). Triadic IBD coefficients and applications to estimating pairwise relatedness. *Genetical Research*, *89*(3), 135–153. doi: 10.1017/S0016672307008798

Waples, Robin S, & Do, C. (2010). Linkage disequilibrium estimates of contemporary N e using highly variable genetic markers: a largely untapped resource for applied conservation and evolution. *Evolutionary Applications*, *3*(3), 244–262. doi: 10.1111/j.1752-4571.2009.00104.x

Waples, Robin S, Grewe, P. M., Bravington, M. W., Hillary, R., & Feutry, P. (2018). Robust estimates of a high Ne/N ratio in a top marine predator, southern bluefin tuna. *Science Advances*, *4*(7), eaar7759. doi: 10.1126/sciadv.aar7759

Waples, Robin S. (2006). A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. *Conservation Genetics (Print)*, *7*(2), 167–184. doi: 10.1007/s10592-005-9100-y

Waples, R S. (1998). Separating the wheat from the chaff: patterns of genetic differentiation in high gene flow species. *Journal of Heredity*, *89*(5), 438–450. doi: 10.1093/jhered/89.5.438

Ward, R. O., Elliott, N. G., & Innes, B. H. (1997). Global population structure of yellowfin tuna, Thunnus albacares, inferred from allozyme and mitochondrial DNA variation. *Oceanographic Literature Review*.

Weir, B. S., & Goudet, J. (2017). A unified characterization of population structure and relatedness. *Genetics*, *206*(4), 2085–2103. doi: 10.1534/genetics.116.198424

Wei, T., & Simko, V. (2021). R package “corrplot”: Visualization of a Correlation Matrix (Version 0.92) [Computer software]. GitHub. Retrieved from https://github.com/taiyun/corrplot

Wells, R. J. D., Rooker, J. R., & Itano, D. G. (2012). Nursery origin of yellowfin tuna in the Hawaiian Islands. *Marine Ecology Progress Series*, *461*, 187–196. doi: 10.3354/meps09833

Weng, K. C., Stokesbury, M. J. W., Boustany, A. M., Seitz, A. C., Teo, S. L. H., Miller, S. K., & Block, B. A. (2009). Habitat and behaviour of yellowfin tuna Thunnus albacares in the Gulf of Mexico determined using pop-up satellite archival tags. *Journal of Fish Biology*, *74*(7), 1434–1449. doi: 10.1111/j.1095-8649.2009.02209.x

Whitlock, M. C., & Lotterhos, K. E. (2015). Reliable Detection of Loci Responsible for Local Adaptation: Inference of a Null Model through Trimming the Distribution of F(ST). *The American Naturalist*, *186 Suppl 1*, S24-36. doi: 10.1086/682949

Wiley, G., & Miller, M. J. (2020). A highly contiguous genome for the Golden-fronted Woodpecker (*Melanerpes aurifrons*) via a hybrid Oxford Nanopore and short read assembly. *BioRxiv*. doi: 10.1101/2020.01.03.894444

Wortley, A. H., Rudall, P. J., Harris, D. J., & Scotland, R. W. (2005). How much data are needed to resolve a difficult phylogeny?: case study in Lamiales. *Systematic Biology*, *54*(5), 697–709. doi: 10.1080/10635150500221028

Ye, C., Hill, C. M., Wu, S., Ruan, J., & Ma, Z. S. (2016). DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Scientific Reports*, *6*, 31900. doi: 10.1038/srep31900

Ye, C., Ma, Z. S., Cannon, C. H., Pop, M., & Yu, D. W. (2012). Exploiting sparseness in de novo genome assembly. *BMC Bioinformatics*, *13*(S6), 8. doi: 10.1186/1471-2105-13-S6-S1

Zagaglia, C. R., Lorenzzetti, J. A., & Stech, J. L. (2004). Remote sensing data and longline catches of yellowfin tuna (Thunnus albacares) in the equatorial Atlantic. *Remote Sensing of Environment*, *93*(1–2), 267–281. doi: 10.1016/j.rse.2004.07.015

Zhang, Y., Cheng, C., Li, J., Yang, S., Wang, Y., Li, Z., … Lou, Q. (2015). Chromosomal structures and repetitive sequences divergence in Cucumis species revealed by comparative cytogenetic mapping. *BMC Genomics*, *16*(1), 730. doi: 10.1186/s12864-015-1877-6