DNA BARCODING AOTEAROA NEW ZEALAND'S RAY-FINNED FISHES (ACTINOPTERYGII): A REFERENCE DATABASE AND USE CASE

This thesis is completed in part of a Masters of Biological Science Degree.

Maddie MacLean Stones | Masters of Biological Science | January 24th, 2022

ACKNOWLEDGEMENTS

Firstly, I would like to acknowledge my supervisor, Libby Liggins. Thank you for all your help and guidance throughout this research, for all the time you have spent towards analysing and reading my work, for your advice and for keeping me on track. I am so thankful for this awesome project I have been able to be a part of and how much I have learnt and grown over the past two years. None of it would have been possible without you!

Secondly, I would like to acknowledge my amazing support system. My wonderful Mum, Sarah, and Ian for helping me through my thesis and always showing the ultimate support, belief and interest in what I do. Thank you for allowing me to talk endlessly about biology and for always checking in and taking the time to listen, understand and help me in every aspect, I am so lucky to have you both.

I would like to also acknowledge the awesome people around me who have kept me going along the way: Elise, Aaron, Aorthi, Luc, Lizzy, Millie, Mitchell, James, Emily, Tom and Karon for the endless laughter, loyalty, support and odd drink that I have needed over the past couple of years. Everything I do is made up from small pieces of each of you so thank you for being a part of this journey.

I would also like to acknowledge and thank Taylor, Flash and the countless canine companions who have kept me moving and grounded. You always turn the chaotic and stressful times throughout completing this thesis into so much fun, thank you for always having my back.

Finally, I would like to acknowledge all the others that have helped me and my work along the way, Iggy Carvajal, Irene Middleton, Vanessa Arranz and David Aguirre, thank you for all your help both in and out of the lab, answering all my questions and showing me the way to ultimately end up where I am now. I would have been completely lost without all your insight and expertise!

This thesis is dedicated to Ruth Maclean, Rosa MacLean and Oliver Clark

Forever present in what I do, even though you are no longer by my side.

GENERAL ABSTRACT

DNA barcode reference databases have been created for the fish biodiversity of many nations, providing a resource to facilitate rapid species identification, biodiversity assessment, and ultimately greater awareness and understanding of freshwater and marine fish fauna. Aotearoa New Zealand (NZ) has a wide diversity of marine, estuarine, and freshwater habitats that comprise a diverse fish fauna, and a high proportion of endemic fish species. Even so, a DNA barcode reference database for the fishes of NZ has not yet been created. In this thesis, I curated a DNA barcode reference database for NZ fishes based on the Cytochrome Oxidase I (COI) gene region using previously published sequences from open-access repositories (i.e., Nucleotide Sequence Database Collaboration, and the Barcode of Life Data System) and novel sequences generated for species not previously sequenced (Chapter 2). To demonstrate the utility of this database, I then provide a use case to genetically identify larval fishes collected off the Northeast Coast of the North Island of NZ and compare these identities to those based on morphology (Chapter 3). To ensure representativeness and integrity of the sequence data within the NZ Fish Barcode Database, I preferentially generated sequences from fish specimens that had been identified by an expert taxonomist or held in museum collections. Furthermore, for widespread species that I did not have sequences for, I sought sequences from specimens collected as geographically close to NZ as possible. In Chapter 2, I was able to generate and retrieve sequences for 965 of the 1320 fish species recorded in NZ (73%); I provide a summary of our progress toward generating a DNA barcode reference database for NZ's fish biodiversity and report interspecies genetic divergences (based on Kimura-2-pairwise distance) between species. In Chapter 3, the database use case, I found that larvae were often able to be identified to their taxonomic family based on morphological features, but in some cases their taxonomic affinities were unknown. DNA barcoding enabled us to identify the species identity of the larval fishes with reference to the NZ Fish Barcode Database. Overall, our generated DNA barcode reference database is a practical resource of value for future environmental DNA studies, biodiversity monitoring, and managing fisheries and commercial fisheries derived products.

TABLE OF CONTENTS:

CHAPTER ONE: GENERAL INTRODUCTION

1.1 DEFINITIONS OF BIODIVERSITY

Biodiversity has been defined in a variety of ways since it was first used by conservationist Raymond F. Dasmann in 1968 (Dasmann, 1968). However, it is essential to have a clear definition of what biodiversity is and how we intend to measure it in order to initiate effective biodiversity management and measure success. One of the earliest publications using the term 'Biodiversity' since Dasmann was Edward O. Wilson in *The Diversity of Life* (1992) describing biodiversity in a way to attract attention to the loss of species that is occurring at an accelerated rate, especially in regards to human activity (McLaurin, 2008). However, the idea of biodiversity cannot be limited to just species numbers, biodiversity and the functioning of biological systems depends on the kinds and combinations of organisms present (McLaurin, 2008). A comprehensive definition of biodiversity was given by DeLong in 1996 "Biodiversity is an attribute of an area and specifically refers to the variety within and among living organisms, assemblages of living organisms, biotic communities, and biotic processes, whether naturally occurring or modified by humans. Biodiversity can be measured in terms of genetic diversity and the identity and number of different types of species, assemblages of species, biotic communities, and biotic processes, and the amount (e.g., abundance, biomass, cover, rate) and structure of each. It can be observed and measured at any spatial scale ranging from microsites and habitat patches to the entire biosphere." Biodiversity is essential for human society through the provisioning of goods and services for human life, therefore, having important economic values (Gamefeldt, Hillerbrans, & Jonsson, 2008). New Zealand (NZ) has its own unique biodiversity which in turn must be measured and managed accordingly.

Although 'species' as a unit cannot encapture biodiversity entirely, it is still a common and important unit for measuring biodiversity. One of the most common ways to measure biodiversity is through species richness and species evenness, both of which are effective for assessing biodiversity in a range of ecosystems (Lakicevic & Srdevic, 2018). International and Domestic legislation relies upon knowing the extinction risk of each individual species as opposed to a larger ecosystem threat measurement, e.g. The IUCN Red list for endangered species (IUCN, 2020), CITES (CITES., 2019) and the New Zealand's Department of Conservation: 'New Zealand's Threat Classification System manual' (Townsend et al., 2008). The utility of species as a unit in biodiversity requires an accurate definition of what a species is and how we distinguish taxa from one another.

The definition of 'species' varies among concepts, all requiring differing rules and boundaries to describe a species (Aldhebiani, 2018). Some modern ways to describe species include: the Biological species concept (as stated by Mayr, 1942) "groups of actually or potentially interbreeding natural

populations which are reproductively isolated from other such groups"); the Morphological species concept (defined by Regan, 1926) as: "a species is a community or a number of related communities, whose distinctive morphological characters are sufficiently different to entitle them as a different species."); the Ecological species concept (a species is a lineage or a closely related set of lineages which occupies an adaptive area minimally different from that of any other lineage in its range and which evolves separately from all lineages outside its range, Colinvaux, 1986); the Evolutionary species concept (Wiley, 1981): "a single lineage of an ancestor- descendant populations of organisms which maintains its identity from other such lineages [in space and time] and which has its own evolutionary tendencies and historical fate"); the Phenetic species concept (Park & Allaby, 2017) defined as "a species is a set of organisms that look similar to each other and distinct from other sets"); the Pluralistic species concept which incorporates many different concepts attempting to recognise that the factors that form a species varies (Mishler & Donoghue, 1982) and the Phylogenetic species concept, simply defines species as a group of organisms that share the same ancestor (Aldhebiani, 2018). Different species concepts preferentially use genetics, morphology or both to define a species. It can be difficult, yet essential, to decide on what a species is in order to recognise and describe all species that make up biodiversity, preventing any integral information from being neglected.

1.2 DEFINING SPECIES

Prior to the development of genetic identification, the only way to identify and categorise species was morphologically. Morphological identification attempts to classify specimens by using physical traits such as meristic measurements and counts, body shape and pigmentation (Ko et al., 2013). The advantages of this form of identification are that DNA is not required for identification which may be limited in museum or fossil specimens with low quality or quantity of DNA (Hillis, 1987). Morphological identification also requires a lower cost in comparison to genetic identification (Hillis, 1987). Although morphological identification can be a useful and essential tool, there are limitations to its use.

Morphological methods of identification can be restricted due to an array of reasons. One limitation of morphological identification is that taxa often having very similar morphological traits, especially in egg and juvenile stages or in cryptic or rare species (Matarese, Spies, Busby, & Orr, 2011) making it difficult to accurately identify species from morphology alone. Larvae are also small and delicate and can be damaged upon collection, making them more difficult to distinguish (Pegg, Sinclair, Briskey & Aspden, 2006). Morphological identification has also proven to be inconsistent between different taxonomists with a very low percentage of species being identified correctly from morphology alone (Ko et al., 2013) especially in juvenile stages of development (Victor, Hanner, Shivji, Caldow &

Hyde, 2009). Where morphological identification is restricted, genetic identification can be successfully utilised.

Genetics within biology includes a range of different uses. Some of these include: molecular identification to classify species, including DNA Barcoding (DeSalle & Amato, 2004; Hebert, Cywinska, Ball, & deWaard, 2003), identification of individual species within bacterial communities (Nimnoi & Pongsilp, 2020), the creation of phylogenetic trees between species (Felenstein, 1988) and analysis of diet through scat samples (Rodríguez-Castro, Saranholi, Bataglia, Blanck, & Galetti, 2018) or stomach contents (Aguilar et al., 2016). Another use of genetics includes utilising eDNA from aquatic based environments (Rees et al., 2014) in order to measure biodiversity (Thomsen et al., 2012), identify species in an environment (Rees et al., 2014), identification of parasitic fish (Trujillo-Gonzalez, Edmunds, Becker, & Hutson, 2019), biosecurity of exotic fish aquatic diseases (Trujillo-González, Becker, Huerlimann, Saunders, & Hutson, 2019), and identification of invasive fish species in order to prevent their establishment and spread through live trade (Roy, Belliveau, Mandrak, & Gagné, 2017; Dejean et al., 2012). In addition to the use of genetics in species identification, genetics within conservation is important for understanding the patterns that are occurring within a population of a species (DeSalle & Amato, 2004). Molecular identification can help identify genetic relationships between both individuals and populations (Lawrence, 2008) and can play an important role in successfully managing a country's biodiversity and ecosystems.

There are a number of advantages when using genetic identification of species over morphological identification. Genetic identification advantages involve being able to accurately identify cryptic species and link early life stages to adult stages in situations where morphological identification is not possible (Trivedi, Aloufi, Ansari, & Ghosh, 2016). Genetic identification does not require a morphological expert to identify specimens and can be useful for identifying species of meat, eggs or carcasses on the market (Trivedi, Aloufi, et al., 2016). Although there are benefits to morphologicalfcomm identification, cryptic, larval and distorted specimens are most accurately identified using molecular identification.

1.3 DEFINING SPECIES USING DNA BARCODING

DNA Barcoding is a useful, modern tool for identifying and verifying the identity of species. DNA Barcoding involves using short genetic sequences, found in every diploid cell of a species, that can act as a 'barcode' (Herbert, Ball, DeWaard, & Cywinska, 2003). Ideally a single barcoding gene sequence can be used to distinguish most, if not all, species for use in barcoding and metabarcoding (Herbert et al. 2003). Barcoding has been used over a range of taxa including invertebrates, amphibians, reptiles, fishes, mammals, fungi and plants (Weigt, Driskell, Baldwin, & Ormos, 2012).

Barcoding can be used to identify species in order to further research within ecology, evolution and conservation (Kress, Garcia-Robledo, Uriarte, & Erickson, 2015). For example, metabarcoding involves barcoding a collection of individuals sampled from the environment, using universal primers, in order to identify all species within a sample (Duke & Burton, 2020). This method can be used to quickly and efficiently measure community composition and diet, as well as detect invasive, foreign, cryptic, rare or threatened species (Duke & Burton, 2020). However, these studies rely on complete barcode sequence reference libraries in order to characterise the species diversity within a sample from the environment. In order to have a complete barcode reference library individual barcodes for each species need to be sequenced and made readily available.

There are a number of regions of the genome that can be used for species identification. Within mitochondrial DNA some of these regions include cytochrome oxidase I (COI) and cytochrome b (cyt b) or 12S, 16S and 18S regions within ribosomal DNA (Hebert et al., 2003). In order to distinguish species, a DNA sequence must be long enough to individually identify a species but short enough for quick and efficient use (Stoeckle Mark & Hebert 2008). Different gene regions in mitochondrial DNA have different evolutionary rates and therefore differing levels of genetic variation (Paine, McDowell, & Graves, 2007). Although variation is required in order to accrue interspecific difference, too much variation can cause issues with hindering primer design (Paine, et.al. 2007). Due to this, using a more conserved DNA region is beneficial when distinguishing between species (Paine, et.al. 2007).

One of the most common and effective barcoding regions for animals is the COI Barcode region. The COI region is a beneficial genetic barcode due to being a mitochondrial gene that has limited recombination, lack of introns and is maternally inherited (Saccone, Pesole, Gissi, & De Chirico, 1999). The array of diversity within a genetic region can be an issue when genetically identifying species so it is essential to select an ideal genetic barcoding region. The COI gene is one of the most conserved protein coding genes in mitochondria (Brown, 1985 as quoted by Paine et.al. 2007). This is due to the COI gene being an essential part of cellular energy production in the mitochondria (Rawson & Burton, 2002). Even minor disruption to the COI gene can cause severe effects for the organism, which suggests why it is so conserved in comparison to other mitochondrial regions (Rawson & Burton, 2002). Although the COI gene is able to distinguish most fishes at the species level, some freshwater fishes have been difficult to distinguish using the COI region (Hulley et al., 2018). Overall, the different factors combined has led to the COI region being the most reliable region for DNA barcoding animals, including fishes.

1.4 FISH BIODIVERSITY OF NEW ZEALAND

NZ has a unique fish biodiversity. NZ has been isolated from other land masses for 70-80 million years, and has developed a high degree of endemic fauna (Roberts, Stewart, & Struthers, 2015). NZ's large latitudinal range (Roberts et al., 2015) allows for waters varying from subtropical to sub-Antarctic temperatures, providing a diversity of habitats for fishes (O'Callaghan, Stevens, Roughan, Cornelisen, Sutton, Garrett, Giorli, Smith, Currie, Suanda, Williams, Bowen, Fernandez, Vennell, Knight, Barter, McComb. Oliver, Livingston, Tellier, Maissner, Brewer, Gall, Nodder, Decima, Souza, Forcen-Vazquez, Gardiner, Burke, Chiswell, Roberts, Hayden, Biggs & MacDonald, 2019). NZ's entire biodiversity of fishes is relatively poorly studied (Roberts, Stewart, & Struthers, 2015) and on-going research suggests that there are still several species undescribed (Matsuura & Middleton, 2017), and several new species still being recorded in NZ (Liggins, Sweatman, Trnski, Duffy, Eddy & Aguirre, 2020). A current total of 1270 fish species have been recorded and described by Roberts, Stewart and Struthers (2015) with a number of these species being recreationally and commercially fished.

NZ has one of the greatest marine Exclusive Economic Zones (EEZ) proportional to land size compared to other Pacific states (Gordon, Beaumont, MacDiarmid, Robertson, & Ahyong, 2010). NZ's large EEZ provides a vast marine reservoir (Gordon et al., 2010) which, for a country with a relatively small population, limits the total biodiversity exploration and provides a challenge for its management (Gordon et al., 2010). NZ has many circum-global species as well as a relatively high number of endemic species (Roberts, Stewart and Struthers, 2015) it is therefore important to accurately identify species in order to measure and manage NZ's marine biodiversity. The majority of marine organisms have morphological traits that vary drastically between life stages including variation between eggs, juveniles and adult life stages (Costello et al., 2015; Ko et al., 2013). New species continue to be identified in both well researched and new locations of NZ (Liggins et al., 2020) especially in ocean environments (Costello et al., 2010; Mora, Tittensor, Adl, Simpson, & Worm, 2011). There is a high proportion of cryptic species in marine environments (De Brauwer et al., 2018) that are difficult to identify and/or distinguish from other species leading to late discoveries and identifications. The ongoing influx of new species records requires accurate identification of species to continue an up-to-date reference of endemic and native organisms within a country's marine environment.

1.5 INTERNATIONAL CAMPAIGNS TO DNA BARCODE ALL FISHES

The Fish Barcode of Life Campaign (FISH-BOL) is an international collaboration that aimed to create a DNA Barcode within the COI region of all fish species (Ward, Hanner, & Hebert, 2009). This enables accurate species genetic identification of any fish larvae, eggs, fillet or part of fish by retrieving COI barcodes that have been uploaded to the Barcode of Life Data System (BOLD) (Ward et al., 2009). BOLD is an online open DNA data repository which facilitates the collection, storage, analysis and publication of DNA barcodes that reach the databases required standards (Ratnasingham & Hebert, 2007) There are several COI barcodes that have been researched for use on fishes. The COI-2 primer cocktail has been identified as the most successful primer cocktail for barcoding fishes in the COI region, followed by COI-3 primer cocktail (Ivanova, Zemlak, Hanner, & Hebert, 2007). A primer cocktail is a combination of primers with the same sequence 'tail' on each primer to allow high throughput sequencing (Ivanova, Zemlak, Hanner, & Hebert, 2007). Other DNA Barcode regions have been researched, including *HRVI,* but the COI Barcode region should preferentially be used in order to align with other countries database as well as the COI database for Fishes (FISH-BOL) that has been, and continues to be, built upon (Pegg, Sinclair, Briskey & Aspden, 2006).

The COI region is the most beneficial genetic barcode region for identifying fishes. Fishes are one of the easiest groups to generate DNA barcodes and the most reliable and modern way of identifying fish species (Ko et al., 2013). Having a DNA barcode reference library is essential for identifying species via barcoding. Due to FISH-BOL and other collaborators utilising online repositories to upload sequences, many fish species sequences are currently readily available. The usefulness of having a complete DNA Barcode reference library for fish species in a country can be seen by the large number of countries that have attempted to create their own barcode library.

Many countries have begun or completed DNA barcoding their entire marine fish biodiversity in the COI region. DNA Barcoding of all Australia's fish species using the COI gene was completed in 2005 (Ward, Zemlak, Innes, Last, & Hebert, 2005). Indian marine fish COI barcodes were completed in order to help identify cryptic species in 2011 (Lakra, Verma, Goswami, Lal, Mohindra, Punia, Goplakrishnan, Singh, Ward & Herbert, 2011). Many parts of China have begun barcoding (Wang, Wu, Liu, Liu, Zhao, Liu, & You, 2018; Xu, Van Damme, Li, Ji, Wang & Du, 2019; Zhang & Hanner, 2012; Zhang, Qin, Zhang, Wang, Lin, 2017; Zhang, 2011), as well as Portugal (Costa, Landi, Martins, Costa, Costa, Carneiro, Alves, Steinke, & Carvalho, 2012), the Canadian Pacific (Steinke et al. 2009), Japan (Zhang & Hanner, 2011), Argentina (Mabragana, de Astarloa, Hanner, Zhang, & Castro, 2011), Brazil (Ribeiro, Caires, Mariguela, Garcia Pereira, Hanner, & Oliveira, 2012), Bangladesh (Ahmed, Datta, Saha, & Hossain, 2021), Saudi Arabia (Rabaoui et al., 2019) and Israel (Shirak, Dor, Seroussi, Ron, Hulata, Golani, 2016). These DNA Barcodes have been useful in

identifying ambiguous species (Ward, Zemlak, Innes, Last, & Hebert, 2005) and partial or digested remains of species (Jeon, Choi, & Suk, 2012; Paine, McDowell, & Graves, 2007).

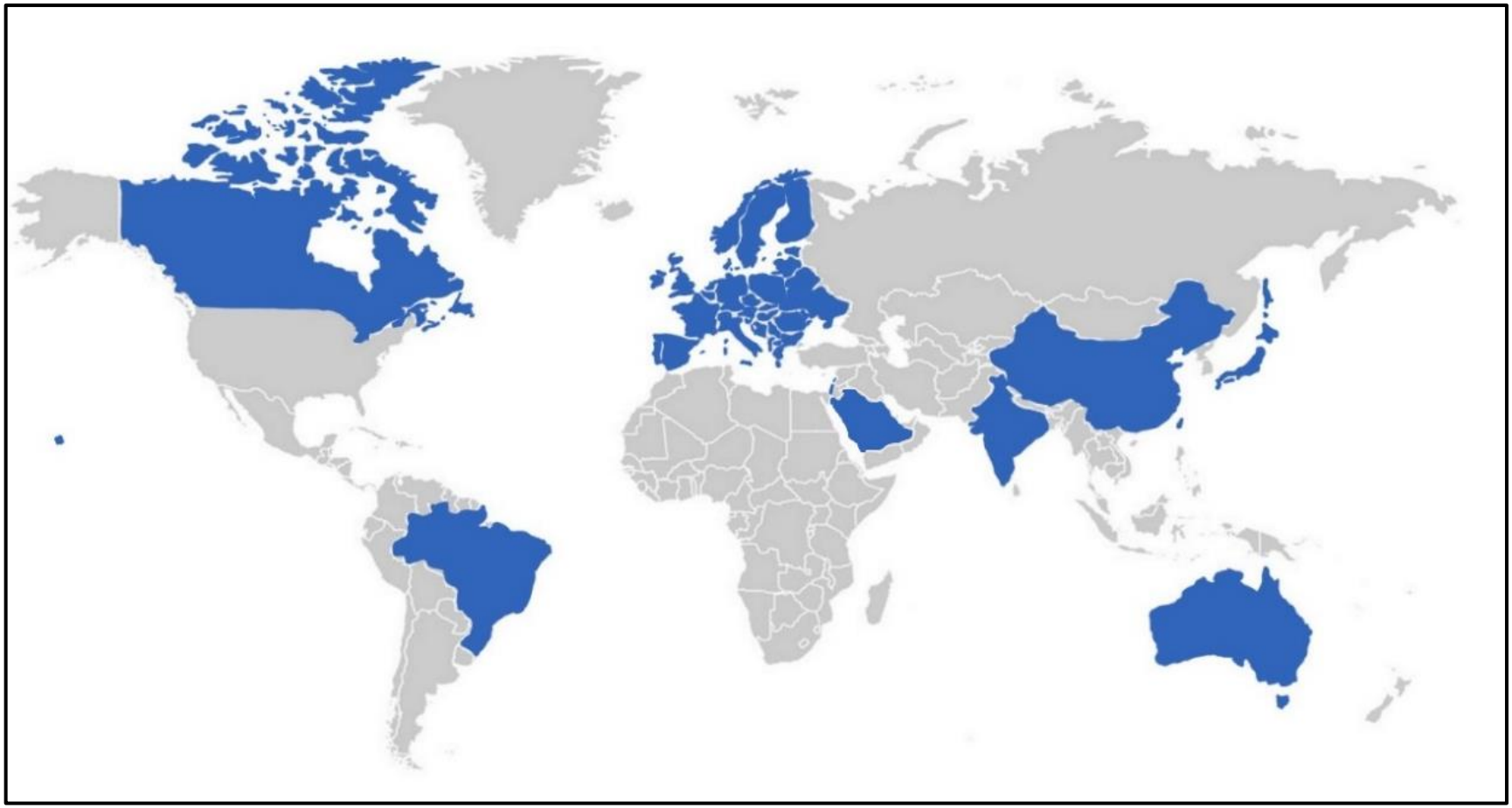


Figure 1. *Every country that has fully or partially attempted to DNA Barcode their marine fish biodiversity*.

1.6 OUTLINE OF THE THESIS

In this thesis I used DNA barcoding of the COI region to provide a resource and case-study for how a DNA barcode reference library could benefit the conservation and management of NZ's fishes. This thesis is composed of four chapters. Chapters two and three are written in manuscript format and I anticipate submitting these for publication soon after receiving my thesis examiner's comments. Due to this, the pronoun 'we' is used to include other collaborators, and there is some repetition of the content and methodological detail among the four chapters presented in this thesis.

In Chapter Two we attempted to create a complete DNA barcode reference library for all NZ's Actinopterygii fishes. We used open access resources to obtain available barcodes of NZ's marine fish fauna, and determine for how many species, and which species, DNA barcodes would need to be generated for providing a full DNA barcode reference database for NZ's fishes. This involved compiling a current list of every marine fish species living in NZ, including new discoveries. We then located and retrieved the barcode regions available for these species on the International Nucleotide Sequence Database Collaboration (known as NCBI, an online DNA repository, Cochrane et al. 2016) or BOLD, and then identified which species do not currently have a barcode sequence. Specimens for these non-barcoded species were then searched for within universal tissue databases. This will establish the likelihood of being able to barcode every marine fish species in NZ which can aid in fish species identification, improving management of fisheries (Paine et al., 2007) as well as ecological and conservation research (Wang et al, 2018).

In Chapter Three, we used DNA barcoding to genetically identify ambiguous larval fish species collected off the Northeast coast of NZ. Initially specimens were identified by morphological identification, however, the identification of larval fish can be very challenging (Ko et al., 2013). The utilisation of genetics aided in accurate identification of ambiguous species and discovering the identity of unknown larval fish species. Identifying these species aids in identifying the spatial ecology of early life history stages of native fishes and may bring more understanding on the locations of spawning and movements of species which are required for successful ecological monitoring (Ko et al., 2013).

In the General Discussion, Chapter four, I discussed the results of Chapters two and three and expand on what these results mean and how they may influence our understanding and management of NZ's marine ecosystems. I discuss the limitations of my study with reference to data collection and the boundaries of plausible assumptions that can be made from my study. Finally, I suggest future directions for developing and maintaining the NZ Fish Barcode Database.

Initially, I had planned to generate DNA Barcodes for species that are not currently in any online database repository and are found in NZ. I had also planned to DNA barcode a greater number of collected samples, to gain a broader understanding of juvenile fishes off the East Coast of North Island NZ. However, due to the lockdowns and restricted travel during the COVID-19 pandemic, this plan was unable to advance. Nonetheless, I feel I progressed an ambitious research plan and look forward to the examiner's suggestions for improvement.

CHAPTER TWO: A DNA BARCODE REFERENCE DATABASE FOR AOTEAROA NEW ZEALAND'S RAY-FINNED (ACTINOPTERYGII) FISH BIODIVERSITY

2.0. ABSTRACT

DNA barcode reference databases have been created for the fish biodiversity of many nations, providing a resource to facilitate rapid species identification, biodiversity assessment, and ultimately greater awareness and understanding of marine fauna. Aotearoa New Zealand (NZ) has a wide diversity of marine, estuarine, and freshwater habitats that comprise of a diverse fish fauna, and a high proportion of endemic fish species. Even so, a DNA barcode reference database for the fishes of NZ has not yet been created. In this study, we curated a DNA barcode reference database for NZ fishes based on the Cytochrome Oxidase I (COI) gene region using previously published sequences from open-access repositories (i.e., NCBI and BOLD) and novel sequences generated for species not previously sequenced. To ensure representativeness and integrity of the sequence data, we preferentially generated sequences from fish specimens that had been identified by an expert taxonomist or held in museum collections. Furthermore, for widespread species that we did not have sequences for, we sought sequences from specimens collected as geographically close to NZ as possible. We were able to generate and retrieve sequences for 965 of the 1320 fish species recorded in NZ (73%). Here we provide a summary of our progress toward generating a DNA barcode reference database for NZ's fish biodiversity and report interspecies genetic divergences (based on Kimura-2-pairwise distance) between species and genera. Overall, our generated NZ Fish Barcode Database is a practical resource of value for future environmental DNA studies, biodiversity monitoring, and managing fisheries and commercial fisheries-derived products.

2.1. INTRODUCTION

It is the mandate of taxonomists and biodiversity scientists to have a complete catalogue of the world's biodiversity. Over recent decades, this pursuit has included the generation of genetic information for each taxon which has aided our understanding of the evolutionary history of species, their taxonomic relationships, and has also provided another means by which we can classify or identify species. In particular, the use of a relatively short DNA sequence called a DNA barcode (Herbert et al. 2003), has proven effective in the identification and distinction of species within several taxonomic groups (Trivedi, Ansari, Ghosh, & Rehman, 2016). Using these genetic barcodes is preferable over morphological identification in many instances, such as when taxa are unable to be

easily distinguished based on morphology or other characters (Hebert, Cywinska, Ball, & deWaard, 2003; Wang et al., 2018). DNA barcoding is also useful in cases where the whole adult specimen is not attainable, such as when there are only remains of a specimen or they are in an unfamiliar life stage (Paine, McDowell, & Graves, 2007; Ward, Zemlak, Innes, Last, & Hebert, 2005); or the specimen only contains shed tissue, hair or faeces (Jeon, Choi, & Suk, 2012; Paine et al., 2007). Based on the efficacy of DNA barcodes in helping to identify species, metabarcoding – which aims to characterize all species present in an area using an environmental DNA sample (eDNA) – has become a popular method by which the biodiversity of a region can be classified, and subsequently monitored (Ruppert, Kline, & Rahman, 2019; Miya, 2021). However, the accuracy of the DNA barcoding method and the utility of a metabarcoding approach is contingent on whether DNA barcodes have been characterized for the biodiversity within a region (Arranz, Pearman, Aguirre, & Liggins, 2020).

Genetic barcoding of marine fish biodiversity has been successfully executed and utilised in many countries. The COI gene is a commonly used barcode region to identify species (Hebert et al., 2003; Ivanova, Zemlak, Hanner, & Hebert, 2007; Ward et al., 2005) including fishes. For instance, the COI barcode region has been used in barcoding marine fish species of Australia (Ward et al., 2005), India (Lakra et al., 2011), China (Wang et al., 2018; Xu et al., 2019; Zhang, 2011; Zhang & Hanner, 2012; Zhang, Qin, Zhang, Wang, & Lin, 2017), Portugal (Costa et al., 2012), the Canadian Pacific (Steinke, Zemlak, Boutillier, & Hebert, 2009), Japan (Zhang & Hanner, 2012), Argentina (Mabragana, Diaz de Astarloa, Hanner, Zhang, & Gonzalez Castro, 2011), Brazil (Ribeiro et al., 2012) and Israel (Shirak et al., 2016). The DNA barcodes generated by these efforts have been useful in identifying ambiguous species (Ward et al., 2005), discovering cryptic species (Hebert et al., 2003; Lakra et al., 2011), identifying partial or digested remains of species (Paine et al., 2007) and in linking life stages of a species, such as larvae to adults (Alcantar-Escalera, Garcia-Varela, Vazquez-Dominguez, & Perez-Ponce de Leon, 2013). These databases have also been used in studies that use environmental DNA (eDNA) and s have detected rare fish (Pfleger, Rider, Johnston, & Janosik, 2016), helped describe the fish fauna of locations based on the DNA barcodes recovered in the eDNA sample (Stoeckle et al. 2021), and been used in the monitoring of marine protected areas (Gold, Sprague, Kushner, Zerecero Marin, and Barber, 2021; Miya, 2021).

Aotearoa, NZ, is an archipelago of islands in the South Pacific, stretching from the subtropics to the sub-Antarctic (Roberts, Stewart, & Struthers, 2015). Over 97% of the nation's Exclusive Economic Zone is marine comprising of 1296 marine fish species and a total of 1320 fish species (Roberts, Stewart, Struthers, Barker, & Kortet, 2020). Although not particularly speciose relative to neighbouring nations, owing to its geographical isolation, a relatively high proportion of NZ's marine fishes are endemic (Roberts, Stewart, & Struthers, 2015). Commercial fishing in NZ is also of great economic importance, contributing over $4 billion-dollars to the economy each year and providing 0.7% of the nation's gross domestic product (Williams, Stokes, Dixon, & Hurren, 2017). Despite this

unique fish fauna being of commercial fishing interest, NZ's overall fish diversity has been relatively poorly studied (Roberts, Stewart, & Struthers, 2015) and on-going research suggests that there are still several undescribed species (Liggins et al. 2021), new species still being recorded (Francis, 2019; Liggins et al., 2020; Matsuura & Middleton, 2016), and several more species are expected to be naturalising in NZ as ocean climate continues to change (Middleton et al., 2021). Barcoding all NZ's fishes would provide a useful resource that may then be used to develop alternative means to study and monitor the nation's fish biodiversity, and manage fisheries and commercial fisheries derived products, such as has been done with cetaceans (Baker et al. 2007).

Creating a sequence reference database can be difficult and time-consuming, which is why many researchers use pre-existing public databases (Arranz, Pearman, Aguirre, & Liggins, 2020). However, using inappropriate, outdated, or un-curated sequence reference databases can compromise study results and mislead inferences (Jin, Kim, Kim & Park, 2020). Open-access sequence databases, such as the National Centre for Biotechnology Information (NCBI, part of the International Nucleotide Sequence Database Collaboration, Cochrane et al. 2016) and the Barcode of Life Database (BOLD, Hebert & Ratnasingham, 2007) frequently have multiple sequences for each species but can also include inconsistencies or errors causing issues in data integrity (Chen, Zobel, & Verspoor, 2017). Recent studies have shown that several sequences in these reference databases are of poor quality, and/or the result of contamination, and can lead to misleading species identifications (as described in Tang, Stiassny, Mayden, and DeSalle (2021); Pentinsaari, Ratnasingham, Miller & Hebert, 2020). In contrast, a curated reference database, cross-referenced with a species checklist for a given region, can potentially provide higher confidence in species identification (Gold et al., 2021; Yang et al., 2017). Such an approach removes the opportunity for an individual sequence to be incorrectly identified as a species that is out of geographical range but is closely genetically related (such is the case for many sister taxa of NZ anti-tropical species that are in the Northern Hemisphere). Furthermore, a sequence reference database for which there are accompanying vouchered specimens can contain sequences for taxa that have not yet formally been described, but can be updated as species are named and/or taxonomically reassigned (Schoch et al., 2020).

In this study, we present a curated reference database of cytochrome oxidase 1 (COI) DNA barcodes for 965 of NZ's 1320 known marine and freshwater fish fauna. With a focus on marine fish fauna in particular, we generated and compiled COI barcodes for 388 species from voucher specimens based in NZ from Te Papa Tongarewa, Auckland Museum Tāmaki Paenga Hira, and Massey University Auckland and used open access sequence repositories to retrieve a further 577 COI barcodes for marine fish species found in NZ, preferentially choosing sequences within, or nearest to, NZ. Our DNA barcode reference database provides a quality assured resource for characterising the marine and freshwater fish fauna of NZ and identifying fish species.

2.2. METHODS

2.2.1. Checklist of New Zealand fishes

We considered all marine and freshwater Actinopterygii within NZ's Exclusive Economic Zone. Although marine fishes were the focus of our DNA barcode generation, we opted to also include freshwater fishes as several have a marine phase within their lifecycle. The most comprehensive checklist of NZ fishes is provided by 'The Fishes of New Zealand' (Roberts et al. 2015) and subsequent updates (Roberts, Stewart, Struthers, Barker, & Kortet, 2017; Roberts, Stewart, Struthers, Barker, & Kortet, 2019; Roberts et al., 2020). Here we used the Roberts et al. 2020 checklist, and updated it according to published new species records and taxonomic changes (e.g., Middleton et al., 2021; Stewart, Knudsen, & Clements, 2021; Short & Trnski, 2021) and unpublished records (Middleton I., PhD thesis in prep.). The total list of marine and freshwater Actinopterygii fishes in NZ is comprised of 1,320 taxa.

For these taxa, we attempted to generate or retrieve COI barcodes from vouchered specimens or specimens identified by an expert taxonomist. Previous sequencing using this approach generated COI barcodes for 139 taxa to date (Eme et al. 2019;Eme et al. 2020). Other studies since this time have additionally generated COI sequences for 5 more taxa, using vouchered specimens, or specimens identified by an expert taxonomist (Conway, Stewart, & Summers, 2018; Delrieu- Trottin et al., 2018; Liggins et al., 2021; Short & Trnski 2021). For the remaining 1,176 taxa, we generated COI sequences where possible, or retrieved existing sequences from NCBI and BOLD as described below.

2.2.2. Generation of novel DNA barcodes for New Zealand's fishes

We accessed tissue from within curated collections with associated specimens where possible. These fishes were collected on several research expeditions undertaken by the Museum of New Zealand: Te Papa Tongarewa, Auckland Museum Tāmaki Paenga Hira, and Massey University Auckland, over recent years. Tissue samples (white muscle or fin clips) were preserved in 80-98% ethanol and subsequently stored at -80°C, -20°C or 5°C. All DNA extraction, PCR, and sequencing preparation for the focal taxa was carried out at Massey University, Auckland. Genomic DNA was extracted using the Promega Salt Extraction kit or DNeasy Blood and Tissue kits (Qiagen, Valencia, CA), or the Chelex 100 chelating resin extraction protocol (described in Walsh et al. 1991) if the first extraction had low DNA yield. To amplify a portion of the barcode region of the COI gene, we used the primer combination named Fish COI-2 Cocktail (as described in Ivanova et al. 2007). All PCRs were conducted using either the MyTaq™ or MyFi™ DNA polymerase kits (Bioline, Australia Pty Ltd, Alexandria, NSW), as per the kit instructions. For the Fish COI-2 primer cocktail, PCR was performed with a denaturation at 94°C for 1 min, followed by an initial 5 cycles (94°C for 30 secs,

50°C for 40 secs, 72°C for 1 min), followed by 35 cycles (94°C for 30 secs, 54°C for 40 secs, 72°C for 1 min), then a final extension at 72°C for 10 mins (as described in Ivanova et al. 2007).

Following PCR, a 1% agarose gel was run using 2µl of PCR product and 1µl of GelRed, alongside a BioLabs Quick Loading DNA Ladder to ensure PCR products were of the right size and sufficient concentration. PCR products were then purified using the ExoSap reagents and protocol (Thermo Fisher Scientific, West Palm Beach, FL) and sent for forward and/or reverse sequencing-using M13-F and M13-R primers, respectively (sequenced by Macrogen, Korea). Quality control of the received sequences was carried out using Geneious version 9.0.5. All sequence chromatographs were inspected by eye and poor-quality nucleotide bases and primer sequences were trimmed prior to alignment of the forward and reverse sequence (where available) to check for consensus. To ensure that none of the consensus sequences were derived from pseudogenes, nucleotide codons were converted into amino acids using the 'Vertebrate Mitochondrial' genetic code and any remaining stop codons trimmed. The taxonomic identity and relationships of the consensus sequence was sanity checked by submitting to BOLD (to check for any existing BIN), and using NCBI's nucleotide BLAST.

2.2.3. Retrieving existing DNA barcodes for New Zealand's fishes

For taxa that we did not have tissues for, or were unsuccessful in generating a COI sequence, DNA barcodes were retrieved using the R package *regPhylo* (https://github.com/dvdeme/regPhylo, Eme et al.2019a) following Tutorial 1 (Eme et al., 2019). To ensure compatibility of species names with the NCBI taxonomic database we excluded taxa with unspecified species status and taxa with a cf. (confere status) or with a question mark in front of the species epithet were assumed to belong to the nominal species (note that these taxa were retained in our species list, and list of sequences, where we generated a sequence from a vouchered specimen as discussed above). We extracted the unique NCBI taxonomic identifiers (taxid) and checked the species names listed as NCBI synonyms using the NCBI taxonomy database (www.ncbi.nlm. nih.gov/Taxonomy/TaxIdentifier/tax_identifier.cgi). For all species with a unique NCBI taxid, COI sequences and associated metadata (Taxa Name, Accession Number, sequence length, Definition, Organism Classification, Source, Title, Authors, Journal, Year Published Location sequence was retrieved from, Geographical coordinates, Collection date, Date Extracted) were extracted from NCBI using the function GetSeqInfo_NCBI_taxid (relying on the R package "Rentrez", Winter et al. 2017) and from BOLD using the GetSeq_BOLD function built from the R package "bold" (Chamberlain, 2021).).).

The geographic distribution of several species within our checklist of marine fishes in NZ extends outside of NZ's Exclusive Economic Zone. For all sequence retrievals when multiple COI sequences were available for a given species, we used *regPhylo* to help us preferentially choose a sequence from

within NZ or closest to NZ, thus minimising the influence of phylogeographic structure on inferences of interspecies genetic distances, and/or allopatric or cryptic species. We used the function SelBestSeq to select the sequence for each species that was closest to the centroid of NZ (-41.3355° S, 174.7976° E). If multiple sequences were equally close to NZ's centroid, or if none of them were georeferenced, we chose the sequence closest in length to the median sequence length to avoid alignment problems caused by long chimeric sequences overlapping other gene regions.

2.2.4. Characterising the sequence reference database of New Zealand marine fishes

We quantified the number of taxa represented in our compiled NZ Fish Barcode Database relative to the total known diversity of marine fishes in NZ, and illustrated the results according to family. The geographic origin of the sequences, where known, were mapped using ggplot in the package ggplot2 (Wickham, 2016). To investigate the level of sequence divergence based on the COI region for NZ fishes, we calculated pairwise sequence divergences among species and genera. We used the Ape R package (Paradis, Claude, & Strimmer, 2004) to calculate the pairwise distances using the Kimura-2-Parameter (K2P; Kimura (1980) substitution model, and used the reshape2 R package (Wickham, 2007) to 'melt' the pairwise distance matrix into a single column of data for analysis. We calculated the mean (and range) of the sequence divergence among all species, and within genera. In this way, the calculated values were intended to help define upper and lower thresholds, helping to signal where further taxonomic work may be required. In cases where we recovered 0% sequence divergence among nominally different taxa, we attempted to trouble shoot whether this was likely a real biological result, or reflected the acquisition of sequences not representative of those species (i.e., because of laboratory error, or mis-labelled sequenced by us or contributors to open-access repositories). In some cases, based on what we could resolve we elected to select a different sequence to represent one, or both species, for inclusion on our NZ Fish Barcode Database.

2.3. RESULTS

The final checklist of NZ marine Actinopterygii is comprised of 1320 species and 254 families. Overall, we compiled DNA barcodes for 965 species (73.1%), including 291 already provided for NZ fishes from vouchered specimens by previous publications (Eme et al. 2019; Eme et al. 2020; etc.) and 97 novel COI barcodes generated for this study from vouchered specimens. An additional 577 sequences were retrieved from NCBI and BOLD. The COI sequence lengths varied from 205 bp to 1593 bp, with a median length of 652 bp. For 355 species we were unable to generate or retrieve COI sequences from these genetic repositories. The 355 species for which we were not able to retrieve or generate COI sequences for showed no bias among different families within Actinopterygii (Fig. 2).

Of the 254 families there were 118 families (46%) with a sequence for every species and 52 families with no sequence (20%) (Fig.2). Among the families with larger numberss of species: Carangidae, Myctophidae, Scombridae and Seramidae had a sequence for each species whereas Arynchobtidae. Carcharihimidae, Centrophidae, Chimareidae and Zoarchidae had no species with a sequence, and thus require more attention in future sequencing efforts.

There were 35 sequences (6.1% of the 577 sequences) retrieved through *regPhylo* with an unknown geographic origin. Of those with location metadata, they varied in the precision of their geographic information, and had the following origins: 1 from Bangladesh, Jamaica, Rarotonga, the North Atlantic Ocean, Portugal, the Solomon Islands, South Korea, Turkey, Vanuatu and Vietnam; 2 from Brazil, Haitii and the Philippines; 3 from Belize; 4 from Antarctica, Canada and China; 5 from Fiji and Mexico; 6 from the Atlantic Ocean, Japan, South Africa and Tonga; 11 from Indonesia; 12 from the Southern Ocean; 13 from Taiwan; 14 from France; 16 from the Pacific Ocean; 17 from French Polynesia and the Indian Ocean; 20 from New Caledonia; 32 from the United States of America; 39 from the Tasman Sea; 60 from NZ; and 218 from Australia, (Fig. 3).

On average, retrieved sequences were 21.0% divergent among species (range = 0.00% – 36.94%; Fig. 1; Table 1). Several putatively distinct taxa had 0% pairwise divergence at the examined COI gene region. Within genera, species were on average 8.62% divergent (range = 0.00% – 11.49%; Table 1).

Number of taxa	**Taxonomic level**	**Minimum (%)**	**Mean (%)**	**Maximum (%)**	**Standard Error**
965	Species	0	0.210297	0.3694	3.03×10^{-5}
174	Genus	0	0.08618	0.114914	0.007

Table 1. Summary of pairwise genetic divergences (K2P pairwise distance) based on Cytochrome Oxidase I among all species, and species within genera. Data was collected from 965 species and 174 genera.

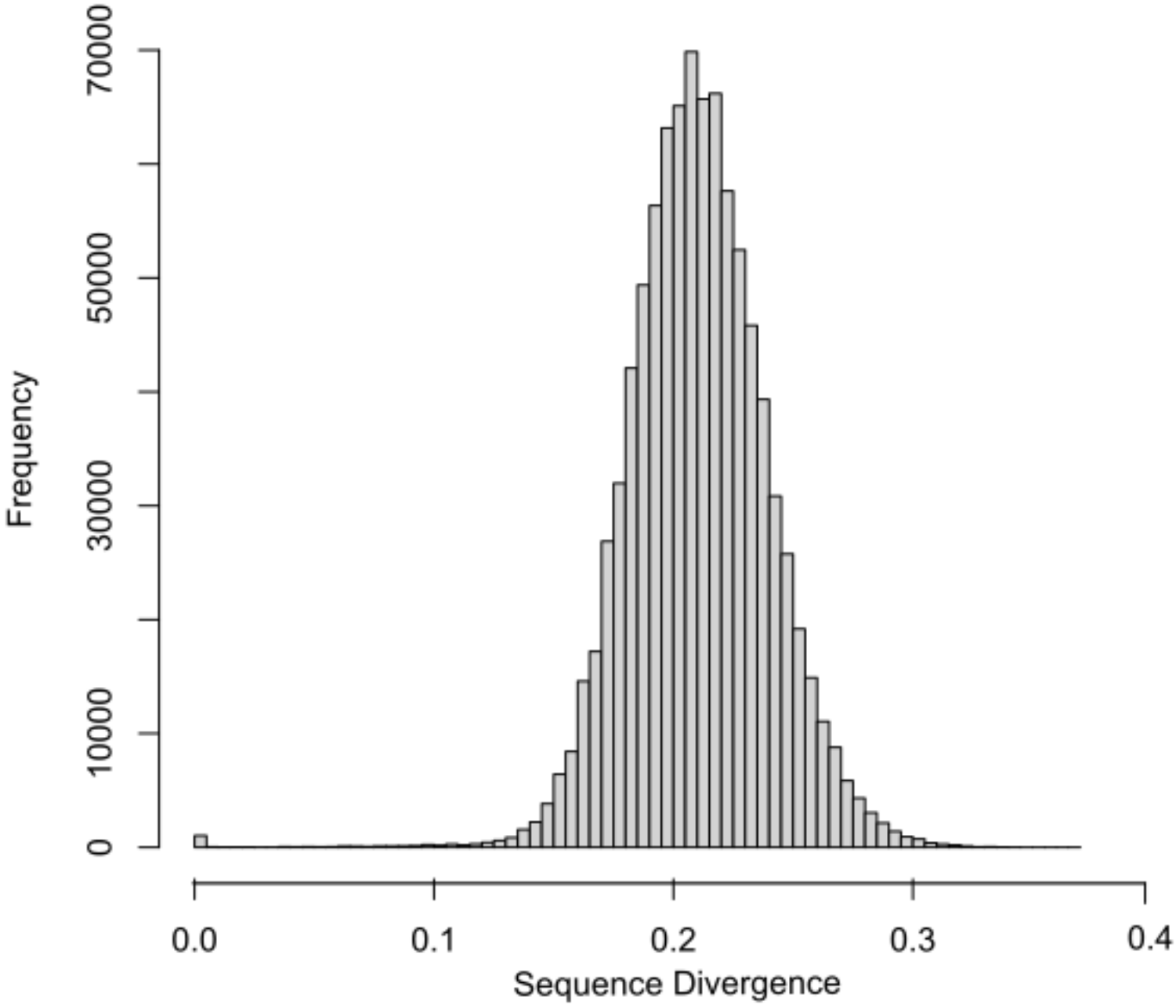


Figure 1. The distribution of pairwise genetic divergences (K2P pairwise distance) based on Cytochrome Oxidase I among all 965 fish species.

Figure 2. Number and proportion of fish species within each family that had a COI barcode sequence available and where it was retrieved from: NZ Fish Barcode Database, BOLD/ NCBI or if there is no sequence available arranged in alphabetical order of family.

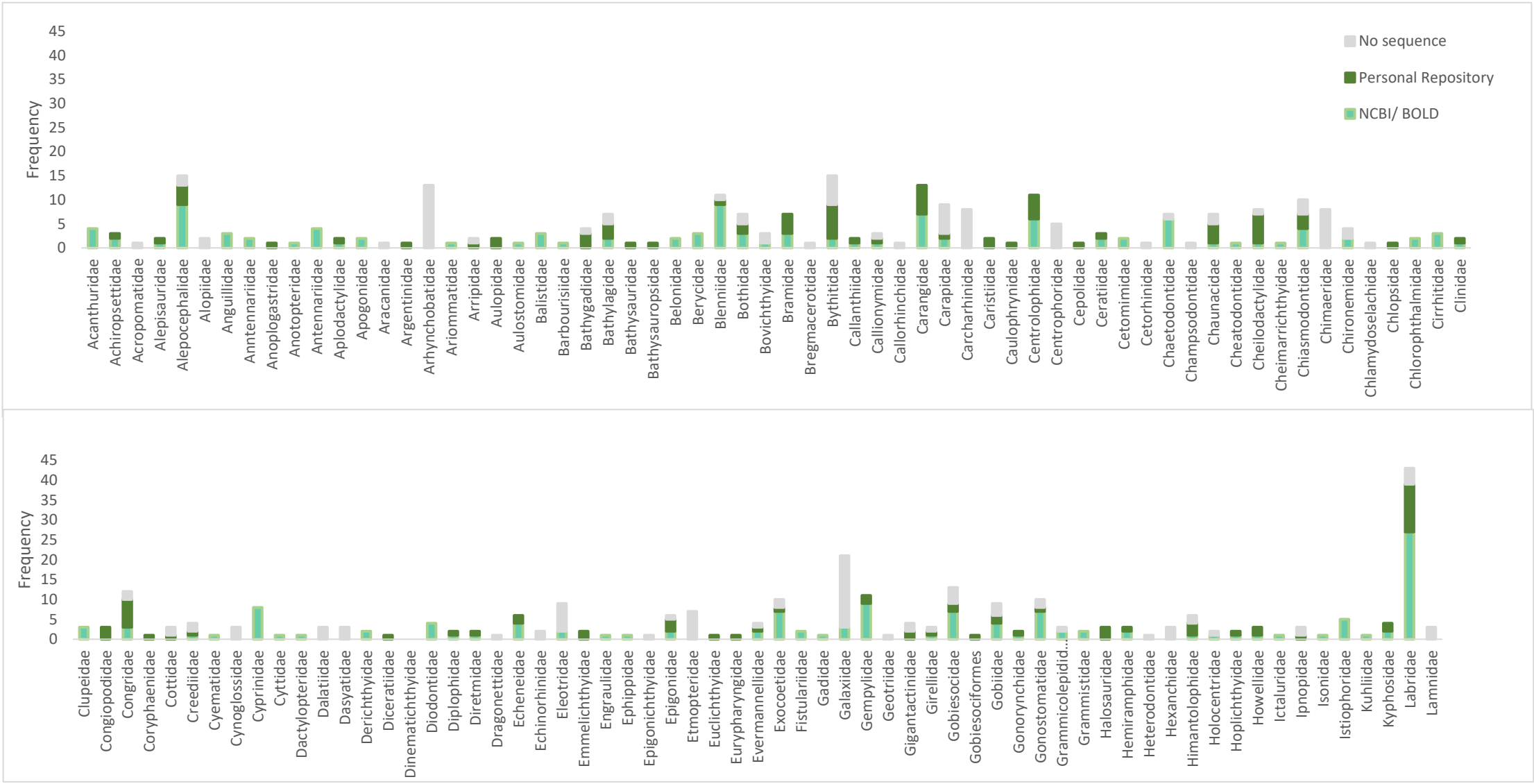
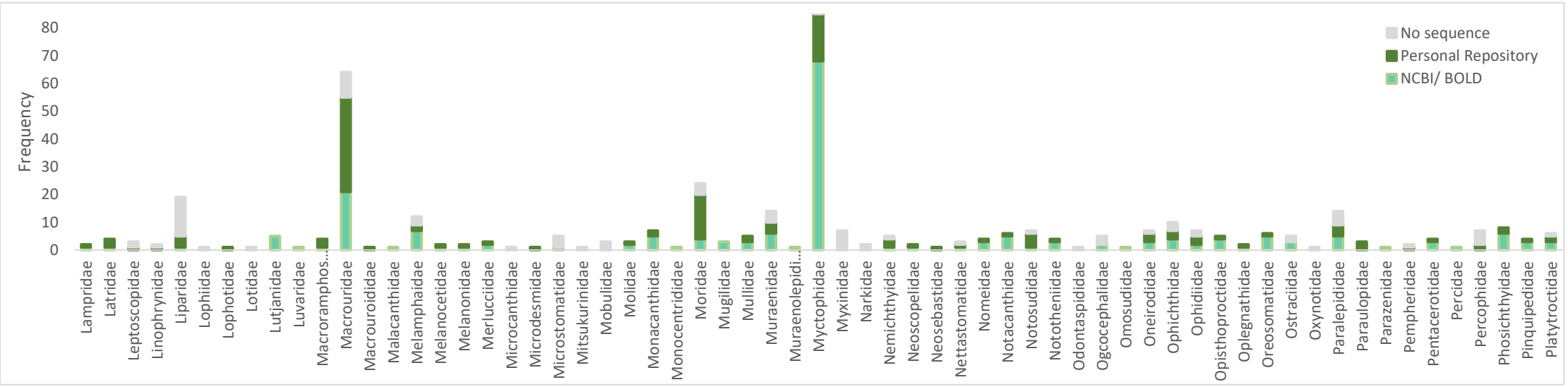
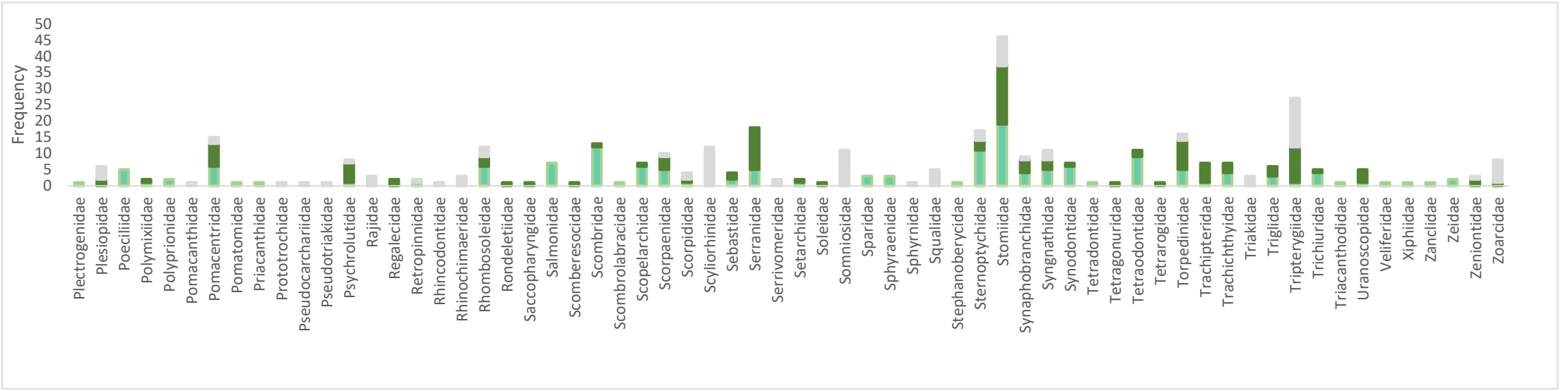


Figure 2, continued





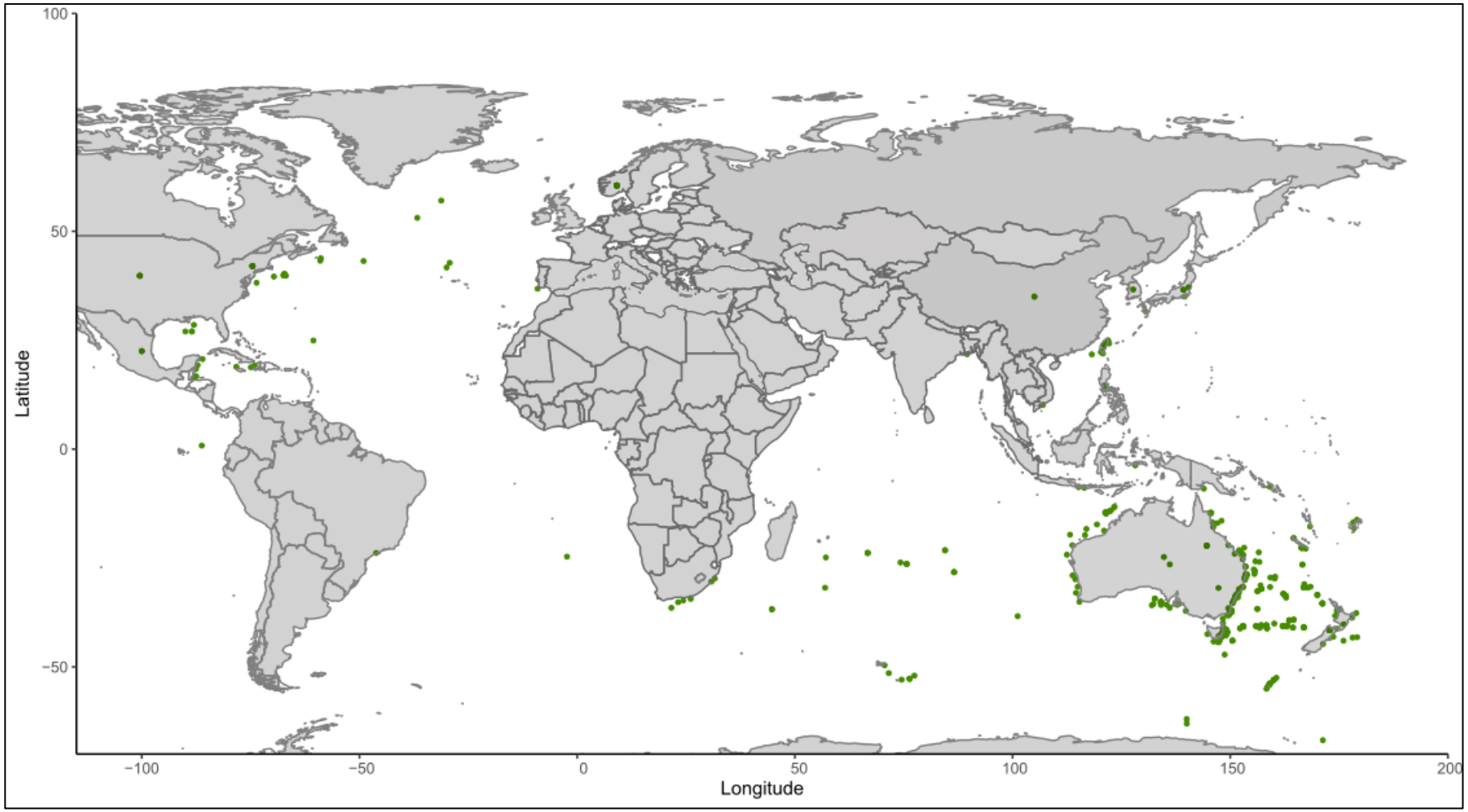


Figure 3. Geographic origin of Cytochrome Oxidase I (COI) sequences retrieved from the BOLD/NCBI for the New Zealand Fishes DNA barcode reference database. A number of these location points are imprecise and are based on general location information provided.

2.4. DISCUSSION

The objective of this study was to create a local DNA barcode reference database for NZ's marine and freshwater Actinopterygii based on the cytochrome oxidase I (COI) gene region. This project benefited from the nation's premiere fish taxonomistsfrom Te Papa Tongarewa and the Auckland Museum, who provided access to their tissue collections, including the National Fish Tissue Collection, to access the National Fish Tissue Collection, as well as Auckland Museum to make use of their tissue collection. The resulting DNA barcode reference database represented 965 of the 1320 fish species found in NZ, based on novel sequences generated for this study and previously published sequences from specimens identified by expert taxonomists or curated collections, and the retrieval of sequences from open-access repositories (i.e., NCBI and BOLD). As a result, we have compiled the most comprehensive and quality assured DNA barcode reference database for NZ fishes to date.

Our NZ Fish Barcode Database includes 73% of the known fish species in NZ. Despite our efforts to preferentially use COI sequence data generated from specimens identified by an expert taxonomist or from curated collections, these sequences contributed only 40% of the species within the reference database. Most of the sequences were retrieved from open-access repositories, with the majority retrieved from Australian waters. Although NZ and Australia share a large proportion of their marine fish fauna, we had hoped to preferentially retrieve sequences from NZ through the use of functions of the regPhylo R package (Eme et al. 2019) that make use of spatial metadata attached to sequences, we had hoped to preferentially retrieve sequences from NZ. Nonetheless, in executing our approach to building a curated DNA barcode reference database – based on sound species identification, vouchered specimens, and geographically relevant populations – we have provided a pathway for future studies and enhancement of this DNA sequence resource.

Our study identified a further 355 fish species, 254 families, for which no COI sequence exists, and would need to be generated in order to have a complete DNA barcode reference database for NZ fishes. For all NZ fish families, we characterised the proportion of species that have a sequence in our current database, where that sequence was sourced from, and how many sequences are still needed to be generated for species within each family (Fig. 2). The families that lack sequences, including 52 families represented by no sequences, could be due to their absence in curated collections. For instance, many deep-sea, cryptic, and rare species are not well represented in collections as they are difficult to acquire (Steinke, Zemlack, Boutillier, Hebert, 2009). In addition to this, some families (including: Gobiidae, Ophichthidae, and Tripterygiidae) are difficult to identify morphologically and therefore are difficult to accumulate COI sequence data for (Ko et al., 2013). Alternatively, several species remain unrepresented in genetic biodiversity studies because commonly used primers (such as the primers used for this study)

do not work well for them (Ivanova et al. 2007), or because the successful extraction and amplification of their DNA relies upon specialist biochemical or mechanical treatments (Zhang & Hewitt, 1998). To help increase the representation of these missing species within a national DNA barcode reference database may require targeted collection efforts to attain elusive species, and use of more specialised laboratory protocols for DNA extraction and taxon-specific primers for amplification.

Our study provides a baseline for future studies looking to complete the COI barcode reference database for NZ's fishes. In pursuit of this, it would be beneficial aligning with other national research institutions and Museums in NZ to coordinate sampling, share specimen collections, and undertake a combined sequencing effort, especially if the required specimens are already within existing collections. Even so, the collection of NZ-based specimens for all species known to occur in NZ is unlikely due to the rarity of some species. For example, a unique species of boxfish, *Kentrocarpus* sp., has only been sighted once, and although accurate description of the taxon was possible from the high-quality photographs provided, there has not yet been the opportunity to attain a specimen (Matsuura & Middleton, 2016). Furthermore, several tropical marine species that are very abundant elsewhere, and for which there are specimens and sequences available in overseas collections, are very rare in NZ, only occurring as juveniles or a few surviving adults (Francis et al. 1999, Middleton et al. 2021). In such cases, the ethics of collecting such locally rare specimens must be weighed up with the information gains.

In addition to identifying which species require barcodes generation, based on our calculation of genetic divergence among species based on the COI sequences (Fig. 1), we identified species pairs that require further investigation. On average, the species in our COI reference database were 21% divergent, but several taxa (15 species pairs) were identified as having 0% sequence divergence between them. In these cases, it is possible that the specimen sequenced was misidentified, there was a tissue sample mix-up, cross-contamination in the laboratory, or there has been taxonomic over-splitting of the group (Landi et al., 2014). For instance, both *Polymixia japonica* and *P. busakhini* are both within the NZ fish list and have 0% sequence divergence between them. It may be that these species have geographically disparate populations (in the Northern and Southern Hemispheres) that were described independently, but may in fact be the same species. It is also known that the COI gene region is not always effective in capturing the known taxonomic distinction among taxa. For example, in several Chondrichthyes groups, species distinction is often not supported by divergence at COI, despite other compelling evidence for species distinction (Naylor et al., 2012; Wong, Shivji, & Hanner, 2009). Accordingly, in our NZ Fish Barcode Database *Carcarhinus obscurus* and *C. galapagensis* were indistinguishable based on COI. In Actinopterygians, we also found the very recently diverged species of *Chrysiptera notialis* and the Kermadec *Chrysiptera demoiselle* (formerly *C. rapanui* and now considered an undescribed species) were

indistinguishable based on COI, despite them being distinguishable based on morphology and the mitochondrial control region (Liggins et al. 2021; Liggins pers. comm.). In future, communication among institutions who hold collections and further sequencing of vouchered specimens will help resolve any of these issues, caused by the misidentification of specimens, cross-contamination, or taxonomic issues. In addition, and particularly in the cases of recently diverged taxa, the addition of other gene regions as "barcodes" would also be beneficial (e.g., 12S, Milan et al., 2020).

The NZ Fish Barcode Database should be maintained alongside the NZ fishes list as an up-to-date record in order to be a working tool for future research. This includes updating the database as the taxonomy of the specimens is revised or new species are discovered ensuring the database covers the full representation of NZ fishes. The benefit of our approach, in preferentially using vouchered specimens from the NZ Fish Barcode Database, is that it can contain sequences generated for vouchered specimens that have not yet formally been described (e.g., *Nemadactylus* n. sp., *Hypoplectrodes* sp. C), and can be changed and updated according to specific changes in the identification and description of the specimens. For example, the Kermadec clingfish *Aspasmogaster sp.* was reassigned to *Flexor incus* after taxonomic research by Conway et al. (2018) allowing for specimens and corresponding sequences for *Aspasmogaster sp.* to be updated to the appropriate taxonomic name. Other recent taxonomic work, describing a cryptic *Halargyeus* species (de Carlos et al., 2020) and therefore the existence of two congenerics in NZ, will also likely lead to the re-assignment of several specimens in NZ collections. We recovered 0% sequence divergence between a specimen named *Halargyeus* sp. and a sequence named "*Halargyeus johnsonii*" retrieved from an open access repository, but from a specimen collected in NZ. In this case, understanding whether both sequences derive from the new cryptic species, or are in fact *H. johnsonii*, based on sequences alone would require considerable retrospective curation of existing sequences in open-access repositories. Having a primary NZ Fish Barcode Database will reduce the chance of synonymous taxa name errors caused by name changes from taxonomists as new information is revised (Schoch et al., 2020) in addition to being a current and reliable research tool.

A limiting factor in our NZ Fish Barcode Database generation includes a lack of geographic location information of retrieved sequences. Lack of geographic information in public databases is a growing issue in molecular ecology. A large proportion of published sequences have inaccurate or no geolocation information (Pope, Liggins, Keyse, Carvalho, and Riginos (2015); Toczydlowski et al. (2021). Lack of geographic information for fishes includes imprecise geolocation, when a location name has not been provided, or when the centroid of that location is not in the ocean from where a specimen was retrieved (such as just a country name). A lack of geographic references hinders our ability to select sequences that are from NZ, or nearby locations (Fig. 3). Continuing to generate sequences from NZ specimens, with

accurate geolocation, and retention of this information as metadata alongside the genetic sequence information would be beneficial for not only our NZ Fish Barcode Database but also, future research.

In regards to future steps and research, utilisation of the NZ Fish Barcode Database in environmental DNA studies will be a beneficial tool. In aquatic environments, studies have shown a high degree of accuracy in identifying species within a habitat with little DNA dispersion despite water movement between domains, therefore providing an accurate assessment of biodiversity in a habitat (Jeunen, Knapp, et al., 2019; Jeunen, Lamare, et al., 2019). The advantage of using the NZ Fish Barcode Database in eDNA studies is that researchers do not need to generate their own database in order to identify species in their sample, The NZ Fish Barcode Database is up-to-date with accurate sequences of species and is therefore able to accurately identify taxa within an eDNA sample.

In summary, our current NZ Fish Barcode Database contains sequences for 73% of NZ's fish species, with 355 species sequences still to be generated where possible. Our database has identified where species may require taxonomic reassessment and which species still require generation of sequences (Appendix 1). The inaccurate collection coordinates of many species also provide an incentive to regenerate sequences where possible from NZ caught specimens with the recording of accurate geolocation. We have created the most comprehensive barcode database for NZ fishes to date, providing a reliable research tool for future eDNA, conservation, and marine research studies.

CHAPTER THREE: IDENTIFICATION OF LARVAL FISHES IN NORTH EAST NEW ZEALAND: A COMPARISON OF MORPHOLOGICAL AND DNA BARCODE METHODS

3.0. ABSTRACT

The identification of larval fishes can be challenging due to their differentiation from the adult life stage, their small size, and their distinguishing characters being easily damaged during collection. Accordingly, genetic methods of identification, such as DNA barcoding, have increasingly been used to identify larval fishes alongside morphological methods. New Zealand (NZ) has a diverse fish fauna comprising of subantarctic, temperate, and tropical species, including several endemic species and some recently colonizing species. This study aimed to genetically and morphologically identify larval fishes collected off New Zealand's North Island East Coast and compare identification methods. In this study, we strategically sampled larval fishes off of the North East Coast of North Island NZ – the most biodiverse fish fauna around mainland NZ. We morphologically identified each fish larva to the lowest taxonomic designation possible, and then amplified and sequenced the Cytochrome Oxidase I ("barcode") gene region of each individual to provide a genetic-based identification. Taxonomic identification of larvae based on the DNA barcodes was performed against a curated reference database of NZ fish sequences, as well as public databases BOLD and NCBI. We found that DNA barcoding enabled the identification of fish larvae of unknown taxonomic identity based on morphology, and increased the precision of taxonomic identification for specimens only identified to the family-level based on morphology alone. Utilization of the NZ Fish Barcode Database generated the most accurate (based on percentage pairwise identity) and reliable specimen identifications in contrast to other online repositories (as they came from specimens of verified identity). Some larvae were unable to be genetically identified to species-level due to missing COI sequences, a limitation of the current databases that should be addressed in future studies.

3.1. INTRODUCTION

The early lives of many organisms are difficult to study in the marine environment. Early life stages, such as eggs and larvae, are often small and are sparsely or patchily distributed across vast oceans. Therefore, new knowledge gains regarding the early life of marine fishes have arguably been less frequent and more opportunistic than for other life stages (Antonio & Pineda, 2007). Nonetheless, what we now know about the early life of marine species has been transformational for our understanding of their biology and spatial ecology. For instance, it is now apparentthat many commercially and recreationally important deepsea fishes have long planktonic larval phases and spend extended periods near the waters' surface before settling into deep-water habitats (e.g., Orange roughy, *Hoplostethus atlanticus*, White, Stefanni, Stamford & Hoelzel, 2009; Hapuka, Bass, Warehous, Rubyfish, and Tripod fishes, Roberts, Stewart & Struthers, 2015; Bradbury & Snelgrove, 2001). We now know that some species of freshwater fishes will temporarily disperse in the ocean as larvae following spawning (e.g., Giant Kōkopu, *Galaxias argenteus,* Banded Kōkopu, *Galaxias fasciatus*; Shortjaw Kōkopu, *Galaxias postvectis*; Franklin, Smith, Baker, Bartels & Reeve, 2015; Poulin et.al, 2012). Strategic sampling of certain habitats could be used to help form a cohesive understanding of the spatial ecology of larval fishes (Li et al., 2017), what habitats and seasons are most important for certain species, and to help monitor changes in the distribution of planktonic larvae and species. The study of larval fish dispersion is an important topic within which New Zealand would benefit from not only to reveal more information about spatial ecology and early life stages but also to identify what juvenile fish may be dispersing to New Zealand, potentially creating a biosecurity risk.

Once sampled, the identification of eggs and even larval fishes to the species-level can be very challenging (Ko et al., 2013). The larval stages of fishes often differ greatly in morphology from the adult forms we are familiar with, and for several species, these life stages have never been observed (Ko et al., 2013; Shin, Jeong, Yoon, Choi, & Kim, 2018). Furthermore, larval fish often disperse much more widely than their adult life stages (Antonio & Pineda, 2007; Franklin, Smith, Baker, Bartels & Reeve, 2015; Putman, 2016) and so it cannot be assumed that the sampled larvae will correlate to the regional pool of adult species (Levin, 2006). Morphological identification includes using features such as fin ray counts to identify larvae after sampling (Ko et al., 2013). However, larval fishes are small and delicate and the sampling methods often cause mechanical damage during collection compromising our ability to identify morphological features (De Battisti et al., 2014). Due to a lack of definitive morphological traits and damage upon collection, a low percentage (13.5%) of species are identified correctly based on the morphology of juvenile stages alone (Ko et al., 2013; Victor, Hanner Shivji, Hyde & Caldrow, 2009) and several studies have looked to use genetic methods to aid in their identification (Valdez-Moreno,

VàsquezYeomans, Elìas- Gutièrrez, Ivanova & Hebert, 2010; Victor, Hanner, Shivji, Hyde & Caldow, 2009; Wibowo, Wahlberg, & Vasemāgi, 2017).

Molecular genetic markers have proven to be very useful in helping to understand the early life biology, developmental stages, and habitat use of many fish species. For instance, through genetically verifying the identity of larval fishes, researchers have been able to link these early life stages of fishes to the adult species (e.g., for 181 marine fish species in Mexico, Valdez-Moreno et al. 2010). In the characterization of the species composition of larval fish samples, genetic-based identification has been suggested to be more accurate and less time-consuming than morphological identification due to the small size and high tissue degradation rates (Fost et al., 2020), leading to a lack of definitive morphological traits (BattaLona, Galindo-Sanchez, Arteaga, Robles-Flores, & Jimenez-Rosenberg, 2019). The genetic marker typically used to distinguish among fish species, is the mitochondrial gene region called Cytochrome Oxidase subunit I (COI) commonly referred to as the 'barcoding' region (Stepanovic, Kosovac, Krstic, Jovic & Tosevski; 2016; Ward, Hanner, & Hebert, 2009). Several studies have evidenced the suitability of this gene region due to its detectable level of genetic differentiation among fish species and its ability to be amplified across a large range of taxonomic groups within fishes (Ward et al., 2009). This region has also been utilised in providing a reference database of fishes to which new sequences can be compared to verify species identities (Ward, Hanner, & Hebert, 2009).

Aotearoa NZ has ~1300 Actinopterygii (ray-finned) fish species recorded within its Exclusive Economic Zone (EEZ) (Roberts, Stewart, & Struthers, 2015). Although not species-diverse relative to tropical regions of the Pacific Ocean, the fish fauna of NZ includes species of tropical to sub-Antarctic affinity, several circum-global species as well as a relatively high proportion of endemic species (Roberts, Stewart, & Struthers, 2015). Recent efforts have worked to characterize the genetic relationships (Eme, Anderson, Myers, Roberts & Liggins, 2020; Eme et al., 2019) and DNA barcodes for the entire NZ fish fauna (Chapter Two). However, fish species are still being discovered (e.g., Matsuura and Middleton, 2016) and taxonomically described (e.g. Stewart, Knudsen and Clements, 2021) and there are several others that represent new arrivals (Liggins et al., 2020), recent range extensions and human-mediated introductions (Middleton et al., 2021). For these reasons, it is important to continue to build the genetic resources for NZ's Actinopterygii, and to monitor the species composition of our marine ecosystems.

The aim of this study was to compare accuracy of morphological larval fish identification against genetic databases NCBI, BOLD and NZ Fish Barcode Database to determine which type of identification and database was the most accurate in identifying NZ larval fishes. We strategically sampled larval fishes off the North East coast of NZ – the most fish species rich region around mainland NZ (Gordon, Beaumont,

MacDiarmid, Robertson, & Ayhong, 2010) and a known hotspot for new species records (Francis, Worthington, Saul, & Clements, 1999; Middleton et al., 2021).We then used DNA barcodes, alongside morphological identifications, to verify the species identities and to compare the identities provided by the two methods. To identify larval fishes based on DNA barcodes, we referenced their sequences against the National Centre for Biotechnology Information (NCBI, part of the International Nucleotide Sequence Database Collaboration, Cochrane et al. 2016) and the Barcode of Life Database (BOLD, Hebert & Ratnasingham, 2007) and our own curated the NZ Fish Barcode Database (Chapter Two). We were interested in the precision and accuracy of the fish species identities recovered using each of the reference databases, where we expected that the molecular identification would be more precise than the morphological identification, and that BOLD would be the best performing reference database across all species, but that our curated database might have more precise and accurate identities for a few species for which the DNA barcodes are not yet publicly available.

3.2. METHODS

3.2.1. Retrieving existing DNA barcodes for New Zealand's fishes

Fish larvae were collected at sampling sites up to 30km offshore from three North Eastern, NZ coastal locations (Whangaroa, Tutukaka and Whangamata). We used paired neuston nets (1000mm x 400mm, 3500mm long (Fig 1.), towed at 2 knots to collect replicate samples at set target water depths (nearshore at 1040m, shelf at 60-120m and deep water at 150m+). At the completion of the tow the samples from the paired nets were combined into one sample and the fishes were separated from the invertebrates (Fig 2.) and any marine debris, and stored in chilled seawater before being morphologically identified and preserved in 100% ethanol back on shore.

Figure 1. Aerial view of paired neuston nets towed at 2 knots 30 km offshore used in collection of larval fishes.

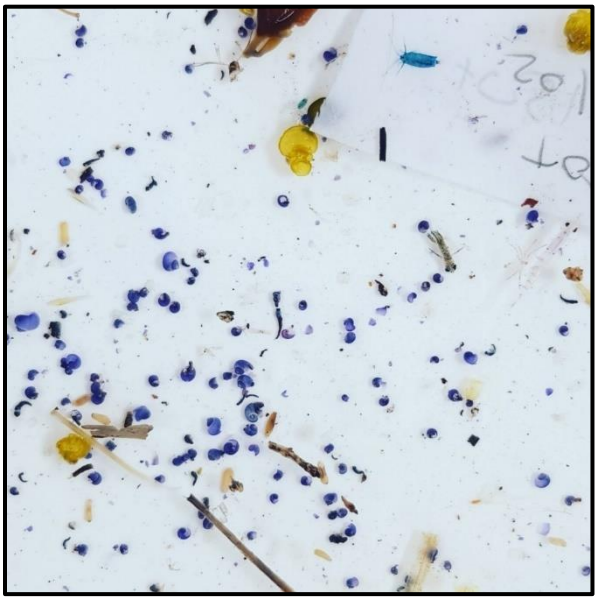


Figure 2. An example sample from our paired neuston nets comprising invertebrates, algae, land-derived debris and larval fishes.

3.2.2. Morphological Identification of Specimens

Fish specimens were morphologically identified to family, genus and species where possible using the following literature: The Fishes of New Zealand (Roberts, Stewart, & Struthers, 2015); Reef and Shore Fishes of the South Pacific (Randall, 2005) and Larvae of Temperate Australian Fishes: Laboratory guide for Larval Fish Identification (Neira, Miskiewicz, & Trnski, 1998). For those that were able to be morphologically identified to species, a level of confidence in the taxonomic identification of the specimen was assigned as: species identified with low confidence (SL), species identified with medium confidence (SM), or species identified with high confidence (SH). If a family identity was assigned and the genus and/or species identity was unknown, but specimens within the same taxonomic level could be distinguished from each other, individual specimens were given the suffix 'unknown 1', 'unknown 2' etc. following the lowest level of taxonomic classification (i.e., either family or genus). These specimens were classified as: family known, but species unknown (FU). Specimens that were unable to be taxonomically assigned to a family based on morphology were labelled as 'unknown 1', 'unknown 2' etc. to distinguish them from other specimens that were also unknown and classified as: unknown (U).

3.2.3. Molecular Sequencing of Specimens

To test the utility of molecular identification of specimens using cytochrome oxidase 1 (COI) we selected a subset of 122specimens across different tows and families for DNA extraction and analysis based on our level of confidence in their classification, including 25 U, 52 FU, 16 SL, 24 SM, and 5 SH. Where possible, several specimens representing the same taxon were selected to assess the replicability of results (including identifications of varying confidence), and specimens from different families and species to test the performance of the method across the full taxonomic diversity found in the larval tows (e.g., 11 *Aldrichetta fosteri* classified as SM, and 1 classified as SL; and 17 Tripterygiidae 'unknown 1', 'unknown 2' etc. classified as FU).

DNA was extracted from 2-5mm^3 of tissue taken from each specimen following the manufacturers protocol for either the Wizard® Genomic DNA Purification Kit or Qiagen DNeasy Blood and Tissue Kit. We amplified and sequenced a portion of the COI gene region. Specifically, ~645 bp of DNA was amplified from the 5' COI region using the COI-2 primer cocktail (Forward primers: LepF1_t1, VF1_t1, VF1d_t1, VF1i_t1, and Reverse primers: LepR1_t1, VR1_t1, VR1d_t1 and VR1i_t1, with M13 tails for

sequencing; Ivanova, Zemlack, Hanner, and Hebert 2007). PCR was conducted as per the MyTaq™ DNA polymerase kit (Bioline, Australia Pty Ltd, Alexandria, NSW, 1435) instructions using 1.5 µl of extracted DNA and a total reaction volume of 20µl, including 10µl of MyTaq Mix s (Bioline, Australia Pty Ltd, Alexandria, NSW), 0.8µl of Forward primer cocktail (10 mM), 0.8µl of Reverse primer cocktail (10 mM), and 6.9µl of Ultrapure® H_2O. If this protocol failed, PCR was conducted following the MyFi Taq DNA polymerase kit instructions using the following amounts per reaction: 1µl of extracted DNA, 4µl of Buffer solution (10 mM), 0.8µl of Forward primer cocktail (10 mM), 0.8µl of Reverse primer cocktail (10 mM), and 12.4 µl of Ultrapure® H_2O for a total reaction volume of 20µl.

PCR was performed with: an initial denaturation at 94°C for 1 min; five cycles of 94 °C for 30 s, 50°C for 40s, and 72°C for 1 min; followed by 35 cycles of 94°C for 30s, 54°C for 40s, and 72°C for 1 min; and a final extension at 72°C for 10 min. Negative controls were included in PCR to ensure there was no contamination of samples.

Following PCR, a 1% agarose gel was run using 2µl of PCR product and 1µl of GelRed, alongside a BioLabs Quick Loading DNA Ladder to ensure PCR products were of the right size and sufficient concentration. PCR products were then purified using the ExoSap reagents and protocol (Thermo Fisher Scientific, West Palm Beach, FL) and sent for forward and/or reverse sequencing (depending on quality of prior sequences)-using M13-F and M13R primers, respectively (sequenced by Macrogen, Korea).

Quality control of the received sequences was carried out using Geneious version 9.0.5. Primer sequences, stop codons, and poor-quality ends of each sequence were trimmed. Using Geneious, bases were called on a confidence scale ranging from dark blue for confidence <20 (1 in 1000/ 10^{-3} of calling a base error), blue for 20-40 and light blue for >40 (1 in 1,000,000 10^{-6} probability of calling a base error). Bases that were ranked with confidence <20, or dark blue, within the main sequence were replaced with 'N' and this was recorded within the sequence metadata. For specimens that had been sequenced in both the Forward and Reverse directions, their sequences were aligned, checked for consistency, and a consensus sequence created. To ensure that none of the sequences were derived from pseudogenes, nucleotide codons were converted into amino acids using the genetic code 'Vertebrate Mitochondrial' in Geneious version 9.0.5.

3.2.4. Molecular Identification of Sequenced Specimens

To identify what species the quality-controlled sequences corresponded to, we queried two public databases, National Center for Biotechnology Information (NCBI) and Barcode of Life Data System (BOLD), and our own database of COI sequences for NZ fishes (including several unpublished sequences, see Chapter Two). From NCBI, the top identified best match (according to percentage

pairwise identity) was recorded along with the Description, E value, Query Coverage, Pairwise Identity, Accession Number and Maximum Length of our query. From BOLD, the top recorded species was recorded along with Similarity (%). The NZ database was added to Geneious by selecting 'Add/ Remove databases' and setting up a 'Custom BLAST' service (hereafter referred to as the NZ Fish Barcode Database). Our sequences were queried against the NZ Fish Barcode Database using the Program Megablast and to produce a 'Hit table' that provided the highest percentage pairwise identity matched Hit result for each individual including: Species Name, Percentage of Identical Sites, Percentage Pairwise Identity, Date sequence was created, E Value, Grade, Min Sequence, Max Sequence, Accession, Molecule, Sequence Length, were exported into a CSV file for analysis.

In all cases: a species-level identification was only accepted if the best match was equal to, or greater than 95% pairwise identity or similarity; a genus-level identification was accepted if it was equal to, or greater than 90%; and a family-level identification was accepted if it was equal to, or greater than 85%. Although there is no universally applicable threshold for these taxonomic assignments, our threshold values were just below the upper third quartile for species threshold and fourth quartile for genera and family threshold values for pairwise COI sequence divergence between fish species, genera, and families of other studies (Kartavtsev & Lee, 2006). The threshold of 95% for species level identification was lower than several previous studies (Lakra et al., 2011; Zhang & Hanner, 2011; Kundu, Rath, Laishram, Pakrashi, Das, Tyagi, Kumar & Chandra, 2018). We elected to use a higher threshold as several species of fish that are resident in New Zealand are allopatric populations of widespread species, and because of New Zealand's oceanographic isolation, these populations may be more genetically differentiated from individuals of the same species that have been sequenced in foreign nations.

3.2.5. Qualifying Morphological and Molecular Identifications of Specimens

For each specimen, we recorded results and the performance of the molecular method at several steps, including: whether a viable sequence was attained from DNA extraction, PCR, and sequencing; the taxonomic identification based on the best matched sequence according to NCBI, BOLD, and our NZ Fish Barcode Database; whether the pairwise identity or similarity was above our stipulated threshold for taxonomic assignment at the species, genus, or family level; whether the accepted taxonomic assignment matched the morphological identification or provided more or less information (i.e. resolution) about the identity of the specimen than the morphological identification. Last, we rationalized which of the taxonomic identifications provided by NCBI, BOLD, and our own NZ Fish Barcode Database, could resolve the identification of the specimen most precisely (based on the highest percentage pairwise

identity match and therefore highest taxonomic assignment), and which was most accurate based on how true the identification was when other factors were considered, such as the known distributions of the taxonomic group, and any other detectable errors in the species assignment.

3.2.6. Statistical Analysis and Graphical Representation of Results

Analysis of our results was conducted in the R Statistic Environment Version 1.4.1106 (R Core Team, 2021) using RStudio (RStudio Team, 2021). We constructed an alluvial plot using the alluvial R package (Bojanowski & Edwards, 2016) to visualize the varying resolution of taxonomic identification of specimens (categorized as U, FU, SL, SM, SH) across the study design, based on morphological identification, and then molecular identification when querying sequences against NCBI, BOLD, and our NZ Fish Barcode Database. Venn diagrams (constructed using the ggVennDiagram R package, Chun-Hui Gao, 2021) were used to illustrate which sequence reference database/s provided the best molecular identification, and whether the best molecular identification or the morphological identification provided a higher resolution taxonomic identification for all specimens, and at what taxonomic levels (i.e., family, genus, species) the best molecular identification and morphological identification matched. To visualize the performance of molecular identification for the different categories of uncertainty in the morphological identification, we used a stacked bar chart (using the R package ggplot2, Wickham, 2016) to illustrate the taxonomic resolution of the best molecular identification for each of the categories. Finally, bubble plots (using the R package ggplot2, Wickham, 2016) were used to show the relative performance of the best molecular identification with the resolution of the morphological identification, according to the included fish families.

3.3. RESULTS

3.3.1. Morphological Identification of Specimens

A total of 122 collected larval fishes were morphologically identified and had a confidence level assigned to their identified family and species, including: 5 SH, 24 SM, 16 SL, 52 FU and 24 U (Appendix 2).

3.3.2. Molecular Identification of Specimens

Of the 122 specimens that had their DNA extracted and amplified, viable sequences were received for 99 specimens (Fig. 3, Appendix 3.1). Failed sequencing (following a positive PCR result) occurred for specimens morphologically identified as the following species (SH, SM and SL): *Scorpis violacea* (3

specimens), *Mugil cephalus, Girella tricuspidae, Gonorynchus fosteri*; families (FU): Anguillidae (2 specimens), Tripterygiidae (6 specimens), Mugilidae (2 specimens), Stigmatopora, Sygnathidae, Bramidae; and two complete unknowns (U) suggesting molecular identification may not be an easily applied method (due to poor amplification with these specific primer cocktails, and/or difficulty sequencing) for these taxonomic groups (Fig. 4A). For 35 specimens we received forward and reverse sequences. These were always in agreement. Sequences ranged from to 377bp - 729bp in length and 13 sequences retained some ambiguous nucleotide sites following quality control.

There were 99 specimens that were able to be identified using molecular methods to varying taxonomic levels. In most cases the molecular identification using NCBI, BOLD and NZ Fish Barcode Database matched specimens to the same species at the species-level (54), but in some cases they differed (Fig. 5). In 5 cases, the BOLD database provided a more precise identification (1), or a more accurate identification (4); in 1 case NCBI provided a more accurate identification; and in 22 cases the NZ Fish Barcode Database provided a more precise identification (1), or a more accurate identification (21). Precision was based on the specimens highest percentage pairwise identity for the same specimen between different reference databases and accuracy was determined based on how true the identification was when other factors were considered i.e. mislabelling of sequences in a database made specimen identification lack accuracy

3.3.3. Qualifying Morphological and Molecular Identifications of Specimens

In 33 instances the best molecular identification of the specimen matched the morphological identification at species level, and was of the same taxonomic precision (Fig. 6). Of these molecularly verified identifications there were 3 nominated to have SH confidence in their morphological identification, 22 of SM, 5 of SL, 2 of FU, and 1 of U. In 58 instances the morphological identification of the taxa was less precise than the molecular identification (34 of FU, and 22 of U) or less accurate (1 of SH, 1 of SL) and in 8 instances morphological identification was more precise than molecular identification (2 of FU, and 6 of SL). In many cases there was a disagreement between the morphological and molecular identifications (42 total; 2 of SH, 1 of SM, 1 of SL, 30 of FU; note that Unknowns, U, were not considered) and in some cases it was unknown which was more accurate (8 of SL). However, in 34 cases it was verified that the molecular identification was more accurate (using other supporting information; 1 of SH, 1 of SL and 33 of FU), and in one case the molecular identification was deemed to be dubious, likely owing to an erroneously labelled sequence, or contaminated sequence uploaded to BOLD (1 of FU). Overall, the NZ Fish Barcode Database produced the most accurate and precise identification for specimens (Fig. 7).

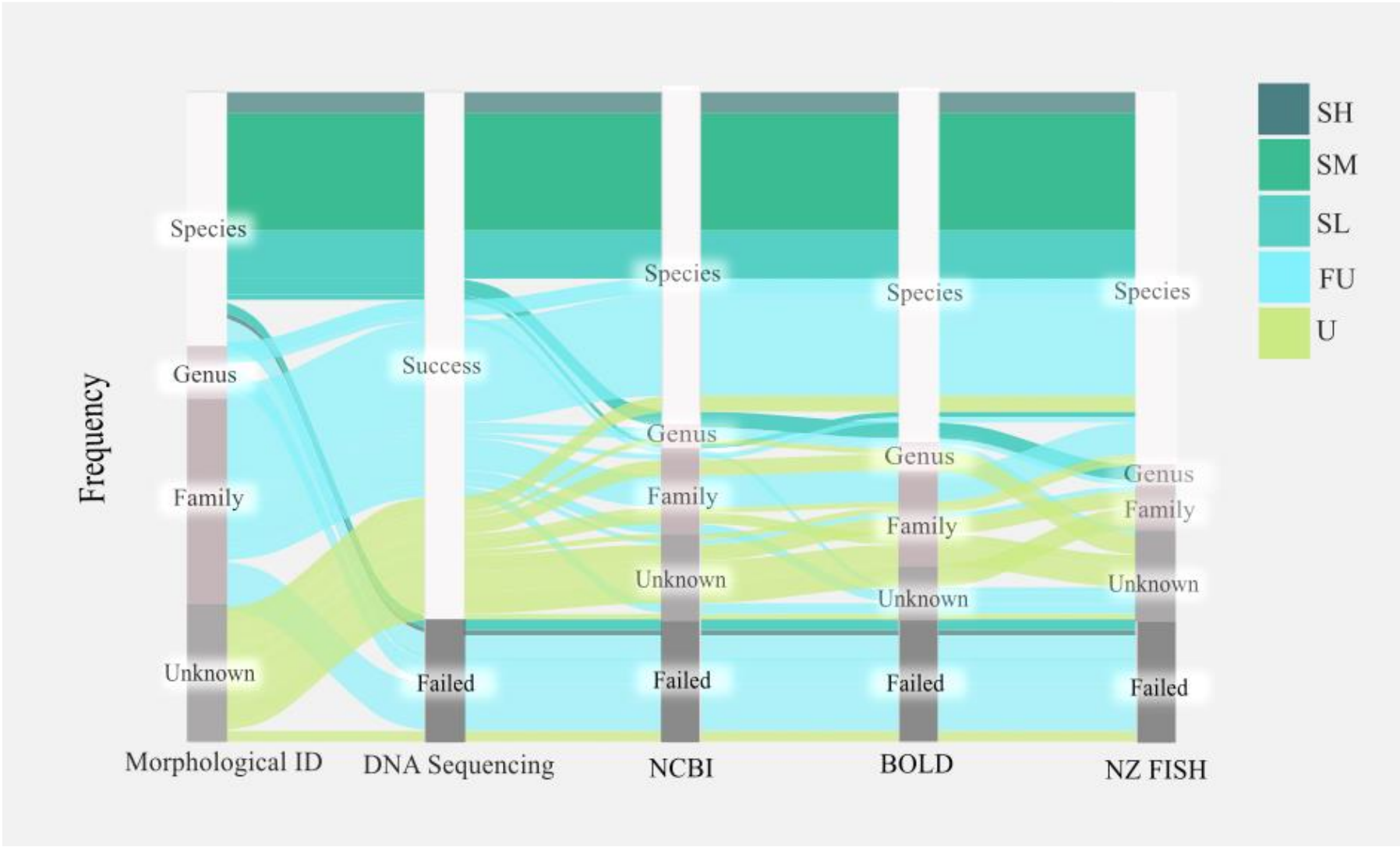


Figure 3. The proportion of individuals within each confidence level of morphological identification (see colour key for SH, SM, SL, FU and U), what taxonomic level they were assigned based on morphology (leftmost stacked bar in greyscale), the outcome of DNA Sequencing (Failed or Success), and what taxonomic level they were assigned to based on the molecular databases: NCBI, BOLD and our NZ Fish database.

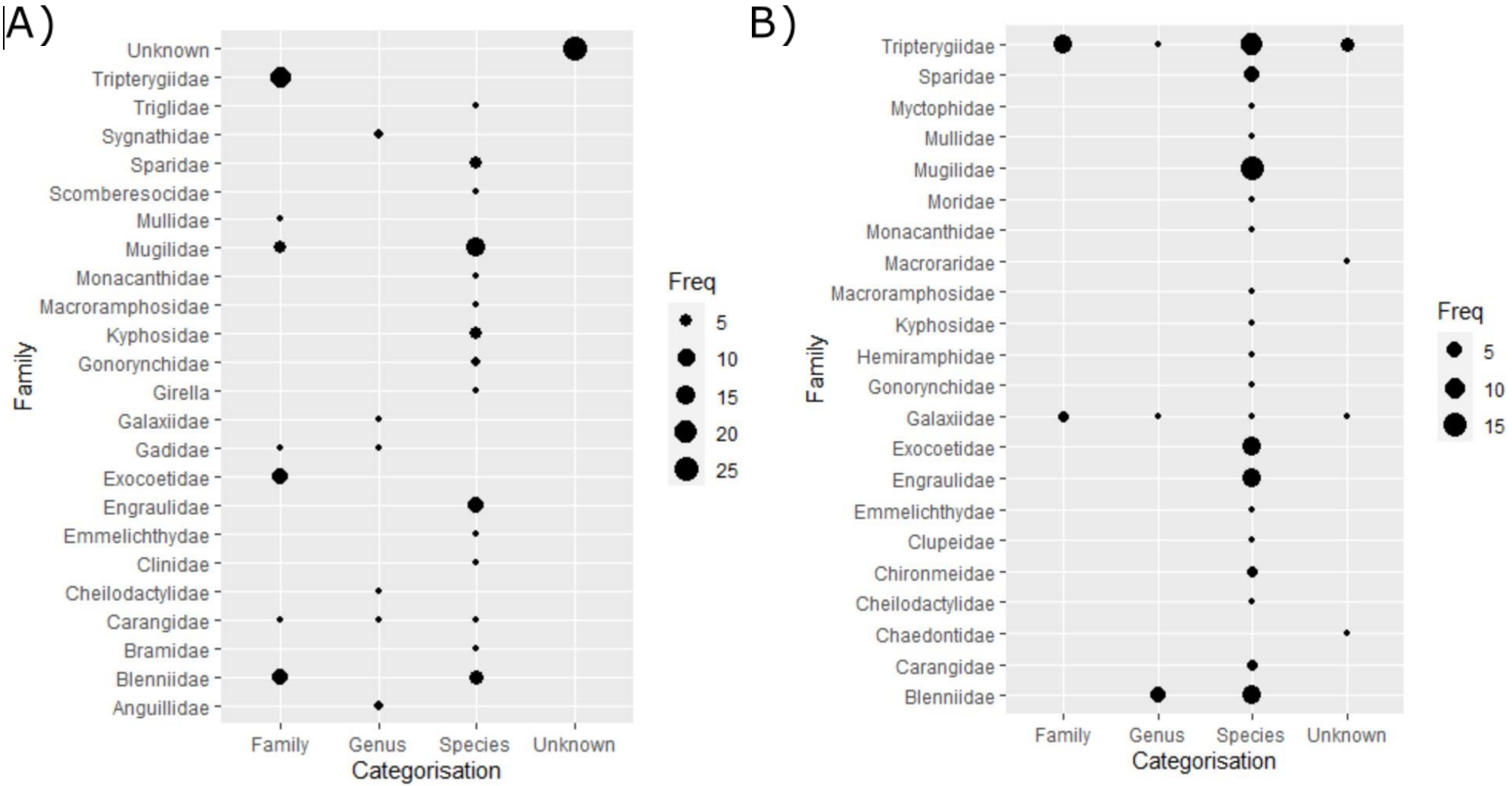


Figure 4. A) The frequency of morphologically identified individuals classified to each taxonomic level, according to family. "Unknown" were species unable to be classified to the family-level based on morphology. B) The frequency of individuals molecularly identified to each taxonomic level, according to family. "Unknown" were individuals that could not be confidently classified to family molecularly (i.e. <85% pairwise identity). Note that 2 specimens of Anguillidae, 6 specimens of Tripterygiidae, 2 specimens of Mugilidae and one specimen from Stigmatopora, Sygnathidae, Bramidae and two complete unknowns could not be successfully DNA sequenced and are not presented here.

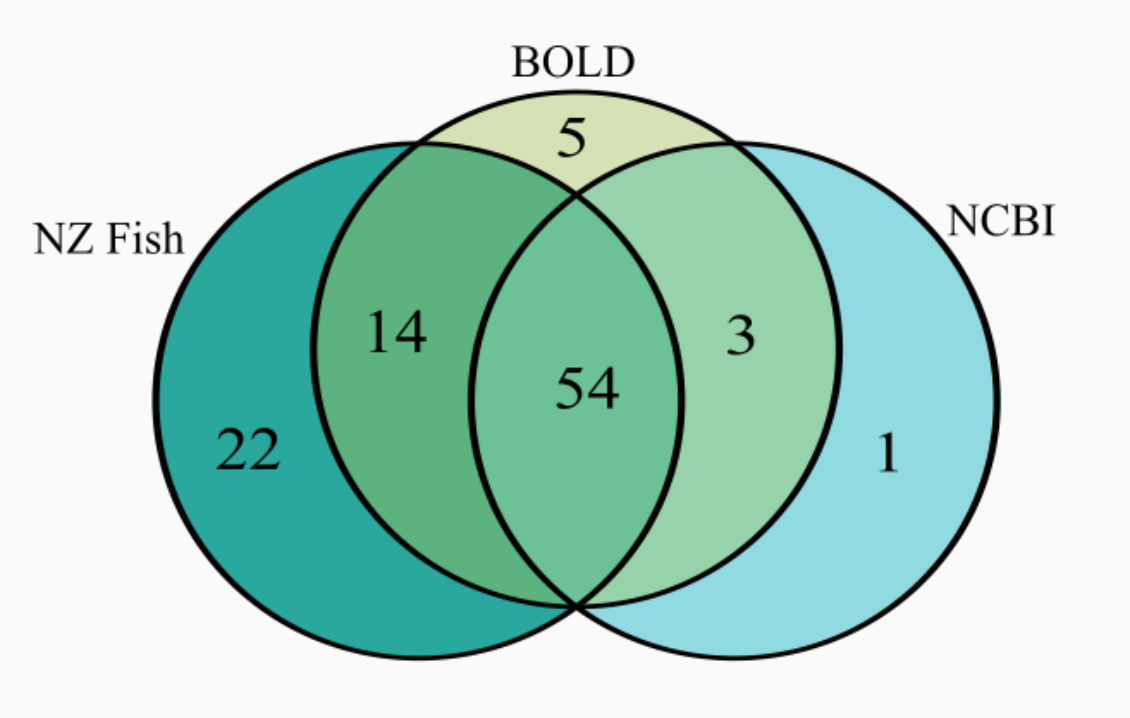


Figure 5. The number of specimens (n=99) most precisely and accurately identified by each of the three sequence reference databases: NCBI, BOLD and NZ Fish

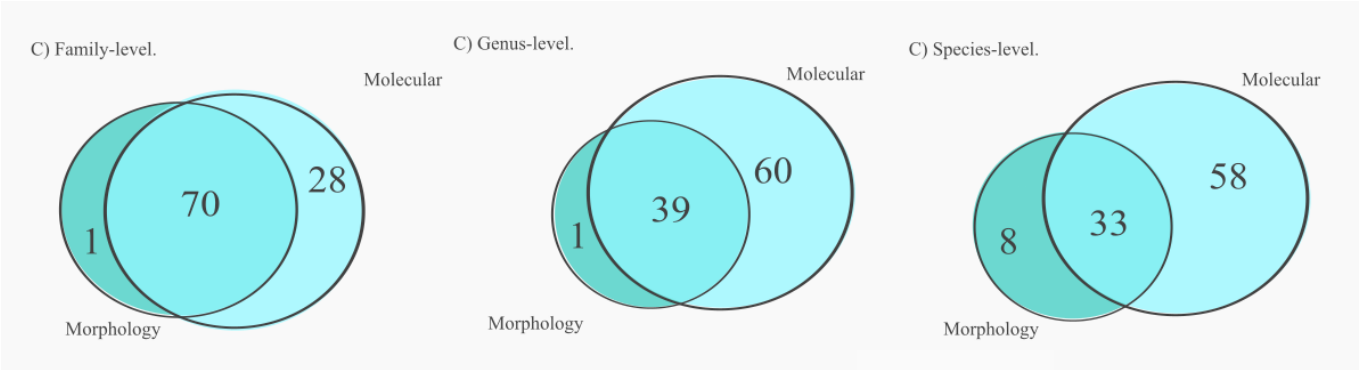


Figure 6. The number of individuals that were most accurately identified based on morphology, based on their DNA barcode ("Molecular"), or equally by both methods of identification at: A) Family-level, B) Genus-level, C) Species level

Figure 7. Number of individuals in each category of morphological classification confidence (SH, SM, SL, FU, and U) that received the highest Percentage Pairwise Identity to their reference barcode for each sequence reference database: NCBI, BOLD and NZ Fish. Individuals for which all three databases gave an equal Percentage Pairwise Identity result were represented as 'Same', and 'Excluded' includes those individuals for which a barcode could not be generated.

3.5. DISCUSSION

In this study, we strategically sampled larval fishes collected off the North East Coast of the North Island of NZ, morphologically identified each specimen, and then genetically extracted, amplified and sequenced the COI Barcode region of each specimen. The genetic barcode was then compared against our NZ Fish Barcode Database as well as public databases BOLD and NCBI. We were interested in whether molecular identification of collected larval fishes may be more precise than morphological identifications (i.e., resolved to a lower taxonomic level), and which sequence reference database provided the most precise and accurate identification. We found that molecular methods enabled the identification of individuals with an unknown taxonomic identity and increased precision in taxonomic identification of specimens only morphologically identified to the family-level.

Overall, the combined use of morphological identification and molecular identification provided the highest number and resolution taxonomic identifications in our study. We found that molecular methods had a greater identification precision at the genus and species level but similar identification precision at family level, except for specimens that were unable to be morphologically identified at any taxonomic level(complete Unknowns (U)) which were better identified molecularly (Fig. 3shows the progression of identification for Unknown specimens to a more accurate taxonomic identification and Fig. 4 illustrates the increase in taxonomic identification within different families and the decrease in specimens that had unknown classification in molecular identification compared to morphological identification. Previous studies have similarly found that correct morphological identification of larval fishes to the family-level ranges from 70% (Rathnasuriya et al., 2021) to 80% occurrence (Ko et al., 2013), (with our data finding 70.1% correct morphological identification at family level) and that the correct identification based on morphology alone is poorer at the species and genus level (Isari et al., 2017; Ko et al., 2013). This suggests that molecular methods are a more reliable means of identification for clarifying an individual's taxonomy with higher resolution than family-level, and a combination of morphology and molecular identification at the family-level is best for total accuracy and precision in identification.

A large proportion of individuals were molecularly identified as the same species among the three sequence reference databases. However, where identities differed, the NZ Fish Barcode Database provided the most accurate identification compared to the other two public databases, NCBI and BOLD. The accuracy of the identification took into consideration the percentage pairwise identity match, as well as other knowledge regarding the likely spatial distribution of nominated species. Hence, the better performance on the NZ Fish Barcode Database was likely due to the preferential retrieval of barcodes generated for NZ fish species, based on NZ specimens, in the creation of the NZ Fish Barcode Database. Having a specifically curated reference database has the advantage of

avoiding potential biases in determination of species and allowing for future unbiased comparisons between studies (Arranz, Pearman, Aguirre, & Liggins, 2020).

Open access databases can contain multiple instances of the same species that may be similar or identical but can also contain inconsistencies and errors that can impact study results (Chen, Zobel, & Verspoor, 2017). For example, the impact of duplicates and errors in NCBI was recently discussed in the systematics of Damselfishes where it is likely sequences from erroneously identified specimens have mislead the inferences of phylogenetic relationships in this group of fishes (Tang, Stiassny, Mayden, & DeSalle, 2021). Within our findings we were able to identify a potential error within the BOLD database. For one specimen, it's highest pairwise identity match in NCBI was *Mendosoma linneatum* (85.70%), whereas in BOLD and the NZ Fish Barcode Database the closest pairwise identity match was *Chironemus marmoratus*. (100% and 99.80%, respectively). Although NCBI has a COI sequence putatively from *Cheilodipetrus macrodon*, this was not the highest matched percentage pairwise identity which suggests that the COI sequence for *Cheilodipetrus macrodon* may be erroneously labelled in that database and a correctly labelled COI sequence is required for *Cheilodipetrus macrodon* in NCBI. As well as potential misidentification of study species, other potential issues in public sequence repositories include human error in annotating gene sequences upon upload, and uploading of sequences that have resulted from unidentified cross-contamination during laboratory work. Our study showed the value of having a curated database, NZ Fishes Barcode Database, based on voucher specimens or specimens identified by an expert taxonomist, for highest accuracy in identification in comparison to online genetic repositories, NCBI and BOLD (Fig. 5) for our specimens.

The sequence reference databases performed similarly in supporting or supplementing the morphological identities, regardless of the confidence in those identifications (Fig. 5). Specifically, for individuals morphologically identified to the species-level with low, medium, or high confidence, all three databases performed relatively equally. For larval fish specimens completely unknown based on morphological identification, BOLD was the most useful sequence reference database for resolving the specimen's taxonomic identity. For those with Family identified only based on morphology, the NZ Fish Barcode Database was particularly helpful.

In total there were 22 different families and 28 different species identified using molecular methods as compared to 23 families, and 15 species based on the morphological identification. Such increased precision of molecular identifications has also been seen in previous studies of larval fishes' identification (Ayala, Riemann, & Munk, 2016; Isari et al., 2017; Ko et al., 2013; Rathnasuriya et al., 2021). Using genetic identification in our study, taxonomy was resolved to the species level for all but five families: Blennidae, Macroraridae, Galaxiidae, Chaedontidae and Tripterygiidae (Fig. 6). Lack of taxonomic resolution using genetic methods has historically been an issue for Tripterygiidae due to an

absence of accumulated COI data for this family (Ko et al., 2013). It is possible that Blennidae, Macrouridae, Chaedontidae Galaxiidae also lack resolution to species level due to absence of COI barcode references in this family that are publicly available. Future research would involve continuing to update COI reference sequences for species and families, in particular Blennidae, Galaxiidae, Macroraridae, Chaedontidae and Tripterygiidae where COI reference sequences are lacking in the NZ Fish Barcode Database as discussed in Chapter Two.

In a number of cases, specimens were incorrectly genetically identified as their sister taxa. This was the case for 14 specimens that were morphologically identified as *Parablennius laticlavius* and genetically identified as *Parablennius tasmanianus* (~94% pairwise identity in BOLD and NZ Fish databases) which has not previously been recorded in NZ. Two congenerics for *P. tasmanianus* are found in NZ; *Parablennius laticlavius* is native to North Eastern NZ and *Parablennius intermedius* has recently been recorded in the region, and is likely to be a human-mediated introduction (Middleton et al. 2021). Although there is not yet an available COI barcode sequence for *P. laticlavius,* there is a reference barcode for *P. intermedius* to which these specimens did not match, meaning these 14 specimens are indeed likely to be *P. laticlavius.* In another example, NCBI and BOLD returned the identification of eight specimens as *Engraulis japonicus* (99.0% and 100%, respectively) and two specimens as *Trachurus japonicus* (99.83% and 98.69%, respectively) despite these species not being found in NZ. The NZ Fish Barcode Database returned identification for the same specimens as *Engraulis australis* (98.40% pairwise identity) and *Trachurus novaezelandiae* (99.80% pairwise identity) which were determined to be the correct species due to knowledge of species distribution in NZ. This suggests that the taxonomic assignment for these species may need to be revisited due to not being significantly genetically distinguishable, or that the COI gene region lacks sufficient variation to distinguish them.

There were 23 specimens that failed to amplify or sequence from nine different families: Anguillidae, Tripterygiidae, Kyphosidae, Triglidae, Mugilidae, Stigmatopora, Gonorynchus, Sygnathidae, and Bramidae (and two specimens of unknown family affinity). To gain a molecular identification for these specimens may require the use of a different DNA extraction protocol, or different set of PCR primers. Several other primer sets are used for fishes, for example alternative primer cocktails COI-1 or COI-3 or 16S primers (Ivanova, Zemlack, Hanner, & Hebert, 2007) or the primers that target the 12S region of fishes such as MiFish- U and MiFish-E (Miya et al., 2015). Although a primer cocktail was used in our study to alleviate issues associated with primer specificity across a broad taxonomic group, future studies might consider using several primers sets and targeted gene regions. This would require the curation of sequence reference databases comprising multiple gene regions in addition to COI.

Overall, this study found that identification of individuals was enabled by molecular methods for specimens with unknown taxonomic identity and increased precision in taxonomic identification of specimens that could only be morphologically identified to family level. Further additions of accurate COI sequences to NCBI, BOLD and NZ Fish Barcode Database will return a higher degree of species ID accuracy for future studies. Taxonomic assignment of some species should be revisited as well as erroneously labelled sequences in NCBI.

CHAPTER FOUR: GENERAL DISCUSSION

4.0. DEFINITIONS OF BIODIVERSITY

This thesis aimed to create a New Zealand (NZ) Fish Barcode Database for Actinopterygii fishes and to use this database in a case study involving the taxonomic identification of New Zealand larval fishes. In my second chapter, I compiled the NZ Fish Barcode Database. This was generated from newly generated sequences, previously published sequences from specimens identified by expert taxonomists or from curated collections, and the retrieval of sequences from open-access repositories (i.e., NCBI and BOLD). Overall, the database represented 965 of the 1320 fish species found in NZ. In my third chapter, I generated COI barcodes for larval fishes collected off the East Coast of the North Island and compared their genetic and morphological identification. I observed that the NZ Fish Barcode Database was most accurate for genetic identification of fishes, and that genetic identification was more accurate in the identification of specimens to species and genus level than identification based on morphology. Overall, the NZ Fish Barcode Database is the most current genetic database for NZ fishes and is a practical resource of value for future environmental DNA studies, biodiversity monitoring, and managing fisheries and commercial fisheries-derived products.

4.1. CHAPTER TWO CONCLUSIONS

The objective of Chapter Two was to create a DNA barcode database for New Zealand's marine and freshwater Actinopterygii based on the Cytochrome Oxidase I (COI) gene region. This was aided by collaboration with the nation's premiere fish taxonomists and access to Te Papa Tongarewa, National Fish Tissue Collection, as well as Auckland Museum's tissue collection. The resulting DNA barcode reference database represented 965 of the 1320 fish species found in New Zealand. This was produced utilising novel sequences generated for this study; sequences that had been previously published by expert taxonomists or from curated collections; and the retrieval of sequences from open-access repositories (i.e., NCBI and BOLD). As a result, I have compiled and produced the NZ Fish Barcode Database: the most comprehensive, quality assured and current reference database for New Zealand fishes to date.

The NZ Fish Barcode Database includes 73% of the known fish species in New Zealand. Within the NZ Fish Barcode Database, 40% were specimens identified by an expert taxonomist or from curated collections. Most of the sequences were gained from open-access repositories, with specimen barcodes sourced and published closer to New Zealand waters preferentially being retrieved (utilising

the regPhylo R package, Eme et al. 2019). This resulted in the majority of sequences being retrieved from Australian waters with which we share a large proportion of our marine fish fauna. In executing our approach to creating a curated DNA barcode reference database –utilising sound species identification, vouchered specimens, and geographically relevant populations – we have provided a pathway for future studies and enhancement of this DNA sequence resource.

This study identified a lack of COI sequences for 355 fish species from 45 families that need to be generated in order to have a complete NZ Fish Barcode Database. Several species remain unrepresented in biodiversity studies with a lack of sequences due to the most common fish primers (such as the primers used in Chapter Three) not working well for those species (Ivanova et al. 2007). It may also be due to their absence of specimens in curated collections or that the successful extraction and amplification of their DNA relies upon specialist biochemical or mechanical treatments (Zhang & Hewitt, 1998). To help increase the representation of these missing species within the NZ Fish Barcode Database, targeted collection efforts to attain elusive species, and use of more specialized laboratory protocols for DNA extraction and taxon-specific primers for amplification may be required.

In addition to identifying which species require sequence generation, we identified fifteen species pairs that were 0% genetically divergent and may require further investigation. In these cases, it is possible that the specimen sequenced was misidentified, there was a tissue sample mix-up, cross-contamination in the laboratory, or there has been taxonomic over-splitting of the group (Landi et al., 2014). For example, both *Polymixia japonica* and *P. busakhini* are on the List of New Zealand Fishes but have 0% sequence divergence between them. It may be that geographically independent populations (in the Northern and Southern Hemispheres) were described separately, but may in fact be the same species. It is also known that the CO1 gene region is not effective in capturing the known taxonomic distinction among taxa. For example, in several Chondrichthyes groups, species distinction is often not supported by divergence at CO1, despite other compelling evidence for species distinction (Naylor et al., 2012; Wong, Shivji, & Hanner, 2009). Accordingly, in our NZ Fish Barcode Database *Carcarhinus obscurus* and *C. galapagensis* were indistinguishable based on COI. In Actinopterygians, we also found the very recently diverged species of *Chrysiptera notialis* and the Kermadec *Chrysiptera* demoiselle (formerly *C. rapanui* and now considered an undescribed species) were indistinguishable based on CO1, despite them being distinguishable based on morphology and the mitochondrial control region (Liggins et al. 2021; L. Liggins pers. comm.). In future, communication among institutions who hold collections and further sequencing of vouchered specimens will help resolve any of these issues caused by the

misidentification of specimens, cross-contamination, or taxonomic issues. In addition, and particularly in the cases of recently diverged taxa, the addition of other gene regions as "barcodes" would also be beneficial (e.g., 12S, Milan et al., 2020).

4.2. CHAPTER THREE CONCLUSIONS

Chapter Three investigated the use of morphological and molecular methods in the identification of juvenile fishes. Morphological identification and molecular identification in combination provided the most accurate taxonomic identifications in our study. A greater taxonomic identification precision was provided via molecular identification at the genus and species level but similar identification precision as morphological identification at the family level, except in the case for specimens that had completely unknown higher-level taxonomic affinities which were better identified molecularly. Previous studies have similarly found that correct morphological identification of larval fishes to the family-level ranges from 70% (Rathnasuriya et al., 2021) to 80% (Ko et al., 2013), and that the correct identification based on morphology alone is poorer at the species and genus level (Isari et al., 2017; Ko et al., 2013). This suggests that molecular methods are a more reliable means of identification for clarifying an individual's taxonomy with higher resolution than family-level, and a combination of morphology and molecular identification at family level is best for total accuracy and precision in identification.

A large majority of individual specimens were molecularly identified as the same species over the three databases (NCBI, BOLD and the NZ Fish Barcode Database). However, where identities differed, the NZ Fish Barcode Database provided the most accurate identification compared to the other two public databases. Determining the accuracy of identification took into consideration the percentage pairwise identity match, as well as the likely spatial distribution of species. Hence, the better performance on the NZ Fish Barcode Database was likely due to the barcode retrieval of preferentially generated New Zealand fish species in the creation of the NZ Fish Barcode Database. Having a specifically curated reference database has the advantage of avoiding potential biases in determination of species and allowing for future unbiased comparisons between studies (Arranz, Pearman, Aquirre, & Liggins, 2020).

4.3. LIMITATIONS OF THE RESEARCH

Although there was a high-standard of quality control applied to the data where possible, the data and results presented in this thesis have some limitations. First, despite attempting to retrieve all spatial and geographic reference information for sequences included in the generation of the NZ Fish Barcode Database, there is still a notable lack of geographic location information for many of the

retrieved sequences. Lack of geographic information in public databases is a growing issue in molecular ecology. A large proportion of published sequences have inaccurate or no geolocation information (Pope, Liggins, Keyse, Carvalho, and Riginos, 2015; Toczydlowski et al., 2021). This includes imprecise geolocation, when a location name has not been provided, or when the centroid of the location is not in the ocean from where a specimen was retrieved, and a generic geolocation is given such as a country name. A lack of geographic references hinders our ability to preferentially retrieve sequences that are from New Zealand, or nearby locations.

Secondly, the utilisation of the COI Barcode as opposed to the 12S barcode region, or several gene regions, does not enable our reference database to be as universally applicable across Actinopterygii, and to future studies. For instance, the 12S barcode is a region of the 12S RNA gene (Gold et al., 2021) that is commonly used in eDNA studies (Miya et al., 2015; Shu, Ludwig, & Peng, 2021). This is due to ribosomal DNA having more conserved regions and priming sites compared to protein coding genes (such as COI) which results in their amplification detecting a wider proportion of target species and increased specificity in primers (Bylemans, Gleeson, Hardy, & Furlan, 2018; Collins et al., 2019). However, 12S regions can be susceptible to detection biases and do not produce the same level of taxonomic resolution as the COI gene (Duke & Burton, 2020; Miya et al., 2015). Also, the 12S fish reference database is fairly incomplete in comparison to the COI database for fishes (Ardura, Planes, & Garcia-Vazquez, 2013; Duke & Burton, 2020; Ward, Hanner, & Hebert, 2009). Ideally, a national reference sequence database would comprise all gene regions (and even whole genomes) generated from expert identified, and vouchered specimens.

4.4. POTENTIAL FUTURE DIRECTIONS OF THE RESEARCH

The next steps to ensure maximum reliability of the NZ Fish Barcode Database is to generate sequences for the remaining 355 species without a COI sequence in the database, where possible. The database has identified which species still require generation of sequences and where species may require taxonomic reassessment. The inaccurate collection coordinates of many species also provide an incentive to regenerate sequences from New Zealand caught specimens with accurate recording of specimen geolocation. In conjunction with completing the database it is also important that the database is continually maintained to include any new species habituating in New Zealand or update any species that have been taxonomically reallocated or renamed. Overall, a completed database could be used to not only identify individual species, but also reference barcodes and individuals from specific geolocations.

The creation of the NZ Fish Barcode Database provides a New Zealand specific repository against which species can be identified when utilising environmental DNA in local studies, benefiting future research within this field. Fish biodiversity around NZ would be better quantified by utilising eDNA

i.e., detecting pest fishes or fishes extending their range to NZ waters. Many studies involving the use of eDNA have been used in order to identify species and protect biodiversity. Expanding the NZ Fish Barcode Database to include more barcode regions, or entire genomes, would benefit a greater range of eDNA studies beyond the commonly used COI barcode region.

Following the generation and updates of the NZ Fish Barcode Database there is potential to further expand the database to include a greater range of barcode regions. This may include generating and compiling DNA Barcodes from ribosomal 12S and 16S gene regions (Milan et al., 2020). Increasingly, eDNA studies have been utilising 16S barcode regions to genetically identify fishes in a sampled environment (Jeunen et al., 2019) which accompanied with COI barcode regions would assist a greater range of future research. In addition to other barcode regions, efforts have been made to generate whole organelle genomes of species, known as 'genome skimming', to be used in species identification (Bohmann, Mirarab, Bafna, & Gilbert, 2020; Coissac, Hollingsworth, Lavergne, & Taberlet, 2016). In order to obtain useful information from 'genome skimming' DNA genome reference databases are required (Coissac et al., 2016) and would be a logical and beneficial next step to further develop the NZ Fish Barcode Database to include organelle genomes in addition to ribosomal and mitochondrial barcodes.

REFERENCES

Aguilar, R., Ogburn, M. B., Driskell, A. C., Weigt, L. A., Groves, M. C., & Hines, A. H. (2016). Gutsy genetics: identification of digested piscine prey items in the stomach contents of sympatric native and introduced warmwater catfishes via DNA barcoding. *Environmental Biology of Fishes, 100*(4), 325-336. doi:10.1007/s10641-016-0523-8

Ahmed, M. S., Datta, S. K., Saha, T., & Hossain, Z. (2021). Molecular characterization of marine and coastal fishes of Bangladesh through DNA barcodes. *Ecology and Evolution, 11*(9). doi:10.1002/ece3.7355

Aldhebiani, A. Y. (2018). Species concept and speciation. *Saudi Journal of Biological Sciences, 25*(3), 437-440.

Antonio, J., & Pineda, C. (2007). Larval Transport and Dispersal in the Coastal Ocean and Consequences for Population Connectivity. *Oceanography, 20*(3), 22-36.

Ardura, A., Planes, S., & Garcia-Vazquez, E. (2013). Applications of DNA barcoding to fish landings: authentication and diversity assessment. *Zookeys* (365), 49-65. doi:10.3897/zookeys.365.6409

Arranz, V., Pearman, W. S., Aguirre, J. D., & Liggins, L. (2020). MARES, a replicable pipeline and curated reference database for marine eukaryote metabarcoding. *Scientific Data, 7*(1), 209. doi:10.1038/s41597-020-0549-9

Ayala, D., Riemann, L., & Munk, P. (2016). Species composition and diversity of fish larvae in the Subtropical Convergence Zone of the Sargasso Sea from morphology and DNA barcoding. *Fisheries Oceanography, 25*(1), 85-104. doi:10.1111/fog.12136

Baker, C. S., Cooke, J. G., Lavery, S., Dalebout, M. L., Ma, Y. U., Funahashi, N., Carraher, C., Brownell, R. L. (2007). Estimating the number of whales entering trade using DNA profiling and capture-recapture analysis of market products. *Molecular Ecology, 16*(13), 2617-2626. doi:10.1111/j.1365-294X.2007.03317.x

Batta-Lona, P. G., Galindo-Sanchez, C. E., Arteaga, M. C., Robles-Flores, J., & Jimenez-Rosenberg, S. P. A. (2019). DNA barcoding and morphological taxonomy: identification of lanternfish (Myctophidae) larvae in the Gulf of Mexico. *Mitochondrial DNA A, 30*(2), 375-383. doi:10.1080/24701394.2018.1538364

Bohmann, K., Mirarab, S., Bafna, V., & Gilbert, M. T. P. (2020). Beyond DNA barcoding: The unrealized potential of genome skim data in sample identification. *Molecular Ecology, 29*(14), 2521-2534. doi:10.1111/mec.15507

Bojanowski M, Edwards R (2016). _alluvial: R Package for Creating Alluvial Diagrams_. R package version: 0.1-2, <URL: https://github.com/mbojan/alluvial>.

Bradbury, I. R., & Snelgrove, P. V. R. (2001). Contrasting larval transport in demersal fish and benthic invertebrates: the roles of behaviour and advective processes in determining spatial pattern. *Canadian Journal of Fisheries and Aquatic Sciences, 58*(4), 811-823. doi:10.1139/f01-031

Bylemans, J., Gleeson, D. M., Hardy, C. M., & Furlan, E. (2018). Toward an ecoregion scale evaluation of eDNA metabarcoding primers: A case study for the freshwater fish biodiversity of the Murray-Darling Basin (Australia). *Ecology and Evolution, 8*(17), 8697-8712. doi:10.1002/ece3.4387

Chen, Q., Zobel, J., & Verspoor, K. (2017). Duplicates, redundancies and inconsistencies in the primary nucleotide databases: a descriptive study. *Database (Oxford), 2017*. doi:10.1093/database/baw163

Chun-Hui Gao (2021). ggVennDiagram: A 'ggplot2' Implement of Venn Diagram. R package version 1.1.2. https://github.com/gaospecial/ggVennDiagram

Cochrane, G., Karsch-Mizrachi, I., Takagi, T., & Database, T. I. N. S. (2016). The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Research, 44*(1), 48-50. doi:10.1093/nar/gkv1323

Coissac, E., Hollingsworth, P. M., Lavergne, S., & Taberlet., P. (2016). From barcodes to genomes: extending the concept of DNA barcoding. *Molecular Ecology, 25*, 1423 - 1428.

Colinvaux, P. (1986). *Ecology*. New York, USA: John Wiley and Sons.

Collins, R. A., Bakker, J., Wangensteen, O. S., Soto, A. Z., Corrigan, L., Sims, D. W., Genner, M. J., Mariani, S., Yu, D. (2019). Non-specific amplification compromises environmental DNA metabarcoding with COI. *Methods in Ecology and Evolution, 10*(11), 1985-2001. doi:10.1111/2041-210x.13276

Convention on International Trade in Endangered Species of Wild Fauna and Flora. (2021). Retrieved from https://cites.org/eng.

Conway, K. W., Stewart, A. L., & Summers, A. P. (2018). A new genus and species of clingfish from the Rangitahua Kermadec Islands of New Zealand (Teleostei, Gobiesocidae). *Zookeys*(786), 75-104. doi:10.3897/zookeys.786.28539

Costa, F.O., Landi, M., Martins, R., Costa, M.H., Costa, M.E., Carneiro, M., Alves, M.J., Steinke, D., & Carvalho, G.R. (2012). A ranking system for reference libraries of DNA barcodes: application to marine fish species from Portugal. *PloS ONE, 7*(4), 1-9. doi:10.1371/journal.pone.0035858

Costello, M. J., Claus, S., Dekeyzer, S., Vandepitte, L., Tuama, E. O., Lear, D., & Tyler-Walters, H. (2015). Biological and ecological traits of marine species. *PeerJ, 3*, 1-29. doi: 10.7717/peerj.1201

Costello, M. J., Coll, M., Danovaro, R., Halpin, P., Ojaveer, H., & Miloslavich, P. (2010). A census of marine biodiversity knowledge, resources, and future challenges. *PLoS One, 5*(8), 1-15. doi: 10.1371/journal.pone.0012110

Dasmann, R. F. (1968). *A different kind of country*. New York: Macmillian.

De Battisti, C., Marciano, S., Magnabosco, C., Busato, S., Arcangeli, G., & Cattoli, G. (2014). Pyrosequencing as a tool for rapid fish species identification and commercial fraud detection. *Journal of Agricultural and Food Chemistry, 62*(1), 198-205.

De Brauwer, M., Hobbs, J. A., Ambo-Rappe, R., Jompa, J., Harvey, E. S., & McIlwain, J. L. (2018). Biofluorescence as a survey tool for cryptic marine species. *Conservation Biology, 32*(3), 706-715. doi: 10.1111/cobi.13033

de Carlos, A., Bañón, R., Cobo-Arroyo, S., Arronte, J. C., del Río, J. L., & Barros-García, D. (2020). DNA barcoding flags the existence of sympatric cryptic species in the slender codling Halargyreus johnsonii Günther, 1862 (Gadiformes, Moridae). *Marine Biodiversity, 50*(4). 1-7. doi:10.1007/s12526-020-01074-8

Dejean, T., Valentini, A., Miquel, C., Taberlet, P., Bellemain, E., & Miaud, C. (2012). Improved detection of an alien invasive species through environmental DNA barcoding: the example of the American bullfrog, *Lithobates catesbeianus*. *Journal of Applied Ecology, 49*(4), 953-959. doi:10.1371/journal.pone.0023398

DeLong, D.C. Jr. (1996). Defining Biodiversity. *The Wildlife Society, 24*(4).

Delrieu-Trottin, E., Liggins, L., Trnski, T., Williams, J. T., Neglia, V., Rapu-Edmunds, C., Planes, S., Saenz-Agudelo, P. (2018). Evidence of cryptic species in the blenniid *Cirripectesal boapicalis* species complex, with zoogeographic implications for the South Pacific. *Zookeys* (810), 127-138. doi:10.3897/zookeys.810.28887

DeSalle, R., & Amato, G. (2004). The expansion of conservation genetics. *Nature Reviews Genetics, 5*(9), 702-712.

Duke, E. M., & Burton, R. S. (2020). Efficacy of metabarcoding for identification of fish eggs evaluated with mock communities. *Ecology and Evolution, 10*(7), 3463-3476. doi:10.1002/ece3.6144

Eme, D., Anderson, M. J., Myers, E. M. V., Roberts, C. D., & Liggins, L. (2020). Phylogenetic measures reveal eco-evolutionary drivers of biodiversity along a depth gradient. *Ecography, 43*(5), 689-702. doi:10.1111/ecog.04836

Eme, D., Anderson, M. J., Struthers, C. D., Roberts, C. D., Liggins, L., & Davies, J. (2019). An integrated pathway for building regional phylogenies for ecological studies. *Global Ecology and Biogeography, 28*(12), 1899-1911. doi:10.1111/geb.12986

Felsenstein, J. (1988). Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics, 19*, 445-471.

Fost, B., Morey, K., Ferguson, B., Bourque, D., Naaum, A., Bradley, D., & Hanner, R. (2020). Rapid cooling via dry ice preserves the genetic and morphological integrity of fish embryos. *Journal of Fish Biology, 96*(3), 820-824. doi:10.1111/jfb.14248

Francis, M. (2019). *Checklist of the coastal fishes of Lord Howe, Norfolk and Kermadec Islands, southwest Pacific Ocean.* 10.6084/m9.figshare.c.4428305.

Frankham, R. (2005). Genetics and Extinction. *Biological Conservation, 126*(1), 131-140.

Franklin, P. A., Smith, J., Baker, C. F., Bartels, B., & Reeve, K. (2015). First observations on the timing and location of giant kōkopu (*Galaxias argenteus*) spawning. *New Zealand Journal of Marine and Freshwater Research, 49*(3), 419-426. doi: 10.1080/00288330.2015.1045004

Gamefeldt, L., Hillerbrand, H., & Jonsson, R. (2008). Multiple functions increase the importance of biodiversity for overall ecosystem functioning. *Ecology, 89*(5), 1223-1231.

Gold, Z., Curd, E. E., Goodwin, K. D., Choi, E. S., Frable, B. W., Thompson, A. R., Walker, H. J. Jr., Burton, R.S., Kacev, D., Martz, L. D., Barber, P. H. (2021). Improving metabarcoding taxonomic assignment: A case study of fishes in a large marine ecosystem. *Molecular Ecology Resource*, 21(7), 1-19. doi:10.1111/1755-0998.13450

Gordon, D. P., Beaumont, J., MacDiarmid, A., Robertson, D. A., & Ahyong, S. T. (2010). Marine biodiversity of Aotearoa New Zealand. *PLoS One, 5*(8), 1-17. doi:10.1111/1755-0998.13450

Hebert, P. D., & Ratnasingham, S. (2007). BOLD: The Barcode of Life Data System (www.barcodinglife.org). *Molecular Ecology Notes, 7*, 355-364.

Hebert, P. D., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of The Royal Society of London, 270*(1512), 313-321. doi:10.1098/rspb.2002.2218

Hillis, D. M. (1987). Molecular Versus Morphological Approaches to Systematics. *Annual Review of Ecology and Systematics, 18*(1), 23-42.

Hulley, E. N., Taylor, N. D. J., Zarnke, A. M., Somers, C. M., Manzon, R. G., Wilson, J. Y., & Boreham, D. R. (2018). DNA barcoding vs. morphological identification of larval fish and embryos in Lake Huron: Advantages to a molecular approach. *Journal of Great Lakes Research, 44*(5), 1110-1116. doi: 10.1016/j.jglr.2018.07.013

Isari, S., Pearman, J. K., Casas, L., Michell, C. T., Curdia, J., Berumen, M. L., & Irigoien, X. (2017). Exploring the larval fish community of the central Red Sea with an integrated morphological and molecular approach. *PLoS One, 12*(8), 1-24. doi:10.1371/journal.pone.0182503

IUCN. (2020). The IUCN red list of threatened species. 2020-2. Retrieved from https://www.iucnredlist.org.

Ivanova, N. V., Zemlak, T. S., Hanner, R. H., & Hebert, P. D. N. (2007). Universal primer cocktails for fish DNA barcoding. *Molecular Ecology Notes, 7*(4), 544-548. doi:10.1111/j.1471-8286.2007.01748.x

Jeon, H., Choi, S., & Suk, H. Y. (2012). Exploring the utility of partial cytochrome c oxidase subunit 1 for DNA barcoding of Gobies. *Animal Systematics, Evolution and Diversity, 28*(4), 269- 278. doi:10.5635/ased.2012.28.4.269

Jeunen, G. J., Knapp, M., Spencer, H. G., Lamare, M. D., Taylor, H. R., Stat, M., Bunce, M., Gemmell, N. J. (2019). Environmental DNA (eDNA) metabarcoding reveals strong discrimination among diverse marine habitats connected by water movement. *Mol Ecol Resour, 19*(2), 426-438. doi:10.1111/1755-0998.12982

Jeunen, G. J., Lamare, M. D., Knapp, M., Spencer, H. G., Taylor, H. R., Stat, M, Bunce, M., Gemmell, N. J. (2019). Water stratification in the marine biome restricts vertical environmental DNA (eDNA) signal dispersal. *Environmental DNA, 2*(1), 99-111. doi:10.1111/1755-0998.12982

Jin, S., Kim, K. Y., Kim, M.-S., & Park, C. (2020). An assessment of the taxonomic reliability of DNA barcode sequences in publicly available databases. *Algae, 35*(3), 293-301. doi:10.4490/algae.2020.35.9.4

Kartavtsev, Y. P., & Lee, J. S. (2006). Analysis of nucleotide diversity at the cytochrome b and cytochrome oxidase 1 genes at the population, species, and genus levels. *Russian Journal of Genetics, 42*(4), 341-362. doi:10.1134/s1022795406040016

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative Studies of nucleotide sequences. *Journal of Molecular Evolution, 16*, 111-120.

Ko, H.L., Wang, Y.T., Chiu, T.S., Lee, M.A., Leu, M.Y., Chang, K.Z., Che, W.Y., Shao, K.T. (2013). Evaluating the accuracy of morphological identification of larval fishes by applying DNA barcoding. *PLoS One, 8*(1). 1-7. doi:10.1371/journal.pone.0053451

Kress, W. J., Garcia-Robledo, C., Uriarte, M., & Erickson, D. L. (2015). DNA barcodes for ecology, evolution, and conservation. *Trends in Ecology and Evolution, 30*(1), 25-35.

Lakicevic, M. & Srdevic, B. (2018). Measuring Biodiversity in forest communities- a role of biodiversity indices. *The Serbian Journal of Agricultural Sciences, 67*(1), 65-70.

Lakra, W.S., Verma, M.S., Goswami, M., Lal, K.K., Mohindra, V., Punia, P., Gopalakrishnan, A., Singh, K.V., Ward, R.D., & Herbert, P. (2011). DNA barcoding Indian marine fishes. *Molecular Ecology Resources, 11* (1), 60-71. doi:10.1111/j.1755-0998.2010.02894.x

Landi, M., Dimech, M., Arculeo, M., Biondo, G., Martins, R., Carneiro, M., . . . Costa, F. O. (2014). DNA barcoding for species assignment: the case of Mediterranean marine fishes. *PLoS One, 9*(9), 1-9. doi:10.1371/journal.pone.0106135

Lawrence, H.A. (2008). *Conservation genetics of the world's most endangered seabird the Chatham Island Tāiko.* [Doctoral dissertation]. Massey University.

Levin, L. A. (2006). Recent progress in understanding larval dispersal: new directions and digressions. *Integrative and Comparative Biology, 46*, 282-297. doi:10.1093/icb/icjO24

Li, Z., Ye, Z., Wan, R., Chen, Y., Tian, Y., Ren, Y., Liu, H., Haisheng, H., Boenish, R. (2017). Evaluating the relationship between spatial heterogeneity and temporal variability of larval fish assemblages in a coastal marine ecosystem (Haizhou Bay, China). *Marine Ecology, 38*(6), 1-10. doi:10.1111/maec.12446

Liggins, L., Kilduff, L., Trnski, T., Delrieu-Trottin, E., Carvajal, J. I., Arranz, V., Planes, S., Saenz-Agudelo, P., Aguirre, J. D. (2021). Morphological and genetic divergence supports peripheral endemism and a recent evolutionary history of Chrysiptera demoiselles in the subtropical South Pacific. *Coral Reefs*. doi:10.1007/s00338-021-02179-7

Liggins, L., Sweatman, J. A., Trnski, T., Duffy, C. A. J., Eddy, T. D., & Aguirare, J. D. (2020). Natural history footage provides new reef fish biodiversity information for a pristine but rarely visited archipelago. *Sci Rep, 10*(1), 3159. doi:10.1038/s41598-020-60136-w

Marbagana, E., de Astarloa, J.M.D., Hanner, R., Zhang, J., & Castro, M.G. (2011). DNA barcoding identifies Argentine marine fishes from marine and brackish waters. *PloS ONE, 6*(12), 1-11.

Matarese, A. C., Spies, I. B., Busby, M. S., & Orr, J. W. (2011). Early larvae of *Zesticelus profundorum* (family Cottidae) identified using DNA barcoding. *Ichthyological Research, 58*(2), 170-174.

Matsuura, K., & Middleton, I. (2016). Discovery of a larva of the Aracanidae (Actinopterygii, Tetraodontiformes) from New Zealand. *Ichthyological Research, 64*(1), 151-154. doi:10.1007/s10228-016-0533-8

Mayr, E. (1942). *Systematics and the origin of species, from the viewpoint of a zoologist.* New York: Columbia University Press.

McLaurin, J. & Sterelny, K. (2008). *What is biodiversity.* University of Chicago Press

Middleton, I., Aguirre, J. D., Trnski, T., Francis, M., Duffy, C., Liggins, L., & Leroy, B. (2021). Introduced alien, range extension or just visiting? Combining citizen science observations and expert knowledge to classify range dynamics of marine fishes. *Diversity and Distributions, 27*(7), 1278-1293. doi:10.1007/s10228-016-0533-8

Milan, D.T., Mendes, I. S., Damasceno, J. S., Teixeria, D. F., Sales, N. G., & Carvalho, D.C. (2020). New 12S metabarcoding primers for enhanced Neotropical freshwater fish biodiversity assessment. *Scientific Reports, 10*(1), 17966. doi:10.1038/s41598-020-74902-3

Mishler, B. D., & Donoghue, M. J. (1982). species concepts: a case for pluralism. *Systematic Zoology, 31*(4), 491-503.

Miya, M. (2021). Environmental DNA Metabarcoding: A Novel Method for Biodiversity Monitoring of Marine Fish Communities. *Annual Review of Marine Science, 11*(31), 1-25. doi:10.1146/annurev-marine-041421-082251

Miya, M., Sato, Y., Fukunaga, T., Sado, T., Poulsen, J. Y., Sato, K., Minamoto, T., Yamamoto, S., Yamanaka, H., Araki, H., Kondoh, M., Iwasaki, W. (2015). MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: detection of more than 230 subtropical marine species. *Royal Society Open Science, 2*(7), 1-33. doi:10.1098/rsos.150088

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B., & Worm, B. (2011). How many species are there on Earth and in the ocean?. *PLoS Biology, 9*(8), 1-8.

Naylor, G. J. P., Caira, J. N., Jensen, K., Rosana, K. A. M., White, W. T., & Last, P. R. (2012). A DNA Sequence–Based Approach to the Identification of Shark and Ray Species and Its Implications for Global Elasmobranch Diversity and Parasitology. *Bulletin of the American Museum of Natural History, 367*, 1-262. doi:10.1206/754.1

Neira, F. J., Miskiewicz, A. G., & Trnski, T. (1998). *Larvae of Temperate Australian Fishes - Laboratory Guide for Larval Fish Identification*. Nedlands, WA: University of Western Australia Press.

Nimnoi, P., & Pongsilp, N. (2020). Marine bacterial communities in the upper gulf of Thailand assessed by Illumina next-generation sequencing platform. *BMC Microbiol, 20*(1), 19.

O'Callaghan. J., Stevens, C., Roughan, M., Cornelisen, C., Sutton, P., Garrett, S., Giorli, G., Smith, R.O., Currie, K.I., Suanda, S.H., Williams, M., Bowen, M., Fernandez, D., Vennell, R., Knight, B.R., Barter, P., McComb, P., Oliver, M., Liningston, M., Tellier, P., Meissner, A., Brewer, M., Gall, M., Nodder, S.D., Decima, M., Souza, J., Forcen-Vazquez, A., Gardiner, S., Paul-Burke, K., Chiswell, S., Roberts, J., Hayden, B., Biggs, B., MacDonald, H. (2019). Developing an integrated ocean observing system for New Zealand. *Frontiers in Marine Science, 6*(143), 1-7.

Paine, M. A., McDowell, J. R., & Graves, J. E. (2007). Specific identification of Western Atlantic Ocean Scombrids using mitochondrial DNA cytochrome c oxidase subunit I (COI) gene region sequences. *Bulletin of Marine Science, 80*(2), 353-367.

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics, 20*(2), 289-290. doi:10.1093/bioinformatics/btg412

Park, C., & Allaby, M. (2017). *A Dictionary of Environment and Conservation* (3rd Edition ed.): Oxford University Press.

Pegg, G., Sinclair, B., Briskey, L., & Aspden, W. J. (2006). MtDNA barcode identification of fish larvae in the southern Great Barrier Reef, Australia. *Scienta Marina, 70*, 7-12.

Pentinsaari, M., Ratnasingham, S., Miller, S.E., & Hebert, P.D.N. (2020). BOLD and GenBank revisited- Do identification errors arise in the lab or in sequence libraries?. *PloS ONE, 15*(4).

Pope, L. C., Liggins, L., Keyse, J., Carvalho, S. B., & Riginos, C. (2015). Not the time or the place: the missing spatio-temporal link in publicly available genetic data. *Molecular Ecology, 24*, 3802- 3809.

Poulin, R., Closs, G. P., Lill, A. W., Hicks, A. S., Herrmann, K. K., & Kelly, D. W. (2012). Migration as an escape from parasitism in New Zealand galaxiid fishes. *Oecologia, 169*(4), 955-963. doi:10.1007/s00442-012-2251-x

Putman, N. F. (2016). An ecological perspective on the migrations of marine fishes. *Environmental Biology of Fishes, 99*(10), 801-804.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rabaoui, L., Yacoubi, L., Sanna, D., Casu, M., Scarpa, F., Lin, Y. J., Shen, K.N., Clardy, T.R., Arculeo, M., Qurban, M. A. (2019). DNA barcoding of marine fishes from Saudi Arabian waters of the Gulf. *J Fish Biol, 95*(5), 1286-1297. doi:10.1111/jfb.14130

Randall, J. E. (2005). *Reef and Shore Fishes of the South Pacific: New Caledonia to Tahiti and the Pitcairn Islands.* Honolulu, HI: University of Hawaii Press.

Rathnasuriya, M. I. G., Mateos-Rivera, A., Skern-Mauritzen, R., Wimalasiri, H. B. U., Jayasinghe, R. P. P. K., Krakstad, J. O., & Dalpadado, P. (2021). Composition and diversity of larval fish in the Indian Ocean using morphological and molecular methods. *Marine Biodiversity, 51*(2), 1-15. doi:10.1007/s12526-021-01169-w

Rawson, P.D., & Burton, R.S. (2002). Functional coadaptation between cytochrome c and cytochrome c oxidase within allopatric populations of marine copepod. *Proceedings of the National Academy of Sciences of the United States of America, 99*(20), 12955- 12958.

Rees, H. C., Maddison, B. C., Middleditch, D. J., Patmore, J. R. M., Gough, K. C., & Crispo, E. (2014). REVIEW: The detection of aquatic animal species using environmental DNA - a review of eDNA as a survey tool in ecology. *Journal of Applied Ecology, 51*(5), 1450-1459.

Regan, C. T. (1926). Organic Evolution. *Report of the British Association for the Advancement of Science, 1*, 75-86.

Roberts, C. D., Stewart, A. L., Struthers, C. D., Barker, J. J., & Kortet, S. (2020). Checklist of the Fishes of New Zealand. Online version 1.2., from Museum of New Zealand Te Papa Tongarewa, Wellington. https://collections.tepapa.govt.nz/document/10564

Roberts, C.D., Stewart, A.L., & Struthers, C.D. (2015). *The fishes of New Zealand: Part one*. Wellington, New Zealand: Te Papa Press.

Rodríguez-Castro, K. G., Saranholi, B. H., Bataglia, L., V. Blanck, D., & Galetti, P. M. (2018). Molecular species identification of scat samples of South American felids and canids. *Conservation Genetics Resources, 12*(1), 61-66.

Roy, M., Belliveau, V., Mandrak, N. E., & Gagné, N. (2017). Development of environmental DNA (eDNA) methods for detecting high-risk freshwater fishes in live trade in Canada. *Biological Invasions, 20*(2), 299-314. doi:10.1007/s10530-017-1532-z

RStudio Team (2021). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Ruppert, K. M., Kline, R. J., & Rahman, M. S. (2019). Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Global Ecology and Conservation, 17*(2019), 1-29. doi:10.1016/j.gecco.2019.e00547

Saccone, C., Pesole, G., Gissi, C. & De Chirico, A. (1999). Nucleotide substitution rate of mammalian mitochondrial genomes. *Journal of Molecular Evolution, 48*(4), 427-434.

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., McVeigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sulllivan, J.P., Sun, L., Turner, S., Karsch-Mizrachi, I. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford), 2020,* 1-21. doi:10.1093/database/baaa062

Scott Chamberlain (2021). bold: Interface to Bold Systems API. R package version 1.2.0. https://CRAN.R-project.org/package=bold

Shin, U. C., Jeong, Y. K., Yoon, S. C., Choi, K. H., & Kim, J. K. (2018). Genetical identification and morphological description of the larvae and juveniles of *Porocottus leptosomus* (Pisces: Cottidae) from Korea. *Fisheries and Aquatic Sciences, 21*(1), 1-10. doi:10.1186/s41240-018-0115-y

Shirak, A., Dor, L., Seroussi, E., Ron, M., Hulata, G., & Golani, D. (2016). DNA barcoding fish species from the Mediterranean coast of Israel. *Mediterranean Marine Science, 17*(2), 459-466. doi:10.12681/mms.1384

Short, G. A., & Trnski, T. (2021). A New Genus and Species of Pygmy Pipehorse from Taitokerau Northland, Aotearoa New Zealand, with a Redescription of *Acentronura Kaup*, 1853 and *Idiotropiscis* Whitley, 1947 (Teleostei, Syngnathidae). *Ichthyology & Herpetology, 109*(3), 806-835. doi:10.1643/i2020136

Shu, L., Ludwig, A., & Peng, Z. (2021). Environmental DNA metabarcoding primers for freshwater fish detection and quantification: In silico and in tanks. *Ecol Evol, 11*(12), 8281-8294. doi:10.1002/ece3.7658

Steinke, D., Zelmak, T.S., Boutillier, J.A., & Hebert, P.D.N. (2009). DNA barcoding of Pacific Canada's fishes. *Marine Biology: International Journey of Life in Oceans and Coastal Waters, 156*(12), 2641-2647. doi:10.1007/s00227-009-1284-0

Stepanovic, S., Kosovac, A., Krstic, O., Jovic, J., & Tosevski, I. (2016). Morphology versus DNA barcoding: two sides of the same coin. A case study of *Ceutorhynchus erysimi* and *C. contractus* identification. *Insect Science, 23*(4), 638-648. doi:10.1111/1744-7917.12212

Stewart, A. L., Knudsen, A. W., & Clements, K. D. (2021). A new species of deep-water triplefin (Pisces: Tripterygiidae) in the genus *Ruanoho* from coastal New Zealand waters. *Zootaxa, 4891*(3).

Stoeckle, M. Y., & Hebert, P. D. N. (2008). Barcode of Life. *Scientific American, 299*(4), 82-89. doi:10.1093/icesjms/fsaa225

Tang, K. L., Stiassny, M. L. J., Mayden, R. L., & DeSalle, R. (2021). Systematics of Damselfishes. *Ichthyology & Herpetology, 109*(1), 258-318. doi:10.1643/i2020105

Thomsen, P. F., Kielgast, J., Iversen, L. L., Moller, P. R., Rasmussen, M., & Willerslev, E. (2012). Detection of a diverse marine fish fauna using environmental DNA from seawater samples. *PLoS One, 7*(8), 1-9. doi:10.1371/journal.pone.0041732

Toczydlowski, R. H., Liggins, L., Gaither, M. R., Anderson, T. J., Barton, R. L., Berg, J. T., Beskid, S.G., Davis, B., Delgado, A., Farrell, E., Ghoojaei, M., Himmelsbach, N., Holmes, A.E., Queeno, S.R., Trinh, T., Weyand, C.A., Bradburd, G. S., Riginos, C., Toonen, R. J., Crandall, E. D. (2021). Poor data stewardship will hinder global genetic diversity surveillance. *PNAS, 118*(34), 1-3.

Townsend, A. J., de Lange, P. J., Duffy, C. A. J., Miskelly, C. M., Molloy, J., & Norton, D. A. (2008). *New Zealand Threat Classification System manual*. Wellington, New Zealand: Science & Technical Publishing

Trivedi, S., Aloufi, A. A., Ansari, A. A., & Ghosh, S. K. (2016). Role of DNA barcoding in marine biodiversity assessment and conservation: An update. *Saudi Journal of Biological Sciences, 23*(2), 161-171.

Trujillo-González, A., Becker, J. A., Huerlimann, R., Saunders, R. J., & Hutson, K. S. (2019). Can environmental DNA be used for aquatic biosecurity in the aquarium fish trade? *Biological Invasions, 22*(3), 1011-1025. doi:10.1007/s10530-019-02152-0

Trujillo-Gonzalez, A., Edmunds, R. C., Becker, J. A., & Hutson, K. S. (2019). Parasite detection in the ornamental fish trade using environmental DNA. *Sci Rep, 9*(1), 5173. doi:10.1038/s41598-019-41517-2

Valdez-Moreno, M., Vásquez-Yeomans, L., Elías-Gutiérrez, M., Ivanova, N. V., & Hebert, P. D. (2010). Using DNA barcodes to connect adults and early life stages of marine fishes from the Yucatan Peninsula, Mexico: potential in fisheries management. *Marine and Freshwater Research, 61*, 665-671.

Victor, B. C., Hanner, R., Shivji, M., Hyde, J., & Caldow, C. (2009). Identification of the larval and juvenile stages of the Cubera Snapper, *Lutjanus cyanopterus*, using DNA barcoding. *Zootaxa, 2215*, 24-36.

Wang, L., Wu, Z., Liu, M., Liu, W., Zhao, W., Liu, H., & You, F. (2018). DNA barcoding of marine fish species from Rongcheng Bay, China. *Peer J, 6*(1), 1-19. doi:10.7717/peerj.5013

Ward, R. D., Hanner, R., & Hebert, P. D. (2009). The campaign to DNA barcode all fishes, FISH-BOL. *J Fish Biol, 74*(2), 329-356. doi:10.1111/j.1095-8649.2008.02080.x

Ward, R.D., Zemlak, T.S., Innes, B.H., Last, P.R., Hebert, P.D.N. (2005). DNA Barcoding Australia's fish species. *Philosophical transactions of the Royal Society, 360*(1462), 1847-1857. doi:10.1098/rstb.2005.1716

Weigt, L. A., Driskell, A. C., Baldwin, C. C., & Ormos, A. (2012). *DNA Barcoding Methods and Protocols*.

White, T. A., Stefanni, S., Stamford, J., & Hoelzel, A. R. (2009). Unexpected panmixia in a long-lived, deep-sea fish with well-defined spawning habitat and relatively low fecundity. *Molecular Ecology, 18*(12), 2563-2573. doi:10.1111/j.1365-294X.2009.04218.x

Wibowo, A., Wahlberg, N., & Vasemägi, A. (2017). DNA barcoding of fish larvae reveals uncharacterised biodiversity in tropical peat swamps of New Guinea, Indonesia. *Marine and Freshwater Research, 68*(6), 1079- 1087. doi:10.1071/mf16078

Wickham, H. (2007). Reshaping Data with the reshape Package. *Journal of Statistical Software, 185*, 137-144.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York.

Wiley, E. O. (1981). Remark on Willis' Species Concept. *Systematic Zoology, 30*(1), 86-87.

Williams, J., Stokes, F., Dixon, H., & Hurren, K. (2017). *The economic contribution of commercial fishing to the New Zealand economy*. Retrieved from https://www.seafood.co.nz/fileadmin/Media/BERL_report/BERL_Report_August_2017.pdf

Wilson, E.O. (1992). *The Diversity of Life*. Belknap Press of Harvard University Press.

Winter, D. J. (2017) rentrez: an R package for the NCBI eUtils API The R Journal 9(2): 520-526

Wong, E. H., Shivji, M. S., & Hanner, R. H. (2009). Identifying sharks with DNA barcodes: assessing the utility of a nucleotide diagnostic approach. *Molecular Ecology Resources, 9*(1), 243-256. doi:10.1111/j.1755-0998.2009.02653.x

Xu, L., Van Damme, K., Hong, Li., Yingying, J., Xuehui, W., Feiyan, D. (2019). A molecular approach to the identification of marine fish of the Dongsha Islands (South China Sea). *Fisheries research, 213*(1), 105-112. doi:10.1016/j.fishres.2019.01.011

Yang, J., Zhang, X., Zhang, W., Sun, J., Xie, Y., Zhang, Y., Burton, G.A, Jr., Yu, H. (2017). Indigenous species barcode database improves the identification of zooplankton. *PLoS One, 12*(10), 1-15. doi:10.1371/journal.pone.0185697

Zhang D, X., Hewitt G.M. (1998) Special DNA Extraction Methods for Some Animal Species. In: Karp A., Isaac P.G., Ingram D.S. (eds) Molecular Tools for Screening Biodiversity. Springer, Dordrecht. https://doi.org/10.1007/978-94-009-0019-6_6

Zhang, J. (2011). Species identification of marine fishes in China with DNA barcoding. *Evidence based Complementary & Alternative Medicine, 8*(1), 1-10. doi:10.1016/j.bse.2010.12.017

Zhang, J., & Hanner, R. (2011). DNA Barcoding is a useful tool for the identification of marine fishes from Japan. *Biochemical Systematics and Ecology, 39*(1), 31-42.

Zhang, J., & Hanner, R. (2012). Molecular approach to the identification of fish in the south China sea. *PloS ONE, 7*(2), 105-112. doi:10.1371/journal.pone.0030621

Zhang, Y., Qin, G., Zhang, H., Wang, X, & Lin, Q. (2017). DNA Barcoding reflects the diversity and variety of brooding traits of fish species in the family Syngnathidae along China's coast. *Fisheries Research, 185*(1), 137-144. doi:10.1016/j.fishres.2016.09.015

APPENDICES

Appendix 1. The list of all Actinopterygii species known to occur in New Zealand to date, whether there is a Cytochrome Oxidase I sequence available for that species, and where that sequence was retrieved from to create the NZ Fish Barcode Database. The "DESCRIPTION" tab provides an explanation of the column headers used in all other tabs. The "NZ Fishes" provides a list of all species, relevant metadata, and whether the NZ Fish Barcode Database comprises a sequence for that species or not. The "PersRep_Metadata" tab provides metadata for sequences generated within our research laboratory at Massey University. The "RegPhylo_Metadata" tab provides metadata for sequences retrieved from public sequence repositories using the RegPhylo R package. Available here:

https://masseyuni-my.sharepoint.com/:x:/g/personal/lliggins_massey_ac_nz/EXtmJTGMzNtFj77dgT5OnvIBswpkiSWZAl_iwMo7utXB2w

Appendix 2. The full list of larval fish specimens sampled, including their morphological identification, and molecular identification, where possible. The "DESCRIPTION" tab provides an explanation of the column headers used to retain metadata for the samples, and report relevant results. Available here:

https://masseyuni-my.sharepoint.com/:x:/g/personal/lliggins_massey_ac_nz/EdcE-39mxaJJoqs5q3CZRUIBAvSMwPREH6KXxieSHUeh0g?e=9XOMKw