A genomic investigation of the marine teleost tarakihi
(*Nemadactylus macropterus*) provides insights into its
genome evolution and population structure

By

Yvan Papa

VICTORIA UNIVERSITY OF
WELLINGTON
TE HERENGA WAKA

A thesis submitted to the Victoria University of Wellington in
fulfilment of the requirements for the degree of Doctor of Philosophy

Victoria University of Wellington
School of Biological Sciences

2022

Primary Supervisor:

Dr Peter A. Ritchie

Victoria University of Wellington

Secondary Supervisors:

Dr Maren Wellenreuther

Plant & Food Research, Nelson

The University of Auckland

Dr Mark A. Morrison

NIWA, Auckland

Abstract

Understanding the patterns of connectivity and stock structure of a fishery is an essential prerequisite for effective and sustainable management. Genetic markers can assist with the delineation of fish stock boundaries. The ongoing progress in high-throughput DNA sequencing technologies has transformed the field of genetics forward into genomics. Genome-wide marker sets allow for an unprecedented level of resolution for the investigation of population differentiation, levels of gene flow, and the functional loci associated with adaptation. They can provide significant insights into the evolutionary processes that influence populations and link these to environmental conditions. Equally important is the application of genetic information to improve population management decisions. However, genomic analyses have not yet been widely used for fisheries research in New Zealand.

Tarakihi (*Nemadactylus macropterus*) is a marine fish widely distributed around the inshore areas of New Zealand and the southern parts of Australia. It supports an important commercial fishery with annual landings in New Zealand averaging over 5,000 tonnes over the past 40 years. As with many other New Zealand fish, very little is known about its stock structure as well as the characteristics of its genome. Previous studies using low-resolution population genetic markers did not detect any significant level of population structure. Genome-wide markers now offer the opportunity to obtain a more complete picture of tarakihi population genetic variation. The aims of this thesis were to (1) study the evolution of the tarakihi genome and to (2) assess the population structure of tarakihi in New Zealand. This was achieved by producing the first tarakihi genome assembly, compare it with genomes of other fish species and analyse a population sample of tarakihi low-coverage genomes to test for genetic differentiation and identify presumably adaptive genetic variation.

This thesis presents an in-depth review of the studies that have been conducted on genetic stock structure of fishes in New Zealand and of the applications of genomics for fisheries management. This provides an important framework and establishes why there is a strong need to carry out population genomic research on New Zealand fisheries species.

A *de novo* genome sequence of tarakihi was assembled using high-molecular-weight DNA obtained from a vouchered specimen. DNA sequencing was performed using 92x high-quality Illumina short-reads and 122x Oxford Nanopore Technology long-reads. Two genome assembly algorithms were used. An assembly based on a trial run of low-coverage PacBio HiFi data from another specimen was also produced. The polished Illumina + Nanopore assembly performed in MaSuRCA with the Flye algorithm gave the best results. The final reference genome is 568 Mb long and comprised of 1,214 scaffolds with an N50 of 3.37 Mb. Genome completeness was high, with 97.8% of complete Actinopterygii BUSCOs retrieved. An additional 250 Gb of PacBio Sequel II Iso-Seq RNA reads were obtained from four tissue types to assist with gene annotation. Approximately 30.5% of the tarakihi genome was composed of repetitive elements and 20,169 protein-coding genes were annotated. Iso-Seq analysis found 91,313 unique transcripts and the most common alternative splicing event was intron retention. This highly contiguous genome assembly will be a valuable genomic resource to assist the study of population genomics and comparative genomics; its value is demonstrated in the subsequent analyses of this study.

The *de novo* short-read genomes of five additional New Zealand fish species were assembled: barracouta (*Thyrstites atun*), blue moki (*Latridopsis ciliaris*), butterfish (*Odax pullus*), kahawai (*Arripis trutta*), and king tarakihi (*Nemadactylus* n.sp.). One specimen of each species was sequenced for at least 33x coverage with paired-end Illumina short reads. The resulting draft genome assemblies ranged in size from 532 Mb (butterfish) to 714 Mb (barracouta), with the number of scaffolds ranging from 58,102 to 150,595, N50 from 10,031 to 30,942 bp, and BUSCO completeness scores from 70.2% to 89.1%. The tarakihi genome was added to the dataset for a comparative genomic analysis using all six species. While the proportion of repeat elements was highly correlated with genome size ($R^2 = 0.97$, $P < 0.01$), most of the metrics for the genic features (e.g. number of exons or intron length) were significantly correlated with assembly contiguity ($|R^2| = 0.79\text{--}0.97$). A phylogenomic tree of Percomorpha including eight additional high-quality fish genomes was reconstructed from sequences of shared gene families. Based on the branch-site model, evidence of positive selection was found in 65 genes in tarakihi and 209 genes in Latridae: most of these were involved in the ATP binding pathway and the integral structure of membranes. These results can be used to inform future studies when considering the strengths and weaknesses of

scaffold-level assemblies for comparative genomics. This also provided useful insights into the evolutionary patterns and processes of genome evolution in bony fishes.

A population genomics analysis was conducted using 175 wild-caught tarakihi specimens obtained from around New Zealand and Tasmania (Australia) and an additional 12 king tarakihi specimens. All individuals were whole-genome sequenced for c. 12x coverage using Illumina short read data. A dataset of 7.5 million high-quality single-nucleotide polymorphisms (SNPs) was obtained. Variant filtering, F_{ST} -outlier analysis, and redundancy analysis (RDA) were used to evaluate population structure, presumably adaptive structure, and locus-environment association. King tarakihi displayed high levels of genetic divergence, differences in heterozygosity, and differences in linkage disequilibrium patterns that were all consistent with reproductive isolation from tarakihi from New Zealand and Tasmania. This provided additional evidence that king tarakihi is genetically distinct and a different species. While the tarakihi population sampled from Tasmania showed a significant level of neutral genetic differentiation from the New Zealand populations ($F_{ST} = 0.0054\text{--}0.0073$, $P \leq 0.05$), there was no clear genetic sub-structuring that could be detected among New Zealand tarakihi samples ($\Phi_{ST} < 0.001$, $P = 0.77$). However, presumably adaptive variation based on outlier SNPs suggested fine-scale adaptive structure between locations around central New Zealand off the east (Wairarapa, Cape Campbell, and Hawke's Bay) and the west coast (Tasman Bay/Golden Bay and Upper West Coast of South Island). Locus-environment association analysis found that 47 loci were correlated with mean depth temperature and projection of the RDA indicated that tarakihi in New Zealand are distributed following a North-South temperature cline. Taken together, these findings indicate that tarakihi are composed of one genetically panmictic stock in New Zealand, but the presence of presumably adaptive variation suggests that the latitudinal pattern of tarakihi migration could be influenced by water temperature or some other environmental feature with a distribution linked to temperature.

The tarakihi genome presented here is the first genome assembly for a species in the Cirrhitioidei superfamily and one of the first to be reported for a New Zealand fisheries species. This will provide a valuable genetic resource for analyses of genome-wide patterns of evolution. The additional five draft fish genomes establish the much-needed genomic

resource for studies of the fishes of New Zealand and are available as reference genomes for population genomic studies. The results of the tarakihi population structure analysis can be used to support stock assessment models and improve fisheries management. Overall, the findings of this thesis research can be built upon with long-term temporal and spatial genomic sampling studies and incorporate the results as part of an integrative evidence-based approach to fisheries management. This will be a crucial step toward sustainable fisheries as wild-harvest resources face strong pressures from increasing commercial demands, a changing climate, and a marine ecosystem with an uncertain future.

Acknowledgements

It is hard to describe how grateful I am to have been able to conduct this PhD thesis in Wellington, New Zealand. This is such a beautiful country, full of great people. These four years really felt like an adventure and I had the chance to discover and learn so much, both as part of my thesis and my travels.

I want to start by thanking my supervisors Peter Ritchie, Maren Wellenreuther and Mark Morrison. You have been so supportive during my thesis, providing guidance, resources and always being available but at the same time leaving me the space to grow as a scientist. Thank you, Pete, for having entrusted me with this project. Your professionalism and expertise coupled with your laid-back and friendly attitude make you an awesome supervisor and a trustworthy person to work with.

Thanks to Jeremy McKenzie for having been instrumental in the sample collection. This study would not have been nearly as complete without your help. Thanks to all the great people at Victoria, NIWA, VUCEL, PFR and the NSW DPI and all the commercial and recreational fishers who contributed either directly or indirectly to make this study happen.

Thank you, Tom Oosting, for having been an awesome partner in crime during these four years; we started together, went through the highs and lows of a PhD together, and finished (almost, kind of!) together. You've been so helpful during all this time, from sharing ideas, lab protocols and bioinformatics scripts to driving in your van to get some fish specimens in somewhat random places late at night. Having you around both in and out of the office is always a blast. I agree with you that we work very well together, and I also hope we will keep doing so. Noemie Valenza-Troubat, we also went through our PhDs together and it was always such fun to see you in Wellington, Nelson, or at a conference. I wish you all the best for your next big adventure.

Thank you, Alex Halliwell, for going through the whole tarakihi sampling ordeal with me during the first couple of years, going on dives together and teaching me the fundamentals of kiwi slang! You are a super fun and hard-working bloke. I wish you all the best for the future (just try not to break any more bones!). Huge thanks to Guillermo, Imogen, Holly, Thomas,

Leah, Seamus and Amber for simply making the Peter Ritchie (or Ritcher Petie?) lab group the best lab group.

Thanks Tom Dawes for having been such a good mate. Thanks to the whole Ian crew for the bi-weekly laughs. Thank you Jemima, Alice, and Ellie for being such fun and supportive flatmates. Thank you Jenny for the time we had together, I will always cherish it. Thank you to all of you back home (you know who you are!) who kept in contact during these past years, I miss you all and I am sincerely looking forward to seeing you again soon.

À toute ma famille, merci pour votre soutien et tout votre amour. À la plus courageuse de toutes les mamans. À Papa, Daniel, Laura, Jo, Celia, Julia et le petit nouveau/la petite nouvelle.

À Sum.

External contributions

This research was supported by the programme "Juvenile fish habitat bottlenecks" (CO1X1618), funded by the New Zealand Ministry of Business, Innovation, and Employment (MBIE) Endeavour Fund. Laboratory work, analyses and writing were conducted by the author with the help of supervisors Drs Peter Ritchie, Maren Wellenreuther and Mark Morrison. Additional contributors are mentioned below.

Plant & Food Research (Nelson) provided the captive-bred tarakihi specimen used for genome assembly. The tarakihi specimen used for transcriptome sequencing and the barracouta specimen were captured by Nick Johnston (recreational fisherman). Dive Wellington provided the opportunity to collect tarakihi and other fish samples (blue moki, butterfish, and kahawai) during their 2019 spear fishing competition.

Collection of the tarakihi and king tarakihi specimens from phase 1 (2017–2018) was supervised by Cameron Walsh (Stock Monitoring Services Limited). Specimens were collected by fishing companies (Gisborne Fisheries, Star Fish Supply, Egmont Seafoods, Moana New Zealand, Hawke's Bay Seafoods, United Fisheries, Talley's Seafood, and Wellington Trawling) and Peter Young (Cruise Fiordland). Alex Halliwell provided assistance for the tissue collections and DNA extractions.

The majority of samples from phase 2 (2019–2020) were collected by NIWA staff including Jeremy McKenzie, Dan MacGibbon, Helena Armiger, Jade Arnold, and Caoimhghin Ó Maolagáin. Samples from Australia were collected by Anne-Marie Hegarty and Matt Taylor (New South Wales Department of Primary Industries).

Additional tarakihi tissue samples were collected by Tom Oosting (Victoria University of Wellington) at Gisborne Tatapouri Sports Fishing Club and Hawkes Bay Sports Fishing Club fishing competitions.

Table of contents

Chapter 1. General Introduction	1
1.1 Evolutionary genetics and population structure	1
1.1.1 The evolutionary process.....	1
1.1.2 Population structure	4
1.2 Population genomics	5
1.3 Genomic evidence of selection and adaptation	6
1.3.1 Local adaptation.....	7
1.3.2 Testing for selection among species.....	9
1.4 Genome assemblies	10
1.5 The application of genomics in fisheries management	11
1.6 Study species: Tarakihi.....	13
1.6.1 Tarakihi biology and dispersal	13
1.6.2 Tarakihi fisheries and management	17
1.6.3 Population structure and genetics of tarakihi: brief review of current knowledge.....	19
1.6.4 The case of the king tarakihi	20
1.7 Aims and thesis structure	21
 Chapter 2. Genetic stock structure of New Zealand fish and the use of genomics in fisheries management: an overview and outlook	 27
2.1 Abstract.....	27
2.2 Introduction	27
2.3 The value of genetics in fisheries stock delineation and fisheries management	30
2.4 The current state of New Zealand fisheries genetics	32
2.4.1 Hoki or blue grenadier (<i>Macruronus novaezelandiae</i> , Merlucciidae)	37
2.4.2 Orange roughy (<i>Hoplostethus atlanticus</i> , Trachichthyidae).....	45

2.4.3	Australasian snapper (<i>Chrysophrys auratus</i> , Sparidae)	47
2.4.4	Ling or pink cusk-eel (<i>Genypterus blacodes</i> , Ophidiidae).....	49
2.4.5	Blue cod (<i>Parapercis colias</i> , Pinguipedidae)	50
2.4.6	Other fishery species	51
2.5	Moving forward: Use of genome-wide sequencing techniques.....	55
2.6	Other issues in fisheries management that will benefit from a more systematic use of genetics and genomics.....	56
2.6.1	Loss of genetic diversity and adaptive potential	57
2.6.2	Fishing-induced evolution.....	58
2.6.3	Climate change.....	59
2.6.4	Measuring abundance and connectivity.....	60
2.7	Conclusion.....	61
2.8	Supplementary Material	63

Chapter 3. Contiguous genome assembly and transcriptome of a highly heterozygous New Zealand fisheries species, the tarakihi (*Nemadactylus macropterus*). 77

3.1	Abstract.....	77
3.2	Introduction	77
3.3	Materials and Methods.....	80
3.3.1	Tissue collection and nucleotide extraction	80
3.3.2	Genome size estimation pre-sequencing	81
3.3.3	Library preparations and sequencing	82
3.3.4	Illumina reads: Quality and contamination filtering.....	83
3.3.5	Illumina reads: Mitogenome assembly and exclusion.....	83
3.3.6	Genome size estimation post-sequencing.....	84
3.3.7	Nanopore reads sequencing and filtering	85
3.3.8	Illumina + Nanopore hybrid assembly	85
3.3.9	HiFi sequencing and assembly	86
3.3.10	Quality assessment and comparison of assemblies	86
3.3.11	Estimation of heterozygosity	88

3.3.12	Genome repetitive elements detection	88
3.3.13	Iso-Seq analysis	89
3.3.14	Genome annotation.....	90
3.3.15	General bioinformatics tools	91
3.4	Results.....	92
3.4.1	Genome sequencing	92
3.4.2	Assemblies comparison and quality assessment.....	93
3.4.3	Final assembly stats	99
3.4.4	Estimation of heterozygosity	101
3.4.5	Repetitive elements and genes annotation.....	102
3.4.6	Iso-Seq analysis	104
3.4.7	Genome size.....	106
3.5	Conclusion.....	107
3.6	Supplementary Material	108

Chapter 4. Comparative genomics of six New Zealand fish species: assembly contiguity affects the identification of genic features but not transposable elements 115

4.1	Abstract.....	115
4.2	Introduction	116
4.3	Materials and Methods.....	118
4.3.1	Tissues collection and DNA extraction	119
4.3.2	Library preparation and sequencing.....	120
4.3.3	Quality and contamination filtering.....	121
4.3.4	Mitogenome assembly and exclusion	121
4.3.5	Genome assembly.....	123
4.3.6	Genome annotation.....	123
4.3.7	Gene family identification and phylogenetic tree construction.....	124
4.3.8	Tests for selection	125
4.3.9	General bioinformatics tools	126
4.4	Results.....	126

4.4.1	Genome sequencing and assembly	126
4.4.2	Repetitive elements and genes.....	132
4.4.3	Genomic comparative analysis	136
4.5	Discussion.....	142
4.5.1	Repetitive elements in fish genomes.....	142
4.5.2	Impact of genome size and fragmentation on genome annotation quality	144
4.5.3	Positive selection in ATP binding genes.....	147
4.6	Conclusions	147
4.7	Pictures credits.....	148
4.8	Supplementary Material	149

Chapter 5. Genomic stock structure of tarakihi (*Nemadactylus macropterus*): evidence of fine-scale adaptation and a temperature-associated cline despite panmixia in New Zealand 163

5.1	Abstract.....	163
5.2	Introduction	163
5.3	Materials and Methods.....	167
5.3.1	Sampling and DNA extraction	167
5.3.2	Whole-Genome Sequencing	170
5.3.3	Quality control and pre-processing	171
5.3.4	Genotyping.....	172
5.3.5	Variant filtering	172
5.3.6	SNP data analysis	175
5.3.7	Genotype-environment association analysis.....	177
5.3.8	General bioinformatics tools	178
5.4	Results.....	178
5.4.1	Sequencing.....	179
5.4.2	SNP datasets	179
5.4.3	Linkage Disequilibrium.....	181

5.4.4	Population structure	182
5.4.5	Genotype-environment association analysis.....	190
5.5	Discussion.....	191
5.5.1	Genetic diversity	192
5.5.2	Linkage disequilibrium	193
5.5.3	Neutral genetic differentiation with king tarakihi and Australia.....	194
5.5.4	Neutral genetic structure in New Zealand tarakihi	195
5.5.5	Adaptive genetic structure	196
5.5.6	Temperature-associated selective cline	198
5.5.7	Fisheries management implications	200
5.6	Conclusion.....	202
5.7	Supplementary Material	204
Chapter 6.	General Discussion	223
6.1	Main findings	223
6.1.1	There is a general lack of studies on the population genetic structure of fishery species in New Zealand.....	223
6.1.2	Genomics is a promising but under-utilized tool for fisheries management.....	224
6.1.3	High heterozygosity and prevalence of intron retention events in the genome of tarakihi.....	225
6.1.4	The proportion of repeat elements in percomorph fishes is correlated to the size of their genome	226
6.1.5	Genes in the ATP binding pathway are especially subject to selection among species.....	227
6.1.6	Tarakihi in New Zealand are a genetically panmictic population with a distribution that follows a temperature-associated cline	228
6.2	Relevance and prospects	228
6.2.1	Implication for tarakihi fisheries in New Zealand	229
6.2.2	Comparative genomics: towards a better understanding of genomic evolution	231

6.3 Recommendations for future research	234
6.4 Conclusions	235
References	238

Data availability statement

Custom made scripts were used throughout this thesis. When mentioned in the text, reported scripts can be found on https://github.com/yvanpapa/PhD_thesis

Glossary

Allele: One of two or more forms of a given gene variant.

Allozyme: Allelic variants of proteins that can be separated by electrophoresis based on differences in charge or structure.

Alternative splicing: The process by which exons can be either excluded or included in or from a pre-mRNA resulting in multiple mRNA isoforms

bp (Kb, Mb, Gb): base pairs. Unit to count the number of nucleotides in a DNA or RNA sequence. If diploid, only one strand is counted.

BUSCO: Benchmark Universal Single-Copy Orthologue. An assessment tool which provides quantitative measures of genomic data completeness in terms of expected gene content. Used to measure the completeness, as a proxy for quality, of genome assemblies and their annotations.

Coding sequence (CDS): Designate the protein-coding segments of exons (i.e. do not include UTR regions).

Circular consensus sequencing (CCS): Long-read, highly accurate sequencing technology. Sequencing is performed on a circular DNA template in which sub-reads are generated during several sequencing passes around the template. These sub-reads are aligned to each other to generate a single high-accuracy consensus read.

Contig: Set of overlapping DNA or RNA fragments that together represent a consensus region of the DNA or RNA sequence. **Contiguity** is a proxy for assembly quality and is usually measured with the N50 metric. A contiguous genome assembly consists of long and few contigs.

Coverage: Sequencing read depth per locus, per individual. Number of times the same DNA / RNA fragment is expected to have been sequenced.

Exon: Segment of a gene that contain the information for coding a protein. Includes the CDS and the UTR regions.

F_{ST} : Measure of population genetic differentiation. Average levels of gene flow based on allele frequencies, under the infinite-sites model.

Gene: Region of DNA that contains the coding information for RNA and protein products. Includes exons (CDS and UTR) and introns.

Gene Ontology (GO): Ontology of defined terms representing gene product properties, notably cellular components, molecular function and biological processes. Each gene can be associated with as many different GO terms as needed. These genes and their associated GO terms are accessible in curated databases like e.g. UniProt.

Genetics: The study of specific and limited numbers of genes, parts of genes, or non-coding sequences that may have a known function. Typically include < 100 genetic markers from specific regions of the genome.

Genome: All of the genetic material of an organism. Includes both coding and non-coding regions.

Genomics: The study of the DNA sequences of an organism at the whole-genome level. Typically including thousands to billions of genetic markers from the quasi-entirety of the genome.

Hi-C: A type of chromosome conformation capture technique. Uses high-throughput sequencing to find the nucleotide sequence of ligated fragments that are physically close in the chromosome molecule. Allows the anchoring of scaffold-level genome assemblies to chromosome regions, resulting in chromosome-level genome assemblies.

Intron: Non-coding regions of a gene that are eliminated by splicing before the translation of mRNA.

Linkage disequilibrium: Non-random association of alleles at different loci in a given population. Influenced by genetic linkage, i.e. the tendency of DNA sequences that are close together on a chromosome to be inherited together by avoiding recombination during chromosomal crossing-over. Also influenced by selection, the rate of genetic recombination, mutation rate, genetic drift, the system of mating, and population structure.

Locus (loci): Region(s) of the genome. Can be a gene or a molecular marker.

L50 / L90: Given a set of contigs, designates the count of the smallest number of contigs whose length sum makes up 50% / 90% of the total genome size. A low L50 / L90 is an indicator of a highly contiguous genome assembly.

N50 / N90: Given a set of contigs, designates the length of the shortest contig at 50% / 90% of the total genome length. . A high N50 / N90 is an indicator of a highly contiguous genome assembly.

Orthologs: Genes which evolved from a common ancestral gene by speciation that usually have retained a similar sequence and function in different species. An **orthogroup** is the set of orthologs derived from the same ancestral gene in all the species under consideration.

Panmixia: random mating (or genetic signature consistent with random mating) of individuals within a population.

PCR: Polymerase chain reaction. Laboratory technique used to amplify short DNA sequences to a large amount of copies. Used as a first step to provide enough material for Sanger sequencing.

Sanger sequencing: Method of direct DNA sequencing. Based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during in vitro DNA replication. Used to obtain the nucleotide sequence of short (typically hundreds to a couple of thousands bp) DNA regions.

Scaffold: A set of contigs separated by gaps of known length. **Scaffolding** of contigs can be performed with various sequencing (e.g. mate-paired reads or long inaccurate reads) and assembly techniques.

SNP: Single Nucleotide Polymorphism. The most common type of genetic variation among organisms, each SNP is a specific nucleotide position that is known to vary among individuals.

UTR: Untranslated region. Either of two sections (5' or 3') that are on each side of the coding sequence of mRNA strand. In a gene, they are considered as part of an exon, the other part being the CDS.

Whole-genome sequencing: DNA sequencing of the entirety, or nearly the entirety, of the genome of an organism. Usually produces thousands to billions of sequencing fragments (reads), of size and accuracy varying depending of the technique that need to be assembled into genome assemblies with bioinformatics. **Whole-genome re-sequencing** consists in mapping fragments of low-coverage whole-genome sequences from a population sample to a contiguous reference assembly in order to detect variants like SNPs, insertions, deletions, copy number variations, etc.

Chapter 1. General Introduction

1.1 Evolutionary genetics and population structure

1.1.1 The evolutionary process

Genetic variation (genetic diversity, genetic polymorphism) designates the differences in DNA sequences between individuals of a given population, or species (Ellegren & Galtier, 2016). Phenotypic variation is determined in part by genetic variation and the evolutionary forces that can act on it to cause changes in the genetic composition of populations (Casillas & Barbadilla, 2017; Charlesworth, 2010). The four key evolutionary processes that will influence genetic diversity are mutations (and recombination), gene flow, genetic drift and selection (Figure 1.1). According to the Hardy-Weinberg principle, allelic frequencies do not change over generations in a population if it is large, randomly mating, and that mutation, selection, and migration with gene flow are not significant. If any of these conditions are not met, the frequencies of alleles and/or genotypes will change from one generation to the next, leading to evolution.

Mutation and recombination are the fundamental sources of new genetic variation. At each generation, the imperfect mechanisms of chromosome replication, strand breakages and faulty repairs result in a small fraction of alleles being different in the offspring than the ones that would have been otherwise identical to the parents. Changes in nucleotides related to protein-coding genes can be non-synonymous or synonymous, depending on whether they result in changes in translated amino acids or not. Synonymous mutations are generally considered neutral, in that they do not affect the phenotype of the organism. Non-synonymous mutations can be either neutral, deleterious, or advantageous. According to the theory of neutral evolution, most non-synonymous mutations are deleterious and will be eliminated from the population, while only a very small fraction will provide an adaptive advantage (M. Kimura, 1968; Ohta, 1992). As a result, most mutations that persist in a population are neutral or nearly neutral. This means that they either have no influence on the phenotype, or if they do, that they do not affect the fitness of individuals in their current environment (Barrett & Schluter, 2008). Nucleotide mutations are not the only intrinsic

molecular source of new genetic variation in organisms. In sexually reproducing organisms, the process of chromosomal recombination leads to the generation of new combinations of genetic variation by recombining genetic material from both parents. Genetic recombination can also create novel alleles through intragenic recombination (Drummond et al., 2005) or duplication of DNA regions (Hastings et al., 2009). DNA duplication can lead to copy-number variation, which is a major source of nucleotide material for the apparition of new genes (Long et al., 2003). Events of DNA slippage, deletions, fusion and fission can also alter genetic variation through generations. Transposable elements (TE) are also another important molecular source of novel genetic variation (Biémont, 2010; Biémont & Vieira, 2006). These are all forms of mutations and recombination.

Genetic drift is the change in allele frequencies in a population due to random sampling of alleles and individuals (Masel, 2011). Since alleles in each offspring are a random sample of the alleles in the parents (in the absence of selection), allele frequencies will fluctuate through generations due purely to chance. Changes in allele frequencies due to genetic drift can lead to the permanent loss or fixation of some alleles (Hamilton, 2009). Especially, rare alleles will tend to disappear quickly due to genetic drift if no other evolutionary forces are acting on them. Genetic drift is always reducing the number of alleles over time and counteracting the effects of mutations. The strength of genetic drift is inversely related to the effective size of a population (N_e). In small populations, the loss of genetic variation occurs much more quickly than in a large population.

Selection is the evolutionary force that can lead to adaptation (the better fit of an organism to its environment). It happens when individuals with a particular allele have a higher chance of reproduction because of an associated phenotypic trait. Differentiation in reproductive success in the context of selection is called fitness. Selection acts on phenotypes but the consequences are in the genotypic changes that occur from one generation to the next. In the case of directional selection, mutations that increase the fitness of individuals tend to be preserved (positive selection) and rise to fixation, and mutations that decrease fitness tend to be eliminated (negative or purifying selection). Through this process, selection can lead to *adaptation*: some individuals, populations, or species will be better adapted to a given environment (e.g. have a better fitness) than others. Selection can be a major driver of

allele frequency change. As opposed to genetic drift which will affect all parts of the genome randomly, selection tends to act only on targeted parts of the genome. Adaptive alleles spread through a population and reduce variation in the genomic area near the locus under selection (Vitti et al., 2013). This produces regions of the genome with allele frequency patterns that deviate from what is expected under neutrality. This is due to the propensity of loci that are located physically close by to be inherited together, resulting in linkage disequilibrium. This effect is sometimes referred to as “genetic hitchhiking” (Nosil & Feder, 2013). Although “hitchhiker” loci are not necessarily subject to selection, they can often show that same pattern as a locus experiencing selection. In small populations, the strength of selection can be overcome by genetic drift. In some situations, strong genetic drift could result in either advantageous alleles being eliminated from a population, or (mildly) deleterious alleles rising to fixation. Selection differs from the other evolutionary forces because it can fix alleles (positive selection), eliminate alleles (negative selection), and maintain alleles (balancing selection) in a population.

Gene flow is the consequence of migration of individuals between populations that results in reproductive success. This results in the exchange of alleles and genotypes between populations (Slatkin, 1987). Gene flow can alter allele frequencies and bring new genetic variation to a population, or it can maintain genetic homogeneity between two populations. The consequences of gene flow are determined by how genetically different populations are from each other and what other evolutionary forces are also operating. The level of gene flow between groups of organisms is the basis of the concept of *population structure*. A lack of gene flow between two groups of individuals will eventually lead to genetic divergence, as mutations, genetic drift, and selection will operate independently within each group. This produces two different evolutionary paths and ultimately the formation of two distinct populations (Morjan & Rieseberg, 2004). If gene flow between populations is significantly limited for a long enough period of time, this can ultimately lead to speciation due to reproductive isolation (Coyne & Orr, 1998). Conversely, high levels of gene flow will result in low genetic divergence among populations (Lenormand, 2002). Two populations that become genetically homogenised by gene flow can be considered to be functioning as one genetically panmictic population (Bohonak, 1999).

1.1 Evolutionary genetics and population structure

Figure 1.1. Schematic view of the four evolutionary processes acting on a population.

1.1.2 Population structure

The levels of gene flow (and thus structure) among natural populations depend on the propensity and the capacity of individuals to migrate from their natal population to another area (Slatkin, 1987). The capacity for dispersal and ultimately migration with gene flow depends on both environmental conditions and life-history traits (Bradbury et al., 2008). As such, lower levels of gene flow are expected to be more common when habitats are fragmented across a land- or sea-scape (Vrijenhoek, 1997). As an example, disruptions to gene flow typically occur when the Earth enters periods of global cooling: associated drops in sea levels can expose land that can become migration barriers to marine populations connectivity (Wallis & Trewick, 2009). The resulting reduction in effective population size can reduce the level of genetic variation and the lack of gene flow increases the genetic structuring of species. These patterns of population disruption can be still be detected in the way genetic variation is distributed in modern populations, even if the populations have since moved or re-established gene flow (Hewitt, 2000, 2004).

The genetic population structure of organisms in the wild can be roughly classified into three types (Laikre et al., 2005) (Figure 1.2): (I) panmixia: all individuals are equally likely to mate regardless of geographic distance, resulting in no genetic differentiation; (II) isolation by distance: the population is more or less continuous and gene flow between two locations decreases as a function of distance (Wright, 1943); (III) spatially distinct populations: past or

current low level of gene flow has resulted in significant genetic differentiation among groups of individuals. Additionally to these three main classes as defined in Laikre et al. (2005), some specific cases of genetic structure can also be defined as temporal (two or more genetic lineages can occur at the same place but not at the same time) and sympatric (where two or more genetic lineages occur at the same place and at the same time).

Figure 1.2. Schematic representation of the three main classes of genetic population structure. Colours represent variation in allele frequencies. In a panmictic population (I), gene flow is strong and uniform throughout the whole range, which results in homogenization of allele frequencies. This can happen even if the populations are demographically uncoupled, as long as gene flow is strong enough. Isolation by distance (II) results in a somewhat continuous gradient where individuals separated by more distance are more likely to show differences in genetic variation. Populations can be considered genetically and spatially distinct (III) when little to no gene flow happens between individuals living in specific geographic ranges. Note that for simplicity, in this figure, the New Zealand islands (in grey) are not considered barriers to gene flow: in reality, land would increase the actual geographic distance between marine individuals.

1.2 Population genomics

Testing for genetic differentiation requires a set of allele polymorphisms that can be used as genetic markers for describing the genetic variation. For example, traditional genetic markers of allele polymorphism include allozymes and microsatellites. Directly determining DNA sequences (via Sanger direct sequencing) has enabled variants to be detected to the level of single nucleotide polymorphisms, or SNPs. Although SNP screening studies have traditionally only involved a small number of genetic markers (or in this context, “loci”), due to the constraints of PCR amplification and Sanger sequencing, recent developments in

massive parallel sequencing technology have enabled SNP detection to be scaled to easily produce extremely large quantities of data points. These methods allow studies to use SNPs from thousands to a virtually unlimited number of loci and on a genome-wide scale. This technical advance has effectively transitioned the field of genetics to what is now called genomics (Bernatchez et al., 2017; Charlesworth, 2010). Reduced-representation genome-wide sequencing methods, such as genotyping-by-sequencing (GBS) or restriction site associated (RAD) markers have provided cost-effective ways to obtain a consistent sample of SNPs from large genomes. However, the resulting sample of polymorphisms is still somewhat biased since it is limited to the restriction enzyme cut sites (Casillas & Barbadilla, 2017). When a reference genome assembly is available, a lower coverage whole-genome resequencing approach can be used. Whole-genome resequencing potentially allows the detection of hundreds of thousands of SNPs from genomic regions experiencing a wide variety of molecular and demographic influences.

This technological transition from genetics to genomics is of particular interest to the study of population structure. Patterns of population differentiation or isolation by distance can be very weak due to different factors such as recent divergence, high gene flow, or low genetic drift due to large effective population size (Attard, Beheregaray, et al., 2018; Palsbøll et al., 2007) and might be undetectable using a small number of genetic markers such as allozymes or microsatellites or small scale SNP methods (Teske et al., 2018). This is particularly an issue for marine organisms because they often have large population sizes and high dispersal potential. These factors can result in high levels of genetic diversity but weak genetic differentiation (Sandoval-Castillo et al., 2018). The amount of information provided by genome-wide sequencing datasets offers an unprecedented level of resolution and has already proven to be highly effective for detecting such weak patterns of structure that may have been undetected by conventional genetic markers (Benestan et al., 2015; Lal et al., 2016; Vendrami et al., 2017). Genetic structure can even be detected in very few individuals as long as the SNP density is high, thus facilitating the data acquisition effort (Jeffries et al., 2016; Nazareno et al., 2017).

1.3 Genomic evidence of selection and adaptation

Genome-wide sequencing makes it possible to distinguish patterns of neutral variation from presumably adaptive variation. Before genomics, that task was difficult or impossible because of the stochastic effects of the small number of genomic regions observed and the impossibility to determine the baseline for neutral expectation (Ellegren, 2008; J. Zhang et al., 2005), which could result in incorrect conclusions. The ability to detect selection can provide important new insights into the fundamental processes of evolution and can improve our understanding of how populations function within their environments. It has been suggested that both neutral and adaptive genetic markers may be used for a better approach to the delineation of conservation units (Funk et al., 2012) and can provide guidance to biodiversity management (Bernatchez et al., 2017; E. E. Nielsen et al., 2009; Russello et al., 2011). A better understanding of micro-evolutionary processes can help with predictions about changes in the abundance and distribution of populations, especially if the environment is rapidly changing due to anthropogenic pressures such as climate change (Gonzalez et al., 2012).

1.3.1 Local adaptation

Local adaptation occurs when organisms have a higher relative fitness in their local habitat than organisms originating from other habitats (Kawecki & Ebert, 2004; Williams, 1966). If the selective pressures are stronger than the homogenising effects of gene flow and the random effects of genetic drift, allele frequencies of the selected traits will differ in their distribution and pattern when compared to neutral alleles (Hauser & Carvalho, 2008; Hoban et al., 2016). Adaptive and neutral variations have different patterns of genetic variation. Genetic drift causes “neutral differentiation” randomly but evenly across the genome, while selection causes “adaptive differentiation” on areas that contain or are linked to genes coding for the trait under selection. Marine species will be likely to show adaptive genetic variation because their population sizes are often large, so drift is weak and selection can operate efficiently, and they are often widely distributed across heterogeneous environments (Attard, Brauer, et al., 2018; Oleksiak & Rajora, 2020; Sandoval-Castillo et al., 2018).

With the availability of genome-wide data, presumably adaptive regions of the genome can be detected using “genome-scan” methods. Although the only way to directly

test for adaptation is through controlled experiments (e.g. reciprocal transplants or common garden experiments) (Kawecki & Ebert, 2004), these genome scans can be used as indirect tests of adaptation to pinpoint genomic locations of interest and draw hypotheses (Ahrens et al., 2018; R. Nielsen, 2005). These “outlier” regions can then be considered “presumably adaptive” (Funk et al., 2012). Two of these methods that have been the most widely used to detect putative evidence of selection, and particularly local adaptation in wild populations, can be broadly divided into two main categories: outlier analyses and environment-association analyses (Ahrens et al., 2018; Dayan, 2020; Hoban et al., 2016).

Outlier analyses use empirical data to simulate null distributions of population genetic parameters under several demographic scenarios. These parameters can include descriptors of, for example, linkage disequilibrium or population differentiation (F_{ST}). The underlying assumption is that neutral genetic differentiation will affect these parameters uniformly across most of the genome. On the other hand, selection will act only on the targeted genes and the surrounding genomic areas, which will skew the parameter statistics in these regions. The “outlier” loci deviating from the neutral distribution are thus expected to appear in the extreme tails of the distribution of the statistic: in these cases, they are flagged as candidates for adaptation (Beaumont & Nichols, 1996; Dayan, 2020). This type of approach has however two challenges. First, many of them rely on demographic models that might not fit the reality (Joost et al., 2013). However, recent advances in the algorithms have at least partially overcome this problem (e.g. the OutFLANK method (Whitlock & Lotterhos, 2015)). The second main issue is that outlier analyses require that separate lineages are defined *a priori* (M. R. Jones et al., 2013; Joost et al., 2013).

The second approach, environmental association analyses, consists of selecting a set of environmental variables that are assumed to drive spatially varying selection and identifying genetic variation that is highly correlated with these variables (Dayan, 2020). These types of methods are also referred to as gene-environment, genotype-environment, or locus-environment association analyses and are the foundation of what is called “landscape genomic” studies (Rellstab et al., 2015). The underlying assumption for this type of analysis is rather intuitive: since in the case of local adaptation, it is by definition the environment that drives adaptive differentiation, loci under selection will display allele frequencies that are

correlated with the environmental variable responsible for the selective pressure. These approaches have several advantages over outlier analyses, in that they do not always rely on demographic models and *a priori* population structure (M. R. Jones et al., 2013; Joost et al., 2013). Especially, environment association analyses using constrained ordination methods (e.g. redundancy analysis (Forester et al., 2016)) do not need to control for neutral genetic structure but take instead a spatially explicit approach, by incorporating geographic distance as a predictor of genetic distance among samples. Constrained ordination allows for the generation of “orthogonal linear combinations of multiple multi-variate explanatory variables (spatial relationships and environmental variables) that explain linear combinations of the response data (allele frequency)” (Dayan, 2020). The two main drawbacks of these methods are the reliance on a set of environmental variables and the necessary assumption that these variables are lowly correlated with each other. Because of the strengths and weaknesses of both methods, the two types of analyses are often used conjointly in population genomics analyses when testing for evidence of adaptation in wild marine populations (Benestan et al., 2016; Sandoval-Castillo et al., 2018)

1.3.2 Testing for selection among species

The d_N/d_S ratio (Z. Yang & Bielawski, 2000) is one of the most widely used methods to detect positive selection among independent lineages (Casillas & Barbadilla, 2017; Kryazhimskiy & Plotkin, 2008). This method is based on the assumption that synonymous nucleotide substitutions (substitutions are mutations that became fixed) in the genome are always neutral. In other words, substitutions that do not change an amino acid do not affect the phenotype, and thus should have no effect on fitness. Non-synonymous mutations that are advantageous are expected to spread through the population and reach fixation faster than neutral mutations. Non-synonymous substitutions in nucleotide sites that are under functional constraints should happen less often than neutral substitutions, due to purifying selection. Thus, selection pressures can be quantified by comparing the rate of nonsynonymous substitutions (d_N) with the rate of synonymous substitutions (d_S) among lineages: $d_N/d_S = 1$ under neutrality, is < 1 under functional constraint (i.e. negative or purifying selection), and is > 1 under positive selection. While broadly used in comparative frameworks, this method is restricted to different species or distantly diverged, independent

lineages because the model's assumptions are violated when the substitutions are actually transient polymorphisms in a population instead of fixed mutations (Kryazhimskiy & Plotkin, 2008). Modern applications of this method can estimate the level of selection pressure on protein-coding genes in a phylogeny by testing hypotheses against specific mathematical models (Z. Yang, 2007). The strength of these models is that they correct for multiple changes and account for the different numbers of synonymous sites and non-synonymous sites among lineages (Jeffares et al., 2015).

1.4 Genome assemblies

Using whole-genome resequencing for the analysis of population structure requires the alignment of low coverage genomes of population samples to a high-quality, contiguous reference genome. Since the first complete human genome was sequenced almost 20 years ago (Collins et al., 2003), the number of *de novo* reference genomes being produced has rapidly increased for a variety of non-model organisms, especially over the last five years (e.g. Austin et al. (2017), Jansen et al. (2017), Liu et al. (2017), Malmstrøm et al. (2017), Tan et al. (2018), Xu et al. (2017)). Several large scale projects exist today that are aiming to assemble the genome of thousands of species (Fan et al., 2020; Feng et al., 2020; Koepfli et al., 2015). Methods such as hybrid assembly of short reads (100s bp) with more error-prone but longer reads (1,000s bp) have shown to yield a substantial improvement in assembly contiguity and completeness (de Lannoy et al., 2017; H. M. Gan et al., 2017; Jansen et al., 2017; Tan et al., 2018; Zimin, Puiu, Hall, et al., 2017). New sequencing technologies and protocols are continually being developed and assembly algorithm pipelines have had to constantly adapt to these changes to the data types (Rice & Green, 2019; Whibley et al., 2020). The latest technologies include the circular consensus sequencing (CCS) PacBio technology that can produce reads that are both thousands of bp long and highly accurate, and scaffolding techniques like chromosome conformation capture methods (e.g. Hi-C technology) that capture the chromatid conformation of DNA in cells and help anchor sequence scaffolds to full-length chromosomes.

A sequenced genome has the most value when it is annotated. Genomic regions can be functionally characterized and mapped (e.g. genes, exons, introns, repeat regions), but this

is far from being a straightforward exercise (Jeffares et al., 2015). A *de novo* reference genome can be annotated using homology-based methods, *ab initio* methods, or most commonly, a mix of both (Abril & Castellano, 2018; Lantz et al., 2018; Yin et al., 2018). While homology-based methods find genes and repeat elements by integrating genetic and/or transcriptomic data from other organisms as evidence, *de novo* methods look for predictive patterns of coding sequences (e.g. codon bias). This task is complicated in eukaryotes due to the complex nature of their genes compared to prokaryotes (Abril & Castellano, 2018). Moreover, some of these patterns can be species-specific. One of the most widely-used annotation methods is thus to train the predicting capacity of *ab initio* algorithms with homology evidence for each newly assembled genome (Holt & Yandell, 2011). Although the assembly and annotation of a genome “from scratch” is still relatively resource-intensive and computationally complex, once it is done it can greatly improve the potential for subsequent analyses of the same species, and in some cases for related species too. The evolution and content of genomes can be studied in a comparative framework (Feng et al., 2020; Randhawa & Pawar, 2021; Shao et al., 2019; Z. Yuan et al., 2018). Polymorphism information of a population acquired from whole-genome resequencing can be traced back to their gene location, allowing to perform more analyses, e.g. the inference of linkage disequilibrium through direct observation of the chromosomal location of SNPs (e.g. Diopere et al. (2018)), and the identification of protein-coding genes with transcriptome data (Austin et al., 2017; Tan et al., 2018).

1.5 The application of genomics in fisheries management

Seafood plays a considerable role in meeting current and future food needs (Bernatchez et al., 2017). It is the only remaining wild animal protein source consumed at such a large scale. Global fish production was estimated to have reached 179 million tonnes in 2018, 156 million tonnes of which were used for human consumption (FAO, 2020). Of these, 84.4 million tonnes came from wild capture marine fisheries, which is very close to the all-time record of 86.4 million tonnes in 1996 (FAO, 2020). Since the advent of globalization and the industrialization of fishery catches in the mid-20th century, fisheries have been declining globally in response to unsustainable catch practices and other anthropogenic pressures, resulting in overfishing (Myers & Worm, 2003; Pauly et al., 2002). Numerous fisheries have collapsed or have become seriously depleted (Mullon et al., 2005). A notable

example is the North Atlantic cod fisheries, which was previously one of the largest fisheries in the world (Myers et al., 1997). Still, a large proportion of the world's marine resources including fisheries is currently overexploited, near collapse, or commercially extinct (McCauley et al., 2015; Pauly & Zeller, 2016). It was estimated by the United Nations Food and Agriculture Organization that 34.2% of the world's fisheries stocks were overfished in 2017 (FAO, 2020). Proper management is thus urgently needed in order to maintain sustainable stocks over time to prevent the dire economic, social and environmental consequences that collapsed fisheries stock could cause (Pullin, 2000; Reiss et al., 2009; Worm et al., 2009).

Marine species typically consist of self-sustaining, reproductively isolated populations. One of the fundamental questions that need to be addressed to manage fisheries efficiently is how to properly delineate these populations into management stocks (Begg et al., 1999; Cadrin, 2020; Kerr et al., 2017; Waples et al., 2008). This is primarily because most management strategies (including those used in New Zealand) rely on biomass harvesting limits that balance catch mortality levels with stock reproduction rates (Beddington et al., 2007; Gibbs, 2008). The success of these approaches relies on the ability to properly delineate the spatial distribution of the biological units so they can be assessed and monitored in a biologically relevant way and quotas can then be allocated (Ward, 2000). Mismatches between management settings and biological units could have disastrous consequences (Benestan, 2019; Reiss et al., 2009). Treating several biological populations as one stock may result in the depletion of one of these populations, due to the differences in abundances among the biological units and/or an uneven distribution of harvesting pressure (Figure 1.3).

In a population genetics framework, the concept of reproductively isolated stocks should correspond to the concept of distinct genetic populations (Altukhov, 1981). Since genetic differentiation is the most definitive evidence of reproductive isolation (Cadrin, 2020), the use of genetic markers has become an important method for delineating fisheries stocks (Benestan, 2019; Bernatchez et al., 2017; Cuéllar-Pinzón et al., 2016; Ovenden et al., 2015; Waples et al., 2008; Ward, 2000). The applications of genetics, and subsequently genomics for the delineation of stock structure and other issues related to fisheries management are reviewed in detail in Chapter 2 of this thesis.

Figure 1.3. Schematic representation of two major issues resulting from a mismatch between biological populations (black outlines) and management units (red outlines). Figure retrieved from Benestan (2019).

1.6 Study species: Tarakihi

1.6.1 Tarakihi biology and dispersal

Tarakihi (*Nemadactylus macropterus* (Forster 1801)), also called Jackass morwong in Australia, and referred to in this thesis as tarakihi, is a demersal marine fish of the Latridae family (K. Kimura et al., 2018; Ludt et al., 2019) that can be visually recognized from other New Zealand “morwongs” (except king tarakihi, see below) by the black saddle across its nape (Figure 1.4). Post-larval fish are pelagic drifters, juveniles are bottom dwellers, and adults form schools over the open bottom in depths ranging from 50 to 250 meters. The species is

widely distributed around the inshore areas of New Zealand and along the southern coast of Australia, including Tasmania (Figure 1.4). Occurrences have also been reported around St. Paul and Amsterdam Island in the Indian Ocean and off the Argentinian coast in South America. However, these reports are considered dubious as they may have been misidentifications of close relative species, respectively *N. monodactylus* and *N. bergi* (Burrige, 1999; Duhamel, 1989; Fricke et al., 2020). In New Zealand, the species occurs all around the continental shelf of the North and South Island and as far as the Snares, Chatham and Three Kings Islands. (Roberts et al., 2015).

Figure 1.4. A tarakihi (*Nemadactylus macropterus*) in Port Davey, Tasmania, February 2013. Photo by courtesy of Ian Shaw, Reef Life Survey. Lower right: Distribution of tarakihi, range in blue (retrieved from wikimedia.org).

Tarakihi are serial summer-autumn spawners (Tong & Vooren, 1972). Three main spawning aggregations have been identified in New Zealand: the East Cape area on the northeast of North Island, the northeast coast of South Island from Pegasus Bay to Cape Campbell, and the lower west coast of South Island (Figure 1.5), although the exact extent of these spawning grounds is not clearly defined and often relies on observations of eggs or ripe, running ripe and spent fish (Hurst et al., 2000; J. R. McKenzie et al., 2017; Morrison et al., 2014). The larval and post-larval stages (also called “paperfish”, Figure 1.6) are pelagic and

drift in offshore waters for 9–10 months. During this phase, the spatial distribution of the fish is thought to be largely influenced by oceanic currents (Figure 1.5) and mixing of larvae originating from different sources likely occurs (Annala, 1987; Bruce, 2001). After c. 10 months, post-larvae morph into juveniles and settle in shallow nursing grounds (Vooren, 1972). The locations of nursing grounds are again poorly known but could potentially be very far from the spawning areas (Figure 1.5). Canterbury Bight and Pegasus Bay on the east coast of South Island are recognised as important nursery regions (Hurst et al., 2000; Langley, 2018; Vooren, 1975). Other areas like the coastal reefs along Napier, Stewart Island and the Golden and Tasman Bays showed evidence for the presence of nurseries.

Figure 1.5. (A) Map of New Zealand showing the locations of the known tarakihi spawning and nurseries areas. The blue arrows indicate the main ocean currents that could potentially influence post-larval dispersal (Annala, 1987). (B) Patterns of migration of adult tarakihi (black arrows) from the east of South Island nursery grounds to the north of North Island based on trends in age and size structure data (Langley, 2018; J. R. McKenzie et al., 2017). Movement of adults on the west coast is still poorly known. Top left figure: tarakihi fisheries management areas (source: Fisheries New Zealand (2021)).

Few studies have been successful in assessing and understanding tarakihi movements, but the tagging data that has been collected showed that adults can migrate for distances up

to 320 km (Annala, 1987; M. K. McKenzie, 1961). Moreover, tarakihi are known to have a lifespan of more than 30 years (Beentjes et al., 2012), which suggests that their dispersal potential over their lifetime is significantly high. Recent studies on trends in age and size structure found evidence that the east coast of North and South Island could be one continuous stock with the Pegasus Bay/Canterbury Bight acting as the main nursery (Langley, 2018). Once they reach maturity, it is thought that a proportion of adult fish migrate northward along the east coast to East Cape, then up to the east Northland area (Langley, 2018). While movements on the west coast are still poorly known, recent observations on the relative strength of individual year classes and growth rates of older fish seem to indicate that Tasman Bay/Golden Bay could act as a nursery area for the entire west coast (Fisheries New Zealand, 2021; Langley & McKenzie, personal communication, April 2021).

Figure 1.6. Two life cycle stages of the tarakihi. In the late-larval stage (top), the individuals display a deep body with a strong ventral keel particularly adapted to pelagic life. This “paperfish” stage is characteristic of species of the Latridae family (Ludt et al., 2019), and is thought to be an important stage for dispersal of tarakihi due to its duration (c. 10 months) and the strong influence of oceanic currents on its spatial distribution. Metamorphosis into the juvenile stage (bottom) occurs when the fish are 7–12 months old and 7–9 cm long. At this stage, the individuals settle in “nurseries” and become more sedentary. Photos courtesy of the New Zealand Institute for Plant and Food Research Limited.

1.6.2 Tarakihi fisheries and management

Tarakihi are usually caught by commercial vessels using bottom trawling but also by set nets off the Kaikoura coast (Langley, 2017), either as a target species or as a bycatch. It is one of the most important commercial finfish species in New Zealand and supports national scale fisheries with annual landings averaging over 5,000 tonnes for the past 40 years (Ministry for Primary Industries, 2017). Yearly harvested biomass has been limited by Total Allowable Catch (TAC) since the national implementation of the Quota Management System (QMS) in the 1980s. More recently concerns were raised when it was estimated that the spawning biomass of stocks along the East Coast of New Zealand has been below the soft limit, corresponding to 20% of the unexploited, equilibrium spawning biomass, since the mid-2000s (Langley, 2018) (Figure 1.7). According to the stock performance indices used by Fisheries New Zealand, this means that this stock is “considered to be overfished or depleted and needs to be actively rebuilt”. Consequently, the total allowable commercial catch (TACC) was reduced in 2018 and again in 2019 for these areas (between 2017 and 2019, TAR1: from 1,447 t to 1,045 t, TAR2: from 1,796 t to 1,350 t, TAR3: from 1,403 t to 936 t). These were the first reductions in tarakihi TACC since the 1980s (Fisheries New Zealand, 2021).

1.6 Study species: Tarakihi

Figure 1.7. Annual trend in spawning biomass (SB) relative to the unexploited, equilibrium spawning biomass (SB_0). The black dashed line represents 40% of SB_0 and the grey dashed line represents the soft limit, which corresponds to 20% of SB_0 . The shaded area is the 95% interval. Figure retrieved from Langley (2018).

Accurate assessment and effective management of tarakihi stocks around New Zealand is thus a high priority, but this relies heavily on an accurate description of their population structure (Hanchet & Field, 2001) and this information has been difficult to obtain. Current doubts about the number and spatial extent of reproductively distinct groups is a major source of uncertainty in the management of tarakihi fisheries (J. R. McKenzie et al., 2017). As such, tarakihi around the main islands are currently considered as one panmictic stock and are managed by fisheries in relation to administrative management boundaries (Figure 1.5) which do not reflect the natural biological units of the species (Fisheries New Zealand, 2021).

1.6.3 Population structure and genetics of tarakihi: brief review of current knowledge

The high potential for long-distance migration and widespread dispersal as well as a presumed lack of barrier to gene flow in the post-larval and adult stages of tarakihi can explain why biological stock boundaries are so difficult to determine. Several studies have attempted to address this question for Australian and New Zealand populations. A complex dynamic of partial mixture between discrete populations was suggested for tarakihi around Australia based on otolith microchemistry of juveniles and adults (Thresher et al., 1994) and modelling of larval dispersal (Bruce, 2001).

In New Zealand, variation in otolith iron content was used to segregate the population into three groups: North, centre, and South, although it was unclear if this result was due to heredity or environmental conditions (Gauldie & Nathan, 1977). Allozyme frequency of one polymorphic allele indicated seven discrete populations of tarakihi around the North and South Island, but the authors could not conclude if this was due to environmental selection as opposed to neutral differentiation, as alleles frequencies correlated with a North-South temperature gradient (Gauldie & Johnston, 1980). Further allozyme and mitochondrial DNA RFLP studies (N. G. Elliott & Ward, 1994; Grewe et al., 1994; Richardson, 1982) did not detect any structure among Australian stocks and found a weak but significant genetic disjunction between Australia and one location in New Zealand. Those results were partially inconsistent with Burrige & Smolenski (2003) who, using seven microsatellite DNA loci, found no divergence among samples from Australian locations but also nothing between Australia and New Zealand. The second and most recent study to investigate the population structure of tarakihi in New Zealand used direct sequencing of the mitochondrial DNA control region from 370 tarakihi collected around New Zealand main islands and the Chatham Islands (Papa, Halliwell, et al., 2021). Again, no overall genetic structure was found among New Zealand and Chatham Island populations. However, two weak genetic disjunctions were detected: one between the west and east coasts of the South Island, and the second was between Hawke's Bay and East Northland (but not between the sampled locations between them).

All of the genetic markers used from these studies were likely to be underpowered for detecting genetic differentiation in a widespread and highly mobile marine species because a small number of markers commonly lacks the power to detect population differentiation in a highly heterozygous species with large population size (Hemmer-Hansen et al., 2014; E. E. Nielsen et al., 2009). The question thus remains whether the use of genome-wide markers could lead to different results.

1.6.4 The case of the king tarakihi

The king tarakihi (Figure 1.8) is a distinct morph of tarakihi caught in the northern areas of New Zealand. Their range also includes some inshore areas of Australia. King tarakihi have recently been proposed as a separate species from tarakihi, with the name *Nemadactylus rex*, based on allozymes and RAPD divergence from tarakihi (P. J. Smith, Roberts, et al., 1996). While this biological distinction between tarakihi and king tarakihi is acknowledged by fisheries managers, fishermen are not required to differentiate them in their catch reports due to the high risk of misidentification and consequent misreporting. It is considered a minor proportion of the TAR1 catch. Subsequent molecular studies brought further evidence of a genetic differentiation level of king tarakihi concordant with what is expected in two separate species (BurrIDGE, 1999; Papa, Halliwell, et al., 2021; P. J. Smith et al., 2008). The scientific name *N. rex* is currently considered invalid with no type designated and from a formal taxonomical point of view, king tarakihi is currently referred to as *Nemadactylus* n.sp. (Roberts et al., 2020).

Figure 1.8. King tarakihi (*Nemadactylus* n.sp). The main visually recognizable difference with tarakihi is the presence of a black band on the pectoral fin. Pictures from Roberts et al. (2015), courtesy of the Museum of New Zealand Te Papa Tongarewa.

1.7 Aims and thesis structure

The two general aims of this thesis were to (1) study the evolution of the tarakihi genome and (2) assess the population structure of tarakihi in New Zealand (Figure 1.9). To achieve these goals, the genomes of the tarakihi and five other New Zealand species were assembled to perform comparative genomic analyses, and DNA samples from tarakihi and king tarakihi around New Zealand and Australia were collected and sequenced. Following a systematic review in Chapter 2, each data chapter is dedicated to addressing one or both of these questions (Chapters 3, 4, and 5).

1.7 Aims and thesis structure

Figure 1.9. Flowchart diagram showing the structure of this thesis research with aims and methods of each chapter.

Chapter 1 (this chapter) is an introduction providing the background and the framework of the conducted research.

The objective of **Chapter 2** was to conduct a literature review of the studies on genetic stock structure of fishes in New Zealand and to outline the use of genomics to address multiple questions related to fisheries. This provided an important framework for understanding the reasons why detecting population structure and selection of a fisheries species like tarakihi was important and how using genomics technologies could help reach this goal, with the benefit of being able to address other questions as well, as opposed to potentially cheaper and resource-demanding alternatives like e.g. traditional genetic markers or non-DNA related methods.

Chapter 3 presents a whole genome sequence of the tarakihi and establishes a draft annotated reference sequence for use in the subsequent chapters. Details are reported for the assembly methodology and annotation that used a combination of several technologies for DNA and RNA sequencing and multiple bioinformatics pipelines. The chapter includes a description and discussion of the features and characteristics of the tarakihi genome and transcriptome.

The main goal of **chapter 4** was to place the genome of the tarakihi in a comparative context with other fish species to identify genes under selection. For this, the genomes of five additional New Zealand fish species were assembled, including the close species king tarakihi and the blue moki (*Latridopsis ciliaris*), which are from the same taxonomic family (Latridae). The genomic features of the six newly sequenced fishes were explored, with a focus on the description of repeat elements and genic features diversity. A phylogeny of ray-finned bony fishes was reconstructed using genomic and proteomic data from the tarakihi, the fishes assembled in this chapter, and eight additional fish species for which well-curated genomic data was available online. The phylogeny was used to test for evidence of selection in the tarakihi and the Latridae lineage.

Chapter 5 addresses the primary goal of this thesis research, which was to assess the population genetic structure of tarakihi in New Zealand. For this, low coverage genomic data

from 161 tarakihi specimens collected from New Zealand, 14 tarakihi from Tasmania and 12 king tarakihi were whole-genome sequenced and mapped to the draft tarakihi reference genome reported in chapter 3. Population structure was inferred using neutral and presumably adaptive loci, and associations of genomic variation with environmental features were explored with a particular focus on water temperature.

Each data thesis chapter is divided into sections including introduction, methods, results, discussion (except Chapter 3) and conclusion. Chapter 3 was formatted for publication as a “data note” for journals that publish genome assemblies. As such, there was no Discussion section in this chapter: the interpretation of results was rather distilled in the Results and Conclusion sections, as is generally the case in this type of article.

Chapter 6 is a general discussion of the thesis that synthesizes all chapters and provides a final conclusion to the study.

The following chapter (**Chapter 2**) is based on the Accepted Manuscript version of the following published review article:

Papa, Y., Oosting, T., Valenza-Troubat, N., Wellenreuther, M., & Ritchie, P. A. (2021). Genetic stock structure of New Zealand fish and the use of genomics in fisheries management: an overview and outlook. *New Zealand Journal of Zoology*, 48(1), 1–31. <https://doi.org/10.1080/03014223.2020.1788612>

YP carried out the literature research from which the review was based on, conceived the figures and tables and wrote the primary and the final manuscript. TO wrote the primary draft of the sections “Australasian snapper” and “Fishing-induced evolution”. NVT wrote the primary draft of the section “Loss of genetic diversity and adaptive potential”. MW and PR reviewed the primary and the final manuscript.

Minor changes have been made for complying with the formatting of this thesis: Figures and tables indexing has been formatted differently. The online Supplementary Table has been divided into three Supplementary Tables to be incorporated in the thesis chapter.

Chapter 2. Genetic stock structure of New Zealand fish and the use of genomics in fisheries management: an overview and outlook

2.1 Abstract

Sustainable management of wild fisheries requires accurate delineation of reproductively isolated stocks to avoid depletion of a commercially and ecologically important resource. However, there is still a lack of reliable information on stock structure for most fishery species in New Zealand. DNA markers can assist in the delineation of stocks, but they also can provide significant insights into other areas related to the genetic diversity and the response to pressures. In this review, we first provide a detailed summary of the population genetic studies of New Zealand fish species, with a particular focus on hoki, orange roughy, snapper, ling, and blue cod. We find that genetic data is uniformly lacking for most species. We then discuss how the global shift from low resolution markers to genomics in fisheries genetics has far reaching consequences for the sustainable management of our aquatic resources, by allowing us to address multiple important pressures that wild fisheries are currently facing, and we introduce some of these briefly. We conclude by emphasising the need for a more systematic and holistic approach for the use of genomics in New Zealand fisheries management, so that the best evidence is available to inform the decisions of policy makers.

2.2 Introduction

The New Zealand Exclusive Economic Zone (EEZ, Figure 2.1) covers more than 4,000,000 km² of ocean and is one of the 10 largest EEZ in the world (Migiro, 2018). Wild capture fisheries in the New Zealand EEZ are made up of more than 100 species, and around 600,000 tonnes are harvested annually. In an effort to ensure sustainable utilisation of this resource, spatial, temporal, and size limits are set using a Quota Management System that divides New Zealand waters into Quota Management Areas. Currently, 98 species or groups

of species of fishes, invertebrates, and seaweeds are managed as one or more stocks under the Quota Management System (Fisheries New Zealand, 2019a). Each stock has an annual Total Allowable Catch, which is supposed to be the upper limit of biomass that can be removed without overfishing the stock. Sustainable catch limits, which Total Allowable Catch are based on, are traditionally set using the results of stock assessments when such data are available (Beddington et al., 2007; Gibbs, 2008). Properly identifying the boundaries of fisheries stocks is crucial because it is a necessary parameter in any stock assessment model (Begg et al., 1999; Cadrin et al., 2014b; Waples et al., 2008). Moreover, discrepancies between biological populations and management units can result in overexploitation of stocks (Benestan, 2019; Reiss et al., 2009) and lack of knowledge of biological spatial structure has already resulted in important fisheries collapses (Cadrin, 2020). Overfishing of the infamous Atlantic northern cod in the late 1980s, depletion of small spawning areas of Atlantic herring in the Northwest Atlantic, and overexploitation of crustacean fisheries in the Gulf of Alaska, which all led to dramatic fisheries collapses, have all been attributed, at least partially, to mismatches between the spatial management or assessment units and the actual population structure (Orensanz et al., 1998; Smedbol & Stephenson, 2001). In spite of those potential issues, the current lack of knowledge of fisheries biological stock boundaries in New Zealand makes it impossible to know with certainty if they are matched with their assessment and management stock boundaries. Current stock hypotheses that are reported from New Zealand fisheries assessment literature (Fisheries New Zealand, 2018a) are either inexistent or are based on various types of observations and scientific data, often sparse, and rarely match the Quota Management Areas (Supplementary Table 2.1 & 2.2).

Figure 2.1 Sampling localities for fish population genetic studies in New Zealand. The dashed line shows the boundary of the New Zealand Exclusive Economic Zone (EEZ), and the continuous line is the 1,000 m bathymetric contour, which is the approximate edge of the continental shelf. New Zealand is bathed from the west by water masses of different temperatures: the warm Subtropical Inflow and the cold Subantarctic Inflow. The warm Tasman Front and Subtropical Front are associated with the main flowing surface currents that might influence fish dispersal around New Zealand. The Subtropical Front separates the warmer northern waters from the colder polar water masses of the Subantarctic Inflow at the Subtropical Convergence (STC, dotted line). South of New Zealand and the Campbell Plateau is the deeper Subantarctic Front and associated cold Antarctic Circumpolar Current. (Currents based on a map from the National Institute of Water and Atmospheric Research).

2.3 The value of genetics in fisheries stock delineation and fisheries management

Over the last decade there have been increasing efforts to use a holistic approach to delineate fish stocks, by combining the results obtained from multiple methods such as physical tagging, morphometrics, parasites as biological tags, life history data (growth, age composition, reproduction, and distribution), otolith chemistry, and genetics (Abaunza et al., 2008; Baldwin et al., 2012; Cadrin et al., 2014a; Corrigan et al., 2018; Espeland et al., 2008; Izzo et al., 2017; Tanner et al., 2014; Zemeckis et al., 2014). The rationale is that these methods capture evidence from different biological processes which can operate over different temporal and spatial scales. As an example, differences in levels of metal ions and trace elements accumulated in otoliths during time spent at a particular locality can be used to infer residency or migratory events on an ecological time scale (months–a few decades). Otolith microchemistry markers thus provide *direct* measures of what an individual has done and where it has been. In contrast, genetic loci are markers that can be used to estimate levels of connectivity at evolutionary time scales (decades–thousands of years). The latter utilises a sample of characters that remain permanent in individuals and are partially passed down from one generation to the next (Tanner et al., 2016). These markers are *indirect* because they measure the consequences of what other individuals (the ancestors) have done and where they have been. Information from genetic markers is said to provide the most definitive inference of reproductive isolation (Cadrin, 2020), although it is important to take into account the intrinsic characteristics of the markers studied, i.e. the mutation rate and the vulnerability to natural selection, in order to interpret results properly.

Using genetic markers either as a standalone method or as part of an integrative approach has led to numerous advances towards establishing the levels of connectivity and structure of fisheries stocks. Recent examples include the discovery of four reproductively isolated stocks of beaked redfish in the North Atlantic (Cadrin et al., 2010), three stocks of Atlantic cod that reproduce at different times of year on the East Coast of North America (Kovach et al., 2010), five global populations of silky sharks based on mitochondrial DNA control region sequences (Clarke et al., 2015), significant genetic population structure among Greenland lumpfish (Garcia-Mayoral et al., 2016) and Greenland halibut (Westgaard et al.,

2017), two mitochondrial lineages of night sharks in the West Atlantic (Domingues et al., 2019), and two genetic stocks of Australasian snapper along the east coast of Australia (Morgan et al., 2019). See also reviews by Cadrin et al. (2014a) and Cuéllar-Pinzón et al. (2016) for more examples. Some of these studies have provided information to fisheries management which could then be used to respond accordingly with practical measures. Examples of these practical applications in North Atlantic fisheries are reviewed in Kerr et al. (2017): they include the setting of temporal and spatial closures of Atlantic cod in North America and the alteration of management stock boundaries of redfish. Detection of morphologically similar mixed stocks through population genetics has also been shown to successfully assist in the management of fragile Atlantic cod populations in Norway. Since allele frequencies of the gene *Panl* can be reliably used to differentiate between the declining Norwegian coastal cod and the stable Northeast Atlantic cod populations (Sarvas & Fevolden, 2005), 'real-time' genotype monitoring of landings could then be routinely undertaken in order to close some fisheries areas when the proportion of Norwegian coastal cods in catches was considered too high (Dahle et al., 2018).

Additionally to stock delineation, genetic tools have other numerous applications for fisheries-related studies and questions. Traditionally, long-term patterns and evolutionary processes are investigated with genetic data to explore population structure, levels of gene flow and connectivity, adaptive differentiation, levels of gene diversity, effective population size (N_e), and inter-species phylogenetic relationships. In the context of fisheries management, the useful applications can extend to age estimation, ecosystem monitoring, and biosecurity, among others. All these well-established and new potential applications for genetics have already been thoroughly reviewed by Ovenden et al. (2015). Now that the levels of resolution have significantly improved with the use of genome-wide DNA sequencing (Bernatchez et al., 2017), that is referred here as 'genomics', the power of these methods has substantially increased. Among all those applications, maintaining levels of genetic diversity in a stock is often an important goal of fisheries management that cannot be addressed without non-genetic methods (Domingues et al., 2018; Hoarau et al., 2005; Pinsky & Palumbi, 2014). However, it is not often monitored and testing for a decline would require long-term sampling that is very rarely undertaken. Despite the importance of both levels of genetic

variation and structure, it has been difficult to include them into fisheries management and planning in New Zealand and globally because there has been limited interest or appropriate support for long-term comprehensive studies (Bernatchez et al., 2017; Waples et al., 2008). Yet, even newer issues related to the genetic integrity of fisheries have started emerging in the scientific literature. As examples, the response of fisheries to new selective pressures due to our current way of size-selective fishing (which might cause a “fishing-induced evolution”) or to the current wave of global warming is still uncertain. Those issues are thus expected to become important topics in the field of fisheries genomics (Benestan, 2019; Bernatchez et al., 2017).

The first part of this review aims at providing an overview of the current state of knowledge on genetics in New Zealand wild fisheries, with a detailed review of the population genetic studies that have been reported for New Zealand fisheries species. The second part is a more general overview of how the field of fisheries genetics is currently evolving towards new methods and tools and how they will be beneficial to New Zealand (and global) fisheries management to address the question of stock structure and some other issues related to genetics.

2.4 The current state of New Zealand fisheries genetics

In 2019, 68 fishes were managed under the Quota Management System with a total allowable commercial catch limit. Several of the fish quotas consist of taxonomic units which do not require separate catch reports, usually because it is difficult to distinguish among the grouped species (however, most of them are still assessed separately whenever it is possible). These 68 fishes thus consist of 86 recognised species *sensu stricto* (Table 2.1). To date, only 23 out of the 86 species have been studied for genetic population structure (Figure 2.1, Table 2.2. See Supplementary Table 2.1 & 2.2 for a more detailed report of all genetic studies that include New Zealand EEZ fishes). Only 11 of them have been studied using microsatellites and/or DNA sequencing (as opposed to allozymes and fragment length polymorphism) and none with genome-wide DNA sequencing methods.

Here we focus on commercially important marine finfish and cartilaginous fish species, for which population genetic data are available. Whenever possible, species were classified into one or more of five categories of population genetic structure: (I) panmixia: no genetic differentiation detected within the population; (II) isolation by distance (IBD): a semi-continuous population where gene flow between two locations decreases as a function of distance (Wright, 1943); (III) spatially distinct populations: two or more areas were shown to have significant genetic differences; (IV) temporal variation: significant genetic differences were detected between samples collected in same localities at different times; (V) sympatric: two or more genetic lineages were detected in the population, without evidence of link with spatial or temporal variation.

Table 2.1. List of 68 fishes corresponding to 86 taxonomic units that are managed under the New Zealand quota management system and for which total allowable commercial catch limits are set.

Common name(s)	Scientific name(s)	Taxonomic units	Reported commercial catch in 2019 (kg)
Hoki	<i>Macruronus novaezelandiae</i> [†]	1	122,404,872
Jack mackerel	<i>Trachurus declivis</i> [†] , <i>T. murphyi</i> [†] , <i>T. novaezelandiae</i>	3	40,734,597
Southern blue whiting	<i>Micromesistius australis</i> [‡]	1	31,887,837
Barracouta	<i>Thyrsites atun</i> [†]	1	18,415,124
Ling	<i>Genypterus blacodes</i> [†]	1	16,740,636
Blue mackerel	<i>Scomber australasicus</i> [†]	1	10,299,535
Oreo	<i>Pseudocyttus maculatus</i> [†] , <i>Allocyttus niger</i> [†] , <i>Neocyttus rhomboidalis</i> , <i>A. verrucosus</i>	4	8,949,515
Silver warehou	<i>Seriola punctata</i>	1	8,607,869
Orange roughy	<i>Hoplostethus atlanticus</i> [†]	1	8,018,783
Snapper	<i>Chrysophrys auratus</i> [†]	1	6,349,644
Tarakihi	<i>Nemadactylus macropterus</i> [†] , <i>Nemadactylus</i> sp. [§] (King Tarakihi)	2	5,149,627
Spiny dogfish	<i>Squalus acanthias</i> [‡]	1	5,053,718
Gurnard	<i>Chelidonichthys kumu</i> [†]	1	4,028,509
Trevally	<i>Pseudocaranx dentex</i> [†]	1	3,089,874
Common warehou	<i>Seriola brama</i>	1	2,928,547
Red cod	<i>Pseudophycis bachus</i> [†]	1	2,904,727
Frostfish	<i>Lepidopus caudatus</i> [§]	1	2,869,061
Giant stargazer	<i>Kathetostoma giganteum</i> [§]	1	2,839,591
School shark	<i>Galeorhinus galeus</i> [†]	1	2,733,332
Redbait	<i>Emmelichthys nitidus</i>	1	2,673,588
Hake	<i>Merluccius australis</i> [†]	1	2,641,532
Alfonsino & Long-finned beryx	<i>Beryx splendens</i> [†] , <i>B. decadactylus</i>	2	2,342,198
Gemfish	<i>Rexea solandri</i> [†]	1	2,194,494

Kahawai	<i>Arripis trutta</i> , <i>Arripis xylabion</i>	2	2,042,298
Flatfish	<i>Colistium nudipinnis</i> , <i>Peltorhamphus novaezealandiae</i> , <i>C. guntheri</i> , <i>Rhombosolea retiaria</i> , <i>R. plebeian</i> , <i>R. leporine</i> , <i>R. tapirine</i> [‡] , <i>Pelotretis flavilatus</i>	8	1,939,211
Blue cod	<i>Parapercis colias</i> [†]	1	1,843,655
Rough skate	<i>Zearaja nasuta</i>	1	1,431,725
Ghost shark	<i>Hydrolagus novaezealandiae</i>	1	1,402,288
Rig	<i>Mustelus lenticulatus</i> [†]	1	1,364,148
Elephant fish	<i>Callorhinchus milii</i>	1	1,353,839
Hāpuku & Bass (Gropers)	<i>Polyprion oxygeneios</i> [†] , <i>Polyprion americanus</i> [‡]	2	1,259,378
Sea perch	<i>Helicolenus percoides</i> [§]	1	1,160,885
White warehou	<i>Seriola lalandi</i>	1	1,027,749
Ribaldo	<i>Mora moro</i>	1	1,020,763
Southern bluefin tuna	<i>Thunnus maccoyii</i>	1	956,979
Grey mullet	<i>Mugil cephalus</i>	1	851,969
Pale ghost shark	<i>Hydrolagus bemisi</i>	1	840,215
Smooth skate	<i>Dipturus innominata</i>	1	701,929
Cardinal fish	<i>Epigonus telescopus</i>	1	696,222
Bluenose	<i>Hyperoglyphe antarctica</i> [‡]	1	671,119
John dory	<i>Zeus faber</i> [§]	1	611,135
Blue moki	<i>Latridopsis ciliaris</i> [§]	1	543,325
Lookdown dory	<i>Cyttus traversi</i>	1	420,470
Short-finned freshwater eel	<i>Anguilla australis</i> [†] , <i>A. reinhardtii</i>	2	346,798
Pilchard	<i>Sardinops sagax</i>	1	343,381
Leatherjacket	<i>Meuschenia scaber</i>	1	320,447
Kingfish	<i>Seriola lalandi</i> [‡]	1	316,421
Swordfish	<i>Swordfish</i> [‡]	1	262,437
Rubyfish	<i>Plagiogeneion rubiginosum</i>	1	225,005
Ray's bream	<i>Brama brama</i>	1	140,150

2.4 The current state of New Zealand fisheries genetics

Butterfish	<i>Odax pullus</i>	1	110,807
Blue shark	<i>Prionace glauca</i>	1	100,183
Parore	<i>Girella tricuspidata</i>	1	81,364
Long-finned freshwater eel	<i>Anguilla dieffenbachii</i> [†]	1	65,971
Trumpeter	<i>Latris lineata</i> [‡]	1	62,499
Porae	<i>Nemadactylus douglasii</i>	1	54,594
Bigeye tuna	<i>Thunnus obesus</i>	1	54,417
Moonfish	<i>Lampris guttatus</i> [‡]	1	45,367
Red snapper	<i>Centroberyx affinis</i>	1	37,739
Porbeagle shark	<i>Lamna nasus</i>	1	36,222
Yellow-eyed mullet	<i>Aldrichetta forsteri</i>	1	33,175
Mako shark	<i>Isurus oxyrinchus</i>	1	26,180
Pacific bluefin tuna	<i>Thunnus orientalis</i> [†]	1	21,629
Garfish	<i>Hyporhamphus ihi</i>	1	16,418
Yellowfin tuna	<i>Thunnus albacares</i>	1	4,919
Anchovy	<i>Engraulis australis</i>	1	3,628
Patagonian toothfish	<i>Dissostichus</i>	1	20
Sprats	<i>Sprattus antipodum</i> , <i>S.</i> <i>muelleri</i>	2	3

(†): Fishes for which genetic population structure hypotheses have already been tested among New Zealand locations. (‡): Fishes that have only been part of a broader population genetic study with one or few sampled locations in New Zealand. (§): Fishes that have only been part of studies at an inter-specific scale with one or few sample locations or specimens in New Zealand

2.4.1 Hoki or blue grenadier (*Macruronus novaezelandiae*, Merlucciidae)

Hoki is a schooling benthopelagic fish distributed throughout the southwest Pacific, most abundant between depths of 200 to 800 m. This is New Zealand's largest and most valuable fishery with a reported commercial catch of more than 122,000 tonnes in 2019, i.e. almost a third of the total commercial catches for the same year (Fisheries New Zealand, 2019b). They are mainly targeted in spawning aggregations during the winter in the Hokitika Canyon on the West Coast of the South Island and in Cook Strait on the East Coast of the South Island. The lifespan of the species is around 25 years, with a long pelagic larval and juvenile phase (4–7 years). Although it only comprises one Quota Management Area around North and South Island, Hoki are effectively managed as separate western and eastern substocks. These were defined based on differences in growth and maturation rates, as well as morphometric characters, between west and east coast stocks (Horn & Sullivan, 1996; Livingston & Schofield, 1996). There is some evidence of spawning stock fidelity for those substocks on the east coast (Cook Strait) and the west coast (off South Island), but no characteristics (otolith rings, mean number of gill rakers or otolith microchemistry) appear to be reliable predictors of spawning ground origin (Francis et al., 2011; Hicks et al., 2003).

No evidence in support of genetic differentiation of *M. novaezelandiae* stocks in New Zealand has been found to date. Allozyme studies of nine locations (P. J. Smith et al., 1981), cytochrome *b* direct sequencing on the west coast, Cook Strait, and Chatham Rise (Baker et al., 1995) and mitochondrial DNA restriction fragment length polymorphism (RFLP) on the west coast and Cook Strait (P. J. Smith, McVeagh, et al., 1996) were unable to detect any significant genetic structure. Microsatellites have only been used on two east coast locations as part of a broader geographical study, again showing no substructure between locations (Takeshima et al., 2011). This means that the hypothesis of two main hoki stocks with spawning ground fidelity has not been confirmed by genetic studies to date. This apparent lack of genetic structure could be explained by a level of migration that is high enough to homogenise sub-populations at neutral sites or that the markers used to date (allozymes, cytochrome *b*) were of too low-resolution for a species with such large population size and thus lacked

2.4 The current state of New Zealand fisheries genetics

Figure 2.2. [Continues next pages]

Figure 2.2. [Continued] Maps of fish population genetic studies in New Zealand. Dots represent sample points. Dashed lines are spatial genetic breaks that have been reported as significant by their respective studies. Gradient arrows represent isolation by distance. T: temporal variation. S: sympatric lineages. Colours correspond to the respective reported study, except in the last map where they represent the respective species from the multi-species allozyme study of Gauldie & Johnston (1980). Half-dots are samples that have been re-used across studies. Solid black

2.4 The current state of New Zealand fisheries genetics

lines represent Quota Management Area boundaries, with names in black capital letters (not represented in the last map for visibility). The black dashed line shows the boundary of the New Zealand Exclusive Economic Zone (EEZ) and the continuous line is the 1,000 m bathymetric contour. Sample points particularly close spatially were merged for visibility. When studies only reported broad sampling areas, points were placed arbitrarily in the centre of the reported area. Maps are scaled to cover their relevant areas. Only studies that have tested genetic population structure hypotheses among at least two New Zealand locations are reported here. Images of fish were used with permission from Seafood New Zealand (<https://www.seafood.co.nz/>) or are in the public domain.

Table 2.2. 23 fish species for which genetic population structure hypotheses have already been tested among New Zealand locations.

Species	Current stocks hypothesis	New Zealand Genetic structure	Trans-Tasman diff.	Markers (no. loci)	Sample distribution (no. sites)	References
Hoki <i>Macruronus novaezelandiae</i>	2 stocks: west coast and east coast	(I) panmictic ^{1,2,3,4}	yes ² /no ³	allozymes (15) mt seq (cyt <i>b</i>) mt RFLP (6) mt seq (cyt <i>b</i> +CR), msat (4)	CR (2), WCSI (2), CS (2), PR(2), BP WCSI, CS, CR, Tas WCSI, CS, Tas CR (2), Chile (3)	[1] Smith et al. (1981) [2] Baker et al. (1995) [3] Smith et al. (1996) [4] Takeshima et al. (2011)
Orange roughy <i>Hoplostethus atlanticus</i>	>8 stocks	(I) panmictic ^{1,2,4,8,9} (II) IBD ⁶ (III) distinct populations ^{5,6,7} (IV) temporal variation ⁶	no ^{3,4,5,8,9}	allozymes (22) VNTR (3) allozymes (11) mt seq (cyt <i>b</i>) mt RFLP (6) allozymes (11) allozymes (11), RAPD (29), mt CAPS (3) mt seq (COI+cyt <i>b</i>) msat (9)	ChP (2), Kai, Wai, CR, Atl (2) CR, LHR (2) [1], Aus (6), Tas (5) PB, CR, ChP, RB, LHR, Tas, SA PB, Wait / WCSI, ChP, RB, CR, Tas BP, EC, RB, Wai, Kai / CR (10) PB / Wait / CR, RB LHR, NNZ, ChP, PB, CR, RB, LR, Aus (2), Chile, Atl (3) ChP, PB, CR, RB, BP (4), NNZ (8), Aus (2), Chile, Atl (3)	[1] Smith (1986) [2] Baker et al. (1992) [3] Elliott & Ward (1992) [4] Baker et al. (1995) [5] Smith et al. (1996) [6] Smith & Benson (1997) [7] Smith et al. (1997) [8] Varela et al. (2012) [9] Varela et al. (2013)
Snapper <i>Chrysophrys auratus</i>	possibly 7 stocks	(III) distinct populations ^{1,2}	yes [†]	allozymes (23)	CE, West NNZ (2) / East NNZ (2), HG, BP, EC / HB, MS, TB, Wel	[1] Smith et al. (1978)

2.4 The current state of New Zealand fisheries genetics

				mt seq (CR), msat (7)	East NNZ, HG, EC / TB / Raglan, HB	[2] Bernal-Ramírez et al (2003)
Ling <i>Genypterus blacodes</i>	5 stocks	(III) distinct populations ^{1,2,3}	yes ³	allozymes (2) allozymes (1) mt seq (cyt <i>b</i> +CR)	EC+BP, CS, CB, WCSI / PR [1], Kai, WCSI, BP, SI, CR (10) / PR (3), AI (2) / CI (2) BP, CR / CP / Tas (2)	[1] Smith (1979) [2] Smith & Francis (1982) [3] Smith & Paulin (2003)
Blue cod <i>Parapercis colias</i>	>8 stocks	(II) IBD ^{1,2} (III) distinct populations ^{1,2}	-	mt seq (CR) mt seq (CR), msat (7)	NNZ, BP, MS (3), Wel, Kai, Banks, Ota, Stl, Fio (2) / Chl [1], GB	[1] Smith (2012) [2] Gebbie (2014)
School shark <i>Galeorhinus galeus</i>	1 stock	(I) panmictic ^{1,2}	no ^{1,2,†}	allozymes (29), mt RFLP (10) mt seq (CR), msat (8)	WCSI (2), ECSI, Atl, Arg, SA, Aus (2), Tas (3) East NNZ, Kai, CR, BRSI, Aus (4), Tas (2), Chile (1)	[1] Ward & Gardner (1997) [2] Hernández et al. (2015)
Hāpuku <i>Polyprion oxygeneios</i>	Unknown, probably 1 stock	(IV) temporal variation ¹ (III) distinct populations ²	-	allozymes (2) mt seq (CR), msat (9)	CS (18), Kai, EC, CR, Wai (2), CE (2) NNZ, HB, CS (2), Kai, Ota, CI / WCSI	[1] Smith & Johnston (1985) [2] Lane et al. (2016)
Short-finned freshwater eel <i>Anguilla australis</i>	1 stock	(III) distinct populations ² (adults only) ¹	yes ²	allozymes (10) msat (7)	Rivers: WCSI (2), ECSI (2) / BP, Raglan (2) Rivers: WCSI, Banks, BP	[1] Smith et al. (2001) [2] Shen & Tzeng (2007)
Long-finned freshwater eel	1 stock	(III) distinct populations (adults only)	-	allozymes (9)	Rivers: WCSI (3), ECSI / BP, Raglan	Smith et al. (2001)

*Anguilla
dieffenbachia*

Alfonsino <i>Beryx splendens</i>	unknown	(III) distinct populations	yes (ECNI)	mt seq (cyt <i>b</i>)	NC (7), Japan, ECNI / CR, Aus (2), West Indian Ocean, / Atl	Lévy-Hartmann et al (2011)
Blue mackerel <i>Scomber australasicus</i>	At least 3 stocks	(V) sympatric?	no [†]	mt seq (CR)	NNZ, HB, Aus	Smith et al. (2005)
Black oreo <i>Allocyttus niger</i>	4 stocks	(I) panmictic	yes [†]	mt CAPS (10), nuc CAPS (11)	ECSI (2), CR (2), PSM (6)	Smith, McMillan, et al. (2002)
Smooth oreo <i>Pseudocyttus maculatus</i>	4 stocks	(I) panmictic	no [†]	mt CAPS (10), nuc CAPS (11)	ECSI (1), CR (2), PSM (6)	Smith, McMillan, et al. (2002)
Rig <i>Mustelus lenticulatus</i>	5 stocks	(I) panmictic	-	allozymes (17), mt RFLP (12)	TB, CS, Raglan, West NNZ (2), BP, HG (2)	Hendry (2004)
Hake <i>Merluccius australis</i>	3 stocks	(I) panmictic	-	allozymes (2)	WCSI, Banks, CR, SI, CP	Smith (1979)
Gemfish <i>Rexea solandri</i>	2 stocks	(I) panmictic	no	allozymes (17), mt RFLP (11)	NNZ, WCSI, Aus (8)	Colgan & Paxton (1997)
Pacific bluefin tuna <i>Thunnus orientalis</i>	1 Pacific stock	(I) panmictic	-	msat (7), mt seq (CR+cyt <i>b</i>)	NNZ, WCSI, Taiwan	Tseng & Smith (2012)

2.4 The current state of New Zealand fisheries genetics

Tarakihi <i>Nemadactylus macropterus</i>	1 stock	(III) distinct populations or adaptive cline	yes [†] / no [†]	allozymes (2)	NNZ, HG / BP / EC (2), CS (2), Kai (2), Wait, Ota / NPFS (2) / BRSI (3), WCSI / KB, GB, TB / WA	Gauldie & Johnston (1980)
Trevally <i>Pseudocaranx dentex</i>	c. 3 stocks, may be more	(III) distinct populations or adaptive cline	-	allozymes (2)	NNZ (2), HG / BP, EC, CE, WA	Gauldie & Johnston (1980)
Barracouta <i>Thyrsites atun</i>	At least 4 stocks	(III) distinct populations or adaptive cline	-	allozymes (2)	WA / HG, BP, EC, Wait / Ota, KB, GB, TB	Gauldie & Johnston (1980)
Red cod <i>Pseudophycis bachus</i>	unknown	(III) distinct populations or adaptive cline	-	allozymes (2)	KB / GB, TB / CS, EC	Gauldie & Johnston (1980)
Jack mackerel <i>Trachurus declivis</i>	unknown	(III) distinct populations or adaptive cline	-	allozymes (2)	CE / BP / EC /	Gauldie & Johnston (1980)
Gurnard <i>Chelidonichthys kumu</i>	unknown	(I) panmictic?	-	allozymes (2)	EC, WA, NNZ (2), HG, BP, EC, Wait, Ota, BRSI, WCSI, KB, GB, TB	Gauldie & Johnston (1980)

Note: Only population genetic studies which have sampled more than one area in New Zealand are reported here. Studies with genetic data from only one area (†) are reported in Supplementary Table 2.1 & 2.2. mt: mitochondrial DNA; nuc: nuclear DNA; seq: direct sequencing; cyt *b*=cytochrome *b*; CR=control region; COI= cytochrome *c* oxidase subunit I; RFLP: restriction fragment length polymorphism; VNTR: variable number tandem repeat; RAPD: random amplified polymorphic DNA; CAPS: cleaved amplified polymorphic sequence; msat: microsatellites. ‘/’ indicates a significant genetic differentiation. New Zealand locations: CR=Chatham Rise, WCSI=West Coast of South Island; CS=Cook Strait, PR=Pukaki Rise, BP=Bay of Plenty, ChP=Challenger Plateau, Kai=Kaikoura/Pegasus Bay, Wai=Wairarapa, PB=Puysegur Bank, RB= Ritchie Bank, Wait=Waitaki/Timaru, EC=East Cape, NNZ=Northern NZ, CB=Canterbury Bight, SI=Snare Islands, AI=Auckland Island, CI=Campbell Island, CP=Campbell Plateau, MS=Marlborough Sounds, Wel=Wellington, Banks=Banks Peninsula, ChI=Chatham Islands, Fio=Fiordland, Ota=Otago, StI=Stewart Island, GB=Golden Bay, ECSI=East Coast South Island, CE=Cape Egmont, HB=Hawkes Bay, PSM= Puysegur-Snares-Macquarie Ridge, TB=Tasman Bay, HG=Hauraki Gulf, NPFS=Nugget Point/Foveaux Strait, BRSI=Bluff/Riverton/Solander Islands, KB=Karema Bight, WA=West Auckland, ECNI=East Coast North Island; other locations: LHR=Lord Howe Rise, LR=Louisville Ridge, Tas=Tasmania, Aus=Australia, SA=South Africa, Atl=Atlantic Ocean, Arg=Argentina, NC=New Caledonia. Current stocks hypotheses are reported from New Zealand fisheries assessment literature and are based on various types of observations and scientific data that sometimes include genetics.

discriminatory power. On a broader geographical scale, comparing New Zealand and Tasmanian populations led to contrasting results. While Baker et al. (1995) detected significant differences in haplotype frequencies between these populations using cytochrome *b* sequences, P. J. Smith et al. (1996) found no structure using RFLP.

2.4.2 Orange roughy (*Hoplostethus atlanticus*, Trachichthyidae)

A large deep-sea fish that is widely distributed in the South Pacific, Indian, south-east, and north-east Atlantic oceans. It mostly occurs along continental slopes, ocean ridges, and seamounts at depths between 700 and 1,500 m and is targeted by trawling. In spite of their extensive distribution, the adults do not appear to be highly migratory. The movement data inferred from seasonal catches only show migration of a few hundred kilometres (Francis & Clark, 1998). Absolute and relative fecundity is low. They are also slow-growing, reaching maturity at 20–30 years of age and evidence from otoliths suggest they can live for more than 100 years (Branch, 2001; Fenton et al., 1991). The low productivity of orange roughy makes it particularly vulnerable to overfishing. Indeed, stocks have rapidly declined globally: orange roughy fisheries have collapsed in several countries, resulting in closures. While New Zealand hosts the last large-scale commercial fishery in the world (> 8,000 tonnes reported commercial catches in 2018), several stocks are depleted (Dunn & Forman, 2011). In the New Zealand EEZ, the Quota Management System divides orange roughy fisheries into eight areas, each one probably containing several discrete stocks (Fisheries New Zealand, 2018a), with the main fisheries being off the southeast of the North Island and the northern South Island, and on the Chatham Rise. There have been numerous attempts to determine the stock structure of New Zealand orange roughy by using several methods, such as genetics, morphometrics, parasite analysis, life history parameters, size structure and timing of spawning. While most of these studies were able to find some differences between areas, it is difficult to infer persistent patterns of geographic variation inside the EEZ based on these, as reviewed in Branch (2001) and Clark et al. (2016)). Evidences of an isolated stock on Challenger Plateau based on parasite analysis were corroborated by an age and length at maturity study, but not by further studies using the same method or genetics. The Challenger Plateau area (ORH7A) is however still managed as a single separate stock, based on evidence (genetics, size

structure, time of spawning, fishery distribution...) of differentiation with nearby fishing areas outside the EEZ (Clark et al., 2016). At least two main stocks are recognized in ORH 3B (Chatham Rise and Puysegur), based on genetic evidence (but see below). Biological stock structure is considered uncertain or unknown across other New Zealand management areas and the separation of stocks for assessment purposes is mainly putative and based on sparse evidence like spawning grounds location, inference of migration and catch-per-unit effort trends (Fisheries New Zealand, 2018a).

The findings reported for population genetic studies of *H. atlanticus* in New Zealand waters have been inconsistent. The first studies used samples from several locations and found no genetic differentiation between western and eastern fishing grounds using 22 allozyme loci (P. J. Smith, 1986), variable number tandem repeat 'fingerprinting' (Baker et al., 1992), and cytochrome *b* sequences (Baker et al., 1995). However, subsequent analyses of mitochondrial DNA RFLP, allozymes, and random amplified polymorphic DNA (RAPD) have detected genetic heterogeneity among populations from the Challenger Plateau, West Coast South Island, Chatham Rise, east coast of North Island, and off the southern coasts of Puysegur and Waitaki (South Island). The detection of genetic differences appears to depend on the type of marker used (P. J. Smith et al., 1997; P. J. Smith, McVeagh, et al., 1996; P. J. Smith & Benson, 1997). Among studies a significant level of genetic differentiation between southern stocks (Puysegur and Waitaki) and all other stocks was reported. This could be due to the fact that these southern stocks lie within the convergence between sub-Antarctic and subtropical water (Figure 2.1), a barrier that appears to be associated with a major faunal disjunction for many fishes (Paul, 1986). Multiple sampling along the Chatham Rise also showed spatial and temporal heterogeneity at two allozyme loci, genetic divergence with the North and South Island coasts and evidence for IBD (P. J. Smith & Benson, 1997). In contrast, more recent analyses using cytochrome *c* oxidase subunit I (COI) and cytochrome *b* sequences as well as microsatellites on samples collected from all around New Zealand did not detect any genetic structure (Varela et al., 2012, 2013). Likewise, some genetic structure has been detected among Australian localities and between Australia and New Zealand (Ovenden et al., 1989; Smolenski et al., 1993) although further studies did not corroborate those findings (Baker et al., 1995; N. G. Elliott & Ward, 1992; Varela et al., 2012, 2013). As a result, patterns of population differentiation in orange roughy have been deemed inconsistent (Ward &

46

Elliott, 2001). Allozymes and RFLP, coupled with non-genetic methods were able to validate the independent stock status of three fishing grounds in New Zealand EEZ and international waters: south-west Challenger Plateau, north-west Challenger Plateau, and Lord Howe Rise (P. J. Smith, Robertson, et al., 2002). On a more global scale, there was a small but significant level of genetic differentiation between north Atlantic and southern hemisphere populations (N. G. Elliott et al., 1994; P. J. Smith, 1986; Varela et al., 2012, 2013). In the Southern Hemisphere, no significant differentiation has been detected among populations from South Africa, Australia, and New Zealand (Baker et al., 1995; Smolenski et al., 1993), while the multi-continental study using DNA microsatellites by Varela et al. (2013) revealed a pattern of IBD at the global scale as well as significant genetic differentiation between stocks sampled from Chile, Namibia, and Australia / New Zealand that was undetectable using mitochondrial DNA sequence data (Varela et al., 2012).

2.4.3 Australasian snapper (*Chrysophrys auratus*, Sparidae)

A coastal species commonly referred to as 'snapper' that inhabits the rocky reef ecosystems around Australia and New Zealand. While recognized as a single species, snapper between New Zealand and Australia are genetically distinct (Tabata & Taniguchi, 2000). Snapper can be found at depths down to 200 m but most commonly occur between 10 and 50 m. Juvenile snapper require sheltered estuaries for development. This species reaches sexual maturity at 3–4 years of age. Areas around North Island such as Parengarenga, Rangaunu, Mahurangi, Whangapoua, Whangarei, Whangape, Kaipara, and Huruhi harbours have been identified as the key locations for snapper recruitment (Parsons et al., 2014). Snapper is one of New Zealand's most important inshore fisheries with c. 6,400 tonnes of total allowable commercial catch and an additional recreational allowance of c. 3,700 tonnes per year (Fisheries New Zealand, 2019b). There are thought to be seven or eight distinct stocks of snapper, which are being managed over four areas around the North Island (three stocks off the North Coast (SNA1), two off the East Coast (SNA2, one of which may be associated with the North Coast), and one off the West Coast (SNA8)) and northern parts of the South Island (SNA7, two stocks). North Coast (Northland, Hauraki Gulf, and Bay of Plenty) and West Coast comprise the majority of the fisheries with 4,500 and 1,300 tonnes of total

allowable commercial catch respectively. Snapper stocks have been heavily fished over the last 100 years. Although the North Coast is thought to have been reduced to less than 20% of its original biomass, current estimates show signs of recovery. The West Coast stock is considered to be depleted below the soft limit of 20%. South Island stocks (Tasman Bay and Marlborough Sounds) were severely depleted (< 10% original biomass) but now show a trend of increasing size. Statistical projections estimate that the South Island stocks will be above its soft limit by 2022 and under current conditions could return to 40% of its original biomass (Fisheries New Zealand, 2018a). The severe population size bottleneck in the South Island stocks is supported by the observed loss of genetic variation (Hauser et al., 2002), likely caused by increased genetic drift following exploitation. Population trends for the East Coast stocks are less clear. Although there is evidence, based on catch at age sampling, for two separate sub-stocks with the boundary at the southern tip of Mahia Peninsula (Walsh et al., 2012), the East Coast is still managed as a single administrative area. Both sub-stocks showed fluctuating trends over the last two decades (Fisheries New Zealand, 2018a).

Genetic analyses support in part stock structure models used for fisheries management. Early work using allozymes identified a North and a West coast population along the North Island (P. J. Smith et al., 1978). Using mitochondrial and microsatellite DNA data, Bernal-Ramírez et al. (2003) found a genetic pattern consistent with population differentiation. Sampling sites along the North Coast of the North Island did not show any significant divergence. The population in Tasman Bay (South Island) was identified as an isolated population, and the Hawkes Bay (East Coast North Island) population was genetically more similar to the West Coast North Island. This led the authors to hypothesise that gene flow was promoted via the D'Urville and Wairarapa coast current. This finding also supports the two recognised sub-stocks along the North Island East Coast, with a gene flow barrier around the Mahia Peninsula. Recent genomic work on snapper involved genome assembly, quantitative trait loci (QTL)-mapping and creation of a 95% optically mapped linkage map (Ashton, Hilario, et al., 2019; Ashton, Ritchie, et al., 2019).

2.4.4 Ling or pink cusk-eel (*Genypterus blacodes*, Ophidiidae)

An eel-like bottom dwelling fish distributed throughout the southern hemisphere, including off the coasts of New Zealand, southern Australia, and Patagonia at depths from 200 to 800 m. They have constantly been in the list of the top five most heavily fished species in New Zealand for the past few years (more than 18,500 tonnes in 2018) and are mainly caught by bottom trawling on the Southern Plateau and the Chatham Rise, although they are also targeted by long line along West Coast South Island and across the North Island coast south of East Cape. Spawning grounds are relatively well documented and distinct but the patterns of larval transport are unknown (Morrison et al., 2014). An integrative review of ling population stock structure around New Zealand concluded that they comprised at least five distinct stocks: Campbell Plateau, Bounty Plateau, Chatham Rise, West Coast South Island, and Cook Strait (Horn, 2005). Those conclusions were mainly based on differences in life history traits like growth rates, length and age at maturity, and timing of spawning.

Only three genetic studies have been conducted on ling from New Zealand (P. J. Smith, 1979; P. J. Smith & Francis, 1982; P. J. Smith & Paulin, 2003) and all of them have detected some level of stock structure. P. J. Smith (1979) used data from the polymorphic enzyme glucose phosphate isomerase, collected from sampling of a few New Zealand locations, to suggest a differentiation between stocks from Pukaki Rise (southern continental shelf) and from mainland (i.e. North and South Island). The subtropical convergence zone (Figure 2.1) was suspected to be acting as a geographical barrier to gene flow and mixing of stocks could potentially occur to the north in the Canterbury Bight. This dataset was expanded with the same allozyme and more samples by (P. J. Smith & Francis (1982). They also did not detect any heterogeneity among mainland locations but found differentiation, based on allele frequencies, between Campbell Island and Pukaki Rise / Auckland Islands samples ($P < 0.10$), as well as some heterogeneity among Chatham Rise samples. These results were interpreted as further evidence for a mainland stock that extends as far south as the Snares, a second stock on the northern Campbell Plateau (Pukaki Rise/Auckland Islands), and a possible third one around Campbell Island, with mixing of stocks along the Chatham Rise. It is important to note that none of the population boundaries suggested in both these studies had strong

statistical support. As a result, evidence for population structure from these studies has been considered weak (Daley et al., 2000; Horn, 2005). The third study by P. J. Smith & Paulin (2003) primarily aimed at re-assessing the taxonomic status of some *Genypterus* species in New Zealand and Australia, but these data from the mitochondrial DNA control region were also able to show a significant genetic differentiation between pink ling populations from the Campbell Plateau and the two northern locations investigated (Bay of Plenty and Chatham Rise), as well as between populations from New Zealand and from Tasmania. Although the results reported in the two first studies lacked strong statistical support and the existence of genetic stock structure has been questioned due to its high mixing potential (Horn, 2005), all genetic studies conducted so far indicate the presence of a barrier to gene flow between ling stocks from the mainland and the southern plateau. This separation is most likely due to the subtropical front acting as a barrier to migration. A more extensive genetic sample of ling fishing grounds and high-resolution genetic markers would be helpful to resolve this uncertainty.

2.4.5 Blue cod (*Parapercis colias*, Pinguipedidae)

A bottom dwelling, territorial reef fish endemic to New Zealand. It is widely distributed in coastal waters that have a rocky substrate and found throughout the country at depths down to 150 m. It is most abundant south of Cook Strait and around the Chatham Islands. The major commercial fisheries are off Southland (lower South Island) and the Chatham Islands, with catches also happening regularly off the east coast of the South Island (Otago, Canterbury) and Cook Strait (Marlborough Sounds and Wanganui). More than 1,800 tonnes of blue cod were commercially caught in 2019, and the recreational fishery is also quite significant with more than 500 tonnes of allowed catch (Fisheries New Zealand, 2019b). Blue cod seem to have a relatively low capacity for dispersal compared with other similar fisheries species. Eggs and larvae have a pelagic phase of only 10 days before settling (Henderson, 2009). Males are territorial and tagging studies have shown that most adults do not travel distances of more than 1 km (Gebbie, 2014). Blue cod are managed in eight Quota Management Areas based on their distribution and abundance. Studies of stock structure are sparse. The findings of tag-recapture studies suggested low dispersal of adults (Carbines & McKenzie, 2004) and spatially structured populations with limited mixing within relatively

small areas, which was a finding also supported by evidence from otolith microchemistry in Fiordland and Banks Peninsula (Beer & Carbines, 2012).

Two unpublished studies have investigated blue cod population genetic structure using an extensive New Zealand-wide sampling. Using the mitochondrial control region, H. M. Smith (2012) found significant genetic differentiation between samples from the Chatham Islands and mainland New Zealand. However, no clear population boundaries were detected among regions from both the North and South Islands, although sampling consisted of catches as far apart as Puysegur Point and Northland. However, pairwise F_{ST} results from mainland samples were positively correlated to geographic distance, albeit very weakly ($R^2 = 0.26572$, $P = 0.08$) which could be a pattern of IBD across the North and South Islands. Identical results were found by Gebbie (2014) when expanding the sampling by one more site and using the same genetic marker. However, investigating microsatellite DNA polymorphisms in samples from a sub-set of those locations did not detect any structured genetic pattern around mainland New Zealand, be it IBD or population boundaries. Blue cod is a good example of a species for which genetic structure should be expected given its low capacity for dispersal (i.e. short egg/larval stage and sedentary adult life). Since there only needs to be a small number of effective migrants exchanged each generation to homogenise genetic variation even with such low levels of movement, genetic differentiation might not be detected using those markers. Although blue cod from the Chatham Islands clearly consist of a separate stock, it is possible that a single nucleotide polymorphism (SNP)-level analysis on a similar sampling scale will provide the needed resolution to detect stronger patterns of neutral and/or adaptive heterogeneity among mainland populations.

2.4.6 Other fishery species

Few other New Zealand EEZ fishes have population genetic information available, with New Zealand-wide studies being very sparse and often restricted to a few locations encompassing only a small portion of their range.

An early study by Gauldie & Johnston (1980) used two allozymes to conduct a New Zealand-wide study on six fish species (tarakihi, trevally, barracouta, red cod, jack mackerel

and gurnard), with a sampling of thousands of individuals that was impressive in quantity and in spatial distribution. This study was able to detect significant population structure based on Z-statistics in all of the species examined, except for the red gurnard. Some of the geographic stock boundaries were shared by some of the species. However, the year-to-year levels of genetic variation in one location for tarakihi and trevally (the two species with temporal samples available) was not significantly different from spatial levels of genetic variation. Moreover, correlations between allele frequencies and water temperature were detected. As a result, the authors strongly suspected that these genetically distinct groups were due to selective clines and not genetic isolation.

School shark (*Galeorhinus galeus*) is a species which occurs in coastal waters of the Pacific, Antarctic, and Indian oceans and is listed as *vulnerable* due to global overexploitation. Two global genetic studies suggest that school shark populations are isolated at the continental level. The first used mitochondrial control region sequences to detect (at least partially) isolated populations from North America, South America, Europe, South Africa, and Australia (Chabot & Allen, 2009), while the second detected three clades (South America, South Africa and Australasia) with microsatellites and nuclear sequencing data (Bester-van der Merwe et al., 2017). New Zealand and Australian school sharks, however, seem to consist of one genetically homogenous group (Bester-van der Merwe et al., 2017; Devloo-Delva et al., 2019; Hernández et al., 2015; Ward & Gardner, 1997), a hypothesis that is supported by evidence of recurring trans-Tasman migration.

Both hāpuku (*Polyprion oxygeneios*) and bass (*P. americanus*) are managed in New Zealand as a single fishery under the name ‘groper’, despite being recognised as different biological species with different distributions. Both species lack reliable information about their stock structure in New Zealand. An early study on hāpuku detected temporal variation of allozyme frequencies in the Cook Strait area (P. J. Smith & Johnston, 1985), while microsatellite DNA analysis showed that individuals sampled from West Coast South Island (Hokitika) were genetically distinct from individuals sampled at all other New Zealand sites (Lane et al., 2016). Bass have been investigated with genetic markers in two global studies using only one New Zealand location. Both studies did not detect any significant lack of gene flow between New Zealand and Australia (Ball et al., 2000; Sedberry et al., 1996).

Short-finned (*Anguilla australis*) and long-finned (*A. dieffenbachia*) freshwater eels show allozyme genetic differentiation between North Island and South Island adults, but no heterogeneity was detected among glass eels (i.e. juveniles). Although treated separately, those results were identical for both species. It was concluded that both species consisted each of a single panmictic population with adults subjected to selection or a 'sweepstake' type reproduction event (P. J. Smith et al., 2001). A significant level of genetic differentiation was found between North and South Island short-finned eels using microsatellite DNA, although the authors were unsure if this was due to geographic or temporal variation (K. N. Shen & Tzeng, 2007).

Mitochondrial DNA control region haplotype frequencies of Alfonsino (*Beryx splendens*) showed that the Chatham Rise and East Coast North Island populations were significantly different (Lévy-Hartmann et al., 2011). Moreover, hierarchical analysis of molecular variance and neighbour-joining tree analysis of the data showed that Chatham Rise alfonsinos were genetically more similar to the ones from Australia and the Indian Ocean, while another cluster consisted of fish sampled from East Coast North Island, New Caledonia, and Japan.

Mitochondrial control region sequences of the blue mackerel (*Scomber australasicus*) were used to build a neighbour-joining tree which showed strong support (bootstrap > 90%) for the presence of two phylogroups in the dataset. However, there was no geographical structure because specimens from all of the sampled areas (North Cape and Hawke's Bay (New Zealand) and South Australia) were found in both phylogroups and there was no detectable differentiation among populations (P. J. Smith et al., 2005).

Complete genetic homogeneity was observed for the pacific bluefin tuna (*Thunnus orientalis*) between New Zealand East Coast North Island, West Coast South Island, and Taiwan based on several genetic markers using a neighbour-joining tree reconstruction (Tseng & Smith, 2012).

Other fishes for which the hypothesis of panmixia has not been rejected by genetic analyses of New Zealand-wide sampling include black oreo and smooth oreo (P. J. Smith, McMillan, et al., 2002), rig (Hendry, 2004), hake (P. J. Smith et al., 1979), and gemfish (Colgan & Paxton, 1997). None of them have yet been investigated using microsatellite DNA, Sanger direct sequencing, or genome-wide sequencing. Some other species have been investigated as part of broader studies with only one location (or none) in New Zealand and are reported in Supplementary Table 2.1, 2.2 & 2.3. Some highly migratory fishes with a wide distribution have shown patterns of structure at the inter-oceanic scale only, thus genetic heterogeneity within New Zealand or Australasia can be considered unlikely (e.g. swordfish (Chow & Takeyama, 2000), spiny dogfish (Veríssimo et al., 2010)). A similar Australasian-wide finding of genetic homogeneity could reasonably be expected from other fish with high dispersal capacity and broad distribution like blue shark, moonfish or most tunas. Conversely, some species with only one New Zealand location sampled have shown evidence of trans-Tasman differentiation (e.g. tarakihi (N. G. Elliott & Ward, 1994; Grewe et al., 1994) and trumpeter (Tracey et al., 2007)), which suggests that within-New Zealand EEZ structure might be found. It is also interesting to note that evidence of trans-Tasman connectivity in a species does not necessarily mean that there is a panmictic “Australasian” stock. Several studies have reported a genetic distinction between west and east Australia but not between east Australia and New Zealand. This is the case for kingfish (P. A. Miller et al., 2011; Nugroho et al., 2001), gemfish (Colgan & Paxton, 1997) and mako shark (Corrigan et al., 2018).

Overall, very few New Zealand fisheries species have good population genetic data which has been used to define their stock boundaries. Snapper is the only species showing a population genetic structure that has been deemed consistent across studies and that is also partially supported with other methods of stock identification (Bernal-Ramírez et al., 2003; Parsons et al., 2014). Findings based on genetic data from other species have either been inconsistent (orange roughy), lacked in statistical support (ling), failed to reject selective neutrality (Gauldie & Johnston, 1980) or did not detect any genetic structure. This lack of genetic heterogeneity found in most fishes could be because: (1) their stocks are genetically panmictic, but because this is typically the null hypothesis it is not easily tested (especially with non-genomic studies); (2) the genetic markers used up until this point in time lacked the

resolution required to detect genetic differentiation; (3) there has been insufficient spatial sampling to properly test for genetic structuring.

2.5 Moving forward: Use of genome-wide sequencing techniques

Genetic differentiation can be undetectable using traditional genetic approaches, such as microsatellite DNA or short stretch DNA sequencing, because these only sample a small portion of the genome. Signals of population genetic difference can be difficult to detect in some cases, because isolation may have only recently been established, low levels of persistent gene flow are homogenizing most loci, or a large N_e makes genetic drift a weak force (Attard, Beheregaray, et al., 2018; Palsbøll et al., 2007). A lack of strong levels of population genetic structure is a common finding for studies of marine organisms. This is particularly pronounced in the many marine species with large populations and hence high levels of genetic diversity, and is further compounded in the many species with a high dispersal power (Sandoval-Castillo et al., 2018). Until recently, genetic markers had to be chosen carefully to give an accurate estimate of the amount and distribution of variation, to properly address the question asked, and be suitable for the practical limitations of data collection (Anne, 2006). These constraints also make the comparison of results from different markers difficult (Portnoy & Heist, 2012).

The development of genome-wide DNA sequencing technology, often referred to as ‘Next Generation Sequencing’, has fundamentally changed the type of population genetic data sets now available. These ‘high-throughput’ methods allow whole genome sequencing data to be collected, and for species with large genomes, reduced-representation methods like restriction site associated DNA markers (RAD-Seq and ddRAD), and genotyping-by-sequencing (GBS), continue to make genome-wide population studies affordable. As DNA sequencing costs continue to decrease and research laboratories switch to these methods, genome-wide DNA sequencing data for fisheries species are slowly becoming more common, as is already the case in farm-based primary production sectors (Kumar & Kocour, 2017; Y.-H. Li & Wang, 2017). The large genome-wide amounts of SNPs obtained from these methods should provide higher levels of resolution to estimate more accurately F_{ST} and $N_e m$ than the

small number of loci microsatellites could sample (Benestan, 2019). Direct comparison of RAD sequencing and microsatellites on great scallop (Vendrami et al., 2017) showed that the genome-wide method could resolve population structure that was undetectable using typical suit of microsatellite loci. Genome-wide sequencing techniques can also provide more accurate estimates of relatedness (Thrasher et al., 2018) and genetic variation at the individual level (Lemopoulos et al., 2019) and are also better at detecting evidence of introgression (Bradbury et al., 2015). In some cases genetic structure can be detected by only sampling a few individuals per population (c. 8–10) as long as the genome-wide SNP density is high (Jeffries et al., 2016; Nazareno et al., 2017). Moreover, genome-wide SNP datasets can detect genomic regions under selection, while there is growing incentive for adaptive diversity to be relevant to fisheries management (Funk et al., 2012; Valenzuela-Quir6nez, 2016). Before genomics made it possible to detect outlier loci, separating neutral from adaptive variation was either impossible (allozymes) or sometimes based on wrong assumptions (e.g. microsatellites (Cadrin, 2020)). Genome-wide SNP data is becoming the preferred approach for addressing important fishery management issues (Allendorf et al., 2010; K. R. Andrews et al., 2016; Portnoy & Heist, 2012). Although studies using genome-wide SNP data to resolve population structure of marine species have been growing in numbers in the last few years (e.g. European lobster (Jenkins et al., 2019), yellowfin tuna (Pecoraro et al., 2018), sea cucumber (Xuereb et al., 2018), and more...) none have been reported for New Zealand fish stocks.

2.6 Other issues in fisheries management that will benefit from a more systematic use of genetics and genomics

As we showed above, although the use of genomics will improve the resolution and confidence in testing genetic structure hypotheses, the advantages of that type of data are not limited to this specific question. The development of genome-wide markers has become an important tool for understanding some of the emerging issues in fisheries (Benestan, 2019; Bernatchez et al., 2017; Valenzuela-Quir6nez, 2016). For example, reliable estimation of stock size and resilience to environmental stressors are critical measures for sustainable fisheries management. In this section, four critical fisheries issues are briefly introduced and discussed, namely 1) the loss of diversity, 2) fishing-induced evolution, 3) the influence of

climate change on stock sizes and range, and 4) the measurement of absolute abundance. These issues are only a few examples among the numerous applications of genetics in fisheries as thoroughly reviewed in Ovenden et al. (2015), but we focussed on what we think might be the main point of interests that could have been overlooked by fisheries managers in New Zealand.

2.6.1 Loss of genetic diversity and adaptive potential

Genetic diversity is the evolutionary potential of a population (Sgrò et al., 2011; Spielman et al., 2004). Higher levels of genetic diversity increase the likelihood that there will be advantageous variants in a population when a new environmental challenge arises. When population size is reduced the force of genetic drift increases, which eliminates alleles from the population. The consequence of lower diversity is reduced adaptive potential. In highly fecund species like most marine fish, N_e may be much smaller than the census population (N_c) due to a bias in reproductive success, large variation in year class strength, and size-dependant fecundity (Waples, 2002). Indeed, harvesting often targets specific sexes or age classes which reduces the $N_e:N_c$ ratio, thereby increasing the rate of loss of heterozygosity without having any detectable effect on population size. Marine finfish species generally display very large population sizes and are consequently assumed to maintain high levels of genetic diversity. It is therefore reasonable to assume that many fisheries species were originally characterised by a high adaptive potential, making them more resilient (Sgrò et al., 2011; Spielman et al., 2004). However, loss of genetic diversity can be significant in just a few generations in small effective populations where genetic drift is strong. Work by Hauser et al. (2002) demonstrated a dramatic loss of genetic diversity in Australasian snapper *Chrysophrys auratus* in New Zealand, despite an estimated minimum N_c size of more than 3,000,000 fish. A meta-analysis using 140 species and more than 10,000 loci showed that allelic richness and heterozygosity are significantly lower in overharvested fish populations (Pinsky & Palumbi, 2014). Moreover, size-selective fishing has been experimentally shown to significantly reduce genetic diversity over very few generations (Therkildsen et al., 2019).

Over the last century, populations have been heavily exploited and most population sizes have drastically decreased. As a result, a loss of genetic diversity can be expected due to increased genetic drift. In contrast, it can take many generations for genetic variation to return in a population through random mutation, and immigration levels from other populations may not always be high enough for gene flow to supplement the lost genetic variation in a short time scale (Pinsky & Palumbi, 2014). This implies that exploited populations will experience lower genetic diversity for hundreds or thousands of years into the future, and a reduced evolutionary potential compared to their 'baseline' level (i.e. before industrial fishing). This could result in more frequent occurrences of inbreeding depression. In spite of the risks associated with genetic erosion, the target levels of genetic diversity to maintain in a population are difficult to quantify because any advantageous alleles only become relevant when presented with a challenging environment. To understand this better, the extent to which genetic variation is lost in exploited populations could be assessed using temporal sampling, i.e. historic (< 100 year old) and ancient (> 100 year old) DNA when available (Oosting et al., 2019). By reconstructing the genetic variation present in past populations, it could be possible to assess how much genetic variation has been lost and use this information to set diversity baselines for fisheries management. Maintaining genetic diversity should be of concern to fisheries managers because the future of fisheries species will be one with lower adaptive potential, making them more vulnerable to environmental challenges than they have ever been in their recent evolutionary history.

2.6.2 Fishing-induced evolution

The practice of size-selective fishing has been observed to affect the evolution of key life-history traits, such as maturation age and growth rate. In short, intense fishing pressure has favoured the reproductive success of individuals with 'fast life-history traits. Shifts in life-history traits have been observed in a large number of exploited fish species (Heino et al., 2015). A key question is whether these changes in life-history traits have been facilitated through phenotypic plasticity (e.g. density-dependant) processes or directional selective pressure on genetic loci. The distinction between these two processes is crucial because shifts in allele frequencies could theoretically lead to the loss or fixation of alleles that would otherwise be beneficial or detrimental to the population in an "unfished" state.

Experimental studies provide strong evidence for the causal link between changes in life-history traits caused by size-selective fishing and directional selection on genetic loci (Conover & Munch, 2002; Marques et al., 2018; Therkildsen et al., 2019; van Wijk et al., 2013). Strong support for this causal link in wild populations is currently lacking because the genetic data from a pre-exploited population are difficult to obtain. Notably, Therkildsen et al. (2013) reported changes in allele frequencies in loci which correlate with temporal changes in probabilistic maturation reaction norm (an approach commonly used to assess changes in maturation). These changes were observed through temporal sampling of wild Atlantic cod (*Gadus morhua*) using historic samples. With recent advances in genome-wide sequencing and improved extraction protocols for ancient DNA, this question can now be more easily addressed. SNP datasets generated from ancient DNA samples can be compared to fish in contemporary stocks to test whether loci associated with maturation and growth traits have been experiencing selection (Oosting et al., 2019). Overall, the increasing evidence of the effects of size-selection on fishery stocks indicates that it may be time to re-think the way fish catches are regulated in New Zealand (and elsewhere), e.g. by shifting to a balanced harvesting model (Garcia et al., 2012; Zhou et al., 2019). Such a model change could be particularly effective on the small-scale fisheries which represent more than half of the global catch for human consumption (Plank et al., 2017).

2.6.3 Climate change

Global warming is shifting the geographic distribution of many marine species. Modelling has shown that changes in distribution predominate towards the poles and/or to lower depths (Brooks, 2019; Morley et al., 2018; Stanley et al., 2018). Management approaches need to account for this, but surprisingly only few empirical studies have been conducted on the genetic changes in fish populations associated with climate change (Munday et al., 2017, 2019; Muñoz et al., 2015). We expect that there will be both short- and long-term changes to levels of genetic diversity, disruptions to population structure, and adaptive responses to this type of environmental shift (Potts et al., 2014; Ramos et al., 2018; Stanley et al., 2018). Monitoring and studying the genetic diversity in a fishery that is associated with adaptation to warmer waters (as well as related impacts like acidification and

decreases in oxygen levels) and genetic founder events as populations expand into new areas will be necessary to avoid overfishing of stocks building at new locations and depletion of those in decline elsewhere.

2.6.4 Measuring abundance and connectivity

Other than stock structure and the resilience capacities of harvested stocks, fisheries managers are mostly concerned with estimating absolute abundance and inter-stock movement. Abundance is traditionally investigated through measures of catch per unit effort, which is only an indicator of change in relative abundance, or through age- and length-composition data, which rely on a lot of assumptions (Maunder & Piner, 2015). Stock connectivity and movements are usually investigated with tagging via mark-recapture. However, this method requires fish be brought to the surface, potentially causing barotrauma and high levels of mortality. Tagging is considered impossible on deep-water species, causing large information gaps for species such as orange roughy and hoki. Classical genetic tagging (e.g. (L. M. Miller et al., 2015)) faces the same issues. Genetic tagging by using biopsy hooks to sample individuals *in situ* (Mace et al. 2020) is a relatively new method that would allow to bypass those constraints, and could simultaneously allow for the estimation of absolute abundance when combined with methodologies like close-kin mark-recapture (Bravington, Skaug, et al., 2016). Using this method, absolute abundance and demographic parameters could even be estimated without the need to resample the same individual (as long as juveniles can also be sampled in the case of close-kin mark-recapture). Genetic tagging coupled with close-kin mark-recapture has already been used to infer the absolute population size of southern bluefin tuna on a large spatial and temporal scale (Bravington, Grewe, et al., 2016) and subsequently the estimation of N_e (Waples et al., 2018), as well as the absolute abundance of white sharks in eastern Australia and New Zealand (Hillary et al., 2018).

A genetic-based tagging programme would open up a range of new possibilities for stock assessment (Mace et al. 2020). If genetic data from enough specimens is collected, this could allow for a more reliable estimation of N_e , an important metric that has suffered from sub-optimal sample size so far (Marandel et al., 2019). Genetic tags are permanent, and the offspring of previously tagged individuals will be detectable for several generations.

Depending of the life expectancy of the species, genetic tagging data can remain relevant for many decades, or even for a century or more for long-lived species like orange roughy.

2.7 Conclusion

Our review showed that the genetic population structure of fisheries species in New Zealand has been largely under-studied. Prior to the advent of genomics, population genetic markers often lacked the level of resolution needed to study marine species that typically have large populations and are highly mobile. New genome-wide sequencing approaches offer the opportunity to more comprehensively resolve a range of previously intractable stock structure issues. Following are several recommendations for promoting a genomics-enabled approach to New Zealand fisheries management.

1. Long-term sampling and data handling strategy: Develop a plan for regular spatial and temporal sample collection for each fishery and in a way that can be easily integrated into a tagging program. Tissues will need to be preserved in a way that enables genomic sequencing. Overall, the genetic data collection (including meta-data) should be standardized as much as possible at a national scale to insure its efficient usability for current and future studies. This requires good coordination among collectors, fisheries data management, and laboratory capabilities. A centralized genomic data storage facility is needed for a range of strategic interests beyond fisheries and is a common need across New Zealand's primary industries and conservation groups. A well-maintained national archive would secure the long-term value of genomics.

2. Prioritise species: A limit on the resources available to support genomic research means priority should be given to the most valuable fisheries that are experiencing the greatest management concerns. For example, deep-water species such as hoki and orange roughy lack good stock assessment information. Genomic markers could be used to enable an in situ deep-water tagging programme to be developed for the first time. Genomic tools could be developed for species (e.g. snapper, trevally, hapuku) that are also good candidates for aquaculture. The dual support of fisheries and aquaculture interests would require a more joined up seafood genomics approach among government, industry and researchers.

The strong pressures that wild fishes are currently facing both nationally and globally require more detailed understanding about the current status of stocks and what changes are occurring, or are likely to occur. Genomics tools fit into the management toolbox as part of a holistic integrative approach to a fishery. There is little doubt that genomics will be a vital component of the response to the challenges fisheries management will face this century

2.8 Supplementary Material

Supplementary Table 2.1. All fish species for which genetic population structure hypotheses have already been tested among New Zealand locations, detailed table, part 1 [Continues next pages]

Common names	Species	EEZ QMAs	Current stock hypothesis	FNZ reported commercial catch (kg) and position in compared to other NZ fishery species	Type of genetic structure in EEZ	Trans-Tasman differentiation	Australasia - World differentiation
Hoki (Blue grenadier, blue hake, New Zealand whiptail, whiptail, whiptail hake)	<i>Macruronus novaezealandiae</i>	2*	2 stocks: west coast and east coast	135,417,657 (1st)	(I) panmictic	-	-
					(I) panmictic	yes	-
					(I) panmictic	no	-
					(I) panmictic	no	no
					(I) panmictic	-	yes
Orange roughy (red roughy, slimehead, deep sea perch)	<i>Hoplostethus atlanticus</i>	8*	>8 stocks: Mercury- Colville, East Cape, Mid-East Coast, Northwest Chatham Rise, East and South Chatham Rise, Puysegur, Challenger Plateau, WCSI + others (MPI 2017)	8,073,415 (9th)	-	-	no
					(I) panmictic	-	yes
					(I) panmictic	-	-
					(I) panmictic	no	-
					-	partial	partial
					(I) panmictic	no	no
					(III) distinct populations	no	-
					(II) isolation-by- distance	-	-
					(III) distinct populations	-	-
					(III) distinct populations	-	-
					(I) panmictic	no	no

2.8 Supplementary Material

					(I) panmictic	no	yes
Snapper	<i>Chrysophrys auratus</i>	6*	possibly 7 stocks (2 SNA1, 2 SNA2, 2 SNA7, 1 SNA8)	6,490,213 (11th)	(III) distinct populations	-	-
					-	yes	yes
					(III) distinct populations	-	-
Ling	<i>Genypterus blacodes</i>	8*	5 stocks	18,694,021 (4th)	(III) distinct populations	-	-
					(III) distinct populations	-	-
					(III) distinct populations	yes	
Blue cod	<i>Parapercis colias</i>	8*	Probably more stocks than QMAs	2,048,577 (23rd)	(II) isolation-by-distance and (III) distinct populations	-	-
					(II) isolation-by-distance and (III) distinct populations	-	-
					(I) panmictic	(no)	-
School shark (Tope shark, snapper shark, soupfin shark)	<i>Galeorhinus galeus</i>	8*	One stock	3,013,366 (17th)	(I) panmictic	no	yes
					-	no	yes
Hapuku (whapuku) and	<i>Polyprion oxygeneiosus</i>	8*	Unknown, probably one stock each	1,249,856 (32th)	unclear	-	-

Bass (Atlantic wreckfish)	<i>nd P. americanus</i>				-	no	yes (North Atlantic vs Southern Hemisphere)
					-	no	yes
					(III) distinct populations	-	-
Short-finned freshwater eel	<i>Anguilla australis & A. reinhardtii</i>	12*	One stock	432,439 (45th)	(III) distinct populations	-	-
					(III) distinct populations	yes	-
Long-finned freshwater eel	<i>Anguilla dieffenbachii</i>	11	One stock	86,438 (54th)	(III) distinct populations	-	-
Alfonsino	<i>Beryx splendens & B. decadactylus</i>	6*	Unknown	2,530,987 (20th)	(III) distinct populations	partial	partial
Blue mackerel	<i>Scomber australasicus</i>	5*	At least 3 stocks in 3 northern QMA	11,399,659 (6th)	-	no	yes
					(I) panmictic	no	-
					-	no	-
Black oreo and smooth oreo	<i>Allocyttus niger</i> and <i>Pseudocyttus maculatus</i>	5*	1 stock / mainland QMA, species assessed separately	8,783,204 (7th)	-	yes (black) / no (smooth)	-
					(I) panmictic	-	-
Rig (Spotted estuary smooth-hound)	<i>Mustelus lenticulatus</i>	6*	Current QMA's reflective of stock structure	1,459,402 (28th)	(I) panmictic	-	-
Hake	<i>Merluccius australis</i>	4*	3 stocks	4,703,587 (13th)	(I) panmictic	-	-
					-	-	yes

2.8 Supplementary Material

Gemfish	<i>Rexea solandri</i>	5*	2 stocks	1,611,955 (27th)	(I) panmictic	no	-
Pacific bluefin tuna	<i>Thunnus orientalis</i>	1	a single Pacific-wide stock	20,706 (64th)	(I) panmictic	-	no
Tarakihi (Jackass morwong)	<i>Nemadactylus macropterus</i>	8*	1 stock	5,741,806 (12th)	(III) distinct populations	-	-
					-	yes	-
					-	yes	-
					-	no	-
Trevally	<i>Pseudocaranx dentex</i>	5*	3 stocks, may be more	3,758,613 (15th)	(III) distinct populations	-	-
Barracouta	<i>Thyrssites atun</i>	5*	at least 4 stocks	29,070,850 (3rd)	(III) distinct populations	-	-
Red cod	<i>Pseudophyciss bachus</i>	5*	unknown	3,367,373 (16th)	(III) distinct populations	-	-
Jack mackerels (greenback horse mackerel, yellowtail horse mackerel and Murphy's mackerel / Chilean jack mackerel)	<i>Trachurus declivis</i> , <i>Trachurus novaezealandiae</i> and <i>Trachurus murphyi</i>	4*	Unknown	45,301,555 (2nd)	(III) distinct populations	-	-
					-	-	no
Gurnard	<i>Chelidonichthys kumu</i>	6*		3,855,378 (14th)	?	-	-
Kingfish (Yellowtail amberjack,	<i>Seriola lalandi</i>	7*	2 stocks: west and east	255,272 (49th)	-	no	yes
					-	no	-

yellowtail
kingfish or great
amberjack)

Flatfish	8 species	5*	many localised stocks	2,013,664 (24th)	-	yes	-
Bluenose (blue eye trevalla)	<i>Hyperoglyph e antarctica</i>	6*	Probably a single stock	653,307 (40th)	-	no	-
Southern blue whiting	<i>Micromesisti us australis</i>	5	4 stocks	15,626,253 (5th)	-	-	yes
Spiny dogfish	<i>Squalus acanthias</i>	7*	5 stocks	6,506,086 (10th)	-	-	yes
Swordfish	<i>Xiphias gladius</i>	1	1 stock	432,439 (44th)	-	-	yes (North Pacific vs Atlantic and Southern Hemisphere)
					-	-	yes (Indo-Pacific vs Medit & Atlantic)
Trumpeter	<i>Latris lineata</i>	10*	unknown	52,422 (59th)	-	yes	yes
Mako shark (shortfin mako)	<i>Isurus oxyrinchus</i>	1	prob a Sth Pacific stock	37,185 (61st)	-	partial	yes

Supplementary Table 2.2. All fish species for which genetic population structure hypotheses have already been tested among New Zealand locations, detailed table, part 2 [Continues next pages]

Common names	Genetic marker	No. loci	Locations (no. sites)	No. specimens collected	Study
Hoki (Blue grenadier, blue hake, New Zealand whiptail, whiptail, whiptail hake)	allozymes	15	NZ-wide (9)	783	P. J. Smith et al. (1981)
	mtDNA seq (cytb)	1	WCSI (1) + CS (1) + CR (1)	19 NZ + 8 Tasman	Baker et al. (1995)
	mtDNA RFLP	6 restr enzymes	2 NZ + 1 Tasman	105	P. J. Smith et al. (1996)
	mtDNA seq (cytb)	1	Baker et al. (1995) +Chile	44	Olavarria et al. (2006)
	mtDNA seq (cytb+CR) + microsat	2 + 4	2 NZ + 3 Chile	69 NZ + 118 Chile	Takeshima et al. (2011)
Orange roughy (red roughy, slimehead, deep sea perch)	mtDNA seq (COI+cytb)	2	Olavarria et al. (2006)+ Sth Africa	1 Sth Africa	Leslie et al. (2018)
	allozymes	12 loci	5 NZ + 2 British Isles	~10-100 / loc	P. J. Smith (1986)
	VNTR	3 clone probes	3 NZ locs in 2 areas		Baker et al. (1992)
	allozymes	11 loci	6 Aus + Smith (1986) NZ	~10-200 / loc	Elliott & Ward (1992)
	mtDNA RFLP	13 restr enzymes	5 Aus + 1 NZ + 1 Sth Africa	286 samples	Smolenski et al. (1993)

	mtDNA seq (cytb)	1	4 NZ + 1 IW + 1 Tasmania + 1 Sth Africa	24 + 4 + 4	Baker et al. (1995)
	mtDNA RFLP	6 restr enzymes	6NZ + 1 Tasmania	246	P. J. Smith et al. (1996)
	allozymes	11 loci	5 ECNZ + 10CR	~100 fish/site	P. J. Smith & Benson (1997)
	allozymes + RAPD + mtDNA CAPS	variable	4 NZ	50-100 fish/site	P. J. Smith et al. (1997)
	allozymes + microsat + mtDNA CAPS	variable	1 EEZ + 3 Tasman Sea IW)	~20-110 fish/site	P. J. Smith, Robertson, et al. (2002)
	mtDNA seq (COI+cytb)	2	7 NZ + 2 Aus + 1Nam +1Chile + 2 NE Atl	25-52 fish / site	Varela et al. (2012)
	microsat	9	16 NZ + 2 Aus + 1Nam + 1Chile + 2 NE Atl	22-48 fish/site	Varela et al. (2013)
Snapper	allozymes	15 enzymes	12 NZ	78-168 fish/site	P. J. Smith et al. (1978)
	mtDNA seq (CR+tRNA) + RFLP	1 + 6 enzymes?	2 NZ, 1 AUS, 2 CHINA, 2 JAPAN	52-95 fish/site	Tabata & Taniguchi (2000)
	mtDNA seq (CR) + microsat	1 + 7	6 NZ	47-50 fish/site	Bernal-Ramírez et al. (2003)
Ling	allozymes	2	5 NZ: NE North Island (Cook Strait, WCSI, Canterbury Bight, Pukaki Rise	60-95 fish/loc	P. J. Smith (1979)

2.8 Supplementary Material

	allozymes	1		25 sampling points around NZ	25-163 fish/loc	P. J. Smith & Francis (1982)
	mtDNA seq (cytb+CR)			3 NZ	48NZ + 5Tas	P. J. Smith & Paulin (2003)
Blue cod	mtDNA seq (CR)	1		14 NZ	475 fish	H. M. Smith (2012)
	microsat + mtDNA seq (CR)	7		Smith (2012) + 1 NZ area	529 fish	Gebbie (2014)
School shark (Tope shark, snapper shark, soupfin shark)	allozymes + mtDNA RFLP	29 enzyme loci + 10 restr enzymes		1 North Atlantic + 1 Argentina + 1 South Africa + 2 Aus + 3 Tasmania + 3 NZ	11-143 fish/loc	Ward & Gardner (1997)
	microsat + mtDNA seq (CR)	8+1		4 NZ + 6 Aus + 1 Chile	10-47 fish/loc	Hernández et al. (2015)
	microsat + mtDNA seq (ND2)	10+1		1 Chile + 1 Argentina + 7 South Africa + 1 Aus + 1 NZ	9-124 fish/loc	Bester-van der Merwe et al. (2017)
Hapuku (whapuku) and Bass (Atlantic wreckfish)	allozymes	2		6 NZ areas sampled one to 18 times	~20-180 fish/sample	P. J. Smith & Johnston (1985) (<i>P. oxygeneios</i>)
	mtDNA CAPS	7 restr enzmes		4 North Atlantic + 1 SE Brazil +1 Aus+ 1 NZ	~10-100 fish/loc	Sedberry et al. (1996) (<i>P. americanus</i>)

	microsat	6	7 North Atlantic + 1 SE Brazil +1 Aus+ 1 NZ	16-118 fish/loc	Ball et al. (2000) (<i>P. americanus</i>)
	microsat + mtDNA seq (CR)	9+1	8 NZ-wide locs	294 Hapuku and 25 Bass	Lane et al. (2016) (<i>P. oxygeneios</i>)
Short-finned freshwater eel	allozymes	10	6 (glassfish) + 4 (adults) NZ rivers	41-200 ind/river	P. J. Smith et al. (2001)
	microsat	7	6 Au + 4 NZ	24-60 ind/river	Shen & Tzeng (2007)
Long-finned freshwater eel	allozymes	9	4 (glassfish) + 4 (adults) NZ rivers	21-72 ind/river	P. J. Smith et al. (2001)
Alfonsino	mtDNA seq (cytb)	1	1 Atlantic + 2 Indian Ocean + 1 JP + 1 Aus +2 NZ + 7 New Caledonia	5-27 fish/loc	Lévy-Hartmann et al. (2011)
Blue mackerel	mtDNA RFLP + mtDNA seq (cytb)	12 restr enzymes + 1	1 NZ + 1 Aus + 1 Red Sea + 1 Japan + 1 Mexico	15-21 fish/loc	Scoles et al. (1998)
	mtDNA seq (CR)	1	2 NZ + 1 Aus	60 fish	P. J. Smith et al. (2005)
	mtDNA seq (CR)	1	3 Aus + 1 NZ	13-23 fish/loc	Schmarr et al. (2011)
	allozyme + mtDNA RFLP	variable	2 Aus + Tasm + 1 NZ (CR)	~100-200 fish/loc	R. D. Ward et al. (1998)

2.8 Supplementary Material

Black oreo and smooth oreo	mtDNA CAPS + nuclear CAPS	variable	4 EEZ (The four main management areas)	100bo+100so /area	P. J. Smith, McMillan, et al. (2002)
Rig (Spotted estuary smooth-hound)	allozymes + mtDNA RFLP	17+12	8 NZ locs	1-38 fish/loc	Hendry (2004)
Hake	allozymes	2	5 NZ locs	36-125 fish/loc	P. J. Smith et al. (1979)
	microsat + mtDNA seq (COI+CR)	9 + 2	Chile + Argentina + NZ	39-54 fish/loc	Machado-Schiaffino et al. (2009)
Gemfish	allozymes + mtDNA RFLP	17 loci + 11 restr. Enz	12 Aus + 2 NZ	3-47 fish/loc	Colgan & Paxton (1997)
Pacific bluefin tuna	microsat + mtDNA seq (CR+cytb)	7 + 2	2 NZ (NENI and WCSI) + 2 Taiwan	53 Taiwan + 70 NZ	Tseng & Smith (2012)
Tarakihi (Jackass morwong)	allozymes	2	21 NZ-wide locs	25-159 fish/loc	Gauldie & Johnston (1980)
	allozymes	8	9 Aus + 1 NZ loc	9-138 fish/loc	Elliott & Ward (1994)
	mtDNA RFLP	9 restr enzymes	8 Aus + 1 NZ loc	14-71 fish/loc	Grewe et al. (1994)
	microsat	7	3 Aus + 1 NZ	49-59 fish/ara	Burridge & Smolenski (2003)

Trevally	allozymes	2	7 NZ-wide locs	70-150 fish	Gauldie & Johnston (1980)
Barracouta	allozymes	2	9 NZ-wide locs	68-159 fish/loc	Gauldie & Johnston (1980)
Red cod	allozymes	2	5 NZ-wide locs	35-50 fish/loc	Gauldie & Johnston (1980)
Jack mackerels (greenback horse mackerel, yellowtail horse mackerel and Murphy's mackerel / Chilean jack mackerel)	allozymes	2	3 NZ-wide locs	31-123 fish/loc	Gauldie & Johnston (1980)
	microsat + mtDNA seq (CR)	4 + 1	3 Chile + 1 Eastern Pacific + 1 NZ	26-52 fish/loc	Cárdenas et al. (2009) (<i>T. murphyi</i>)
Gurnard	allozymes	2	14 NZ-wide locs	43-230 fish/loc	Gauldie & Johnston (1980)
Kingfish (Yellowtail amberjack, yellowtail kingfish or great amberjack)	microsat + mtDNA CAPS	3+1	3 Aus + 1 JP + 3ENthNZ (pooled)	25-80 fish/area	Nugroho et al. (2001)
	microsat + mtDNA seq (NADH-ND5)	7+1	5 Aus + Nugroho et al. (2001)	15-111 fish/loc	P. A. Miller et al. (2011)

2.8 Supplementary Material

Flatfish	allozymes	30	5 Tasmania + 1 Aus + 1 NZ	36-100 fish/loc	van den Enden et al. (2000) (<i>Rhombosolea tapirina</i>)
Bluenose (blue eye trevalla)	mtDNA seq (CR)	1	10 Aus + 1 NZ	320 fish	Robinson et al. (2008)
Southern blue whiting	microsat	6	1 NZ + 1 EC South America	25-50 fish/loc	Ryan et al. (2002)
	mtDNA seq (cytb + CR) + microsat	2 + 4	3 NZ (very close) + 2 Chile	81 NZ + 74 Chile	Takeshima et al. (2011)
Spiny dogfish	microsat + mtDNA seq (ND2)	8	9 worldwide + 1 NZ	18-56 fsh/loc	Veríssimo et al. (2010)
Swordfish	mtDNA CAPS	4	1 Aus&NZ + 12 worldwide	11-63 fish/area	Chow et al. (1997)
	nuclear CAPS + mtDNA CAPS	1 + 2	1 Aus&NZ + 18 worldwide	19-68 fish/area	Chow & Takeyama (2000)
Trumpeter	mtDNA seq (CR)	1	1 St Paul Island + 5 NZ + 3 Tasmania + 1 Foundation seamount	1-61 fish/area	Tracey et al. (2007)
Mako shark (shortfin mako)	microsat + mtDNA seq (CR)	10 + 1	Worldwide + 9 NZ locs pooled	389 fish	Corrigan et al. (2018)

Supplementary Table 2.3. Studies focused on inter-specific phylogenies with very few locations or specimens in New Zealand or with insufficient data to infer significant population structure

Common name	Scientific name(s)	References
Blue moki	<i>Latridopsis ciliaris</i>	P. J. Smith et al. (2003)
Freshwater eels	<i>Anguilla australis</i> , <i>Anguilla dieffenbachii</i> , <i>Anguilla reinhardtii</i>	Dijkstra & Jellyman (1999) Minegishi et al. (2005) Minegishi et al. (2009)
Frostfish	<i>Lepidopus caudatus</i>	Ward et al. (2008)
Giant stargazer	<i>Kathetostoma</i> sp.	P. J. Smith et al. (2006)
Hake	<i>Merluccius australis</i>	Deli Antoni et al. (2015)
John dory	<i>Zeus faber</i>	Ward et al. (2008)
Rig	<i>Mustelus lenticulatus</i>	Gardner & Ward (2002) Boomer et al. (2012)
Sea perch	<i>Helicolenus</i> spp	P. J. Smith et al. (2009)
Southern blue whiting	<i>Micromesistius australis</i>	Halvorsen et al. (2012)
Tarakihi	<i>Nemadactylus macropterus</i>	Burridge & Smolenski (2004)

Chapter 3. Contiguous genome assembly and transcriptome of a highly heterozygous New Zealand fisheries species, the tarakihi (*Nemadactylus macropterus*).

3.1 Abstract

Although being some of the most valuable and heavily exploited wild organisms, few fisheries species have been studied at the whole-genome level. This is especially the case in New Zealand, where genomics resources are urgently needed to assist fisheries management to ensure sustainable levels. Here we generated 55 Gb of short Illumina reads (92× coverage) and 73 Gb of long Nanopore reads (122×) to produce the first genome assembly of the marine teleost tarakihi (*Nemadactylus macropterus*), a highly valuable fisheries species in New Zealand. An additional 300 Mb of Iso-Seq RNA reads were obtained from four tissue types of another specimen to assist in gene annotation. The final genome assembly was 568 Mb long and consisted of 1,214 scaffolds with an N50 of 3.37 Mb. The genome completeness was high, with 97.8% of complete Actinopterygii BUSCOs retrieved. Heterozygosity values estimated through both *k*-mer counting (1.00%) and bi-allelic SNPs (0.64%) were high when compared to the same values reported for other fishes. Repetitive elements covered 30.45% of the genome and 20,169 protein-coding genes were annotated. Iso-Seq analysis recovered 91,313 unique transcripts (isoforms) from 15,515 genes (mean ratio of 5.89 transcripts per gene), and the most common alternative splicing event was intron retention. This highly contiguous genome assembly along with the isoform-resolved transcriptome will provide a useful resource to assist the study of population genomics, as well as comparative eco-evolutionary studies in other teleost and related organisms.

3.2 Introduction

The tarakihi or jackass morwong (*Nemadactylus macropterus*, Centrarchiformes: Cirrhitioidei, NCBI Taxon ID: 76931) is a species of demersal marine teleost fish that is widely distributed around all inshore areas of New Zealand and along the southern coasts of

Australia. It is distinguishable from other New Zealand “morwongs” by the black saddle across its nape (Roberts et al., 2015) and displays a single elongated pectoral fin ray that is characteristic of *Nemadactylus* species (Ludt et al., 2019). The species and its genus have been recently moved from the Cheilodactylidae to the Latridae following extensive revision of the taxonomy of both families, which until then was poorly understood (K. Kimura et al., 2018; Ludt et al., 2019). Tarakihi is an important commercial and recreational inshore fishery, especially in New Zealand, where more than 5,000 tonnes are harvested every year (Fisheries New Zealand, 2018b). Like many other fisheries species, tarakihi stocks have been heavily fished over the past century. As a result, the spawning biomass is now concerningly depleted to numbers below the fisheries management soft limit of 20% on the east coast of New Zealand, where fishing effort is highest (Langley, 2018). Low effective population size and spawning biomass are of concern for the long term sustainability of this species, particularly with added and increasing environmental pressures due to global warming. Climate change is already impacting marine ecosystems and is expected to affect the distribution and productivity of many fisheries species (Babcock et al., 2019; Burrows et al., 2011; Ramos et al., 2018).

The application of genome-wide markers for tarakihi fisheries management has been limited by fundamental knowledge gaps, most notably, the lack of a reference genome. Consequently, the first step in developing new genomic resources for this species is to assemble a high-quality reference genome that can be used to develop high-resolution studies of genetic stock structure. This would offer the potential for detecting adaptive genetic variation. Incorporating adaptive genetic variation, along with neutral variation, will greatly enhance the management application of fisheries genetic data (Benestan, 2019; Bernatchez et al., 2017; Papa, Oosting, et al., 2021; Thomson et al., 2021). While the neutral markers can detect reproductively isolated stocks, the adaptive loci can detect differentiation in reproductive success of migrant fish moving to locally adapted stocks. Using high-resolution markers sets for both neutral and adaptive variation will revolutionise the way genetic markers are used to define fisheries stocks.

As sequencing technology is rapidly changing and improving, a range of sequencing data types has been used to produce genome assemblies, thus providing a range of genome

qualities, contiguity and completeness. While short-read Illumina sequencing produces highly accurate reads, their short length (usually less than 200 bp) makes them computationally difficult to assemble. This is particularly problematic for regions that span highly repetitive segments of the genome. Complex genomes often result in highly fragmented assemblies (Koren et al., 2012; Rice & Green, 2019). The development of less accurate long-read sequencing technologies from Oxford Nanopore and Pacific Biosciences (PacBio) has improved the assembly process by combining them with short-read data to create “hybrid”, more contiguous genome assemblies (Austin et al., 2017; Dhar et al., 2019; Jiang et al., 2019; Tan et al., 2018; Wiley & Miller, 2020; Zimin, Puiu, Luo, et al., 2017; Zimin, Stevens, et al., 2017).

The rapid improvements in sequencing technologies have also improved the ability to collect RNA sequence (RNA-seq) data. Short read RNA-seq data has been used to assist in genome annotation by first assembling a transcriptome and mapping it to the orthologous sequences to find protein-coding genes. A downside of this short read length (c. 100–150 bp) is that it is difficult or impossible to detect and characterise alternative isoforms of the coding sequences, while alternative splicing is known to occur in the vast majority of multi-exon genes (Hardwick et al., 2019). The circular consensus sequencing (CCS) PacBio technology produces reads that are both thousands of bp long and highly accurate (as opposed to the Nanopore and PacBio continuous long reads mentioned above). CCS long-read DNA sequencing can be applied to DNA (i.e. High fidelity, or HiFi reads) and RNA (i.e. isoform sequencing, or Iso-Seq). By capturing the entire sequence length of RNA molecules, Iso-Seq allows for the sequencing of complete, uninterrupted mRNAs, which enables the accurate characterisation of isoforms (An et al., 2018; Byrne et al., 2019; Y. Gao et al., 2019; Hoang & Henry, 2021). Iso-Seq has been used to provide further insights into alternate splicing in the transcriptomes of several organisms, like the human (*Homo sapiens*) (Kuo et al., 2020), the chicken (*Gallus gallus*) (Kuo et al., 2017) or the goldfish (*Carassius auratus auratus*) (W. Gan et al., 2021). Iso-seq is now also used to annotate *de novo* genome assemblies of non-model organisms, like the cave nectar bat (*Eonycteris spelaea*) (Wen et al., 2018), the pharaoh ant (*Monomorium pharaonis*) (Q. Gao et al., 2020), the red-eared slider turtle (*Trachemys scripta elegans*) (Simison et al., 2020), or the sponge gourd (*Luffa* spp.) (Pootakham et al., 2020).

The main goal of this study was to achieve the first tarakihi genome assembly. This was attempted by using a combination of short-read Illumina and long-read Nanopore sequencing data. Four assembly pipelines were compared, three of which used algorithms implemented in MaSuRCA for hybrid assembly, and a fourth pipeline based on a trial run of low-coverage DNA sequence reads (4 Gb) generated using the PacBio HiFi platform. Iso-Seq data was also included to assist with gene annotation and the identification of gene isoforms.

3.3 Materials and Methods

3.3.1 Tissue collection and nucleotide extraction

Tissues for Illumina and Nanopore sequencing were collected from a freshly vouchered *N. macropterus* specimen (standard length: 285 mm, weight: 460 g) identified as male by observation of the gonads. The specimen was a captive-bred from Plant and Food Research, Nelson, New Zealand (Figure 3.1A) and is thereby referred to as TARdn1 (for “tarakihi *de novo*”). A caudal fin clip and a heart piece were stored in 96% EtOH, and a kidney piece was stored in DESS (20% DMSO, 0.25 M EDTA, NaCl saturated solution). Total genomic DNA was extracted from these tissues using a high-salt extraction protocol adapted from Aljanabi & Martinez (1997) that included an RNase treatment and then suspended in Tris-EDTA buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA). The integrity of DNA fragments was assessed by gel electrophoresis in 1% agarose. The purity and quantity of DNA (concentration > 200 ng/μl, A260/280 ≈ 1.8, A60/230 ≈ 2, total weight > 20 μg) were estimated with CLARIOstar spectrometer (BMG Labtech). Purified DNA samples were sent to Annoroad Gene Technology Co. Ltd. (Beijing, China) and NextOmics Biosciences Co., Ltd. (Wuhan, China) for Illumina and Nanopore library preparation and sequencing.

Tissues for HiFi sequencing and Iso-Seq were obtained from a wild specimen captured by a recreational fisherman at Kau Bay, in the Wellington harbour (New Zealand), thereby referred to as TARdn2 (Figure 3.1B). The specimen was collected for tissue sampling after being filleted by the fisherman. It had a standard length of 255 mm and was identified as male by observation of the gonads. Tissues were collected a few hours after capture and flash-frozen in liquid nitrogen. Five pieces of tissues were sent to BGI Tech Solutions Co., Ltd. (Hong Kong, China): one tissue (heart) for DNA extraction and HiFi sequencing and four tissues (liver,

white muscle, brain, and spleen) for RNA extraction and Iso-Seq. DNA and RNA were extracted by BGI using a phenol-chloroform method.

Figure 3.1. Tarakihi specimens used in this study. (A) TARdn1: captive bred specimen used for Illumina and Nanopore sequencing. (B) TARdn2: wild-caught specimen used for HiFi sequencing and Iso-Seq.

3.3.2 Genome size estimation pre-sequencing

To estimate the size of the *N. macropterus* genome and ensure there was a sufficient amount of DNA sequencing for adequate coverage, genome information from closely related species was assessed. As of October 2018, only two other Centrarchiformes genome assemblies were deposited in NCBI at the scaffold level (accession numbers: GCA_002120245.1 (Murray cod, *Maccullochella peelii*), and GCA_003416845.1 (barred knifejaw, *Oplegnathus fasciatus*)), which had genome lengths of 633.24 and 766.3 Mb. Moreover, the species closest to *N. macropterus* for which genome size was estimated on the

Animal Genome Size Database (<http://www.genomesize.com>) was the red morwong *Cheilodactylus fuscus*, with a C-value of 0.72, or approximately 700 Mb. The genome size of *N. macropterus* was thus estimated to be about 700 Mb. The quantity of Illumina and Nanopore bases to be sequenced was tuned for a deep 85× Illumina coverage (c. 60 Gb) and 140× Nanopore coverage (c. 100 Gb), following sequencing provider recommendations.

3.3.3 Library preparations and sequencing

Library preparations, sequencing, and the first filtering step (except for Nanopore reads) were performed by the sequencing providers. For Illumina reads, DNA samples were sheared with Bioruptor® Pico system (Diagenode) for a fragment insert size of 350+/-50 bp, and a PCR-free library was obtained with NEBNext® Ultra™ II DNA Library Prep Kit for Illumina (New England Biolabs). Approximately 200 million of 150 bases pair-end reads were generated using the HiSeq X System (Illumina). Raw Illumina reads were filtered by discarding read pairs if (1) one read contained some adapter contamination for more than five nucleotides, (2) more than 10% of bases were uncertain in either one read, or (3) the proportion of bases with Quality Value ≤ 19 was over 50% in either one read. For Nanopore library preparation, large size DNA fragments were selected by automated gel electrophoresis with BluePippin (Sage Science) followed by enrichment and purification using beads. Fragmented DNA was then end-repaired, A-tailed and purified, and adapter ligation was done using the Ligation Sequencing Kit 1D 108 (Oxford Nanopore Technologies). The resulting DNA library of 20–40 Kb fragments was then loaded into two flow cells for real-time single-molecule sequencing on PromethION (Oxford Nanopore Technologies). Reads were base-called from their raw FAST5 files using Albacore 2.0.1 (<https://community.nanoporetech.com>). HiFi library was prepped with SMRTbell® Express Template Prep Kit 2.0 (Pacific Biosciences) and CCS was performed on one-third of an SMRT Cell 8M with a PacBio Sequel II sequencer. ZMWs were filtered to retain a minimum of three passes and a predicted quality value (RQ) of 99. Four Iso-Seq libraries of 0–5 kb insert sizes (one per tissue) were generated using the SMRTbell® Express Template Prep Kit 2.0. The multiplexed libraries were sequenced on one SMRT Cell 8M with a PacBio Sequel II sequencer, resulting in 3.6 million polymerase reads from which sub-reads were extracted.

3.3.4 Illumina reads: Quality and contamination filtering

Primary quality filtering performed by the provider resulted in 405.2 million Illumina pair-end reads (60.78 Gb). Quality metrics of these filtered reads were visualised with FastQC v0.11.7 (S. Andrews, 2018) before proceeding to the next steps. Kraken v2.0.7-beta (Wood et al., 2019) was used to detect and filter contamination from archaea, bacteria, viral and human sequences based on the MiniKraken2 v2 8GB database (Wood, 2019). The 9.25% of reads that were classified as contaminants were discarded, leading to 367.8 million non-contaminated reads (55.16 Gb) (Table 3.1).

3.3.5 Illumina reads: Mitogenome assembly and exclusion

Illumina reads filtered for quality and contamination were mapped against the Peruvian morwong (*Cheilodactylus variegatus*) complete mitochondrial sequence retrieved from Genbank (accession number: KP704218.1) with Geneious v11.04 (Kearse et al., 2012) using five iterations of the default mapper set to the highest sensitivity. The extracted consensus sequence resulted in a 16,650 bp assembly of the *N. macropterus* mitogenome. The mitogenome was then annotated using the MitoAnnotator web interface (Iwasaki et al., 2013). These two steps above (mitogenome assembly and annotation) were carried out as part of a separate study (Halliwell, 2019). Sequences of mitochondrial origin were then filtered out of the Illumina reads as follows: first, bwa-kit v0.7.15 (H. Li & Durbin, 2009) was used to align the Illumina reads to the indexed *N. macropterus* reference mitogenome with default parameters. Among other aligners, the BWA-MEM algorithm (H. Li, 2013) was selected because it is the most accurate for this type of short-read data (Keel & Snelling, 2018). In the resulting SAM alignment, 0.46% of reads mapped to the mitogenome. Then, all the reads from the alignment that did not map to the mitogenome were extracted to a new mitochondria-free alignment using SAMtools v1.9 (H. Li et al., 2009) `view` with parameters `-b -f 4`, sorted by name and finally converted back to FASTQ paired-end reads with bedtools v2.27.1 (Quinlan & Hall, 2010).

3.3.6 Genome size estimation post-sequencing

Genome size and sequencing coverage based on the Illumina sequence reads was performed with a *k*-mer frequency analysis. Total number of 17, 21, and 27-mers were counted with jellyfish v2.2.10 (Marçais & Kingsford, 2011) command `count` and the resulting histograms were computed with command `histo`. The histograms were analysed with GenomeScope (Vurture et al., 2017), which estimated a genome size of c. 516–520 Mb, with a high heterozygosity level of 1.01–1.07 % and a duplication level of 0.98–1.10 % (

Figure 3.2, Supplementary Figure 3.1). This estimated haploid genome size was consistent, albeit c. 150 Mb lower than the size estimated pre-sequencing. However, it is common for *k*-mer estimated size and genome assembly size to be smaller than the size estimated with C-value (Austin et al., 2017; Feron et al., 2020; Jansen et al., 2017). The heterozygous coverage of 40× was considered sufficient for performing genome assembly.

Figure 3.2. Histogram of 21-mer frequency in Illumina reads. Estimation of genome size, heterozygosity, and duplicated regions. The first and second peaks show the *k*-mer frequency of heterozygous and homologous regions, respectively. See Supplementary Figure 3.1 for 17- and 21-mer models.

3.3.7 Nanopore reads sequencing and filtering

A total of 99.18 Gb was obtained from the raw unfiltered reads, with an average read length above 6 Kb and a maximum length above 1 Mb (Table 3.1). Quality control of the raw reads was performed with NanoPack v1.0.0 (De Coster et al., 2018) using NanoStat on both FASTQ reads and Albacore summary files. Nanopore reads were filtered and trimmed with NanoFilt by applying a minimum length cut-off of 500 bases (Tan et al., 2018), a minimum average read quality score of 7 (c. 80% base call accuracy), and removing the first 50 nucleotides following the author's recommendations (De Coster, 2017). Given that quality values based on summary files were slightly lower overall than when based on reads (as expected, c.f. github.com/wdecoster/nanofilt), the quality filtering was done based on the summary file values to be more stringent. Filter-trimmed reads from both cells were merged into a single FASTQ file.

3.3.8 Illumina + Nanopore hybrid assembly

De novo genome assembly of short and long reads was performed with the Maryland Super-Read Celera Assembler pipeline, MaSuRCA (Zimin et al., 2013; Zimin, Puiu, Luo, et al., 2017). This is one of the most common assemblers for performing short and long reads hybrid genome assemblies of eukaryotes, with consistently good results across studies (Jiang et al., 2019; Tan et al., 2018; Thai et al., 2019). In brief, MaSuRCA typically works as follows: Illumina paired-end short reads are first assembled into non-ambiguous super-reads, which are then mapped to Nanopore reads to further assemble them in long, high-quality pre-mega-reads. If there are gaps between mega-reads in respect to their mapping to the Nanopore reads, these gaps are filled with the Nanopore read sequence only if the Nanopore read stretch meets some minimum criteria of coverage and quality to produce the mega-reads. If there are still gaps that cannot be merged between mega-reads due to poor quality of the Nanopore sequence, regions flanking these gaps are linked together as linking pair mates. The mega-reads and linking pairs are then assembled with either CABOG or Flye (see below).

Before assembly, the filtered Illumina reads were not trimmed or edited as per MaSuRCA author recommendation (<https://github.com/alekseyzimin/masurca>). The hybrid

Illumina + Nanopore assembly was run on MaSuRCA v3.2.9 with recommended parameters, automatic *k*-mer size computation, and a jellyfish has size of 20,000,000,000 (`PE = pe 350 50, NANOPORE, EXTEND_JUMP_READS = 0, GRAPH_KMER_SIZE = auto, USE_LINKING_MATES = 0, USE_GRID = 0, GRID_BATCH_SIZE = 300000000, LHE_COVERAGE=25, MEGA_READS_ONE_PASS=0, LIMIT_JUMP_COVERAGE = 300, CA_PARAMETERS = cgwErrorRate = 0.15, KMER_COUNT_THRESHOLD = 1, CLOSE_GAPS = 1, NUM_THREADS = 32, JF_SIZE = 20000000000, SOAP_ASSEMBLY = 0`). MaSuRCA v3.2.9 uses a modified version of the CABOG assembler (J. R. Miller et al., 2008) for the final assembly of corrected mega-reads. However, later releases of MaSuRCA included the Flye assembler (Kolmogorov et al., 2019) as a supposedly faster and more accurate alternative tool for the same step. To compare both methods, a second assembly was run on MaSuRCA v3.4.1 with the same parameters as above, but this time using `FLYE_ASSEMBLY = 1`. The Flye assembly was subsequently polished with POLCA (Zimin & Salzberg, 2020) as implemented in MaSuRCA v3.4.1 on default settings, using the clean Illumina reads to fix substitutions and indel errors.

3.3.9 HiFi sequencing and assembly

HiFi reads were converted from BAM to FASTA and FASTQ with SMRTLink v9.0 (PacBio, 2020) `bam2fastx`. Assembly was performed with hifiasm v0.13 (H. Cheng et al., 2021) using default parameters. The primary contigs were extracted from the GFA graph and converted to FASTA with command `awk '/^S/{print ">"$2;print $3}'`. Another assembly was also tentatively performed with HiCanu as implemented in Canu v2.1.1 (Nurk et al., 2020), with an estimated genome size of 600 Mb. However, the read coverage estimated (6.68×) was lower than the minimum coverage allowed by HiCanu (10×), so the assembly could not be completed.

3.3.10 Quality assessment and comparison of assemblies

After each assembly, basic contiguity statistics were computed with bbmap v38.31 (Bushnell, 2018) script `stats.sh`. Length, GC content and GC skew of scaffolds in all assemblies were also reported with seqkit v0.10.1 (W. Shen et al., 2016) command `fx2tab`. To assess the completeness of the assemblies, the Benchmarking Universal Single-Copy

Orthologs (BUSCO) tool v3.0.2 (Simão et al., 2015) was used with parameter `-sp zebrafish` on the Actinopterygii odb9 orthologs set, which contains 4,584 single-copy orthologs that are present in at least 90% of ray-finned fish species. Augustus v3.3.1 (Stanke et al., 2004), NCBI blast+ v2.7.1 (Camacho et al., 2009), hmmer v3.2.1 (Eddy, 2011), and R v3.6.0 (R Core Team, 2020) were also required to run the BUSCO shell script.

The quality of the CABOG and Flye assemblies was further compared by mapping clean Illumina reads back to the assemblies themselves with bwa-kit v0.7.15 using `bwa mem -a -M`. The resulting alignment files were also used to plot Feature Response Curves (FRC) (Vezzi et al., 2012b) with FRCbam v5b3f53e-0 (Vezzi et al., 2012a). This allowed comparison of quality of the assemblies without relying on contiguity, by plotting the accumulation of error “features” along the genome (e.g. areas with low or high coverage, number of unpaired reads, misoriented reads). The presence of unmerged haplotigs in the CABOG and the Flye polished assembly was investigated by using minimap v2.16 (H. Li, 2018) with parameters `-ax map-ont --secondary = no` to map the clean Nanopore reads back to the assembly and then analysing the resulting alignment with Purge Haplotigs v1.1.1 (Roach et al., 2018) command `hist`. The presence of trailing Ns in the Flye polished assembly was tested by using seqkit v0.10.1 command `-is replace -p "^n+|n+$" -r ""` and comparing the input and the output.

A last quality check of the CABOG and Flye polished assemblies was done by plotting assemblies against each other and against two chromosome-level fish assemblies using MashMap 2.0 (Jain et al., 2018) with a minimum mapping segment length of 500 bp and a minimum identity of 85% (for comparison between tarakihi assemblies) and 90% (for comparison between different species). To visualise the presence of potential misassemblies on the longest scaffolds, the results from MashMap were used to plot the mappings of these scaffolds between different assemblies with a custom R script (`plot_mashmap_scaffolds.R`). The first fish chromosome-level assembly used for comparison was the mandarin fish *Siniperca chuatsi* (SinChu7, GCA_011952085.1) because it was the phylogenetically closest chromosome-level assembly (Centrarchiformes, Centrarchoidei) available on NCBI at the time this analysis was performed. The second was the Australasian snapper *Chrysophrys auratus*

(SNA1, <https://www.genomics-aotearoa.org.nz/data>), in order to compare with a well-curated specimen from a more evolutionarily distant species.

Final visualisation of contiguity and completeness of the genome assemblies was generated with assembly-stats v17.02 (Challis, 2017) as implemented in the grpiccoli container (Piccoli, 2021).

3.3.11 Estimation of heterozygosity

The heterozygosity of TARdn1 was estimated a second time by calling SNPs from the Illumina reads aligned to the final assembly. The reads were mapped to the polished assembly with bwa-kit v0.7.15 using the command `bwa mem -a -M`. Duplicates were marked with picard v2.18.20 (Broad Institute, 2019) `MarkDuplicates`. SNPs were called using bcftools v1.9 (H. Li, 2011) commands `mpileup (-C50 -q10 -incl-flags 2)` and `call (-m - - variants-only -- skip-variants indels)`. To filter for good quality SNPs, variants depth distribution was plotted. The modal depth of coverage was 82, with an increase in steepness starting at c. 20 and a decrease starting at c. 120 (Supplementary Figure 3.2). Consequently, the final SNP set was filtered with vcftools v0.1.16 (Danecek et al., 2011) for a minimum reference allele frequency of 0.25, a genotype depth of minimum 20 and maximum 120, and a minimum site quality of 20.

3.3.12 Genome repetitive elements detection

Repetitive elements (RE) in the *N. macropterus* genome were identified both by *de novo* modelling and based on repeats homology. RepeatModeler v2.0.1 (Flynn et al., 2020), as implemented in Dfam TE Tools container v1.2 (<https://github.com/Dfam-consortium/TETools>), was used to identify repeat models *de novo* using parameter `-LTRStruct` to include the detection of long terminal repeat retrotransposons. For the homology-based library, RepeatMasker v4.1.1 (Smit et al., 2013) tool `famdb.py` was used to obtain known Actinopterygii repeats from the combined total Dfam v3.3 (Storer et al., 2021) and RepBase RepeatMasker Edition v20181026 (Bao et al., 2015) databases, using parameters `--ancestors -descendants --include-class-in-name --add-reverse-complement`. Both *de novo* and homology-based repeat libraries were then

concatenated in a custom repeat library for *N. macropterus*. The genome assembly sequences were then mapped against the custom repeat library with RepeatMasker v4.1.1 (`-gff -xsmall`) to classify repeat regions, create a repeat annotation file, and produce a “soft-masked” (i.e. masked bases in lower case) genome assembly. An alternate “hard-masked” assembly was also created by converting lower cases in the soft-masked assembly into Ns.

3.3.13 Iso-Seq analysis

Iso-Seq sub-reads were processed with the SMRTLink v9.0 Iso-Seq pipeline. Circular consensus sequences were generated from the sub-reads with command `ccs` using a minimum read quality (RQ) of 0.9. Clontech and NEB primers removal and de-multiplexing were performed using `lima` with parameters `--isoseq --dump-clips --peek-guess`. Poly-A tails were trimmed and concatemers were removed with `isoseq3 refine`. At that point, BAM files containing sequence reads from the four tissues were merged in one. Clustering and polishing of full-length reads were performed with `isoseq3 cluster` and parameter `--use-qvs` to obtain a dataset of high-quality isoforms with a predicted accuracy > 0.99. These high-quality polished isoforms were then aligned to the unmasked *N. macropterus* genome with `pbmm2` (`--preset ISOSEQ --sort`). Subsequently, redundant isoforms were collapsed into non-redundant transcripts loci using the command `collapse`. Non-redundant transcripts were screened for REs against the *N. macropterus* custom repeat library with RepeatMasker v4.1.1. Transcripts with $\geq 70\%$ bases masked were considered REs. Identified REs were discarded from further analyses using a custom bash script for filtering (`Count_filter_N_isoseqrepeats.bash`) and categorized using a custom R script (`R_characterize_transcripts.R`).

Alternative splicing (AS) events in the repeat-cleaned Iso-Seq reads were counted and classified with SUPPA v2.3 (Trincado et al., 2018) with default parameters. These results were compared with reported AS values for other animal species from studies that also used SUPPA on Iso-Seq reads. Results reported were compiled for the zebrafish (*Danio rerio*) (Nudelman et al., 2018), the goldfish (*Carassius auratus auratus*) (W. Gan et al., 2021), the Wuchang bream (*Megalobrama amblycephala*) (Chen et al., 2021), the whiteleg shrimp (*Litopenaeus*

vannamei) (X. Zhang et al., 2019), and the cave nectar bat (*Eonycteris spelaea*) (Wen et al., 2018).

3.3.14 Genome annotation

The unmasked *N. macropterus* genome was annotated using the MAKER v2.31.10 (Holt & Yandell, 2011) pipeline. First, the simple repeats were filtered out of the repeats annotation file with a custom bash script (`rm_simple_repeats.bash`) to retain only complex repeats. Only complex repeats were kept because MAKER will hard-mask every region provided in the repeats annotation file before running, discarding them from the gene detection process. However, the simple repeats should be available for gene annotation because low-complexity regions are expected within many genes. Hard-masking only complex repeats regions as a first step allows MAKER to subsequently identify and soft-mask the simple repeats regions internally. Gene matches that start in a non-masked region but extend in a soft-masked region can then be taken into account in the gene detection process. A first round of MAKER was run on the unmasked genome using the high-quality, non-redundant, non-repetitive Iso-Seq transcripts to infer gene predictions (`est2genome = 1`). For repeat masking during this step, the complex repeats GFF file was provided for hard masking and only simple repeats were annotated (`model_org = simple`). All GFF and FASTA outputs were then merged with `ggf3_merge` and `fasta_merge`. Training files for the *ab initio* gene predictors SNAP v2013.11.29 (Korf, 2004) and Augustus v3.3.1 (Stanke et al., 2004) were generated based on round 1 results. For SNAP, only gene models with a maximum Annotation Edit Distance (AED) of 0.25 and a minimum protein length of 50 were used. For Augustus, all the regions that contain mRNA annotations, including the 1,000 surrounding bp, were extracted to a FASTA file using a custom bash script (`augustus_rndx.bash`). BUSCO v3.0.2 was then run in “genome” mode on the FASTA file using the Actinopterygii odb9 orthologs set, the zebrafish as initial HMM model, and parameter `--long` to self-train Augustus. MAKER was then run a second time using SNAP and Augustus training files, as well as the Iso-Seq transcriptome and repeats alignments as evidence (`est2genome = 0`). For this, all lines containing “est2genome” and “repeat” in the merged GFF from round 1 were extracted and copied in two files that were provided as evidence with the parameters `est_gff` and `rm_gff`, respectively. Additionally, gene predictions were also inferred from protein

homology during this round (`protein2genome = 1`), by using protein sequences of zebrafish (*Danio rerio*), three-spined stickleback (*Gasterosteus aculeatus*), spotted gar (*Lepisosteus oculatus*), Nile tilapia (*Oreochromis niloticus*), medaka (*Oryzias latipes*), Japanese puffer (*Takifugu rubripes*), green spotted puffer (*Tetraodon nigroviridis*), and southern platyfish (*Xiphophorus maculatus*) that were downloaded from Ensembl release version 103 (Kersey et al., 2016). After that, SNAP was trained again using the results from round 2, and a third run was performed by using the *ab initio* training files, as well as the extracted repeats, Iso-Seq, and protein homology GFF files as evidence. Genes were renamed with MAKER `maker_map_ids` and `map_x_ids`.

All proteins predicted from the second round of MAKER were blasted against the NCBI non-redundant protein sequences database (NR) with `blastp (-evalue 1e-6 -max_hsps 1 -max_target_seqs 1 -outfmt 6)` as implemented in blast+ v2.6.0. All putative gene functions based on the best homology matches were annotated in the genome with a custom bash script (`add_blast_annotation_custom.bash`). Protein-coding genes were also searched for protein domains and signatures and annotated for InterPro (IPR), Pfam, and Gene Ontology (GO) terms using InterProScan v5.50-84.0 (P. Jones et al., 2014) and MAKER `ipr_update_gff`. Protein domains were exported as features in a GFF file using MAKER `iprscan2gff3`.

Finally, low-quality genes were identified with AGAT v0.6.0 (Dainat, 2021). These genes were filtered out if they were shorter than 50 amino acids and flagged if they had an incomplete open reading frame (ORF). Gene models produced by the second MAKER round were kept as the final reference dataset based on their higher number, AED distribution, and BUSCO completeness (Supplementary Table 3.2). Genome annotation was also inspected visually with JBrowse v1.1.10 (Skinner et al., 2009).

3.3.15 General bioinformatics tools

After each assembly, scaffolds were sorted by size using `seqkit v0.10.1` command `sort -l -r -2` and renamed with command `replace -p .+ -r "{nr}"` (i.e.

scaffold “1” being the longest, etc.). All alignment files were systematically sorted by leftmost coordinates, converted to BAM and indexed with SAMtools v1.9. Alignment summary reports were produced with BAMtools v2.5.1 (Barnett et al., 2011). FASTQ files were converted in FASTA when needed with seqtk v1.3 (<https://github.com/lh3/seqtk>), and similarly, GFFs were converted to GTF with AGAT v0.6.0. Analyses were performed on Rāpoi, the Victoria University of Wellington high-performance computer cluster. Analyses requiring R scripts were performed in R v4.02 (R Core Team, 2020) on RStudio (RStudio Team, 2020). All bash and R scripts used for this chapter are available on GitHub on the following repository: https://github.com/yvanpapa/tarakihi_genome_assembly.

3.4 Results

3.4.1 Genome sequencing

Illumina sequencing reads filtering (i.e. quality, contamination and mitochondria) resulted in a final dataset of 54.91 Gb short reads (Table 3.1) with a c. 92× depth of coverage. The GC content was 43% and the overall sequence read quality was high. Both forward and reverse reads passed all the FastQC criteria, i.e. they were never flagged for poor quality (Supplementary Figure 3.3). Although there was a small bias in per base sequence contents of the first c. 10 bases, this was expected due to the non-random nature of the hexamer priming step during sequencing (Hansen et al., 2010). This slight deviation from uniformity in sequence content was not considered an issue because there is no quantitative step involved in the analyses based on the short reads. Nanopore sequencing, filtering and trimming resulted in 9.18 million reads (73.39 Gb), or 122× coverage, with an average read length of 8 Kb (Table 3.1), a mean read quality of 7.9, and an N50 length of 9.5 Kb. A total of 285,997 CCS Hi-Fi reads (4.01 Gb) and 91,602 repeat-free, non-redundant, high-quality Iso-Seq transcripts (312.31 Mb) were also obtained.

Table 3.1. Summary of number, base quantity, and length of reads obtained at several steps of the quality filtering pipelines.

Reads	Number of reads	Total number of bases	Minimum read length	Average read length	Maximum read length
Raw Illumina PE reads	425,740,632	63,861,094,800	150	150	150

Quality-filtered Illumina PE reads	405,228,300	60,784,245,000	150	150	150
Uncontaminated Illumina PE reads	367,760,592	55,164,088,800	150	150	150
Final Illumina PE reads	366,065,036	54,909,755,400	150	150	150
Raw Nanopore reads cell 1	8,270,853	52,169,467,195	5	6,307.6	1,029,695
Raw Nanopore reads cell 2	7,229,556	47,015,342,634	5	6,503.2	1,035,919
Final Nanopore reads	9,178,726	73,394,980,774	450	7,996.2	182,445
HiFi reads	285,997	4,009,988,664	49	14,021.1	27,427
Iso-Seq sub-reads (4 tissues)	171,924,197	302,196,904,697	51	2,601.15	278,803
Final Iso-Seq transcripts	91,602	312,308,038	80	3,409.4	10,426

Notes: Reads in bold were the ones used in the final retained (Flye polished) assembly. Final Illumina PE reads have been filtered for quality, DNA contamination, and mitochondrial DNA. Final Nanopore reads have been filtered for quality. Final Iso-Seq CCS transcripts were filtered for quality and repeat transcripts and were non-redundant.

3.4.2 Assemblies comparison and quality assessment

The Flye assembly reduced the number of scaffolds by more than half compared to the CABOG assembly (Table 3.2). The scaffold N50 length of the Flye assembly was almost twice as long and the number of complete BUSCOs was higher. The Flye assembly size was also more consistent with the haploid genome size pre-estimated by *k*-mer counting (c. 520 Mb) than the CABOG assembly. Interestingly, the Flye assembly also corrected a misassembly of the first scaffold of the CABOG assembly (see below). Polishing the Flye assembly resulted in the correction of 43,080 substitutions errors and 42,783 deletion errors. The polished assembly had the same number of scaffold and contigs, but a few hundred fewer bases, and one missing BUSCO was recovered into an additional single-copy BUSCO. The hifiasm assembly performed on the HiFi reads did not produce satisfactory results compared to the Illumina + Nanopore hybrid assemblies, with six to ten times more scaffolds, an N50 length 50 times smaller, and a BUSCO completeness lower than 90%. This was most probably due to the low coverage of HiFi reads (c. 6.5x) used for this sequencing trial.

3.4 Results

Table 3.2. General statistics of the four assemblies produced.

Reads type	PacBio HiFi reads	Illumina + Nanopore reads		
	hifiasm	CABOG	Flye	Flye polished
Genome Assembly				
Scaffold assembly size	778,095,731 bp	608,975,097 bp	567,903,348 bp	567,902,715 bp
Total number of scaffolds	13,511	2,696	1,214	1,214
Longest scaffold	469,394 bp	18,930,378 bp	13,913,512 bp	13,913,694 bp
Scaffold N50 / L50	67.836 Kb / 3,650	1.87 Mb / 69	3.37 Mb / 45	3.37 Mb / 45
Scaffold N90 / L90	30.868 Kb / 10,456	140.52 Kb / 535	437.51 Kb / 219	437.54 Kb / 219
Proportion of gap sequences	0.001%	0.002%	0.001%	0.001%
Contigs size	778.096 Mb	609.964 Mb	567.900 Mb	567.900 Mb
Total number of contigs	13,511	2,809	1,245	1,245
Contig N50 / L50	67.836 Kb / 3,650	1.79 Mb / 74	2.94 Mb / 52	2.94 Mb / 52
Contig N90 / L90	30.868 Kb / 10,456	137.36 Kb / 556	429.99 Kb / 242	429.98 Kb / 242
A / T / G / C / bases (%)	28.17 / 28.14 /	28.06 / 28.13 /	28.10 / 28.15 /	28.10 / 28.15 /
	21.84 / 21.85	21.91 / 21.90	21.87 / 21.88	21.87 / 21.88
GC standard deviation	2.13%	5.87%	3.87%	3.87%
Genome Completeness (4,584 Actinopterygii BUSCOs)				
Complete BUSCOs	88.8%	97.6%	97.7%	97.8%
Complete single-copy BUSCOs	57.3%	92.9%	95.1%	95.2%
Complete duplicated BUSCOs	31.5%	4.7%	2.6%	2.6%
Fragmented BUSCOs	3.5%	0.8%	0.8%	0.8%
Missing BUSCOs	7.7%	1.6%	1.5%	1.4%

Note: The Flye polished assembly (in bold) yielded the best results and was retained for all subsequent analyses.

Approximately 99.7% of Illumina reads could be mapped back to the CABOG assembly, and 99.8% to both Flye assemblies, making the Flye assemblies slightly more accurate according to that metric. The Flye polished assembly had a slightly higher proportion of “proper-pairs” reads mapped (86.23%) than the un-polished assembly (85.7%). FRC curves showed that both Flye assemblies were more accurate than the CABOG assembly (Figure 3.3). Moreover, while both the un-polished and polished Flye assemblies have a very similar curve, for the same genome coverage, the polished Flye assembly always had a slightly lower amount of cumulative errors compared to the un-polished assembly (Supplementary Figure 3.4.).

Figure 3.3. FRC curves for the CABOG, Flye and Flye polished assembly. The Y-axis represents the cumulative size of the assembly and the X-axis is the cumulative number of potential errors (i.e. “features”). Assemblies for which the curves are steeper are considered more accurate.

While there was evidence of the presence of unmerged haplotigs in the CABOG assembly (Figure 3.4A), none were detected in the Flye polished assembly (Figure 3.4B), thus a filtering step was not required. Trailing Ns were not present in the Flye polished assembly either.

Figure 3.4. Read depth histograms of the genome assemblies contigs, obtained by mapping the clean Nanopore reads back to the assembly. A unimodal distribution with a peak equal to the sequencing reads depth is expected for a haplotig-free assembly. Another peak at half of the sequencing reads depth (arrow) is indicative of the presence of unmerged haplotigs. A: CABOG assembly B: Flye polished assembly.

Interestingly, the longest scaffold of the CABOG assembly, scaffold 1, was 5 Mb longer than the longest scaffold of the Flye assembly (Table 3.2). Between-scaffolds alignment scores

obtained from MashMap (Supplementary Figure 3.5) were used to visualise a potential misassembly at that scaffold. The longest scaffold of the CABOG assembly corresponded indeed to the two longest scaffolds of the polished Flye assembly, scaffolds 1 and 2 (Figure 3.5A). The CABOG scaffold 1 is highly likely to have been misassembled, since it also corresponds to two long regions in two different linkage groups (i.e. chromosomes) in both chromosome-level assemblies of *S. chuatsi* and *C. auratus*. This is not the case for scaffold 1 in the polished Flye assembly (Figure 3.5B). This supported the interpretation that the “correct” longest scaffold is the one from the polished Flye assembly.

3.4 Results

Figure 3.5. Scaffolds plotted against total assemblies based on identity results from MashMap with minimum mapping region (i.e. “fragments”) length of 500bp. Each horizontal box is a scaffold of the reference on which the query scaffolds are mapped according to a given identity threshold. Mapped regions are ordered by base coordinate along the query scaffold on the x-axis, and the reference scaffolds on the y-axes. (A) CABOG assembly scaffold 1 mapped to the total polished Flye assembly, with corresponding Flye scaffold numbers reported on the right. (B) CABOG and Flye assemblies scaffold 1 mapped to the *S. chuatsi* and *C. auratus* chromosome-level assemblies.

3.4.3 Final assembly stats

The Flye polished assembly provided the best results and thus was used in all subsequent analyses. This final genome assembly consisted of 567,902,715 bases in 1,214 scaffolds, with a scaffold N50 length of 3.37 Mb and a proportion of gaps of 0.001% (Table 3.2, Figure 3.6). Base composition was A: 28.10%, T: 28.15%, G: 21.87%, C: 21.88% and overall standard deviation of GC content was 3.87%. The BUSCO completeness was very good overall, with more than 95% of the single-copy Actinopterygii orthologs retrieved in the final assembly (Table 3.2, Figure 3.6). The contiguity and completeness were high when compared to other Illumina + Nanopore hybrid assemblies (Table 3.3). The final assembly was named fNemMar1, in accordance with the Earth Biogenome Project sample naming scheme (<https://gitlab.com/wtsi-grit/darwin-tree-of-life-sample-naming>).

Figure 3.6. Visualisation of contiguity and completeness of the final tarakihi assembly. The contiguity is visualised in a circle representing the full assembly length of c. 568 Mb. The longest scaffold was 13.9 Mb. There were very few scaffolds (c. 2%) shorter than 100 Kb in length and the GC content was uniform throughout. See Supplementary Figure 3.6 for a comparison with the three other assemblies that were not retained.

Table 3.3. Comparison of the contiguity and completeness of genomes that were assembled using a hybrid approach including only short Illumina reads and long Nanopore reads.

Species	Genome (total scaffolds) length	Number of scaffolds	Scaffold N50 length	Complete BUSCOs	Protein-coding gene models	Functionally annotated genes
Tarakihi	568 Mb	1,214	3.4 Mb	97.80%	20,327	19,823
Murray cod	633 Mb	18,198	0.1 Mb	94.20%	26,539	25,607
Clownfish	881 Mb	6,404	0.4 Mb	96.30%	27,420	26,211
<i>Danionella translucida</i>	735 Mb	27,639	0.3 Mb	91.50%	24,097	21,491
Snout otter clam	544 Mb	622	2.1 Mb	95.80%	26,380	23,701
Indian blue peacock	915 Mb	15,025	0.2 Mb	not reported	23,153	21,854

Note: All fish genome assemblies that corresponded to the criteria are reported (Murray cod (*Maccullochella peelii*): Austin et al. (2017), clownfish (*Amphiprion ocellaris*): Tan et al. (2018), *Danionella translucida*: Kadobianskyi et al. (2019)) and two selected additional species have been included for comparison with other groups of organisms (Mollusc, snout otter clam (*Lutraria rhynchaena*): Thai et al. (2019); bird, Indian blue peacock (*Pavo cristatus*): Dhar et al. (2019)).

3.4.4 Estimation of heterozygosity

Variant calling of Illumina reads against the polished assembly resulted in a total of 3,654,819 SNPs. By dividing this number by the size of the genome, this corresponded roughly to a heterozygosity level of 0.64%. This is lower than the level estimated by *k*-mer frequency (c. 1.00%). However, it is common for heterozygosity estimated by *k*-mer frequency to be lower compared to called SNPs, because the SNP calling approach is more conservative (Thai et al., 2019). Nevertheless, the heterozygosity estimated for TARdn1 is one of the highest reported for a fish species. To our knowledge, this is the highest heterozygosity estimated for a fish through *k*-mer analysis, with other reported values ranging from 0.1% (Tibetan loach *Triplophysa tibetana* and Murray cod *Maccullochella peelii*) to 0.9% (Java medaka *Oryzias javanicus*) (Austin et al., 2017; Ge et al., 2019; Gong et al., 2018; Lu et al., 2020; Nguinkal et al., 2019; Takehana et al., 2020; Vij et al., 2016; X. Yang et al., 2019; H. H. Zhang et al., 2020; Zheng et al., 2021). Even the heterozygosity estimated through SNPs (0.64%) is high compared to estimations from other fish using the same method (e.g. large yellow croaker: 0.36% (Wu et al., 2014), grass carp: 0.25% (Y. Wang et al., 2015)). This result is even more striking in that the variant analysis was very stringent in our case by retaining only high-quality bi-allelic SNPs. This reinforces the recent findings that *N. macropterus* is a species with a historically large population that displays a particularly high genetic diversity (Papa, Halliwell, et al., 2021).

3.4.5 Repetitive elements and genes annotation

REs represented 30.45% of the genome, or a total of 172,911,032 bp. Although the proportion of REs in fish genomes can vary greatly at scales from 10% to 60% (Z. Yuan et al., 2018), the proportion of repeat elements in *N. macropterus* is on par with the proportion observed in other Centrarchiformes (Largemouth bass (*Micropterus salmoides*): 33.79%, Big-eyed mandarin fish (*Siniperca kneri*): 26.55%) (Lu et al., 2020; Sun et al., 2021) and for Perciformes in general (Z. Yuan et al., 2018). Of the REs known in the databases, interspersed repeats accounted for 27.62% of the genome, including 10.87% of DNA transposons and 6.17% of retro-elements (LINEs, LTR, SINEs, and PLE in that order). The rest of the repeat elements consisted of simple sequence repeats (Supplementary Table 3.1).

After filtering for length, the final predicted gene set included 20,169 protein-coding genes with a mean length of 13,832 bp, among which 95.5% had an AED < 0.5. The mean exon length was 229 bp, and the mean intron length in CDS was 1,184 bp. More than 98% of the genes were functionally annotated by at least one of the two methods used (blastp 98.2%, InterProScan 82.8%). The final genome annotation included the gene models, the Iso-Seq transcripts, the repeat elements, the protein homologies and the InterProScan protein domains (Figure 3.7).

Figure 3.7. Snapshot of the final genome annotation of fNemMar1, as visualised in JBrowse. The region shown spans bases 8,677,011 to 8,691,830 of scaffold 1. Black lines are introns and yellow squares are exons (except for repeat elements) and arrows indicate the 5'-3' strand direction. Some examples of annotation features include InterProScan protein domains (top right), MAKER gene model (feature with the blue exon), Iso-Seq unique transcripts (labelled "PB.XXX|TRPTXXX"), protein homologies from other fishes (lower right), SNAP and Augustus gene models, and repeat regions.

3.4.6 Iso-Seq analysis

Of the 93,949 full-length polished, non-redundant Iso-Seq transcripts, 2,347 were classified as REs and were filtered out from downstream analyses. For each of these RE transcripts, the main RE elements included DNA elements (801), LINEs (639), LTRs (464), SINEs (94), rRNAs (47), low complexity / simple repeats (33), rolling circles (26), satellites (16), and retroposons (2), as well as one LINE/LTR hybrid, and 224 unknown RE.

The final non-RE Iso-Seq dataset included 91,313 unique transcripts from 15,515 genes. The mean transcript per gene ratio was 5.89, with a median of 3 and a maximum of 211 (Figure 3.8A). This is higher than the values recently reported for humans (3.62) and two species of bats (1.92 and 1.49), but lower than pharaoh ants (9) (Q. Gao et al., 2020; Wen et al., 2018). Less than 5% of genes had more than 20 different transcripts. The predicted proteins of both genes that produced the most transcripts (respectively 211 and 164 transcripts) were collagen alpha chains isoforms (XP_006787735.1: collagen alpha-2(I) chain-like isoform X2, XP_020490299.1: collagen alpha-1(V) chain-like isoform X1), implicated in the structural integrity of the cellular matrix (GO:0005201).

Figure 3.8. Alternative transcripts metrics in the tarakihi transcriptome (A) Number of unique alternative transcripts per gene. (B) Classification and frequency of alternative splicing events. A5/A3: Alternative 5'/3' Splice Sites. AF/AL: Alternative First/Last Exons. MX: Mutually Exclusive Exons. RI: Retained Intron. SE: Skipping Exon.

A total of 26,644 AS events were detected in the tarakihi transcriptome (Figure 3.8B). The most frequent AS event was the retention of intron (46%), while “alternative last exons” and “mutually exclusive exons” were the rarest (less than 1% each). Some examples of these AS events were visualised in the tarakihi genome (Figure 3.9).

Figure 3.9. The seven types of alternative splicing events classified in the tarakihi transcriptome, with examples of each event class as visually shown in the annotation of the genome.

Comparison of the frequency of AS events in the tarakihi with other species showed that the trends are globally similar across organisms (Figure 3.10). Most organisms show

relatively high occurrences of RI, A3, A5, AF, and to a lesser degree SE, compared to AL and MX. The figure also shows that tarakihi, goldfish and cave nectar bat may have a better representation of the AS events proportions due to a much deeper coverage compared to the Wuchang bream, zebrafish and whiteleg shrimp (although values for MX and SE were not reported for the goldfish study). While it is the most common AS event in both tarakihi and goldfish, the proportion of RI events is much higher in tarakihi compared to the proportion of other events. While intron retention was thought until recently to be the least prevalent AS form in animals, it is now clear that this is not the case (as shown in the studies in Figure 3.10 but also e.g. Q. Gao et al. (2020); X. Wang et al. (2019)). RI events are widely used across organisms to tune down the levels of transcription of some genes in cells and tissues depending on their function (Braunschweig et al., 2014).

Figure 3.10. Comparison of alternative splicing event counts between tarakihi and five other animal species from other Iso-Seq AS studies. MX and SE events were not reported in the goldfish study.

3.4.7 Genome size

The size of the tarakihi genome was consistent with values for fish genomes that have been reported so far. A recent review of publicly available fish genome assemblies (comprising 244 species) showed that the average genome length of fish is 872.64 Mb but

varies between c. 300 Mb to c. 4.5 Gb (Fan et al., 2020). The genome size of *N. macropterus* (568 Mb) is several hundred Mb shorter than the two other published Centrarchiforme genomes, the largemouth bass *Micropterus salmoides* (964 Mb) and the big-eye mandarin fish *Siniperca kneri* (732.1 Mb) (Lu et al., 2020; Sun et al., 2021). However, *N. macropterus* is still evolutionarily far apart from these two species. The largemouth bass and the big-eye mandarin fish both belong to the Centrarchoidei sub-order, which is thought to have split from Cirrhitioidei at least 70 million years ago (Sanciango et al., 2016).

3.5 Conclusion

Following the advances in sequencing technologies, it has become clear how valuable reference genome assemblies are for the study of biology and conservation, leading to a global effort to assemble the genomes of as many organisms as possible (Fan et al., 2020; Koepfli et al., 2015; Worley et al., 2017). Here we present the first genome assembly of the tarakihi, a valuable commercial fisheries species and the first representative out of the c. 60 species of Cirrhitioidei to have a whole genome sequenced. While performing a hybrid assembly of Illumina and Nanopore reads with the latest tools led to a highly contiguous assembly with high gene completeness, this could be still improved in the future by adding Hi-C data to scaffold it to a chromosome-level assembly (Whibley et al., 2021). Moreover, while PacBio HiFi data was a very new and still relatively expensive technology at the time of this study, it will probably replace the short and long reads hybrid assembly method as the optimal genome assembly strategy by offering the best of both worlds (long reads and high quality) and allowing phasing of genomes. However, the present genome and its accompanying highly accurate transcriptome will still be a valuable resource for future studies, including, but not restricted to comparative genomics, population structure analyses, and the study of adaptive selection.

3.6 Supplementary Material

Supplementary Figure 3.1. Histograms of 17- and 27-mer frequency in clean Illumina reads.

Supplementary Figure 3.2. Distribution of coverage (top) and quality (bottom) of SNPs called from Illumina reads back to the assembly. SNPs were filtered for a minimum genotype depth of 20 according to the increase in steepness starting approximately at this point. Quality was always high, so the default site quality value was used.

Supplementary Figure 3.3. Some of FastQC quality metrics results, for forward (R1) and reverse (R2) reads. Top: Per base sequence quality. Middle: Per base sequence base content. Bottom: GC distribution over all sequences. See main text for the explanation on the slight bias in bases content for the first few bases in all reads. GC content of reverse reads detected a few over-represented sequences, which were most probably harmless sequencing artefacts that should be discarded during the quality control step of the MaSuRCA assembly.

Supplementary Figure 3.4. FRC curves as shown in Figure 3, but with only the Flye and Flye polished assemblies projected for better visualisation. For the same cumulative genome size, the Flye unpolished assembly always accumulates slightly more potential errors (i.e. features).

Supplementary Figure 3.5. Plots of pairwise alignment scores between scaffolds, obtained with MashMap. Each dot represents a match between the query and the reference sequence. Colours correspond to the strand direction (red for positive, blue for negative).

Supplementary Figure 3.6. Visualisation of contiguity and completeness of the four assemblies produced.

Supplementary Table 3.1. Main classes and proportions of repeat elements detected in the tarakihi genome.

Repeat type	No. of elements	Length occupied (bp)	% in the genome
Retroelements	323634	35060271	6.17
SINEs	33606	2627490	0.46
Penelope	8128	793327	0.14
LINEs	214886	24389420	4.29
CRE/SLACS	1	69	0
L2/CR1/Rex	139371	15942708	2.81
R1/LOA/Jockey	6466	805273	0.14
R2/R4/NeSL	5543	659291	0.12
RTE/Bov-B	24021	2686662	0.47
L1/CIN4	12218	1572014	0.28
LTR elements	75142	8043361	1.42
BEL/Pao	6293	705505	0.12
Ty1/Copia	3175	392807	0.07
Gypsy/DIRS1	36293	4302376	0.76
Retroviral	15032	1103666	0.19
DNA transposons	578638	61749831	10.87
hobo-Activator	293706	32369155	5.7
Tc1-IS630-Pogo	80564	7567833	1.33
En-Spm	0	0	0
MuDR-IS905	0	0	0
PiggyBac	13600	1089280	0.19
Tourist/Harbinger	44201	5208743	0.92
Other	11367	1081984	0.19
Rolling-circles	35706	2925989	0.52
Unclassified	458433	60021928	10.57
Total interspersed repeats		156832030	27.62
Small RNA	9364	715919	0.13
Satellites	6240	816959	0.14
Simple repeats	238583	10392751	1.83
Low complexity	28658	1630008	0.29

Supplementary Table 3.2. Quality control statistics of the gene models obtained after different rounds of MAKER.

	Round 1	Round 2	Round 3
Number of gene models	9,008	20,327	19,930
Average gene length	11,455	13,741	14,057
AED $\leq$ 0.5	100%	95.50%	94.00%
Complete BUSCO transcripts	58.70%	76.00%	74.90%

Chapter 4. Comparative genomics of six New Zealand fish species: assembly contiguity affects the identification of genic features but not transposable elements

4.1 Abstract

Comparative analysis of whole-genome sequences can provide valuable insights into the evolutionary patterns of diversification and adaptation of species, including the genome contents and the regions under selection. However, such studies are lacking for teleost fishes in New Zealand. To supplement the recently sequenced genome of tarakihi (*Nemadactylus macropterus*), the genomes of five additional percomorph fish species native to New Zealand (king tarakihi (*Nemadactylus* n.sp.), blue moki (*Latridopsis ciliaris*), butterfish (*Odax pullus*), barracouta (*Thyrsites atun*), and kahawai (*Arripis trutta*)) were determined and assembled. Short read Illumina sequencing was used to achieve a minimum of 33x coverage. The genomes ranged in size from 532 Mb (butterfish) to 714 Mb (barracouta), with the number of scaffolds ranging from 58,102 to 150,595, N50 from 10,031 to 30,942 bp, and BUSCO completeness scores from 70.2% to 89.1%. While the proportion of repeat elements was highly correlated with the genome size ($R^2 = 0.97$, $P < 0.01$), most of the metrics for the genic features (e.g. number of exons or intron length) were significantly correlated with assembly contiguity ($|R^2| = 0.79\text{--}0.97$). A phylogenomic tree including eight additional high-quality fish genomes was reconstructed from sequences of gene families that are shared among the sequenced taxa. The radiation of the Percomorpha clade was estimated to have occurred around c. 112 million years ago (mid-Cretaceous), while the Centrarchiformes (tarakihi, king tarakihi, and blue moki) have diverged from the true Perciformes c. 83 million years ago (late Cretaceous). Evidence of positive selection was found in 65 genes in tarakihi and 209 genes in Latridae: the largest portion of these are involved in the ATP binding pathway and the integral structure of membranes. These results and the *de novo* genome sequences can be used to (1) inform future studies on both the strength and shortcomings of scaffold-level assemblies for comparative genomics and (2) provide insights into the evolutionary patterns and processes of genome evolution in bony fishes.

4.2 Introduction

The analyses of DNA sequences have enabled an extensive number of studies to be conducted into the evolutionary history of organisms. These have been used to investigate the underlying evolutionary mechanisms at both the inter-population and inter-species levels. While DNA studies have previously only been carried out on small regions of the genome (e.g. short mitochondrial or nuclear DNA sequences via Sanger sequencing), recent advances in sequencing technologies have greatly improved the acquisition of large, genome-wide DNA sequence data sets. This technical advance enabled the field of comparative genomics to rapidly expand (Hardison, 2003; W. Miller et al., 2004). At its core, comparative genomics utilizes a range of analyses that align contiguous sequences of long stretches of the genome to identify orthologous regions (i.e. sequences that share a common ancestor) and quantify the amount and type of change that has occurred between them (Ellegren, 2008). Genomic sequence similarities and dissimilarities allow inferences to be made about gene functions and structural variation, and how these might influence the evolutionary process. One of the key goals in the field of molecular evolution is to elucidate how selection operates in the genome (R. Nielsen, 2005). By minimizing the stochastic effects seen in short sequences and variation among different genes, it is possible to detect signatures of selection in specific regions of the genome (Ellegren, 2008; Vitti et al., 2013). This is typically achieved by testing for deviations from neutral expectations (J. Zhang et al., 2005).

Genome assembly and annotation allow for the discovery, description, and comparison of orthologous gene coding regions and other genomic features that play key roles in the evolutionary history of organisms. For example, it is possible to detect and characterize repeat elements (RE) in the genome (Lerat, 2018). REs can either consist of tandem repeats (e.g. satellite DNA, also referred to as “simple” repeats) or interspersed repeats (Lerat, 2018; Richard et al., 2008). Interspersed repeats include tRNA genes, genes paralogues, and transposable elements (TE) (Richard et al., 2008). TEs are stretches of DNA that have the capacity to move from one position to another along the chromosomes. There are several types of TEs, classified according to their transposition intermediate (RNA or DNA), their structural features, and their evolutionary origin (Wicker et al., 2007). There is a growing number of evidence that REs make a significant contribution to genome evolution (Biémont,

2010; Biémont & Vieira, 2006) and can have both deleterious and beneficial effects (Chuong et al., 2017). TEs can drive genetic diversification by e.g. altering the protein coding capacity of genes, inducing structural rearrangement, and providing new material on which natural selection can act on. Moreover, both TEs and simple repeats can have an influence on the size of genomes (Lerat, 2018; Sotero-Caio et al., 2017; Z. Yuan et al., 2018). However, very few studies have used genome-wide data to compare the proportion and diversity of repeat elements in several teleost fishes genomes (Brawand et al., 2014; B. Gao et al., 2016; Shao et al., 2019).

There are other genomic features that are seldom explored. This includes “genic features”, such as the number and diversity metrics of genes and their components (e.g. exons, introns, UTR regions). There is still much to discover about these genic features from a comparative genomics perspective. For example, a synthetic review of the genome content of eukaryotes found that while the number of genes increases with genome size, the proportion of coding elements decreases and the proportion of introns increases (T. A. Elliott & Gregory, 2015). It was also suggested that there might be a positive relationship between the exonization of TEs and intron length in animals (Sela et al., 2010). However, few genome assembly studies report these genic features metrics and usually report only the number of genes. T. A. Elliott & Gregory (2015) emphasized that the lack of reporting of metrics like the number and proportion of coding regions and introns is an issue that should be addressed. A recent study that explored the patterns of size, GC content, number of chromosomes and number of genes in fish genomes (Randhawa & Pawar (2021) highlighted that “surprisingly, no study exists on record that has used the WGS annotation data to defines the trends, effects of taxonomic distribution and interrelation of genome attributes”. Yet, such studies have successfully unveiled inter-lineage genomic characteristic patterns in e.g. rodents (Capilla et al., 2016) and birds (Feng et al., 2020).

Current sequencing technologies (e.g. long reads of thousands of bp and scaffolding data like Hi-C) allow for chromosome level and phased assemblies to anchor the totality of genomic information (e.g. number of chromosomes, structural variation) and minimize the risks of incorrect inferences (e.g. unmerged haplotigs, potential scaffolding errors). However,

the required integrity of DNA can be hard to obtain from sampling wild-caught specimens, particularly when the sampling conditions prevent the rapid conservation of the tissue samples in optimal conditions. Rapid degradation of DNA occurs in the first few hours after harvesting a specimen (Oosting et al., 2020). This is particularly problematic for producing long-read sequences (Klingström et al., 2018) and keeping the chromatin integrity for Hi-C data. However, depending on the application, genomes of sufficient quality can be obtained with less sampling constraints using only short-read (100–1000 bp) technology, for which partial degradation of DNA is less of an issue. While genome assemblies based on short reads only will invariably be fragmented to hundreds or thousands of scaffolds, sufficient read coverage can still lead to high-quality contigs and scaffolds that can be used to carry out diverse comparative genomic analyses (Feng et al., 2020; Malmstrøm et al., 2016, 2017).

Although more than a thousand species of fishes occur in New Zealand waters (Roberts et al., 2020) and its marine ecosystem is highly valuable commercially, recreationally and culturally, genome-wide analyses have seldom been used to study its ichthyofauna (Papa, Oosting, et al., 2021). The few exceptions have focused on genomic variation at the intra-species and population level (Catanach et al., 2019; Koot et al., 2021; Oosting, 2021). The main goal of this study was to investigate the patterns of genome-wide variation and evolution of tarakihi (*Nemadactylus macropterus*) and five newly sequenced New Zealand fish species (king tarakihi (*Nemadactylus* n.sp.), blue moki (*Latridopsis ciliaris*), greenbone butterflyfish (*Odax pullus*), barracouta (*Thyrstites atun*), and kahawai (*Arripis trutta*) (Figure 4.1) in a comparative genomic framework. To achieve this, the five new genomes were assembled using Illumina short read sequences. The high-coverage *de novo* genome assembly and annotation of tarakihi produced in Chapter 3 was combined with these five genomes to explore the diversity of repeat elements and genic features. High-quality genomic data from eight other fishes retrieved from Ensembl were then added to identify genes under positive selection in tarakihi and Latridae in a phylogenomic framework.

4.3 Materials and Methods

4.3.1 Tissues collection and DNA extraction

DNA was isolated from five specimens each belonging to a different fish species (king tarakihi, blue moki, butterfish, barracouta, and kahawai) for the respective *de novo* genome assemblies (Figure 4.1). The fish were sampled opportunistically, based on availability, as part of the sampling campaign for the genome assembly and population genomics of tarakihi (Chapter 3 and Chapter 5). The king tarakihi specimen was collected by a commercial fishing trawler around Three King Islands at a depth between 140 and 250 m. It had a standard length of 477 mm, a weight of 1,900 g, and was identified as a female by observation of the gonads. A piece of muscle tissue (c. 2 cm) from the tail was collected and stored in 99% EtOH at -20 °C. The barracouta (70 cm, male) was captured by a recreational fisherman in the Wellington harbour (New Zealand). The three remaining specimens (blue moki, butterfish, and kahawai, sex undetermined) were caught by recreational spear-fishers off Island Bay, around the south coast of Wellington. For these and the barracouta, a piece of the pectoral fin was collected and immersed in 20% DMSO, 0.25 M EDTA, NaCl saturated solution (DESS) and stored at -20 °C. Total genomic DNA was extracted from all tissues with a modified high-salt extraction protocol (Aljanabi & Martinez, 1997) that included an RNase step. The extracted DNA was re-suspended in Tris-EDTA buffer (10 mM Tris-HCl pH 8.0, 0.1 mM EDTA). The purity of DNA samples was assessed using a NanoPhotometer® NP80 (Implen). The high-molecular-weight DNA was visualised using ethidium bromide-staining gel electrophoresis and the quantity of double-strand DNA was measured using a Qubit™ dsDNA BR Assay Kit. See Chapters 3 and 5 for more details on the DNA extraction and quality control methods for Illumina library preparation.

Figure 4.1. The six fish species for which a *de novo* genome assembly has been produced in Chapter 3 (A) and this chapter (B–F). (A) Tarakihi (*Nemadactylus macropterus*: Centrarchiformes, Latridae) (B) King tarakihi (*Nemadactylus* n.sp.: Centrarchiformes, Latridae) (C) Barracouta (*Thyrsites atun*: Scombriformes, Gempylidae) (D) Blue moki (*Latridopsis ciliaris*: Centrarchiformes, Latridae) (E) Kahawai (*Arripis trutta*: Scombriformes, Arripidae) (F) Greenbone butterfish (*Odax pullus*: Labriformes, Labridae). Pictures from Roberts et al. (2015), courtesy of the Museum of New Zealand Te Papa Tongarewa.

4.3.2 Library preparation and sequencing

Prior to sequencing, the genome sizes of the five fish species were estimated based on previously published studies to be approximately 800 Mb (+/- c. 200 Mb), which corresponds to the average genome size of bony fishes (Fan et al., 2020) and the approximate genome size for most percomorphs (Z. Yuan et al., 2018). This estimation was rounded up to a less conservative 1 Gb and was used to determine the required sequencing coverage. Coverage depth needed for short Illumina reads was estimated to be 20–30×, although a plateau in BUSCO completeness is often reached at around 15× (Malmstrøm et al., 2017). Samples were whole-genome sequenced with the aim of obtaining at least 25 Gb per sample.

The purified DNA samples were sent to the Australian Genome Research Facility (AGRF, Melbourne, Australia) for DNA library preparation and sequence determination. Sequencing was performed during several rounds on different lanes depending on the well availability during sequencing of other species (tarakihi and snapper (*Chrysophrys auratus*)) from other studies. Illumina DNA shotgun libraries were prepared following the Nextera DNA FLEX low volume protocol with Nextera DNA Combinatorial Dual Indexes (Illumina) for fragment sizes 300–350 bp. Short reads sequencing of 150 bp paired-end reads was performed on NovaSeq 6000 (Illumina) with NovaSeq 6000 S4 Reagent Kit and NovaSeq XP 4-Lane Kit for 300 cycles. Base calling and quality scoring were performed with RTA3 software v3.3.3. De-multiplexing of the sequencing data was performed using Illumina bcl2fastq pipeline v2.20.0.422.

4.3.3 Quality and contamination filtering

Quality of reads was assessed and visualised at several steps on the filtering pipeline using FastQC v0.11.7 (S. Andrews, 2018) and results were compiled in MultiQC v1.7 (Ewels et al., 2016). Raw paired-end reads were filtered as follows: First, reads that contained adapter contamination were discarded. For this, Trimmomatic v0.39 (Bolger et al., 2014) was used with parameter `NexteraPE-PE.fa:2:30:10` to trim adapters contamination from the reads. seqkit v0.13.2 (W. Shen et al., 2016) was used to keep only the reads that were not trimmed (i.e. reads with a length of 150bp). Then, reads that contained more than 10% uncertain bases (i.e. Ns) and/or for which the proportion of bases with Quality Value ≤ 19 was over 50% were also filtered out using custom bash scripts. Paired-end reads were both discarded if either the forward or the reverse read did not pass the above filtering criteria. Finally, DNA sequence contamination from archaea, bacteria, viruses or human DNA was detected and filtered out using Kraken v2.0.7-beta (Wood et al., 2019) with the MiniKraken2 v2 8GB database (Wood, 2019).

4.3.4 Mitogenome assembly and exclusion

For each fish sample, quality-filtered and contaminant-free Illumina reads were mapped against the complete mitochondrial sequence of an available close species with

Geneious v11.04 (Kearse et al., 2012) by running five iterations of the default mapper set to highest sensitivity. The mitochondrial genomes used as references and their Genbank accession numbers were the following. Barracouta: silver gemfish, *Rexea solandri* (NC_023952.1); blue moki: Peruvian morwong, *Cheilodactylus variegatus* (KP704218.1), butterflyfish: herring cale, *Olisthops cyanomelas* (NC_009061.1), kahawai: same species, *Arripis trutta* (NC_015787.1). For the king tarakihi, the tarakihi (*Nemadactylus macropterus*) mitogenome produced in Halliwell (2019) and reported in Chapter 3 was used as reference. The assembled mitogenomes were then annotated using the MitoAnnotator web interface (Iwasaki et al., 2013). Sequences that were of mitochondrial origin were then filtered out as follows: bwa-kit v0.7.15 (H. Li & Durbin, 2009) was first used to align the Illumina reads to their respective indexed reference mitogenomes with default parameters. All the reads from the alignment that did not map to the mitogenome were then extracted into a new mitochondria-free alignment using SAMtools v1.9 (H. Li et al., 2009) `view` with parameters `-b -f 4`, sorted by name and finally converted back to FASTQ paired-end reads with bedtools v2.27.1 (Quinlan & Hall, 2010).

Once the mitogenomes were assembled, some specific sections were selected for BLAST analysis on NCBI to rule out the possibility of incorrect morphology-based identification of the specimens at collection or DNA sample contamination with another fish species. A c. 550 bp subset of the cytochrome *c* oxidase I region of barracouta, blue moki, and kahawai were blasted against the NCBI nucleotide database. In all cases, the best matches were the correct species with percent identity scores between 99.82% and 100%, consistent with the morphological identification of the species used in each sample. Since there was no COI sequence available for the butterflyfish on NCBI, the 16S RNA region was used instead and showed a 100% identity match. The king tarakihi specimen was not blasted because there were no sequences available for this species on NCBI. The specimen was already identified with high confidence using morphology and based on the phylogenetic position of the sequence data (see Chapter 5).

4.3.5 Genome assembly

Genome assemblies were performed with the Maryland Super-Read Celera Assembler, MaSuRCA v3.4.1 (Zimin et al., 2013; Zimin, Puiu, Luo, et al., 2017). All assemblies were run with recommended parameters for medium-size genomes based on short paired-end Illumina reads only (`PE = pe 350 50`, `EXTEND_JUMP_READS = 0`, `GRAPH_KMER_SIZE = auto`, `USE_LINKING_MATES = 1`, `USE_GRID = 0`, `GRID_BATCH_SIZE = 300000000`, `LHE_COVERAGE=25`, `MEGA_READS_ONE_PASS=0`, `LIMIT_JUMP_COVERAGE = 300`, `CA_PARAMETERS = cgwErrorRate = 0.15`, `CLOSE_GAPS = 1`, `NUM_THREADS = 32`, `SOAP_ASSEMBLY = 0`). The `JF_SIZE` parameter was set as 20^9 for king tarakihi, 13^9 for barracouta and blue moki, and 10^9 for butterflyfish and kahawai.

Assemblies scaffolds were sorted by size using `seqkit v0.13.2` and renamed with simple numbers (i.e. “1” for the longest scaffold, then “2”, etc.) with command `replace -p .+ -r "{nr}"`. Basic contiguity statistics of all assemblies were computed with `bbmap v38.31` (Bushnell, 2018) script `stats.sh`. Genome completeness of each assembly was assessed with the Benchmarking Universal Single-Copy Orthologs (BUSCO) tool v3.0.2 (Simão et al., 2015) and its dependencies (Augustus v3.3.1 (Stanke et al., 2004), NCBI blast+ v2.7.1 (Camacho et al., 2009), hmmer v3.2.1 (Eddy, 2011), and R v3.6.0 (R Core Team, 2020)). Contiguity and completeness of the genome assemblies were graphically visualised with `assembly-stats v17.02` (Challis, 2017) as implemented in the `grpicolli` container (Piccoli, 2021).

4.3.6 Genome annotation

Repetitive elements in the five genomes were identified using the same method and tools as reported in Chapter 3. During the repeat identification step, the Actinopterygii homology-based repeat library produced in Chapter 3 was combined with each of the *de novo* repeat libraries produced separately for each species. Genome annotation was carried out with the MAKER v2.31.10 (Holt & Yandell, 2011) pipeline using the same method, tools and steps detailed in Chapter 3 (including the way short and complex repeats were handled) with a few minor modifications. For each assembly, the first round of MAKER was run on the

unmasked genome using protein and transcriptome data for gene models prediction (`protein2genome=1`, `est2genome=1`) and the GFF of complex repeats only for hard masking (`model_org=simple`). The same protein data obtained from Ensembl release v103 (Kersey et al., 2016) and used in Chapter 3 was used as protein homology in the first round. The transcriptome data used differed depending on the species but was always sourced from a reasonably closely related species (parameter `altest` was set instead of `est`): for the king tarakihi and the blue moki, the repeat-filtered, non-redundant Iso-Seq transcripts produced in Chapter 3 from the tarakihi were used. The Transcriptome Shotgun Assembly (TSA) dataset from the Atlantic mackerel *Scomber scombrus* (Prefix ID: GHRT01, length: 379.6 Mb) was downloaded from NCBI and used for the barracouta and the kahawai. Similarly, the TSA dataset from the hogfish *Lachnolaimus maximus* (Prefix ID: GFXS01, length: 304.2Mb) was used for the butterflyfish. The SNAP v2013.11.29 (Korf, 2004) and Augustus v3.3.1 (Stanke et al., 2004) gene prediction step, the second round of MAKER, the gene model set quality control step, and the functional annotation were carried out exactly like in Chapter 3, except that no filter was applied for the gene models used to train SNAP and that the GFF protein alignment produced in round 1 was used as evidence (`protein2genome=0`) for round 2. The influence of both genome size and genome assembly fragmentation on the annotation of several genomic features was tested using Pearson correlation tests with R v4.02 (R Core Team, 2020) base function `cor.test`. The genomic features included e.g. the proportion of repeat elements, the number of genes or the total intron length. The BUSCO completeness score was used to represent genome contiguity as it is a strong predictor of this metric (Jauhal & Newcomb, 2021).

4.3.7 Gene family identification and phylogenetic tree construction

Protein and DNA coding sequences of each of the six New Zealand species were retrieved from their respective MAKER annotation pipeline. Protein and DNA coding sequences of zebrafish (*Danio rerio*), three-spined stickleback (*Gasterosteus aculeatus*), spotted gar (*Lepisosteus oculatus*), Nile tilapia (*Oreochromis niloticus*), medaka (*Oryzias latipes*), Japanese puffer (*Takifugu rubripes*), green spotted puffer (*Tetraodon nigroviridis*), and southern platyfish (*Xiphophorus maculatus*) (i.e. the same species that were used for the homology-based genome annotation) were downloaded from Ensembl release v103 (Kersey

et al., 2016). Single-copy gene family orthologs between the 14 species were found using OrthoFinder v2.5.2 (Emms & Kelly, 2019) on the protein sequences with default parameters. The protein sequences in each orthogroup were then aligned with MAFFT v7.480 (Katoh & Standley, 2013). Poorly aligned positions and divergent regions were removed with Gblocks v0.91b (Castresana, 2000). All alignments were concatenated in one single alignment using seqkit v0.13.2 and the best substitution model was inferred with ModelTest-NG v0.1.7 (Darriba et al., 2020). The phylogenomic tree of the 14 species was built with MrBayes v3.2.7 (Ronquist et al., 2012) using the JTT+I+G4+F model with 20,000 generations and the default parameters (three heated chains, one cold chain and a burn-in of 25%). After calibrating some of the nodes with the interval values retrieved from TimeTree (www.timetree.org), the divergence times along the tree were calculated with MCMCTree from PAML v4.9 (Z. Yang, 2007; Z. Yang & Rannala, 2006) using the method for approximate likelihood with protein data. The gene family expansions and contractions along the tree branches were calculated with CAFE v4.2.1 (De Bie et al., 2006). The gene families used for the CAFE analysis were based on the orthogroups defined earlier by OrthoFinder, except that orthogroups were not used if they contained species with more than 100 gene copies due to issues with high variance.

4.3.8 Tests for selection

Genes were investigated for evidence of positive selection separately in two lineages: the tarakihi and the Latridae (tarakihi, king tarakihi, and blue moki). Latridae were of particular interest because it is a speciose but taxonomically contentious clade (K. Kimura et al., 2018; Ludt et al., 2019) restricted to the Southern Hemisphere and no other genome assembly currently exists for this family. Selection was detected in the one-to-one orthologues as follows: the protein sequence alignments were converted to nucleotides codon alignments with PAL2NAL v14.1 (Suyama et al., 2006) using the corresponding DNA coding sequences. The alignments were polished with Gblocks v0.91b (parameter `codon`) and converted to phylip format with seqmagick v0.8.0 (<https://github.com/fhcr/seqmagick>). Positively selected genes were detected with CodeML from PAML v4.9 with the branch-site model test, using one of the two selected lineages (tarakihi only or tarakihi, king tarakihi and blue moki) as the foreground branches. The null model assumed that the substitution rates

at nonsynonymous and synonymous sites (dN/dS ratio) for all codons in all branches must be ≤ 1 . The alternative model assumed that the foreground branches included codons evolving at $dN/dS > 1$, indicating the fixation of advantageous mutations. The analysis was applied on each gene (i.e. single-copy orthogroup) DNA alignment separately, using the tree topology obtained earlier with MrBayes. For each gene, the log likelihoods of the alternative ($\ln L_1$) and null models ($\ln L_0$) were compared with a likelihood ratio test ($\Delta LRT = 2(\ln L_1 - \ln L_0)$). The associated p-values were calculated under the chi-square distribution with 1 degree of freedom ($df = 1$). All p-values were then adjusted for multiple testing with the false discovery rate method. Genes were considered positively selected if the adjusted p-value was ≤ 0.05 and if at least one amino-acid site had a Bayes probability $\geq 95\%$ of being positively selected. All Gene Ontology (GO) terms associated with the selected genes were then obtained from the UniProt database (Bateman, 2019). The zebrafish gene accession numbers were used for associated GO terms because this is a well-curated model species.

4.3.9 General bioinformatics tools

See the method section in Chapter 3 for general bioinformatics tools and commands used. All bash and R scripts used for this chapter are available on GitHub in the following repository: https://github.com/yvanpapa/NZfish_genomes_comparisons.

4.4 Results

4.4.1 Genome sequencing and assembly

The DNA sequencing data sets contained 156–197 million reads after they were filtered for quality, contamination and mitochondrial sequences. This total amount corresponded to 23.43–29.52 Gb per species (Supplementary Table 4.1). The sizes of the five final genome assemblies (Figure 4.2, Table 4.1) ranged from 532 Mb (butterfish) to 714 Mb (barracouta). While the genome sizes of bony fishes cover a very wide range of values (0.34 Gb for *Tetraodon nigroviridis* to 2.97 Gb for *Salmo salar*), these results are concordant with the expected sizes for many percomorphs (Z. Yuan et al., 2018). Contiguity and completeness varied among species, with the number of scaffolds ranging from 58,102 (king tarakihi) to 150,595 (barracouta) and the N50s from 10,031 (blue moki) to 30,492 (king tarakihi). Based

on the genome assembly sizes, the lowest read coverage was 33× (barracouta) and the highest was 52× (butterfish). King tarakihi, blue moki and kahawai coverages were all above 40×. Final genome completeness was also variable, with a BUSCO completeness for single-copy Actinopterygii orthologs ranging from 70.20% (kahawai) to 89.10% (king tarakihi). Both contiguity and completeness of the produced genomes were on par with quality metrics for fish assemblies using short Illumina reads (Malmstrøm et al., 2017).

4.4 Results

Figure 4.2. Visualisation of contiguity and completeness of the five new genome assemblies. The contiguity is visualised in a circle representing the full assembly length (532–714 Mb). Lengths of longest scaffolds ranged from 100.9 to 325.8 Kb. There were very few scaffolds (c. 2%) shorter than 100 Kb in length and the GC content was always uniform throughout. See chapter 3 for assembly statistics of the tarakihi. [Continues next page]

Figure 4.2. [Continued]

4.4 Results

Table 4.1. General statistics of the five assemblies produced.

		King tarakihi	Barracouta	Blue moki	Butterfish	Kahawai
(bp)	Genome Assembly					
	Scaffold assembly size	575,828,338	713,861,692	560,782,505	531,871,215	636,431,267
	Reads coverage	40.70×	33.33×	42.93×	52.14×	46.38×
	Total number of	58,102	150,595	108,286	67,538	111,829
	scaffolds					
L50	Longest scaffold (bp)	325,842	158,436	118,632	290,593	100,923
	Scaffold N50 (bp) /	30,492 /	10,964 /	10,031 /	18,273 /	11,325 /
	5,169	17,677	15,899	7,941	16,202	
	Proportion of gap	0.510%	0.452%	0.412%	0.291%	0.344%
	sequences					
contigs	Contigs size (bp)	572,888,966	710,633,565	558,474,240	530,323,606	634,240,427
	Total number of	106,327	231,192	148,983	94,271	154,681
	Contig N50 (bp) / L50	12,009 bp /	5,899 /	6,943 /	11,807 /	7,607 /
	13,187	33,702	23,019	12,336	24,285	
	A / T / G / C bases (%)	28.30 /	29.59% /	28.39% /	29.12% /	30.33% /
BUSCOs		28.24 / 21.72 / 21.74	29.42% / 20.47% / 20.52%	28.33% / 21.63% / 21.65%	29.03% / 20.92% / 20.93%	30.27% / 19.69% / 19.71%
	Genome Completeness (4,584 Actinopterygii BUSCOs)*					
	Complete BUSCOs	89.1%	72.4%	71.90%	75.10%	71.50%
	Complete single-copy	87.2%	71.2%	70.40%	73.40%	70.20%
	Complete duplicated	1.9%	1.2%	1.50%	1.70%	1.30%
genes	Fragmented BUSCOs	6.7%	11.5%	18.90%	15.90%	18.20%
	Missing BUSCOs	4.2%	16.1%	9.20%	9.00%	10.30%
	Genome annotation					
0.5	No. of protein-coding	22,258	24,378	23,804	24,816	22,840
	% genes with AED <	92%	86%	87%	85%	96%

No. funct. annotated	21,650	23,840	23,269	24,266	22,570
proteins					
Mean gene length	7,664	4,371	3,808	4,163	5,352
(bp)					
Mean exon length	160	170	173	169	166
(bp)					
Mean intron length	861	794	651	757	737
(bp)					

Note: See Chapter 3 for assembly statistics of the tarakihi. *Barracouta BUSCO completeness was assessed with Actinopterygii orthologs database v10 (3,640 BUSCOs) instead of v9 (4,584 BUSCOs) because of version software compatibility at the time of the analysis. The results from both orthologs databases are still comparable.

4.4.2 Repetitive elements and genes

The total proportion of repetitive elements in the six genomes (Figure 4.3, Table 4.2, Supplementary Table 4.2) varied from 24.83% (butterfish) to 39.12% (barracouta). The proportion of repeat elements in the king tarakihi and the blue moki (c. 30%) is similar to the proportion obtained for the tarakihi (Figure 4.3, Table 4.2). These three species are from the same family (Latridae). The proportion of repeat elements was highly correlated to the genome size (Figure 4.3) (see Discussion).

Figure 4.3. Correlation between genome sizes and the proportion of repetitive elements in the genome. Tarakihi, king tarakihi and blue moki are from the same family (Latridae).

In all assemblies, the largest classified repeat element proportion was always DNA transposons (6.92%–14.85%), followed by long interspersed nuclear element (LINE) retrotransposons (3.26%–5.83%) (Figure 4.4, Supplementary Table 4.2). The proportion of repeat elements that were not classified in the Dfam/RepBase databases was 10.53%–13.49%. Long terminal repeat (LTR) retrotransposons (1.29%–1.82%), simple interspersed nuclear element (SINE, 0.45%–0.50%) and simple repeats (1.18%–1.71%) represented a smaller fraction in comparison.

Figure 4.4. Proportions of the main classes of repeat elements in the genomes. Repeat elements are either simple repeats or transposable elements (TE), which include DNA transposons (DNA), long terminal repeat (LTR) retrotransposons, and non-LTR retrotransposons (long and short interspersed nuclear elements, LINE and SINE).

To assess the proportion of repeat elements in more detail, the ten most frequent repeat families were identified for each species (Figure 4.5, Supplementary Figure 4.1). The ten most frequent repeat families were always the same in tarakihi, king tarakihi, blue moki, barracouta and kahawai: they were unclassified elements, simple repeats, DNA hAT-Ac, LINE L2, DNA hAT-Tip100, DNA TcMar-Tc1, LINE Rex-Babar, unknown DNA elements, DNA PIF-Harbinger, and DNA hAT-Charlie (not always in the same order). For butterfish, the ten most frequent repeat families were almost all the same as the other species, with two exceptions: LTR Ngaro and rolling-circle helitrons were part of the ten most frequent repeat elements in that species, instead of the unknown DNA elements and DNA hAT-Tip100.

4.4 Results

Figure 4.5. Proportions of the most represented families of repeat elements in the genomes. Repeat elements families are sorted vertically based on their abundance in tarakihi. “Other” includes all the families that are not in the top ten of the most abundant RE in at least one species. See Supplementary Figure 4.1 for more details on the repeat element included in this group.

The number of genes detected by the annotation pipeline in the five new genome assemblies ranged from 22,258 (king tarakihi) to 24,816 (butterfish), which is higher than the 20,169 genes found in tarakihi (Table 4.2). The mean number of exons per gene ranged from five (blue moki, butterfish, and barracouta) to eight (king tarakihi), both lower mean values when compared to the eleven found for tarakihi (Table 4.2). Many of the other gene feature metrics were comparable across species, with the largest deviations seen for tarakihi and king tarakihi (Table 4.2), see Discussion.

Table 4.2. Gene features of the six fish genome assemblies

	Tarakihi	King tarakihi	Blue moki	Butterfish	Barracouta	Kahawai
Genome length (bp)	567,902,715	575,828,338	560,782,505	531,871,215	713,861,692	636,431,267
Number of genes / mRNA / CDS	20,169	22,258	23,804	24,816	24,378	22,840
Number of exons	217,298	185,789	129,362	131,762	130,618	154,018
Number of introns	197,129	163,531	105,558	106,946	106,240	131,178
Mean exons per gene / mRNA	11	8	5	5	5	7
Total gene / mRNA length	278,993,619	170,627,661	90,669,335	103,342,350	106,564,991	122,281,322
Total exon length	49,834,014	29,734,201	22,409,584	22,298,859	22,288,118	25,630,404
Total CDS pieces length	33,362,898	29,578,314	21,700,488	21,984,588	22,028,601	25,377,594
Total 5' UTR length	1,942,114	122,107	80,708	73,260	79,275	114,997
Total 3' UTR length	14,529,002	33,780	628,388	241,011	180,242	137,813
Total intron length	229,356,734	141,056,991	68,365,309	81,150,437	84,383,113	96,782,096
Mean gene / mRNA length	13,832	7,665	3,808	4,164	4,371	5,353
Mean total CDS length in genes	1,654	1,328	911	885	903	1,111
Mean exon length	229	160	173	169	170	166
Mean 5' UTR length	220	37	56	52	56	33
Mean 3' UTR length	1,465	375	834	554	541	443
Mean CDS pieces length	162	159	172	168	170	165
Mean intron length	1,163	862	647	758	794	737
Proportion of repeat elements in genome (%)	30.45	30.73	29.66	24.83	39.12	34
Proportion of exons in genome (%)	8.78	5.16	4.00	4.19	3.12	4.03
Proportion of introns in genome (%)	40.39	24.50	12.19	15.26	11.82	15.21
Proportion of genes in genome (%)	49.16	29.66	16.19	19.45	14.94	19.23

Note: Gene = Exons + Introns. UTR = untranslated region. CDS = coding sequences = exon - UTR regions.

4.4.3 Genomic comparative analysis

Gene family clustering of the 14 fish species was able to assign 96.3% of all the genes in orthogroups. A total of 19,874 orthogroups were found, among which 1,481 were single-copy orthogroups that were used for phylogenetic reconstruction (Figure 4.6). There were 11 orthogroups only specific to tarakihi, while the number for the five other New Zealand species varied from 4 to 71 (Supplementary Table 4.3). Genes contained only in the tarakihi-specific orthogroups were: coxsackievirus and adenovirus receptor (four copies), endonuclease-reverse transcriptase (three copies), poly(rC)-binding protein, small G protein signalling modulator 1, paired box protein Pax-7, glucagon-1-like, sulfhydryl oxidase 2 (two copies) and a few unannotated genes. Tarakihi, spotted gar, and green spotted puffer were the only three species with a negative average gene expansion, meaning they displayed a net loss of genes per gene family (Supplementary Table 4.4).

Overall the phylogenetic tree (Figure 4.6) was consistent with recently reported molecular phylogenies of Percomorpha (Betancur-R et al., 2017; Sanciangco et al., 2016; Shou & Han, 2021). All nodes were strongly supported (posterior probability = 100%) and orders were retrieved as expected. The Latridae family (which include tarakihi, king tarakihi, and blue moki), representing the order Centrarchiformes, was placed as a sister clade of “true” Perciformes (stickleback), which is consistent with Betancur-R et al. (2017). There was one difference with the phylogenies from Betancur-R et al. (2017), Sanciangco et al. (2016) and Shou & Han (2021): the positions of the Pelagiaria (Scombriformes: Barracouta and Kahawai) and the Ovalentaria clades (which include Cyprinodontiformes, Beloniformes, and Cichliformes) are swapped. In Eupercaria, the Tetraodontiformes (pufferfishes) were placed as a sister clade of Labriformes (wrasses: butterfly) in our phylogeny, which is consistent with Shou & Han (2021) but not with Betancur-R et al. (2017) and Sanciangco et al. (2016). These studies placed pufferfishes as a sister clade of Perciformes + Centrarchiformes instead. This is of particular interest because all the species newly assembled here belong to the clade Percomorpha (Percomorphaceae *sensu* Betancur-R et al. (2017)). This taxonomic clade is historically considered a “bush at the top” of the fish tree of life that includes around 55% of extant bony fish species (Sanciangco et al., 2016). Consequently, both the phylogenetic inter-relationships between the large clades cited above (Ovalentaria, Pelagiaria (Scombriformes),

and Eupercaria) and the position of Tetraodontiformes in Eupercaria are still contentious, with nodes supports being relatively low even in recent studies (Betancur-R et al., 2017; Sanciangco et al., 2016).

According to the present phylogeny, the radiation of Percomorpha has occurred c. 112 million years ago, in the mid-Cretaceous (Figure 4.6). The Centrarchiformes (tarakihi, king tarakihi, and blue moki) have then diverged from the true Perciformes c. 83 million years ago, in the late Cretaceous. Tarakihi and king tarakihi are estimated to have diverged 4.7 million years ago. While still relatively recent, this is older than the minimum time since divergence of 0.3–0.8 million years estimated with nucleotide divergence rate of the mitochondrial control region (Papa, Halliwell, et al., 2021). The confidence interval is, however, comparatively large (95% CI = 1.77–10.12 MYA), and the “true” time since divergence could be situated somewhere in the lower end of that interval.

4.4 Results

Figure 4.6. Phylogenomic tree of the six New Zealand fish species newly assembled (in bold) and eight other fish species. A dot on a node indicates that the clade is monophyletic *sensu* Betancur-R et al. (2017). The red dots in particular have been used for node age calibration with TimeTree. Black numbers indicate the estimated divergence time with 95% confidence intervals. Blue and red numbers represent respectively the expansion and the contraction of gene families along the phylogeny. Green numbers are the posterior probabilities in percentage and indicate the support values of the nodes. See the “pictures credits” section below for fish pictures attributions.

A total of 65 genes in the tarakihi genome showed evidence consistent with positive selection (Supplementary Table 4.5). These 65 genes under selection were associated with 295 GO terms for biological processes. The GO terms that appeared the most often (i.e. more than two times) included terms related to function: ATP binding, metal ion binding, regulation of transcription by RNA polymerase II, integral component of membrane, RNA polymerase II cis-regulatory region sequence-specific DNA binding, RNA binding, zinc ion binding, DNA-binding transcription factor activity (RNA polymerase II-specific), GTPase activator activity, signal transduction, microtubule binding, hydrolase activity, protein dimerization activity, and cell division (respectively). They also included terms related to location: cytoplasm, nucleus, cytosol, endosome, Golgi apparatus (respectively) (Figure 4.7).

A total of 209 genes showed evidence of positive selection in the Latridae lineage, which included tarakihi, king tarakihi and blue moki (Supplementary Table 4.6). They included almost all of the genes detected in tarakihi except for six: cell division cycle 25B, high mobility group 20A, nuclear receptor subfamily 2 group C member 2, pater pan homolog, ribosomal protein L27, and zgc:85936 were positively selected in the tarakihi only (Supplementary Table 4.5). The most frequent GO terms for selected genes functions in Latridae were integral component of membrane, ATP binding, metal ion binding, RNA binding, regulation of transcription by RNA polymerase II, zinc ion binding, RNA polymerase II cis-regulatory region sequence-specific DNA binding, DNA-binding transcription factor activity (RNA polymerase II-specific), respectively. The most frequent location terms were cytoplasm, nucleus, cytosol, plasma membrane, and membrane, respectively (Figure 4.8).

4.4 Results

Figure 4.7. Frequency of gene functions and locations according to GO terms for genes under positive selection in tarakihi. Only terms associated with three genes or more are shown.

Figure 4.8. Frequency of gene functions and locations according to GO terms for genes under positive selection in Latridae. Only terms associated with three genes or more are shown.

4.5 Discussion

This study provided the first genome assemblies for five New Zealand fish species. This contribution fills an important gap in the genomic resources for the South Pacific ichthyofauna. The analyses of the genomes enabled a range of genomic features to be identified and described, in particular the diversity of repeat elements and the most represented repeat families. The phylogenetic analysis of the combined data set of 13 species provided insights into the evolutionary history of fish and the effects of selection on fish genomes.

4.5.1 Repetitive elements in fish genomes

DNA transposons represented the highest proportion of repeat elements in all six genome assemblies, making up more than 10% of the total genome sequence in all species except butterfish (Figure 4.4, Supplementary Table 4.2). DNA transposons differ from retrotransposons in that they do not rely on an RNA intermediate for moving within the genome (Sotero-Caio et al., 2017; Wicker et al., 2007). This finding is consistent with other comparative analyses that reported DNA transposons as the most abundant class of repeat elements in most teleost fish species (Brawand et al., 2014; Chalopin et al., 2015; B. Gao et al., 2016; Shao et al., 2019). While all six species investigated here belong to the large Percomorpha clade, one of the most comprehensive studies in terms of species numbers (35 Actinopterygii from 14 orders) showed that while DNA transposons are often the dominant RE class in ray-finned fishes, including Percomorpha, they can in some cases be outnumbered by LINE or LTR retrotransposons (Shao et al., 2019). This was the case for the three-spined stickleback (a “true” Perciform, which is the sister taxon of Centrarchiformes that include tarakihi, king tarakihi and blue moki) and other Percomorpha with reduced genome size like Japanese puffer and green spotted puffer. Also consistent with our results, short interspersed nuclear element (SINE) transposons usually represent a much smaller fraction of the genome in fishes (Shao et al., 2019) compared to e.g. mammals (Sotero-Caio et al., 2017).

Observing the frequencies of RE at the intra-class level showed that all six species, except butterfish, shared the same most common TEs (Figure 4.5). These were comprised of three TE families from the hAT superfamily of DNA transposons (Ac, Tip100, and Charlie), two

other DNA transposons superfamilies (Tc1/Mariner, mostly represented by the Tc1 family, and PIF/Harbinger), and two superfamilies of LINE retrotransposons (L2 and Rex/Babar). Albeit not always in the exact same order, the relative proportion of each of these families were similar in tarakihi, king tarakihi, blue moki, barracouta and kahawai (Figure 4.5). hAT, Tc1/mariner and PIF/Harbinger are “cut-and-paste” DNA transposons. “Cut-and-paste” DNA TEs move among genomic locations via excision and insertion using a transposase (Y. W. Yuan & Wessler, 2011). L2 and Rex/Babar are long interspersed nuclear elements (LINEs). LINEs are retrotransposons that contain a reverse transcriptase gene but lack long terminal repeats (Finnegan, 2012). These results are consistent with a former extensive comparative study that found that Tc1/Mariner, hAT, and L2 superfamilies are among the most important TE in fish genomes both in terms of genome proportion and representation across taxa (Shao et al., 2019), although they also included L1 and Gypsy elements in that list. Interestingly, butterflyfish was the only one of the six species that showed a different trend in most represented RE elements, with an over-representation of Ngaro elements (a type of Long Terminal Repeats retrotransposon) compared to e.g. hat-Tip100 DNA transposons (Figure 4.5). While Ngaro elements do not seem to be especially dominant in fishes overall in comparison to other LTRs like e.g. Gypsy elements (Shao et al., 2019), they were found to differ highly in terms of abundance and diversity among teleost lineages (B. Gao et al., 2016). Indeed this trend is also observed in these six fish species. Finally, while the helitron elements were among the ten most common TE families in butterflyfish only, they were not especially more abundant in comparison with the five other species, where they also represent a sizeable proportion of all TE (Figure 4.5). Helitrons are a specific kind of DNA transposon that do not move via “cut-and-paste” but rather use a “rolling-circle” replication mechanism (Wicker et al., 2007).

A strong and significant relationship between the proportion of repeat elements and the genome size was found (Figure 4.3). This observation was especially robust since it did not depend on the quality of the genome assembly (Figure 4.9) (see below). This result provides additional evidence for the strong relationship between the genome size and the abundance of REs that has already been observed in a broad number of fish taxa (B. Gao et al., 2016; Shao et al., 2019; Z. Yuan et al., 2018). Current evidence suggests that the genome size of ray-finned fishes is more dependent on repeat elements in general than for tetrapods

(Chalopin et al., 2015; Sotero-Caio et al., 2017). Specifically, it has been suggested that DNA transposons could be significant drivers of the genome size differentiation in teleost fish, whereas LTR and non-LTR retrotransposons seem to dominate the genome expansion of most reptiles and mammals (B. Gao et al., 2016).

4.5.2 Impact of genome size and fragmentation on genome annotation quality

Genomic features were detected through the annotation pipeline in the newly assembled fish genomes (Table 4.2). One of the goals of this study was to explore the genomic features (i.e. number and proportion of both repeat elements and genes and their constituents) in more detail to infer trends across species. However, it was necessary to test for potential biases induced by the genome assembly quality beforehand for the interpretation of the results to be robust. We found that all genomic features explored, with the notable exception of the proportion of repeat elements, correlated strongly with genome completeness (Figure 4.9) as opposed to e.g. genome size (Figure 4.10). Higher fragmentation (i.e. lower contiguity) of the genome assembly reduces the length of genes detected and reduces both the number and length of the genic components, including exons, introns and UTR regions (Figure 4.9). It is possible that a higher fragmentation creates a bias toward the detection of smaller genomic features. However, the number of genes decreased when the assembly contiguity was higher. This could be due to several reasons. First, the higher fragmentation of the genomes might increase the quantity of erroneously unmerged gene duplicates. This result could also be due to differences in the annotation pipeline. The tarakihi and the king tarakihi genomes, which also happen to be the two most highly continuous assemblies, have been annotated using full-length RNA reads (Iso-Seq) instead of short RNA-seq data. The RNA reads used for the tarakihi were also the only ones coming from the same species as opposed to a close species: this led to some slight changes in the annotation pipeline parameters for the five other species compared to tarakihi, which might have influenced the gene discovery process by making it less stringent.

Figure 4.9. Correlation between BUSCO completeness and proportion, number and length of genomic features annotated in the six genome assemblies, with corresponding Pearson correlation coefficient (R^2) and p-value (P). Grey area is the 95% confidence interval. Red asterisks indicate significance, with $P \leq 0.1$ (*), 0.05 (**), and 0.01 (***).

4.5 Discussion

Figure 4.10. Correlation between genome size and proportion, number and length of genomic features annotated in the six genome assemblies, with corresponding Pearson correlation coefficient (R^2) and p-value (P). Grey area is the 95% confidence interval. Red asterisks indicate significance, with $P \leq 0.1$ (*), 0.05 (**), and 0.01 (***).

4.5.3 Positive selection in ATP binding genes

The most frequent GO term for function in tarakihi was “ATP binding”. The term was associated with nine of the positively selected genes detected in tarakihi, which means that more than 10% of its 65 positively selected genes are related to that function. It was also the most frequent function of the selected genes for Latridae just after “integral component of membrane”, although in most cases the two functions were related to the same genes. Genes in the ATP binding pathway provide energy for the active transport of ions across cell membranes (Melkikh & Seleznev, 2012). Genes in the ATP binding pathway were the most common genes found to be under positive selection in a former study based on molecular evolution rates of four model fish species (Steinke et al., 2006) and were also found to be positively selected in the largemouth bass (Sun et al., 2021), another fish of the order Centrarchiforme. Mutations in genes involved in ATP binding, especially the ATP-Binding Cassette transporter genes, are known to be associated with a wide variety of pathologic disorders (Dean & Annilo, 2005). Taken together, these results may imply that ATP binding genes are submitted to both strong purifying selection against deleterious mutations and positive selection favouring beneficial mutations. Further research will be needed to assess if the propensity of these genes to fix novel mutations is related to the speciation process, and if it is a driver or a result of this process.

4.6 Conclusions

The genome sequence of five New Zealand bony fish species, six including the tarakihi, provided several results relevant to shed light on the genomic features associated with the evolution of these species. In addition to their utility in a comparative genomic study, the genome assemblies can be used as references for population genomics studies. While it always depends on the question addressed, if the resources are available, the aim for these types of applications should ideally be to obtain a sex-specific, phased, chromosome-level genome with transcriptome-informed annotation. In these cases, long-read DNA sequencing (e.g. PacBio Hi-Fi reads, ONT reads), supplemented with scaffolding data (e.g. Hi-C), isoform data (e.g. Iso-Seq) and sometimes short high-quality reads (e.g. Illumina short reads, DNBSeg) should be used. However, given the decreasing costs and fewer sampling constraints, the

present pipeline could be an efficient way to easily assemble the genomes of more fish species, especially the ones for which funding is less available because of lower commercial importance.

The study also provided a molecular phylogeny based on whole genomes for the contentious clade Percomorphaceae with strong node supports. Adding more species to the tree, both with newly assembled genomes and from genomic data already available will surely provide further insight into the complicated evolutionary history of this clade that represents a considerable portion of living vertebrates. Finally, we found evidence of positive selection on genes involved in the ATP binding pathway, consistent with other studies of fish. Further studies of the molecular and biological processes involved in these pathways and their putative association with environmental factors will be needed to better apprehend the benefits of maintained novel mutations in these genes.

4.7 Pictures credits

All fish pictures used in the phylogenomic tree are either courtesy of the Museum of New Zealand Te Papa Tongarewa or derived from work in the public domain, except for the following:

Lepisosteus oculatus photo by Brian Gratwicke:

https://commons.wikimedia.org/wiki/File:Lepisosteus_oculatus1.jpg / [CC-BY-2.5](#)

Oryzias latipes photo by NOZO:

<https://upload.wikimedia.org/wikipedia/commons/4/49/Nihonmedaka.jpg> / [CC-BY-3.0](#)

Xiphophorus maculatus photo by vxixiv:

<https://www.flickr.com/photos/21630815@N06/3406554064/> / [CC-BY-2.0](#)

Takifugu rubripes photo by DataBase Center for Life Science (DBCLS):

<https://doi.org/10.7875/togopic.2012.10> / [CC-BY-4.0](#)

4.8 Supplementary Material

Supplementary Figure 4.1. Proportions of the most represented families of repeat elements in the genomes. “Other” includes all the families that are not in the top ten of the most abundant RE in at least one species. Repeat elements families are sorted vertically based on their abundance in the tarakihi.

4.8 Supplementary Material

Supplementary Figure 4.2. Correlation between genome size and proportion, number and length of genomic features annotated in the genome assemblies without including tarakihi, with corresponding Pearson correlation coefficient (R^2) and p-value (P). Grey area is the 95% confidence interval. Red asterisks indicate significance, with $P \leq 0.1$ (*), 0.05 (**), and 0.01 (***).

Supplementary Table 4.1. Summary of number of reads and bases obtained at several steps of the quality filtering pipelines.

		Raw reads	Quality-filtered reads	Uncontaminated reads	Mitochondrial reads (%)	Final Illumina PE reads
King tarakihi	No. reads	188,470,220	170,505,602	156,465,038	0.14997344	156,230,382
	No. bases	28,270,533,000	25,575,840,300	23,469,755,700	0.14997344	23,434,557,300
Barracouta	No. reads	205,858,992	184,329,218	158,680,852	0.034627997	158,625,904
	No. bases	30,878,848,800	27,649,382,700	23,802,127,800	0.034627997	23,793,885,600
Blue moki	No. reads	189,522,608	174,855,726	160,748,294	0.151186674	160,505,264
	No. bases	28,428,391,200	26,228,358,900	24,112,244,100	0.151186674	24,075,789,600
Butterfish	No. reads	219,182,928	198,607,764	186,745,068	0.983231322	184,908,932
	No. bases	32,877,439,200	29,791,164,600	28,011,760,200	0.983231322	27,736,339,800
Kahawai	No. reads	232,301,418	211,988,224	197,198,048	0.201719035	196,800,262
	No. bases	34,845,212,700	31,798,233,600	29,579,707,200	0.201719035	29,520,039,300

Supplementary Table 4.2. Main classes and proportions of repeat elements detected in the five new assembled genomes

	King tarakihi	Barracouta	Blue moki	Butterfish	Kahawai
RNA-mediated class I transposons (retrotransposons)					
SINEs	33677 0.45%	40399 0.50%	31713 0.44%	32533 0.43%	36814 0.48%
Penelope	8207 0.14%	14509 0.26%	8222 0.14%	9252 0.14%	11805 0.22%
LINEs	217366 4.08%	303643 5.83%	198643 3.69%	180766 3.26%	238505 4.80%
LTR elements	76131 1.38%	96989 1.64%	71646 1.29%	110681 1.82%	74939 1.36%
RNA-independent class II DNA transposons					
DNA transposons	591041 11.05%	874958 14.85%	568932 10.81%	388300 6.92%	692761 12.82%
Rolling-circles	34869 0.46%	46474 0.64%	32087 0.42%	24792 0.36%	37629 0.60%
Unclassified transposable elements	502186 11.06%	701196 13.49%	490965 11.17%	466526 10.53%	571345 11.98%
Small RNA	8989 0.11%	12665 0.14%	8695 0.11%	7875 0.10%	10716 0.13%
Satellites	8263 0.51%	9219 0.20%	5925 0.11%	4487 0.08%	6646 0.12%
Simple repeat	211399 1.44%	273196 1.71%	205792 1.45%	159110 1.19%	186620 1.18%
Low complexity	27126 0.25%	30005 0.21%	23797 0.23%	20738 0.20%	26541 0.21%
Total REs	30.73%	39.12%	29.66%	24.83%	33.59%

Supplementary Table 4.3. Statistics of orthogroups per species

	Tarakihi	King tarakihi	Barracouta	Blue moki	Butterfish	Kahawai	Zebrafish	Stickleback	Spotted gar	Nile tilapia	Medaka	Takifugu	Tetraodon	Platyfish
No. of genes	20169	22258	24378	23804	24816	22840	30313	20787	18341	28189	23622	21411	19602	23774
No. of genes in orthogroups	19745	21479	22627	22369	22697	22163	29333	20112	17908	27621	22889	20845	19342	23226
No. of unassigned genes	424	779	1751	1435	2119	677	980	675	433	568	733	566	260	548
% of genes in orthogroups	97.9	96.5	92.8	94	91.5	97	96.8	96.8	97.6	98	96.9	97.4	98.7	97.7
% of unassigned genes	2.1	3.5	7.2	6	8.5	3	3.2	3.2	2.4	2	3.1	2.6	1.3	2.3
No. of orthogroups containing species	14126	14430	14458	14644	14404	14355	14053	13112	13060	14441	14037	13272	12559	14247
% of orthogroups containing species	71.1	72.6	72.7	73.7	72.5	72.2	70.7	66	65.7	72.7	70.6	66.8	63.2	71.7
No. of species-specific orthogroups	11	4	52	39	71	12	318	29	73	207	91	66	33	68
No. of genes in species-specific orthogroups	25	8	112	81	149	26	2463	278	352	1701	772	256	116	346
% of genes in species-specific orthogroups	0.1	0	0.5	0.3	0.6	0.1	8.1	1.3	1.9	6	3.3	1.2	0.6	1.5

Supplementary Table 4.4. Statistics of gene family expansions and contractions

Species	Expanded fams	Genes gained	genes/expansion	Contracted fams	Genes lost	genes/contraction	No change	Avg. Expansion
Barracouta	4226 (4226)	7008	1.66	3637 (3637)	4240	1.17	10930	0.147289
Blue moki	3855 (3855)	6428	1.67	3352 (3352)	3885	1.16	11586	0.135316
Butterfish	4126 (4126)	7058	1.71	3656 (3656)	4256	1.16	11011	0.149098
Kahawai	3240 (3240)	5804	1.79	2950 (2950)	3413	1.16	12603	0.127228
King tarakihi	2726 (2726)	5119	1.88	2966 (2966)	3394	1.14	13101	0.0917895
Medaka	1822 (1822)	5799	3.18	3207 (3207)	3492	1.09	13764	0.122758
Nile tilapia	2157 (1672)	8586	3.98	2711 (188)	2935	1.08	13925	0.300697
Platyfish	1904 (1904)	6172	3.24	2875 (2875)	3111	1.08	14014	0.16288
Spotted gar	2476 (787)	4427	1.79	5890 (2)	5898	1	10427	-0.0782738
Stickleback	1593 (1593)	4695	2.95	4146 (4146)	4610	1.11	13054	0.00452296
Takifugu	1872 (1872)	5241	2.8	4013 (4013)	4511	1.12	12908	0.0388443
Tarakihi	1788 (1788)	3757	2.1	3326 (3326)	3783	1.14	13679	-0.00138349
Tetraodon	1915 (1915)	4684	2.45	4659 (4659)	5225	1.12	12219	-0.0287873
Zebrafish	5780 (2419)	13130	2.27	5046 (0)	5046	1	7967	0.43016

Supplementary Table 4.5. Genes positively selected in the tarakihi [Continues next page]

Gene	Protein name
abhd11	abhydrolase domain containing 11
acmsd	aminocarboxymuconate semialdehyde decarboxylase
apex1	APEX nuclease (multifunctional DNA repair enzyme) 1
arhgap24	Rho GTPase activating protein 24
arl13a	ADP-ribosylation factor-like 13A
cabin1	calcineurin binding protein 1
cdc25b*	cell division cycle 25B
cpeb1b	cytoplasmic polyadenylation element binding protein 1b
cs	citrate synthase
ddx46	DEAD (Asp-Glu-Ala-Asp) box polypeptide 46
dnajc22	DnaJ (Hsp40) homolog, subfamily C, member 22
dop1b	DOP1 leucine zipper like protein B
e2f1	E2F transcription factor 1
endou2	endonuclease, polyU-specific 2
fbxo40.1	F-box protein 40, tandem duplicate 1
fsd1	fibronectin type III and SPRY domain containing 1
gclc	glutamate-cysteine ligase, catalytic subunit
gins2	GIN5 complex subunit 2
gk5	glycerol kinase 5
gkap1	G kinase anchoring protein 1
glmna	glomulin, FKBP associated protein a
hddc2	HD domain containing 2
hectd3	HECT domain containing 3
hgs	hepatocyte growth factor-regulated tyrosine kinase substrate
hltf	helicase-like transcription factor
hmg20a*	high mobility group 20A
itgb4	integrin, beta 4
kif4	kinesin family member 4
letm2	leucine zipper-EF-hand containing transmembrane protein 2
lonp2	lon peptidase 2, peroxisomal
lrrn3b	leucine rich repeat neuronal 3b
mapk15	mitogen-activated protein kinase 15
mlx	MAX dimerization protein MLX
mrpl41	mitochondrial ribosomal protein L41
nr2c2*	nuclear receptor subfamily 2, group C, member 2
paip1	poly(A) binding protein interacting protein 1
pik3c3	phosphatidylinositol 3-kinase, catalytic subunit type 3
pik3r5	phosphoinositide-3-kinase, regulatory subunit 5

4.8 Supplementary Material

ppan*	peter pan homolog
ppp1r12a	protein phosphatase 1, regulatory subunit 12A
psmc5	proteasome 26S subunit, ATPase 5
qars1	glutaminyl-tRNA synthetase 1
rab24	RAB24, member RAS oncogene family
rad9b	RAD9 checkpoint clamp component B
rfng	RFNG O-fucosylpeptide 3-beta-N-acetylglucosaminyltransferase
rfxank	regulatory factor X-associated ankyrin-containing protein
rgs14a	regulator of G protein signaling 14a
rmdn1	regulator of microtubule dynamics 1
rpl27*	ribosomal protein L27
rplp0	ribosomal protein, large, P0
snrnp40	small nuclear ribonucleoprotein 40 (U5)
sreb2	sterol regulatory element binding transcription factor 2
supt7l	SPT7 like, STAGA complex gamma subunit
taf5	TAF5 RNA polymerase II, TATA box binding protein (TBP)-associated factor
tbcd	tubulin folding cofactor D
thoc5	THO complex 5
thtpa	thiamine triphosphatase
trip11	thyroid hormone receptor interactor 11
unc45b	unc-45 myosin chaperone B
wdr13	WD repeat domain 13
zfand4	zinc finger, AN1-type domain 4
zgc:152830	zgc:152830
zgc:162698	zgc:162698
zgc:85936*	zgc:85936
zgc:92518	zgc:92518

Note: Asterisks (*) indicate genes that are **not** also detected as selected in Latridae

Supplementary Table 4.6. Genes positively selected in Latridae [Continues next pages]

Gene	Protein name
abhd11	abhydrolase domain containing 11
abtb1	ankyrin repeat and BTB (POZ) domain containing 1
acer1	alkaline ceramidase 1
acer3	alkaline ceramidase 3
acmsd	aminocarboxymuconate semialdehyde decarboxylase
aff4	AF4/FMR2 family, member 4
alkbh7	alkB homolog 7
ankzf1	ankyrin repeat and zinc finger peptidyl tRNA hydrolase 1
apex1	APEX nuclease (multifunctional DNA repair enzyme) 1
arhgap24	Rho GTPase activating protein 24
arhgap32b	Rho GTPase activating protein 32b
arl13a	ADP-ribosylation factor-like 13A
atp5f1c	ATP synthase F1 subunit gamma
atp6v1d	ATPase H ⁺ transporting V1 subunit D
aup1	AUP1 lipid droplet regulating VLDL assembly factor
babam1	BRISC and BRCA1 A complex member 1
bspry	B-box and SPRY domain containing
c1qbp	complement component 1, q subcomponent binding protein
C3H17orf75	zgc:153240
cabin1	calcineurin binding protein 1
CACFD1	si:ch73-209e20.5
capn10	calpain 10
cars2	cysteinyl-tRNA synthetase 2, mitochondrial
ccdc84	coiled-coil domain containing 84
cdadc1	cytidine and dCMP deaminase domain containing 1
cdk10	cyclin-dependent kinase 10
cdkn2aip	CDKN2A interacting protein
cfap58	cilia and flagella associated protein 58
cmtr1	cap methyltransferase 1
cmtr2	cap methyltransferase 2
comm5	COMM domain containing 5
cpeb1b	cytoplasmic polyadenylation element binding protein 1b
cs	citrate synthase
ddx46	DEAD (Asp-Glu-Ala-Asp) box polypeptide 46
dhh	desert hedgehog signaling molecule
dhx29	DEAH (Asp-Glu-Ala-His) box polypeptide 29
dhx33	DEAH (Asp-Glu-Ala-His) box polypeptide 33
dhx58	DEXH (Asp-Glu-X-His) box polypeptide 58
dnajc22	DnaJ (Hsp40) homolog, subfamily C, member 22
dop1b	DOP1 leucine zipper like protein B
dot1l	DOT1-like histone H3K79 methyltransferase
e2f1	E2F transcription factor 1
e2f3	E2F transcription factor 3
ecsit	ECSIT signaling integrator

4.8 Supplementary Material

EIF3S6IP	eukaryotic translation initiation factor 3, subunit 6 interacting protein
ENDOU2	endonuclease, polyU-specific 2
FADS6	fatty acid desaturase 6
FAM199X	family with sequence similarity 199, X-linked
FBXO40.1	F-box protein 40, tandem duplicate 1
FDFT1	farnesyl-diphosphate farnesyltransferase 1
FGB	fibrinogen beta chain
FLRT3	fibronectin leucine rich transmembrane 3
FOXRED2	FAD-dependent oxidoreductase domain containing 2
FRA10AC1	FRA10A associated CGG repeat 1
FSD1	fibronectin type III and SPRY domain containing 1
GCLC	glutamate-cysteine ligase, catalytic subunit
GINS2	GINS complex subunit 2
GK5	glycerol kinase 5
GKAP1	G kinase anchoring protein 1
GLMNA	glomulin, FKBP associated protein a
GPAM	glycerol-3-phosphate acyltransferase, mitochondrial
GPAT2	glycerol-3-phosphate acyltransferase 2, mitochondrial
GPR135	G protein-coupled receptor 135
GPR143	G protein-coupled receptor 143
GRAMD1C	GRAM domain containing 1c
GTPBP4	GTP binding protein 4
HBEGFA	heparin-binding EGF-like growth factor a
HDDC2	HD domain containing 2
HECTD3	HECT domain containing 3
HERC4	HECT and RLD domain containing E3 ubiquitin protein ligase 4
HGS	hepatocyte growth factor-regulated tyrosine kinase substrate
HIGD2A	HIG1 hypoxia inducible domain family, member 2A
HLTF	helicase-like transcription factor
IFT22	intraflagellar transport 22 homolog (Chlamydomonas)
INPPL1B	inositol polyphosphate phosphatase-like 1b
ITGB4	integrin, beta 4
IYD	si:ch211-286f9.2
KBTBD3	kelch repeat and BTB (POZ) domain containing 3
KIF26BA	kinesin family member 26Ba
KIF4	kinesin family member 4
KLHL30	kelch-like family member 30
KPNB3	karyopherin (importin) beta 3
KRCP	kelch repeat-containing protein
LAPTM4A	lysosomal protein transmembrane 4 alpha
LEO1	LEO1 homolog, Paf1/RNA polymerase II complex component
LETM2	leucine zipper-EF-hand containing transmembrane protein 2
LIMK2	LIM domain kinase 2
LONP2	lon peptidase 2, peroxisomal
LPCAT2	lysophosphatidylcholine acyltransferase 2
LRGUK	leucine-rich repeats and guanylate kinase domain containing
LRN3B	leucine rich repeat neuronal 3b
MAPK15	mitogen-activated protein kinase 15

mapk6	mitogen-activated protein kinase 6
mcoln2	mucolipin 2
mcu	mitochondrial calcium uniporter
meak7	MTOR associated protein, eak-7 homolog
memo1	mediator of cell motility 1
mettl13	methyltransferase like 13
miga2	mitoguardin 2
mknk1	MAPK interacting serine/threonine kinase 1
mlx	MAX dimerization protein MLX
mms19	MMS19 homolog, cytosolic iron-sulfur assembly component
mocs3	molybdenum cofactor synthesis 3
mon1a	MON1 secretory trafficking family member A
mpnd	MPN domain containing
mrpl41	mitochondrial ribosomal protein L41
mthfr	methylenetetrahydrofolate reductase (NAD(P)H)
naa30	N(alpha)-acetyltransferase 30, NatC catalytic subunit
ncbp2	nuclear cap binding protein subunit 2
ndufv2	NADH:ubiquinone oxidoreductase core subunit V2
nelfcd	negative elongation factor complex member C/D
nfat5b	nuclear factor of activated T cells 5b
nfrkb	nuclear factor related to kappaB binding protein
notch2	notch receptor 2
nploc4	NPL4 homolog, ubiquitin recognition factor
nsrp1	nuclear speckle splicing regulatory protein 1
numa1	nuclear mitotic apparatus protein 1
ogfod1	2-oxoglutarate and iron-dependent oxygenase domain containing 1
opa1	OPA1 mitochondrial dynamin like GTPase
paip1	poly(A) binding protein interacting protein 1
pax8	paired box 8
pde12	phosphodiesterase 12
pdia6	protein disulfide isomerase family A, member 6
phgdh	phosphoglycerate dehydrogenase
phyhd1	phytanoyl-CoA dioxygenase domain containing 1
pik3c3	phosphatidylinositol 3-kinase, catalytic subunit type 3
pik3r5	phosphoinositide-3-kinase, regulatory subunit 5
plekha8	pleckstrin homology domain containing, family A (phosphoinositide binding specific) member 8
PLEKHH3	si:ch211-18i17.2
polr3e	polymerase (RNA) III (DNA directed) polypeptide E
pomgnt2	protein O-linked mannose N-acetylglucosaminyltransferase 2 (beta 1,4-)
ppp1r12a	protein phosphatase 1, regulatory subunit 12A
psma1	proteasome 20S subunit alpha 1
psmc5	proteasome 26S subunit, ATPase 5
psmd8	proteasome 26S subunit, non-ATPase 8
qars1	glutamyl-tRNA synthetase 1
qser1	glutamine and serine rich 1
rab24	RAB24, member RAS oncogene family

4.8 Supplementary Material

rad9b	RAD9 checkpoint clamp component B
ranbp3b	RAN binding protein 3b
rapgef3	Rap guanine nucleotide exchange factor (GEF) 3
rbck1	RanBP-type and C3HC4-type zinc finger containing 1
rbm17	RNA binding motif protein 17
rfng	RFNG O-fucosylpeptide 3-beta-N-acetylglucosaminyltransferase
rfxank	regulatory factor X-associated ankyrin-containing protein
rgs14a	regulator of G protein signaling 14a
rhag	Rh associated glycoprotein
rmdn1	regulator of microtubule dynamics 1
rnf113a	ring finger protein 113A
rpl4	ribosomal protein L4
rplp0	ribosomal protein, large, P0
rps2	ribosomal protein S2
rps6	ribosomal protein S6
rps6kl1	ribosomal protein S6 kinase-like 1
sash3	SAM and SH3 domain containing 3
scly	selenocysteine lyase
sik3	SIK family kinase 3
snai1a	snail family zinc finger 1a
snrnp40	small nuclear ribonucleoprotein 40 (U5)
snrpb	small nuclear ribonucleoprotein polypeptides B and B1
snw1	SNW domain containing 1
snx24	sorting nexin 24
snx5	sorting nexin 5
srebf2	sterol regulatory element binding transcription factor 2
ssb	small RNA binding exonuclease protection factor La
stx18	syntaxin 18
stxbp3	syntaxin binding protein 3
suc1g1	succinate-CoA ligase, alpha subunit
sumf2	sulfatase modifying factor 2
supt20	SPT20 homolog, SAGA complex component
supt4h1	SPT4 homolog, DSIF elongation factor subunit
supt7l	SPT7 like, STAGA complex gamma subunit
svild	supervillin d
syf2	SYF2 pre-mRNA-splicing factor
taf5	TAF5 RNA polymerase II, TATA box binding protein (TBP)-associated factor
tbcd	tubulin folding cofactor D
tbrg1	transforming growth factor beta regulator 1
tex261	testis expressed 261
tfip11	tuftelin interacting protein 11
thoc5	THO complex 5
thtpa	thiamine triphosphatase
tm9sf4	transmembrane 9 superfamily protein member 4
tmem135	transmembrane protein 135
tmem161b	transmembrane protein 161B
traf7	TNF receptor-associated factor 7
trip11	thyroid hormone receptor interactor 11

ubap2l	ubiquitin associated protein 2-like
ufd1l	ubiquitin recognition factor in ER associated degradation 1
unc45b	unc-45 myosin chaperone B
usp39	ubiquitin specific peptidase 39
usp4	ubiquitin specific peptidase 4 (proto-oncogene)
usp49	ubiquitin specific peptidase 49
wasf2	WASP family member 2
wdr13	WD repeat domain 13
wdr61	WD repeat domain 61
wnt4	wingless-type MMTV integration site family, member 4
xpc	xeroderma pigmentosum, complementation group C
xrcc5	X-ray repair complementing defective repair in Chinese hamster cells 5
xrn2	5'-3' exoribonuclease 2
zfand4	zinc finger, AN1-type domain 4
zgc:101663	zgc:101663
zgc:103625	zgc:103625
zgc:112294	zgc:112294
zgc:114119	zgc:114119
zgc:152830	zgc:152830
zgc:162698	zgc:162698
zgc:162879	zgc:162879
zgc:92518	zgc:92518
zwilch	zwilch kinetochore protein

Chapter 5. Genomic stock structure of tarakihi (*Nemadactylus macropterus*): evidence of fine-scale putative adaptation and a temperature-associated cline despite panmixia in New Zealand

5.1 Abstract

Tarakihi (*Nemadactylus macropterus*) supports an important fishery in New Zealand. However, little is known about its population structure. Here we conducted whole-genome resequencing of 175 tarakihi from around New Zealand and from Tasmania (Australia) and twelve king tarakihi (*Nemadactylus* n. sp.) to obtain a dataset of 7.5 million high-quality SNPs. Variant filtering, F_{ST} -outlier analysis, and redundancy analysis (RDA) were used to evaluate population structure, adaptive structure, and locus-environment association. A weak but significant level of neutral genetic differentiation was found between tarakihi from New Zealand and Tasmania ($F_{ST} = 0.0054\text{--}0.0073$, $P \leq 0.05$), supporting the existence of at least two separate reproductive stocks. No clustering was detected among the New Zealand populations ($\Phi_{ST} < 0.001$, $P = 0.77$). Outlier-based, presumably adaptive variation suggested fine-scale structure between locations around central New Zealand off the east (Wairarapa, Cape Campbell, and Hawke's Bay) and the west coast (Tasman Bay/Golden Bay and Upper West Coast of South Island). Allele frequencies from 55 loci were found to be associated with at least one of six environmental variables, of which 47 strongly correlated with mean depth temperature. Projection of the RDA indicated the existence of a latitudinal temperature cline. A large-scale, long term, genomic sampling strategy will be needed to fully resolve genomic-level differences and provide a comprehensive picture of the stock structure for management action.

5.2 Introduction

Effective fisheries management relies on the identification and delineation of stocks to enable optimal utilisation and conservation (Begg et al., 1999; Cadrin et al., 2014b; Waples

et al., 2008). Ideally, the ultimate goal of fisheries management is to harvest each separate stock at a rate that matches their level of recruitment when taking into account natural mortality (Beddington et al., 2007; Zhou et al., 2019). Discrepancies between the stock management units and the boundaries of biological units can result in overexploitation (Benestan, 2019; Laikre et al., 2005; Reiss et al., 2009), which if left unchecked, could lead to the collapse of a stock (Cadrin, 2020; Orensanz et al., 1998; Ying et al., 2011). However, biologically accurate information about stock boundaries is still lacking for many fisheries species, particularly those residing in the New Zealand Exclusive Economic Zone (Papa, Oosting, et al., 2021).

Marine environments often contain few physical barriers when compared to freshwater or terrestrial environments. Therefore, the levels of genetic divergence among groups of marine fishes are often expected to be low. This is especially true for marine species with large population sizes and high potential for larval and/or adult dispersal (Lal et al., 2016; Ovenden, 2013; Sandoval-Castillo et al., 2018). Traditional genetic markers (e.g. microsatellites or mitochondrial sequences), that typically represent a very small proportion of the genome, often don't provide the required level of resolution required to detect fine-scale genetic structure. However, not detecting any significant genetic differentiation does not necessarily mean there is a level of migration relevant to fisheries management. Populations could be demographically independent but only recently sundered, there could still be some low level of gene flow, or effective population size could be high, with differentiation caused by drift happening slowly (Attard, Beheregaray, et al., 2018; Benestan et al., 2015). Moreover, when variation is detected with low-resolution markers, it is typically very difficult to determine whether the observed variation is neutral (i.e. due to the accumulation of random mutations through reproductive isolation and genetic drift) or adaptive (i.e. due to natural selection which can lead to e.g. local adaptation) (Allendorf et al., 2010; Carvalho & Hauser, 1994).

In contrast to low-resolution traditional genetic markers, next-generation-sequencing technologies can produce very large genome-wide datasets that have two advantages: (1) a vast increase in the number of available neutral loci, which increases the level of resolution attainable when testing for genetic differentiation, and (2) the ability to detect genomic

regions that are currently, or at some time in the recent past, experiencing selection (Benestan, 2019; Bernatchez et al., 2017; Papa, Oosting, et al., 2021). Previously unknown population structure has been detected in several marine species by using genome-wide SNP datasets, they include the American lobster (Benestan et al., 2015), yellowfin tuna (Pecoraro et al., 2018), silky shark (Kraft et al., 2020), green abalone (Mejía-Ruíz et al., 2020), or California market squid (S. H. Cheng et al., 2021). Even when populations display little to no neutral genomic differentiation, adaptive population structure can be detected through outlier-based methods (e.g. albacore (Vaux et al., 2021)) or environment association methods (e.g. American lobster (Benestan et al., 2016), summer flounder (Hoey & Pinsky, 2018), or greenlip abalone (Sandoval-Castillo et al., 2018)).

Tarakihi (*Nemadactylus macropterus* (Forster 1801)) is a demersal marine fish species which is widely distributed in the inshore areas of New Zealand (Figure 5.1). The species occurs from the Three Kings Islands in the north to the Snare Islands in the south and the Chatham Islands in the east, at depths of 10–250 m (Annala, 1987; Roberts et al., 2015). It is also distributed along the southern inshore areas of Australia, including Tasmania (Roberts et al., 2015). They are broadcast spawners that form serial breeding aggregations during summer and autumn (Tong & Vooren, 1972). Tarakihi late-stage larvae go through an epipelagic “paperfish” larval phase for approximately 10 months (Annala, 1987; Roberts et al., 2015). During this period, their dispersal is mainly driven by oceanic currents, where mixing of individuals from different spawning areas can occur (Bruce, 2001). Adults can live for more than 30 years and have the potential to disperse over large distances, sometimes hundreds of kilometres (Annala, 1987; Hanchet & Field, 2001). Tarakihi are mainly caught by bottom trawling at depths of about 250 m. Commercial catches are around 5,000 tonnes per year over the past 30 years, with a very recent reduction to 4,400 tonnes for the fishing year 2019-2020 (Fisheries New Zealand, 2021). While tarakihi are commercially caught in all the unprotected Quota Management Areas of the New Zealand Exclusive Economic Zone, the majority of catches (c. 80%) occur off the east coast of the North and South Island (Langley, 2018). The spawning biomass in some areas (TAR1, TAR2 and TAR3, Figure 5.1) are thought to be below the fisheries management soft limit (20% of the unexploited, equilibrium biomass) since the early 2000s (Langley, 2018). Consequently, the total allowable commercial

catch (TACC) was reduced in 2018 and again in 2019 for these areas, which is the first reduction in tarakihi TACC since the 1980s (Fisheries New Zealand, 2021). The observed declines highlight the need for evidence based managements strategies that incorporate best knowledge about the biological stock structure of this species. However, both stock structure and the levels of connectivity among fished areas are poorly known for this species.

DNA-based markers are particularly appropriate to provide evidence for stocks boundaries, especially if a population has experienced long-term (total or partial) reproductive isolation (Cadrin, 2020; Ovenden et al., 2015; Waples et al., 2008). Six population genetic studies have been conducted on tarakihi in an effort to resolve its stock structure. Four of these studies investigated samples from around Australia and included only one location from New Zealand, using allozyme electrophoresis (N. G. Elliott & Ward, 1994; Richardson, 1982), mitochondrial DNA restriction fragment length polymorphism (Grewe et al., 1994), and microsatellite DNA markers (Burridge & Smolenski, 2003). No significant genetic structure among Australian stocks was detected. Weak but significant genetic divergence was found between New Zealand and Australia (N. G. Elliott & Ward, 1994; Grewe et al., 1994; Richardson, 1982), except in one study (Burridge & Smolenski, 2003). Two genetic studies have been conducted among New Zealand locations. The first study (Gauldie & Johnston, 1980) was conducted on c. 3,000 samples from around New Zealand. Significant population genetic differentiation was observed based on allelic frequencies at the phosphoglucumutase allozyme locus, which led to the proposition of eight geographical stock boundaries around the North and South Island (see Gauldie & Johnston (1980) for more details on these putative stock boundaries). However, it was not possible to rule out a deviation from selective neutrality for that locus. Indeed, the authors emphasized that the observed differences in allele frequencies might not be indicative of differentiation due to reproductively separated stocks, but were most probably due to an adaptive cline related to water temperature. The second study used direct sequencing of the mitochondrial DNA control region from 370 tarakihi collected around New Zealand main islands and Chatham Islands (Papa, Halliwell, et al., 2021). While again no overall genetic structure was found among New Zealand and Chatham Island populations, two weak genetic disjunctions were detected. The first was between the west and east coasts of the South Island, which would be consistent with biological observations based on age and size structure. The second was

between Hawke's Bay and East Northland (but not the locations between them), which may be indicative of a complex migration pattern along the east coast of North Island.

The overall goal of this study was to determine the population genetic structure of tarakihi sampled from sites around New Zealand and analysed using whole-genome resequencing. This high-resolution data set was used to 1) characterize neutral and 2) presumably adaptive genetic variation across sampling locations, and 3) determine whether there is a relationship between genetic differentiation and environmental factors. The results are compared to the current fishery stock hypotheses for tarakihi.

5.3 Materials and Methods

5.3.1 Sampling and DNA extraction

One hundred eighty-eight samples were used, including 161 tarakihi from New Zealand, 14 tarakihi from Tasmania and 12 king tarakihi (*Nemadactylus* n.sp.) from the north of New Zealand (Figure 5.1, Table 5.1). The king tarakihi phenotype is similar to tarakihi and is managed as part of the same fisheries. Samples from king tarakihi were included to compare the observed levels of diversity with the sister taxon. An additional specimen, caught in a fishing competition in Gisborne (East Cape), was visually identified as a king tarakihi and was added to the dataset (referred to as "GBK" for Gisborne king tarakihi). All 188 samples were sourced from specimens captured during two sampling phases. The first sampling phase took place between October 2017 and April 2018 and aimed at collecting specimens from all around New Zealand for a previous study on the population genetics of tarakihi based on a mitochondrial marker (Papa, Halliwell, et al., 2021). All specimens were sourced from commercial fishing companies (which in some cases conducted specific tows targeting tarakihi) with the exception of Fiordland samples that were provided by recreational fishers. Although most of the tissues collected during the first sampling phase were suitable for Sanger sequencing as applied in the mitochondrial study, DNA in these samples was generally too degraded for Illumina library preparation and whole-genome sequencing. As a result, 46 samples from phase 1 that passed the quality controls criteria were used in this study. The remaining 142 samples came from specimens captured during a second sampling phase from

January 2019 to June 2020. This included fish captured by commercial fishing companies and/or as part of a broader monitoring campaign from the National Institute of Water and Atmospheric Research (NIWA), as well as specimens captured during recreational fishing competitions. An additional 14 specimens captured off Tasmania were purchased at a Sydney market through the New South Wales Department of Primary Industries program.

Figure 5.1. Sampling sites of tarakihi (sites 1–17) and king tarakihi (site 18) used in this study. King TAR: king tarakihi from Three Kings Islands. Red asterisks (*) indicate populations that may be adaptively divergent based on outlier analysis. Blue lines and labels correspond to the New Zealand Exclusive Zone and the quota management areas for tarakihi (TAR 1–5, 7–8, 10). The bathymetric contour is shown in light grey (125 m) and dark grey (1000 m). Bottom left: range of tarakihi in blue area (Kaschner et al., 2019).

Tail muscle (phase 1) or pectoral fin (phase 2) tissues were collected from the specimens and immersed in 99% ethanol (phase 1 and 2) or DESS solution (20% DMSO, 0.25 M EDTA, NaCl saturated) (phase 2) and then stored at -20°C as soon as possible. DESS was found to be more suitable to preserve DNA when tissues are sampled in the field (Oosting et al., 2020), which is generally the case when dealing with wild fisheries specimens. Total genomic DNA was extracted with a rapid salt-extraction protocol (Aljanabi & Martinez, 1997) that included an RNase step, suspended in 30–150 μl Tris-EDTA buffer (10 mM Tris-HCl pH 8.0, 0.1 mM EDTA), and stored at -20°C . The quantity of Tris-EDTA was chosen based on visual assessment of the size of the DNA pellet. A low concentration of EDTA (0.1 mM) was used to

reduce the risks of enzymatic inhibitions during library preparation. The integrity of extracted DNA samples was assessed by visualising the presence of high molecular weight DNA with agarose gel electrophoreses. Electrophoreses were run in TBE-buffered 1% agarose gels at 90 V and 400 mA for 30 min. Samples were visually classified as high, medium, or low weight depending on the intensity of DNA > 20 Kb as measured with a Lambda *Hind*III DNA ladder (Supplementary Figure 5.1). Purity of DNA samples was measured with a NanoPhotometer® NP80 (Implen). Readings of 260/280 were classified as good (1.8–1.95), medium (1.6–2) or bad (any other value). Similarly, 260/230 values were classified as good (≥ 2), medium (≥ 1.4), or bad. Quantity of double-strand DNA was measured using a Qubit™ dsDNA BR Assay Kit. Samples were classified based on dsDNA concentration in a final volume of 100 μ l as follows: high (≥ 50 ng/ μ l), medium (≥ 5 ng/ μ l), or low (< 5 ng/ μ l). DNA samples were selected based on both quality and geographic representation for DNA library preparation and sequencing.

Table 5.1. Sampling sites and samples information.

Management area	<i>N</i>	Sampling location	<i>N</i>	Code	Source of samples	Sampling phase (<i>n</i> samples)
Tarakihi						
TAR1	32	Upper West Coast North Island	12	01.UWCNI	NIWA 2019 survey	2 (12)
		East Northland / Hauraki Gulf	10	14.ENHG	NIWA 2019 survey	2 (10)
		East Northland	10	15.ENLD	Commercial fishing / NIWA 2019 survey	1 (1) / 2 (9)
TAR2	33	Wellington	21	10.WGTN	Commercial fishing / Fishing competition	1 (1) / 2 (20)
		Wairarapa	5	11.WAI	Commercial fishing	1 (5)
		Hawke's Bay	7	12.HB	Commercial fishing / Fishing competition	1 (1) / 2 (6)
		East Cape	21	13.EC	Commercial fishing / Fishing competition	1 (4) / 2 (17)
TAR3	16	Christchurch	16	08.CHCH	NIWA 2020 survey	2 (16)
TAR4	2	Chatham Islands*	2	16.CHAT	Commercial fishing	1 (2)
TAR5	3	Fiordland*	3	07.FRDL	Recreational fishers	1 (3)
TAR7	43	Tasman Bay / Golden Bay	12	03.TBGB	NIWA 2019 survey	2 (12)
		TBGB juveniles*	5	04.TBGBJ	NIWA 2019 survey	2 (5)
		Upper West Coast South Island	9	05.UWCSI	Commercial fishing	1 (9)
		Lower West Coast South Island	11	06.LWCSI	NIWA 2019 survey	2 (11)
		Cape Campbell	6	09.CC	Commercial fishing	1 (6)
TAR8	32	Taranaki	11	02.TARA	Commercial fishing / NIWA 2019 survey	1 (3) / 2 (8)
Australia	14	Tasmania	14	17.TAS	NSW DPI 2019 survey	2 (14)
King Tarakihi						
TAR1	12	Three Kings Islands*	12	18.KTAR	Commercial fishing	1 (12)
TAR2	1	East Cape*	1	GBK	Fishing competition	2 (1)

Note: (*) Samples that were not included in the final neutral and adaptive datasets. NIWA: National Institute of Water and Atmospheric Research. Sampling locations and corresponding codes are plotted on the map in Figure 5.1.

5.3.2 Whole-Genome Sequencing

A total of 188 DNA samples were selected for whole-genome sequencing. An effort was made to sequence at least 10 specimens per sampling location, however, this could not be achieved for seven out of 18 locations. Especially, only a few DNA samples of sufficient

quality could be obtained for remote locations that were sampled only in phase 1 (Chatham Islands and Fiordland). Nevertheless, a good overall representation of tarakihi fishing areas around mainland New Zealand was obtained (Figure 5.1), especially in the most fished management areas (TAR1, TAR2, and TAR3). All DNA samples were diluted to an equal volume of 80 µl with a DNA concentration of 30 ng/µl (sometimes lower when not possible) and sent to the Australian Genome Research Facility (AGRF, Melbourne, Australia) for DNA library preparation and sequencing. The Illumina DNA shotgun library was prepared following the Nextera DNA FLEX low volume protocol with Nextera DNA Combinatorial Dual Indexes (Illumina) for insert sizes 300–350 bp. Sequencing of 150 bp paired-end reads was performed on NovaSeq 6000 (Illumina) with NovaSeq 6000 S4 Reagent Kit and NovaSeq XP 4-Lane Kit for 300 cycles. For each sequencing run, at least one non-*Nemadactylus* species was being sequenced on the same lane (either snapper (*Chrysophrys auratus*) from another study or the fish used in Chapter 4 of this thesis). Shared lanes allowed for a post-sequencing check of the correct labelling of samples throughout the sequencing and de-multiplexing process. Each lane contained 96 wells, and each individual was sequenced in one well. Since the sequencing yield of each lane was 700–800 Gb, it was expected that each individual would be sequenced for c. 8Gb. Based on the genome size of 700 Mb estimated pre-sequencing (cf. Chapter 3), the sequencing coverage was estimated to be c. 11× per sample. Base calling, quality scoring, and de-multiplexing were performed by sequencing provider with RTA3 software v3.3.3 and Illumina bcl2fastq pipeline v2.20.0.422.

5.3.3 Quality control and pre-processing

The quality of the paired-end reads was assessed with FastQC v0.11.7 (S. Andrews, 2018) and MultiQC v1.7 (Ewels et al., 2016) before and after trimming. Raw reads were trimmed for adapter contamination and unpaired reads were filtered out using Trimmomatic v0.39 (Bolger et al., 2014) with the following parameters: paired-end mode, phred 33 base quality encoding, maximum allowed mismatch count = 2, match accuracy between adapter-ligated reads = 30, match accuracy between adapter and read = 10. Nextera adapters were targeted using the file `NexteraPE-PE.fa` provided by Trimmomatic.

5.3.4 Genotyping

Trimmed and filtered paired reads were mapped to the tarakihi reference genome presented in Chapter 3 of this thesis by using the Burrows-Wheeler Alignment (BWA) method with `bwa-kit v0.7.15` (H. Li & Durbin, 2009). In brief, a BWA index was created for the reference genome, as well as a SAM index with `SAMtools v1.9` (H. Li et al., 2009). For each specimen, the forward and reverse reads were mapped to the genome with the command `bwa mem -a -M` to flag all single-end and unpaired reads and mark shorter split hits as secondary. Duplicated reads were marked using `Picard v2.18.20` (Broad Institute, 2019) `MarkDuplicates` with default settings. Genotype likelihoods were produced from the BAM files using `bcftools v1.9` (H. Li, 2011) command `mpileup`. The minimum mapping quality was set to 10, the mapping quality downgrading coefficient was set to 50 for reads with excessive mismatches, and only reads mapped in proper pair were kept (e.g. no secondary alignment or duplicates). SNPs were then called with `bcftools` multi-allelic caller on default parameters, outputting only variant sites and skipping indels. Genotype likelihoods and SNPs calling were both performed separately on each of the 1,214 assembly scaffolds and results were then merged into a genome-wide SNP dataset using `bcftools concat`. The final BCF file was sorted and indexed after having renamed the headers.

5.3.5 Variant filtering

Several SNP datasets were produced and analysed separately in this study (Figure 5.2). The first was a pruned dataset that contained the 188 tarakihi and king tarakihi individuals, where SNPs were filtered for minimum quality criteria and pruned for linkage and Hardy-Weinberg disequilibrium. The second was a dataset of neutral SNPs (hereafter referred to as the “neutral dataset”) that only included the 250 longest scaffolds from locations with ≥ 5 sampled adult tarakihi individuals (king tarakihi, Fiordland, Chatham Islands, Tasman Bay/Golden Bay (TBGB) juveniles, and GBK were discarded). SNP filtering for the neutral dataset was identical to the pruned dataset but with an additional step of filtering out potentially adaptive outliers (see below). The 250 longest scaffolds were retained because (1) they contained more than 90% of the total number of bases in the tarakihi genome (L90 = 219, cf. Chapter 3 of this thesis), (2) for computational efficiency, and (3) because the default

parameters used for the OutFLANK analysis (see below) were not optimal anymore to fit the F_{ST} curve in some scaffolds past that number, which means they would have had to be manually tuned for the c. 1,000 remaining scaffolds. Chatham Islands, Fiordland and the juvenile TBGB samples were discarded at that stage because of their low number of samples (two, three, and five, respectively), which was considered too low to confidently predict population allele frequencies. The GBK specimen was discarded because of its dubious field identification, and the king tarakihi were discarded for being a different species with a different demographic history. The third and fourth datasets were outlier-based (presumably) adaptive datasets containing the same individuals and scaffolds as the neutral dataset, but including only strong candidates for adaptive loci (see below). The fifth, environment-adaptive, dataset was obtained through locus–environment association analysis (see below). Filtering of variant sites was performed with VCFtools v0.1.16 (Danecek et al., 2011). Filtering steps were primarily carried out separately on each scaffold, and resulting VCF files were merged with Picard v2.18.20.

Quality filtering: Only bi-allelic sites with a minimum quality of 600 were retained (`--max-alleles 2 --min-alleles 2 --minQ 600`). The minimum allelic depth for sites in an individual was set to 3 and the mean depth of sites across all individuals was set between eight and twenty-five (`--minDP 3 --min-meanDP 8 --max-meanDP 25`). Sites that were missing in more than 5% of individuals and sites with a minor allele frequency lower than 1% were filtered out (`--max-missing 0.95 --maf 0.01`). For each site, potential allelic imbalance was detected by running a binomial test on the sum of all reference alleles and the sum of all alternative alleles across all sequenced reads of heterozygous individuals with a custom R script, using the genotype (GT) and allelic depth (AD) format information. Sites were filtered out with VCFtools (`--exclude-positions`) if the total proportion of reference and alternative alleles was significantly different from 50% ($P \leq 0.05$ after correction for false discovery rate) according to the binomial test.

Linkage disequilibrium: In order to choose a threshold value for thinning in the pruning step, linkage disequilibrium decay was plotted for the 30 longest scaffolds on the quality-filtered dataset. For this, the squared correlation coefficients between genotypes of sites

separated by a maximum of 50,000 bp were calculated with VCFtools (`--geno-r2 --ld-window-bp 50000`). Nucleotide position and R^2 values for a random subset of a million sites in each scaffold were then plotted with a custom R script (Oosting, 2021). The analysis was run twice, once on the 188 individuals, and once without including king tarakihi specimens.

Pruning: Sites that were significantly deviating from Hardy-Weinberg equilibrium were filtered out, as well as sites occurring within a distance of 1,500 bp from one another in the scaffolds (`--hwe 0.05 --thin 1500`). A minimum allele frequency of 0.01 was applied a second time. The presence of remaining linkage disequilibrium after thinning was detected with PLINK v1.90 (Chang et al., 2015), using a window size of 50 sites, shifting every five sites and using a pairwise R^2 threshold of 0.2 (`--indep-pairwise 50 5 0.2`). These SNPs were then filtered out with VCFtools.

Neutral filtering: Sites that were potentially under selection were detected with OutFLANK v0.2 (Whitlock & Lotterhos, 2015) using default parameters. Loci flagged as outliers with a minimum heterozygosity of 0.1 ($H_e > 0.1$) and a false discovery threshold below 1% ($q\text{values} < 0.01$), were filtered out of the dataset with VCFtools.

Adaptive filtering: In parallel of the neutral filtering, OutFLANK v0.2 was used on the pruned dataset to flag outlier loci with a minimum heterozygosity of 0.1 and a false discovery threshold (q-value) below 0.05. This analysis was run twice: the sampling locations that were detected as differentiated were discarded between the first and second analyses to detect SNPs with finer population structure. After each analysis, a custom R script adapted from Oosting (2021) was used to select only the site with the highest F_{ST} in each non-overlapping 50,000 bp sliding window in order to obtain a set of independent presumably adaptive SNPs (Supplementary Figure 5.2 & Supplementary Figure 5.3).

Figure 5.2. Variant filtering pipeline applied on the raw SNP dataset, resulting in one intermediary dataset (quality-filtered) and five final datasets (pruned, neutral, adaptive 1, adaptive 2, and environment-adaptive).

5.3.6 SNP data analysis

Number of reads per individuals and mean GC content were reported with FastQC v0.11.7 and MultiQC v1.7. Mean read depth of variant sites and proportion of missing sites in each individual were calculated with VCFtools v0.1.16. Observed heterozygosity and number of fixed alleles were obtained with dartR v1.9.6 (Gruber et al., 2018). Analyses of Molecular Variance (AMOVA) were conducted on the neutral dataset with poppr v2.9.2 (Kamvar et al., 2014) using hierarchical models of several groupings based on management areas, New Zealand coasts, and sampling locations. Associated *p*-values were computed with ade4

v1.7.16 (Dray & Dufour, 2007) function `randtest` using 999 replicates. Principal component analyses (PCA) were run on the pruned, neutral, and adaptive SNP datasets with `dartR` v1.9.6 using default parameters and results were projected onto axes with `ggplot2` v3.3.3 (Wickham, 2009). Successive K-means cluster identification and discriminant analyses of principal components (DAPC) were performed with `adegenet` v2.1.3 (Jombart, 2008). See parameters used in Supplementary Table 5.1.

Pairwise weighted F_{ST} values (Weir & Cockerham, 1984) were computed with `StAMPP` v1.6.1 (Pembleton et al., 2013) on the neutral and adaptive datasets. Associated p -values and 95% confidence intervals were generated by running 1,000 bootstraps across loci and a false discovery rate correction (Benjamini & Hochberg, 1995) was applied to the p -values. The pairwise weighted F_{ST} values were plotted onto heat maps clustered via dendrogram using `pheatmap` 1.0.12 (Kolde, 2019). Population structure was tentatively inferred in the neutral dataset with `fastSTRUCTURE` v1.0.3 (Raj et al., 2014). The number of populations (K) was tested from 1 to 10, and the value of K that best explained the observed structure was chosen with the utility tool `chooseK.py`.

A test of isolation by distance (IBD) was performed using a Mantel test (999 replicates) with `ade4` v1.7.16 on the neutral SNP dataset. The analysis was restricted to the samples from New Zealand locations. The geographic distances between samples coordinates were estimated with `gdistance` v1.3.6 (van Etten, 2017) by applying a ‘least-cost distance’ model of geographic dispersal where travel is restricted to the ocean. For this, a shapefile of New Zealand was rasterized with `raster` v3.4.5 (Hijmans, 2019), and each cell grid corresponding to the ocean and the land were assigned movement costs of 0 and 1, respectively. Sampling locations coordinates were obtained by calculating the mean between the start and stop trawl latitudes and longitudes provided by the fishing vessels, or approximated arbitrarily when not provided. Pairwise weighted F_{ST} values between sampling locations were used for the matrix of genetic distances.

5.3.7 Genotype-environment association analysis

Environmental feature data was obtained with the R package *sdmpredictors* v0.2.9 (Bosch, 2020). The complete dataset available from Bio-ORACLE v1, v2.1, and MARSPEC (Assis et al., 2018; Sbrocco & Barber, 2013; Tyberghein et al., 2012) was compiled into a total environmental dataset of 362 variables. Since the locus-environment association method used is a regression-based analysis, it was necessary to check for collinearity (i.e. non-independence) between the environmental variables before proceeding (Dormann et al., 2013). This was done in several steps. First, the 362 variables were manually clustered into groups that measured the same physical, chemical, or biological process (Supplementary Table 5.2). As an example, the group “dissolved oxygen”, included 25 variables, such as sea surface mean concentration, minimum concentration at mean depth, or concentration range at the bottom. For each of these groups, the R package *psych* v2.1.3 (Revelle, 2021) was used to assess the correlation level among variables. A set of weakly correlated variables (Pearson correlation $|R^2| < 0.7$) was manually selected from each group and compiled in a second dataset of 48 variables (Supplementary Figure 5.4). The Pearson correlation and associated significance were computed among these 48 variables with the R package *Hmisc* v4.5-0 (Harrell, 2021). All variables that were significantly correlated to “mean temperature at mean depth” ($P \leq 0.05$) were discarded. Temperature was chosen because it is usually correlated with many other environmental variables as it is associated with, or drives, many physical, chemical, and biological processes, and the mean value at mean depth was chosen to reflect the biology of tarakihi, which is a demersal, inshore species. Variables that were considered to be not biologically relevant for the studied species (cloud fraction, ice cover, East/West aspect and North/South aspect) were also discarded. This led to a third dataset of nine variables (Supplementary Figure 5.5). Three variables (profile curvature, plan curvature, and mean chlorophyll concentration at minimum depth) were further discarded for being correlated to concavity and mean primary production at the bottom ($|R^2| > 0.5$, $P < 0.001$). This reduced the dataset to a final six lowly correlated variables: salinity range at mean depth, mean primary production at minimum depth, concavity (which is indicative of the location being on a slope or in a valley), mean temperature at mean depth, mean diffuse attenuation

coefficient (indicator of water clarity), and mean iron concentration at the sea surface (Supplementary Figure 5.5).

Association between genotypes and environmental variables was assessed using a Redundancy Analysis (RDA). RDA is a multivariate, ordination-based locus-environment association method that is effective at detecting local adaptation on multiple loci under numerous demographical, biological, and sampling scenarios (Forester et al., 2016, 2018; Rellstab et al., 2015). The analysis was performed separately on the eight longest scaffolds using the quality-filtered SNP dataset (Figure 5.2), without including king tarakihi, GBK, specimens from Tasmania, and juveniles from TBGB. Missing genotype data was estimated using the most common genotype at each locus. The RDA was run with the package `vegan` v2.5.7 (Oksanen et al., 2020) on the genotype data and the six environmental variables cited above, with parameter `scale = TRUE` to scale observations to unit variance. The significance of each RDA model, as well as each constrained axis, was assessed with `vegan` v2.5.7 function `anova.cca` using 999 permutations. Loci that were strong candidates for local adaptation were selected by identifying the SNPs with scores ± 3.5 standard deviations from the mean score of each significant constrained axis. The position in the genome and putative associated function of the selected loci were obtained using the tarakihi reference genome assembly annotation produced in Chapter 3 of this thesis.

5.3.8 General bioinformatics tools

All SAM files were converted to sorted BAM files with SAMtools `sort` and BAM files were indexed with SAMtools `index`. Alignments statistics were computed with SAMtools `flagstats` and BamTools v2.5.1 (Barnett et al., 2011) `stats`. VCF files were imported and converted to `genlight` objects in R with `vcfR` v1.12.0 (Knaus & Grünwald, 2017). `dartR` v1.9.6 was used for manipulations of `genlight` objects (e.g. population assignment, filtering of individuals and monomorphic loci).

5.4 Results

5.4.1 Sequencing

Following quality filtering and adapter trimming, a total of 12.1 billion Illumina reads, with an average length of 150 bp, were obtained from 188 individuals. The mean GC content was 44.8%. All sequences were of high quality with a mean Phred Score per base > 35 (Supplementary Figure 5.6) and all individual samples passed all the FastQC criteria. The slight bias in per base sequence contents observed in Chapter 3 was also present, and for the same reasons detailed as in that chapter, that was not considered an issue here either. The number of reads per sample ranged from 17.1 million to 124.0 million (64.1 million on average) (Figure 5.3A). This translated to a mean read sequencing depth per individual of 16.9×, ranging from 4.5× to 32.7×, with only eight individuals below 8×.

5.4.2 SNP datasets

Quality filtering of variants led to a total of 7,536,950 high-quality bi-allelic SNPs with a mean depth per individual of 12.7× (Figure 5.3B). The “pruned” dataset that included all scaffolds and individuals was made up of 183,443 high-quality, independently segregating SNPs. To obtain the neutral dataset, 23 individuals were discarded and the analysis was restricted to the 250 longest scaffolds (see Methods). This final neutral dataset contained 166,022 high-quality, independently segregating, neutrally evolving bi-allelic SNPs (Figure 5.2). The mean SNP depth per individual in that dataset was 12.0× (0.04x-23x) and the number of missing sites per individual ranged from 5 to 164,506, with 151 individuals (92% of total) missing less than 400 sites (Figure 5.3C).

5.4 Results

Figure 5.3. Summary statistics computed on (A) trimmed Illumina reads, (B) quality-filtered SNP dataset, and (C) neutral SNP dataset. N seq: number of Illumina reads. Ho: mean observed Heterozygosity. N missing: number of missing sites. Each circle is an individual on the X-axis.

There was a clear difference in observed heterozygosity between tarakihi and king tarakihi specimens in the quality-filtered dataset, with an average of 0.12 and 0.06, respectively (Figure 5.3B). The GBK specimen had a heterozygosity level typical of a tarakihi. In the neutral dataset, the mean observed heterozygosity was 0.13, ranging from 0.09 to 0.15. The levels of heterozygosity did not significantly vary among locations but were directly related to the mean SNP depth, which in turn was correlated to the number of missing sites

per individual. The relationship between heterozygosity levels and depth of coverage was especially evident in the 14 samples with a heterozygosity < 0.12 , which all had a mean depth $< 7x$.

In the high-quality total SNP dataset, 84,144 allelic differences were fixed between the king tarakihi and the tarakihi specimens (not including GBK). In comparison, the average number of fixed mutations among all tarakihi locations ranged from 1 (East Cape) to 145 (Chatham Islands), with only TBGB juveniles, Wairarapa, Tasmania, Fiordland, and Chatham Islands having more than 20. There was only one fixed mutation between GBK and the tarakihi specimens.

5.4.3 Linkage Disequilibrium

Linkage disequilibrium in tarakihi was low overall, with mean pairwise R^2 values never exceeding 0.2, even between nucleotide sites less than 100 bp apart (Figure 5.4). Linkage disequilibrium decay was also rapid: both R^2 and mean R^2 values plotted on the distance between nucleotides always reached a plateau between 500 and 1,500 bp (Figure 5.4). The threshold limit for the thinning step in the variant filtering pipeline was thus set to a conservative 1,500 bp. Interestingly, when the same analysis was run while including the king tarakihi specimen, the R^2 values were generally higher and more uniform along the length of the scaffolds (Supplementary Figure 5.7).

Figure 5.4. Linkage disequilibrium decay over genetic distance on scaffold 1, calculated on the quality-filtered SNP dataset, minus king tarakihi specimens. The horizontal red dashed line shows the threshold of 0.2 that is commonly applied to identify independent degradation of nucleotide sites. The orange line is the background level of linkage disequilibrium (intercept). The blue line is the trend of linkage disequilibrium decay fitted to the plot (minimum and maximum variance in dashed lines for the mean).

5.4.4 Population structure

Several AMOVAs were run on the neutral and the adaptive datasets to assess the proportion of genetic variance that could be explained by pre-assigned groupings in these

datasets. Groupings of sampling locations included management areas (with or without including Tasmania) and New Zealand coasts (i.e. two groups: west coast and east coast). Overall, there was no significant genetic structure among management areas, coasts, or sampling locations when analysing the neutral dataset Table 5.2. The only significant genetic differentiation ($\Phi = 0.001$, $P = 0.001$), was detected among sample locations without any *a priori* broader grouping, and only when the Tasmania location was included, which means the differentiation between New Zealand and Tasmania is driving that result. Conversely, the AMOVAs based on presumably adaptive SNPs showed sample locations to always be significantly differentiated ($P \leq 0.01$) with the percentage of explained variation ranging from 6.1% to 20.5% depending on the grouping used (Table 5.2), while the majority of the variation was still within individuals (76.5%–93.32%, $P \leq 0.01$). No significant genetic structure was detected among management areas or between the west and east coasts of New Zealand in the adaptive dataset.

5.4 Results

Table 5.2. Results from Analysis of Molecular Variance performed on the neutral and presumably adaptive SNP datasets, with seven *a priori* groupings.

<i>A priori</i> grouping	<i>N</i> ind.	<i>N</i> locs	<i>N</i> groups	Among groups			Between locs within groups			Between ind. within locs			Within individuals		
				%Var	Φ	<i>P</i>	%Var	Φ	<i>P</i>	%Var	Φ	<i>P</i>	%Var	Φ	<i>P</i>
Neutral															
no grouping	165	14					0.084	0.001	0.001	-0.260	-0.003	0.565	100.176	-0.002	0.593
no grouping, NZ only	151	13					-0.011	0.000	0.769	-0.314	-0.003	0.625	100.324	-0.003	0.640
NZ TAR areas + AU	165	14	6	0.104	0.001	0.123	-0.005	0.000	0.725	-0.260	-0.003	0.609	100.161	-0.002	0.586
NZ TAR areas only	151	13	5	0.002	0.000	0.501	-0.012	0.000	0.731	-0.314	-0.003	0.611	100.324	-0.003	0.620
NZ West vs. East	130	12	2	0.002	0.000	0.261	-0.009	0.000	0.671	-0.322	-0.003	0.592	100.329	-0.003	0.642
NZ North Island: West vs. East	76	7	2	-0.001	0.000	0.497	0.010	0.000	0.462	0.061	0.001	0.417	99.930	0.001	0.449
NZ South Island: West vs. East	54	5	2	0.018	0.000	0.109	-0.033	0.000	0.967	-0.861	-0.009	0.654	100.876	-0.009	0.678
Adaptive															
no grouping	165	14					20.459	0.205	0.001	1.176	0.015	0.10	78.365	0.216	0.001
no grouping, NZ only	151	13					6.060	0.061	0.001	1.632	0.017	0.07	92.308	0.077	0.001
NZ TAR areas + AU	165	14	6	16.006	0.160	0.202	6.341	0.075	0.001	1.148	0.015	0.10	76.505	0.235	0.001
NZ TAR areas only	151	13	5	-2.073	-0.021	0.969	7.786	0.076	0.001	1.638	0.017	0.08	92.649	0.074	0.001
NZ West vs. East	130	12	2	-0.634	-0.006	0.970	6.923	0.069	0.001	1.609	0.017	0.08	92.102	0.079	0.001
NZ North Island: West vs. East	76	7	2	-1.816	-0.018	0.960	7.140	0.070	0.001	3.074	0.032	0.06	91.601	0.084	0.001
NZ South Island: West vs. East	54	5	2	-1.436	-0.014	0.909	7.181	0.071	0.001	0.931	0.010	0.32	93.323	0.067	0.001

Note: %Var = variance component in percentage of the total variation. The three 'West vs. East' groupings did not include the Wellington location. Significant *P*-values (≤ 0.05) are in bold.

The PCA performed on the pruned dataset containing all 188 individuals showed that king tarakihi (18.KTAR) and the tarakihi from Tasmania (17.TAS) form two distinct clusters that are separate from all New Zealand tarakihi specimens (Figure 5.5). The differentiation between these two groups drove most of the variation of the first axis and the second axis, which explained 4.61% and 0.72% of the total variation, respectively. No structure was apparent among the New Zealand populations: sampling locations were randomly distributed on the third and fourth axes (Figure 5.5). Similarly, the DAPC conducted on K-means clustering did not infer any groups related to sampling locations among New Zealand tarakihi (Supplementary Figure 5.8 & Supplementary Figure 5.9). In both PCA and DAPC, the GBK specimen was always grouped within the tarakihi individuals and did not display any particular deviation from them, thus challenging its field identification as a king tarakihi. The PCA and DAPC conducted on the neutral dataset gave identical results, in that Tasmania was separated from New Zealand but no structure could be detected among New Zealand locations (Supplementary Figure 5.10-Supplementary Figure 5.12).

Figure 5.5. Principal component analysis of the pruned SNP dataset that includes 183,443 loci from 188 tarakihi and king tarakihi. Ellipses represent the 95% confidence intervals. Top: Axes 1 and 2. Bottom: Axes 3 and 4. Bottom left: Eigenvalues. Sampling location codes as referred to in Table 5.1.

PCA and DAPC performed on the presumably adaptive dataset (389 SNPs) showed a genetic differentiation of Tasmania (17.TAS), Wairarapa (11.WAI), and Cape Campbell (09.CC) populations from the remaining tarakihi in New Zealand (Figure 5.6). Each of these groups explained most of the variation on the first, second, and third axes of the PCA, which accounted for 25.63%, 2.41%, and 1.81% of the total variation, respectively (Supplementary Figure 5.13).

Figure 5.6. Discriminant analysis of principal components of the first presumably adaptive SNP dataset that include 389 loci from 165 tarakihi. Top: results of the K-means clustering analysis, with colours corresponding to inferred groups (inf). Bottom: Projection of the DAPC based on inferred groups. Bottom right: Discriminant analysis eigenvalues. Sampling location codes as referred to in Table 5.1.

The PCA and DAPC analyses were performed a second time on the presumably adaptive dataset, but this time the Tasmania, Wairarapa, and Cape Campbell locations were discarded before the outlier analysis (see Methods), which resulted in a second presumably adaptive dataset of 61 SNPs from 140 individuals. The DAPC discriminated three additional groups that corresponded to sampling locations (Figure 5.7): Upper West Coast of South Island (05.UWCSI), Hawke's Bay (12.HB), and a group from Tasman Bay/Golden Bay (03.TBGB) that also included one individual from Wellington (10.WGTN) and one from the Upper West Coast of North Island (01.UWCNI). These results were partially observable on the PCA (Supplementary Figure 5.14.).

5.4 Results

Figure 5.7. Discriminant analysis of principal components of the second presumably adaptive SNP dataset that include 61 loci from 140 tarakihi. Top: results of the *K*-means clustering analysis, with colours corresponding to inferred groups (inf). Bottom: Projection of the DAPC based on inferred groups. Bottom left: Discriminant analysis eigenvalues. Sampling location codes as referred to in Table 5.1.

Pairwise weighted mean F_{ST} computed on the neutral dataset always showed high and significant values when comparing the Tasmania location from New Zealand locations ($F_{ST} = 0.0054\text{--}0.0073$, $P \leq 0.05$ after false discovery rate correction), indicative of a clear genetic differentiation between tarakihi from Australia and New Zealand Supplementary Figure 5.15. When comparing only populations from New Zealand (Figure 5.8), the F_{ST} values were globally

low ($F_{ST} = 0\text{--}0.0022$), indicating a lack of overall sub-structure. However, some of the pairwise F_{ST} comparisons among locations were significant ($P \leq 0.05$ after false discovery rate correction): in particular for Wairarapa, which was significantly different from four other locations: Upper West Coast of South Island, Taranaki, East Northland, and East Cape ($F_{ST} = 0.0012\text{--}0.0021$). The only other significant differentiation was between Wellington and East Northland/Hauraki Gulf ($F_{ST} = 0.0004$). The dendrogram based on F_{ST} separated Wairarapa from all other New Zealand locations, which were split into two further clades separating Upper West Coast of South Island, Taranaki, and East Northland from the rest of the locations.

Figure 5.8. Heatmap of pairwise weighted F_{ST} estimates (corresponding to the values above and below the diagonal) among sample locations of the neutral SNP dataset, minus Tasmania. The dendrogram shows the inferred relationship between sample locations. Significant P -values (≤ 0.05 after false discovery rate correction) are in bold.

The relatively high divergence between Tasmania and New Zealand locations was also apparent in the presumably adaptive dataset (Supplementary Figure 5.16) with all F_{ST} values being significant and ranging between 0.5232 and 0.5532. The mean pairwise F_{ST} values among New Zealand samples was also higher than in the neutral dataset ($F_{ST} = 0.0251\text{--}0.1931$) and all values were highly significant ($P < 0.01$ after false discovery rate correction) (Figure 5.9). This was expected since the dataset was composed of outlier SNPs with the highest F_{ST} values only. Wairarapa and Cape Campbell were the most divergent from the rest of the locations ($F_{ST} = 0.1096\text{--}0.1931$), followed by East Northland/Hauraki Gulf and Hawke's Bay ($F_{ST} = 0.0435\text{--}0.1832$).

Figure 5.9. Heatmap of pairwise weighted F_{ST} estimates (corresponding to the values above and below the diagonal) among sample locations of the presumably adaptive SNP dataset, minus Tasmania. The dendrogram shows the inferred relationship between sample locations. Significant P -values (< 0.01 after false discovery rate correction) are in bold.

FastSTRUCTURE did not detect any significant groupings in the neutral dataset, even including samples from Tasmania: the number of populations that best explained the structure was one. No significant pattern of isolation by distance was detected among New Zealand tarakihi locations when using a matrix of least-coast distance restricted to ocean travel on the neutral dataset ($R = 0.260$, $P = 0.414$).

5.4.5 Genotype-environment association analysis

Out of the eight longest scaffolds, only scaffold 1 was significant for its respective RDA model ($P = 0.022$). Moreover, for this model, only the first axis was significant ($P = 0.01$, with $P > 0.1$ for all other axes). Interpretations of the results were thus restricted to axis 1 of the RDA of scaffold 1. The adjusted R^2 of the model was 0.05, and the variance inflation factors (vegan v2.5.7 function `vif.cca`) of the six predictor variables were all below 1.6, indicating that there was no multicollinearity among them. The projection of the SNPs, samples, and environmental variables on the first two axes shows that the sampling locations were effectively discriminated (Figure 5.10). Moreover, all samples projected in negative values of axis 1 were either from North Island or TBGB. Conversely, samples projected in positive values of axis 1 included all South Islands locations (except TBGB), all samples from Wairarapa and Chatham Islands, and a few samples from East Cape and Hawke's Bay close to the centre.

Figure 5.10. Redundancy analysis (RDA) performed on the quality-filtered SNP dataset from scaffold 1, restricted to adult tarakihi from New Zealand. This included 108,903 loci from 156 tarakihi. The two constrained axes show samples from 16 localities in relation to six lowly correlated environmental variables (black arrows). All samples on the left of the blue dashed line are from North Island and Tasman Bay / Golden Bay, while all samples on the right are from South Island, Wairarapa and Chatham Islands. Top right: Eigenvalues. Sampling location codes as referred to in Table 5.1.

Selection of outliers on the end tails of the loading distribution resulted in the identification of 55 candidate loci for local adaptation (Supplementary Table 5.3). Out of these 55 loci, 47 most strongly correlated with mean temperature at mean depth, five with mean primary production at minimum depth, two with mean iron concentration at the sea surface, and one with salinity range at mean depth. Thirty-two were located inside gene coding regions (with two pairs of loci being located in the same genes, *mindy3* and LOC111664994), eight were located in simple repeat regions, 14 in unannotated regions (most of them surrounded by highly-repetitive regions), and one in an unannotated, putatively transcribed region.

5.5 Discussion

This study investigated the neutral and adaptive stock structure on an important inshore commercial fishery species and applied gene-environment association analyses to detect environmental drivers associated with this structure.

5.5.1 Genetic diversity

Tarakihi in New Zealand and Tasmania displayed similar levels of heterozygosity across their range (Figure 5.3). The differences in heterozygosity observed in some individuals did not depend on the location but rather the original depth of sequencing, which means that a proportion of the genetic variation might not have been detected in the 14 lower coverage individuals. A plateau in heterozygosity seems to be reached when the mean depth of coverage was $> 7-8\times$, which indicates that the totality of the relevant genetic variation that could be captured with this protocol has been captured in all other individuals. The slightly higher mean heterozygosity for the neutral dataset could indicate that some of the filtered out outlier loci had a higher proportion of homozygotes, with some alleles possibly restricted to only a few populations in the dataset. A small fraction of less variable sites might also have been filtered out during the pruning step, due to deviations from Hardy-Weinberg and/or linkage equilibrium. The observed heterozygosity per individual is simply calculated as the proportion of heterozygous loci for that individual against the background of loci that are polymorphic in the dataset. It is difficult to say if this value is high in absolute terms. Comparison with results reported from other genome-wide SNP studies are not relevant, because contrarily to e.g. mitochondrial markers, the observed loci are not necessarily directly comparable because the proportion of loci that are invariant across all individuals in each dataset is typically unknown (Gruber et al., 2021). A similar study using whole-genome resequencing on Australasian snapper (*Chrysophrys auratus*) in New Zealand found a higher average heterozygosity of c. 0.20 for neutral loci (Oosting, 2021), while the snapper populations in New Zealand are smaller with supposedly lower connectivity among stocks than tarakihi.

Based on the same SNP dataset, it was clear that the heterozygosity in king tarakihi is lower than that of tarakihi ($H_o = 0.06$, Figure 5.3B). This is expected from a population with a much smaller size and is consistent with recent findings of lower genetic diversity and smaller

historical and current population size for this species (Papa, Halliwell, et al., 2021). However, since the reference genome was a tarakihi, it is possible that the heterozygosity values of the king tarakihi specimens were underestimated due to only more conserved regions being mapped. This clear cut difference in heterozygosity was strong evidence that GBK had been misidentified when captured on the field, with a value corresponding to tarakihi and not to king tarakihi ($H_o = 0.12$, Figure 5.3B). The high number of fixed mutations between tarakihi and king tarakihi (84,144 fixed allelic differences) is additional strong evidence of the separate status of these two organisms (Burridge, 1999; Papa, Halliwell, et al., 2021; P. J. Smith et al., 2008; P. J. Smith, Roberts, et al., 1996). As a comparison, there are only three fixed allelic differences between tarakihi from Tasmania and New Zealand. These fixed mutational differences could be used as the basis for an SNP assay to differentiate tarakihi from king tarakihi in the field. The number of fixed SNPs among tarakihi in New Zealand locations was very low overall. Even for Chatham Islands, where the number of mean fixed alleles with all New Zealand populations was 145, there were only five fixed alleles compared to the rest of New Zealand. This result could be due to chance because of the very low number of samples from Chatham Islands ($n = 2$).

5.5.2 Linkage disequilibrium

Linkage disequilibrium values were overall low in tarakihi, with mean R^2 values never exceeding 0.2 (Figure 5.4). The decay rate was also very fast, with a plateau reached between 500 and 1,500 bp in all scaffolds. Fast decay of linkage disequilibrium is expected to be common for marine fishes, due to them maintaining very high effective population sizes over long periods of time (Hemmer-Hansen et al., 2014). Indeed, tarakihi effective population size has been historically very high, with a current estimation of about 100 million individuals (Papa, Halliwell, et al., 2021). However, linkage disequilibrium can also be impacted by admixture, mutation rate, founder effect, inbreeding and selection. Linkage disequilibrium values for genome-wide data sets are seldom available for marine organisms, but the results reported here are in par with values reported for European eel (*Anguilla anguilla*, complete LD below 10 kb) and Atlantic cod (*Gadus morhua*, complete LD decay below 10 centimorgans) by Hemmer-Hansen et al. (2014). This is also consistent with results reported from

Australasian snapper, with a complete LD decay at 1,500 bp (Oosting, 2021). However, mean R^2 values were overall higher for snapper, which could be indicative of its smaller effective population size compared to tarakihi.

Including the king tarakihi specimens in linkage disequilibrium analyses resulted in a slightly steeper disequilibrium decay and higher, more uniform R^2 values along the scaffolds Supplementary Figure 5.7. This could be indicative of an overall higher linkage disequilibrium in king tarakihi, due to a combination of small and recent stable population size. This could also indicate the presence of structural differences in the king tarakihi genome that would result in the physical distance between the bases being both underestimated and overestimated when aligned to the tarakihi genome, resulting in the detection of high linkage disequilibrium between relatively distant nucleotides (and inversely). Comparing results by aligning the king tarakihi SNPs to a contiguous king tarakihi genome assembly would be a good way to test for this.

5.5.3 Neutral genetic differentiation with king tarakihi and Australia

The strong genetic differentiation between tarakihi (*N. macropterus*) and king tarakihi (*N. n.sp.*) is now well established (e.g. Figure 5.3 and Figure 5.5), to the point of king tarakihi being candidate for a formal taxonomic description as a different species (tentatively *N. rex*) (Papa, Halliwell, et al., 2021; Roberts et al., 2015, 2020; P. J. Smith, Roberts, et al., 1996). However, this most likely separate species is still reported and managed as part of TAR1 for practical reasons.

Evidence for (at least partial) lack of connectivity between Tasmania and New Zealand tarakihi is very strongly supported in the results of this study. A significant level of genetic divergence between these two areas ($P \leq 0.01$) was detected by both AMOVA and pairwise F_{ST} analyses (Table 5.2, Supplementary Figure 5.15). Tasmanian individuals were always clustered together and separated from the New Zealand samples. These results support similar findings of trans-Tasman differentiation for tarakihi based on various genetic markers (N. G. Elliott & Ward, 1994; Grewe et al., 1994; Richardson, 1982), which microsatellites likely failed to detect (Burridge & Smolenski, 2003). It can be reasonably extrapolated that this

genetic divergence to New Zealand would also be found in samples analysed from other locations in the Australian range, meaning there are likely two separate reproductive stocks, one in Australia and one in New Zealand. It would be useful to compare samples taken from additional locations in Australia: some genetic studies have reported a genetic distinction between fish stocks from west and east Australia even when the gene flow between Australia and New Zealand was putatively high. This is the case for e.g. gemfish (*Rexea solandri*) (Colgan & Paxton, 1997), kingfish (*Seriola lalandi*), (P. A. Miller et al., 2011; Nugroho et al., 2001), and mako shark (*Isurus oxyrinchus*) (Corrigan et al., 2018).

5.5.4 Neutral genetic structure in New Zealand tarakihi

The analysis of 166,022 neutral genome-wide SNPs conducted in this study indicates that tarakihi have a panmictic genetic population structure throughout their distribution around mainland New Zealand. This may also include Chatham Islands, although the number of samples for this location is too low to test this. No obvious genetic structure related to sampling locations or management areas were detected by any of the methods used (AMOVA, PCA, DAPC, and FastSTRUCTURE) and no significant isolation by distance was detected either. The only significant genetic differences detected were through the pairwise F_{ST} between Wairarapa (south-east of North Island) and four other locations: Upper West Coast South Island, Taranaki, East Northland, and East Cape Figure 5.8.

This is difficult to interpret in terms of geographic stock structure since these four locations are situated at several distant areas around New Zealand and are separated from Wairarapa by other non-significantly divergent locations (e.g. TBGB, Hawke's Bay and Hauraki Gulf) (Figure 5.1). The divergence with East Cape and East Northland could be indicative of a complex, fine-scale migration pattern northward along the east coast of North Island, which would be partially concordant with the results from the mitochondrial study (Papa, Halliwell, et al., 2021). The divergence of Wairarapa compared to Taranaki and Upper West Coast of South Island could be due to a partial lack of connectivity between Wairarapa and the west coast through Cook Strait. However, this seems unlikely since no significant genetic divergence was detected between Wairarapa and Upper West Coast of North Island or Lower

West Coast of South Island. It is worth noting that the genetic divergence found between Wairarapa and these four locations, while still significant after *fdr* correction, is very low ($F_{ST} = 0.0012\text{--}0.0021$). Moreover, while all Wairarapa samples have been sequenced at a good depth and have a “normal” (i.e. expected) level of heterozygosity Figure 5.3, and therefore the relevant individual genetic variation can be reasonably thought to have been captured, the number of individuals from this location is the lowest included in the neutral dataset ($n = 5$ individuals only). The detected genetic divergence might thus be an artefact due to the low number of samples for this location, causing individuals to have a high proportion of fixed allelic differences because the totality of the allelic variation in that location has not been sampled. However, simulation studies have shown that genetic differentiation measured by F_{ST} can still be accurately estimated in populations as small as $n = 4\text{--}6$ when the number of SNPs is above thousands (Willing et al., 2012). No comparison with the study from Gauldie & Johnston (1980) could be done since they did not include samples from around Wairarapa.

This overall lack of population genetic structure for tarakihi around New Zealand is concordant with results from Papa, Halliwell, et al. (2021) where no evidence of genetic structure was detected for the overall New Zealand area, including Chatham Islands. These results appear to consistently suggest that there is one panmictic tarakihi stock in New Zealand, with very high spatial connectivity and no spatial reproductive isolation. This finding is concordant with the characteristics of a species with a high capacity for dispersal, reproduction output, and effective population size, and would also explain the failure to detect any genetic structure among Australian stocks either, despite numerous attempts (Burridge & Smolenski, 2003; N. G. Elliott & Ward, 1994; Grewe et al., 1994; Richardson, 1982).

5.5.5 Adaptive genetic structure

The acquisition of a first and then a second set of putatively adaptive outlier SNPs resulted in the detection of fine-scale genetic structure that was not observed in the neutral tarakihi genetic dataset (Figure 5.6–5.7, Figure 5.9). Interestingly, the presumably adaptive genetic variation detected always discriminated groups at the sampling location level, rather than clustering groups at a broader geographic level (e.g. management area, coast, island).

The first set of 389 presumably adaptive SNPs discriminated Tasmania from New Zealand, meaning that the genetic divergence observed between these two stocks could be due to both physical isolation and local adaptation. They also discriminated Wairarapa and Cape Campbell from all other New Zealand locations (Figure 5.6). While Wairarapa and Cape Campbell are geographically close (south-east of North Island and north-east of South Island, Figure 5.1) it is not obvious if they should be considered together as one or two different adaptive stocks separated from New Zealand: although they formed two different clusters on the PCA and DAPC (Figure 5.6 & Supplementary Figure 5.13), most of the variation of both groups were on the axis 2 of the DAPC (Figure 5.6) and on both the axes 2 and 3 of the PCA (Supplementary Figure 5.13). Moreover, the dendrogram inferred on the pairwise F_{ST} clustered the locations together as outgroup of the rest of New Zealand, even though the F_{ST} value between them ($F_{ST} = 0.1762$) was one of the highest among all pairwise locations (Figure 5.9).

Excluding Tasmania, Wairarapa and Cape Campbell improved the statistical power of the outlier method. With this second dataset of 61 putative adaptive SNPs, the K-mean clustering analysis was able to also discriminate Upper West Coast South Island, Hawke's Bay, and TBGB without any *a priori* grouping. Interestingly, although TBGB and Upper West Coast South Island are both on the west coast of South Island, these three putatively adaptive stocks seem to be actually separated from each other and from the rest of New Zealand (Figure 5.7).

Wairarapa, Cape Campbell, Hawke's Bay, Upper West Coast South Island, and TBGB were found to be genetically divergent from each other and from all of the remaining New Zealand locations, which includes sites as far apart as East Northland, Christchurch, and lower West Coast South Island. If these outlier-based stocks are indeed adaptive, this would indicate that there is no broad separation of adaptive genetic stocks around New Zealand, but rather that the tarakihi population displays some very fine-scale local adaptation to specific areas around New Zealand, that are not, at first look, directly linked to e.g. temperature or depth. This adaptive genetic variation could be school-specific, i.e. representative of genetically adapted small groups, rather than due to large-scale environmental factors. Interestingly, Gauldie & Johnston (1980) also found genetic differentiation between lower and upper west

coast of South Island, although the sampling locations were a bit further apart (Jacksons Bay and Westport). While they also detected a genetic break between TBGB and Cape Campbell, the former was thought to be part of a broader western genetic stock including Upper West Coast South Island and Taranaki and the latter an eastern one including Christchurch and East Cape.

Given these results and the small sample size of some of these locations (especially Wairarapa, $n = 5$ and Hawke's Bay, $n = 7$), it is legitimate to question if these observations are statistical artefacts. The outlier method used here is particularly suited to detect loci under heterogeneous selection and local adaptation (Whitlock & Lotterhos, 2015). Although the two SNP datasets obtained with this method are only putatively adaptive, they are very strong candidates for adaptation because the parameters used to detect them were quite stringent: only loci with $q < 0.05$ and $H_e > 0.1$ were retained, the first to greatly minimize the risk of false positives and the second to discard low-frequency alleles that do not fit the neutral F_{ST} distribution used to find outliers. Moreover, the OutFLANK method does not rely on an *a priori* population model, and is very robust against false positives: in fact, when the number of individuals per location is low, OutFLANK tends to lose power and will produce more false negatives instead of false positives (Whitlock & Lotterhos, 2015). Thus it is likely that the adaptive divergence observed is biologically relevant and not an artefact due to the small sample size of some locations.

5.5.6 Temperature-associated selective cline

Contrarily to the PCA and DAPC performed on the neutral and outlier-based adaptive SNP datasets, the RDA of the reduced quality-filtered SNP dataset was very effective at discriminating samples based on localities (Figure 5.10). Given that only the first axis of the RDA on scaffold 1 was significant, and that 47 out of 55 (85%) of the candidate environmentally-adapted loci were most strongly correlated with mean temperature at mean depth, the variation observed on axis 1 must be mostly driven by adaptation to temperature, or to any other environmental variable that is strongly correlated with temperature (e.g. salinity, pH, dissolved oxygen concentration, among others). Moreover, the samples were ordinated following a latitudinal gradient, which is directly related to temperature on the

continental shelf, where tarakihi occur (Supplementary Figure 5.17). Only TBGB and Wairarapa are not projected with, respectively, the South and the North Island (Figure 5.10). This is explained by the fact that TBGB is actually situated further north than Wairarapa. This means that the observed genetic variation is directly related to temperature (and thus latitude) rather than a theoretical differentiation between the North and the South Island driven by e.g. a dispersal barrier in Cook Strait. This was verified with a linear regression analysis that showed that there was a significant correlation between the ordination on the RDA axis 1 and both the mean temperature and the latitude at sampling locations (Pearson $R^2 = -0.65$, $P < 0.001$ for both variables, Supplementary Figure 5.18). While the next logical step would be to check if some of these SNPs are situated in genomic regions known to be somewhat linked to temperature in fish, a literature search of the associated genes (Supplementary Table 5.3) did not lead to such observations. However, a search of the Gene Ontology terms available for the same genes in zebrafish (*Danio rerio*) showed that four of these genes (transmembrane protein 67, glycoprotein endo-alpha-1,2-mannosidase-like protein, thrombospondin type-1 domain-containing protein 7A, and chondroitin sulfate proteoglycan 5a) are linked to integral components of membranes (three for the plasma membrane and one for the Golgi apparatus membrane).

Finally, to assess if it was possible to explain the presumably adaptive variation observed with the F_{ST} outlier method with environmental adaptation, the outlier dataset 1 (389 SNPs), outlier dataset 2 (61 SNPs), and environment adaptive dataset (55 SNPs) were scanned for overlaps. Only one SNP (base no. 7,132,428, Supplementary Table 5.3) was shared between the outlier dataset 1 and the environment adaptive dataset, and it was also the only SNP from scaffold 1 present in the outlier dataset 2. It was also one of the only five SNPs that correlated most with primary production at the ocean bottom, rather than temperature at mean depth. However, this SNP is situated in an unannotated part of the tarakihi genome, surrounded by several small repeat regions, themselves flanked by an unnamed protein product (blastp:CAF97853.1, ends at base 7,130,362) and a zyxin-coding gene (blastp:XP_006802957.1, start at base 7,137,643). The function of the region associated with this SNP is currently unknown.

5.5.7 Fisheries management implications

It is now well established that the Quota Management Area boundaries for tarakihi in New Zealand do not match the biological stock boundaries (Fisheries New Zealand, 2021; Langley, 2018). The Chatham Islands are usually considered a separate stock from the North and South Island because these two areas are geographically separated by deep water which is not usually inhabited by adult tarakihi (Fisheries New Zealand, 2021; Morrison et al., 2014). Although the main islands are split into several management areas (Figure 5.1), they are not thought to be accurate reflections of the reproductive stock boundaries. Both North and South Islands have been generally considered as one single stock due to the high capacity of post-larvae and adults for dispersal and the lack of evidence of genetic isolation (Fisheries New Zealand, 2021; Morrison et al., 2014). Some specimens tagged in the Kaikoura area on the east coast of South Island have been recaptured on the west coast of North Island (Hanchet & Field, 2001), suggesting that there is at least some level of connectivity between the west and east coasts. Recent studies on trends in age and size structure in TAR1, TAR2 and TAR3 found evidence that the east coast of North and South Island could be one continuous stock (Langley, 2018; J. R. McKenzie et al., 2017). The Canterbury Bight/Pegasus Bay area (east of South Island) seems to be the main nursery area for juveniles of the entire eastern stock. At the onset of maturity, some of the adult fish migrate northward along the east coast to East Cape, then up to the east Northland area. This scenario is backed up by the higher proportion of juveniles in Canterbury Bight/Pegasus Bay and the observed increase in the proportion of older fish from TAR 2, Bay of Plenty and east Northland (Langley, 2018). Subsequent studies have found a similar pattern for the west coast of New Zealand (Fisheries New Zealand, 2021). Observations about the relative strength of individual year classes and growth rates of older fish seem to indicate that Tasman Bay/Golden Bay (TBGB) could act as a nursery area for the entire west coast, which would constitute a single separated stock from the east coast, with a possible lack of connectivity between the upper west coast of North Island and the East Northland fisheries (Fisheries New Zealand, 2021; Langley & McKenzie, personal communication, April 2021).

There is thus strong preliminary evidence that tarakihi are constituted of two main stocks in New Zealand, however, the present analysis of neutral genomic variation does not

support this hypothesis, since no genetic divergence was detected between the west and the east coast, for both islands, even when grouping them using AMOVA. Moreover, the weak genetic divergence that was detected between the west and east coast of the South Island using mitochondrial data (Papa, Halliwell, et al., 2021) could not be corroborated in this study. A similar situation occurred with the Australian stocks: the observation of no genetic structure in Australia reported by Burrridge & Smolenski (2003) was in contrast with two studies that detected three stocks within the south-east of Australia based on otolith microchemistry (Thresher et al., 1994) and larval advection (Bruce, 2001).

When a test for genetic differentiation rejects the model of panmixia, it can usually be used with confidence as a biologically meaningful finding. However, it is more difficult to firmly conclude that stocks are demographically coupled whenever the test fails to reject panmixia (Waples et al., 2008). The demographic structure described by Langley (2018) might not be a reasonable conclusion from the present population genomic analyses because there may be a high enough level of connectivity between stocks to sufficiently homogenise genetic variation, but not enough to provide demographic coupling between areas. Some level of migration between the west and east coasts is highly likely, based e.g. on tagging studies (Hanchet & Field, 2001). Only a few migrants per generation are needed to homogenise most genetic loci, and a large effective population size reduces the speed at which differentiation occurs through the action of genetic drift. Furthermore, any demographic separation between stocks might have arisen too recently for a parallel pattern to also be seen in a test for genetic differentiation. The population genomics findings do not necessarily contradict the findings reported from age and size data. There is still the possibility that there is a western and an eastern stock that may only be partially reproductively isolated, and that some small level of gene flow still occurs and is enough to homogenise genetic variation.

The outlier-based, presumably adaptive differentiation reported here is unlikely to be directly useful for the delineation of fisheries management stocks at this stage, since the genetic variation appears to be highly localized and not reflecting any known broader stock boundaries (Figure 5.1). However, the genotype-environment association analysis hints at the possibility of a North-South adaptive cline in the stock related to water temperature (Figure

5.10, Supplementary Figure 5.18). The pattern of genetic variation found at these loci should be monitored through time to test whether changes in water temperatures have influenced the distribution of alleles over time. The tarakihi population may show an adaptive response to future climate and sea temperature changes, but the consequences of this for the resilience of the fishery is largely unknown.

While the level of population genomic resolution found in this study was high, the number of individuals sampled is relatively low, especially in some locations because very few samples could be obtained. Simulation studies have shown that genetic structure can sometimes be detected by only sampling a few individuals per population (c. 8–10) as long as the genome-wide SNP density is high (Jeffries et al., 2016; Nazareno et al., 2017). However, the power of clustering methods (i.e. K-means clustering for DAPC) can drop sharply when a small number of individuals is used (Waples & Gaggiotti, 2006), and given the high effective population size and the high level of gene flow in New Zealand tarakihi, a much larger DNA sampling and sequencing campaign might be necessary to provide fisheries managers with a definitive answer of the presence or absence of genetic stock structure among tarakihi in New Zealand.

5.6 Conclusion

This study is the first population genomics analysis of tarakihi (*Nemadactylus macropterus*) and one of the very first genome-wide analyses of a New Zealand marine species. The acquisition and subsequent filtering of a large SNP dataset allowed for the detection of a low but highly significant genetic differentiation between Tasmania (and thus putatively the whole of Australia) and New Zealand. No neutral genomic structure was detected among New Zealand locations, which means the genomic data did not support the hypothesis of two separate reproductive stocks on the west and east coast of New Zealand, but it could not reject it either. A latitudinal adaptive cline strongly correlated to water temperature was found. The associated loci are strong candidates for further investigation to identifying their functional role in the adaptation to the environment.

Tarakihi is a commercially important fishery but it has been reported as declining. Implementation of routine genomic sampling could enhance spatial and temporal genomic resolution. This could be implemented using a long-term, standardised, well archived genetic-based tagging program (Mace et al., 2020; Papa, Oosting, et al., 2021). Integrating results obtained from genetic-tagging data with life history information and results obtained from other fisheries assessment methods will prove extremely valuable for the long-term sustainable use of this important fishery species.

5.7 Supplementary Material

Supplementary Figure 5.1. Example of gel electrophoresis containing samples of high (1–5, 11), medium (10 & 12), and low (6–9) quality.

Supplementary Figure 5.2. Plot of outlier SNPs on scaffolds 1-6. The F_{ST} of all SNPs (grey dots) was calculated by OUTFLANK and plotted across each scaffold. Blue circle dots are SNPs that have been selected as outliers after filtering for $q < 0.005$, $H_e > 0.1$ and highest F_{ST} value in a non-overlapping 50,000 bp sliding window.

Supplementary Figure 5.3. Plot of outlier SNPs on scaffolds 1-6, with only SNPs with heterozygosity > 0.01 plotted. The F_{ST} of all SNPs (grey dots) was calculated by OUTFLANK and plotted across each scaffold. Blue circle dots are SNPs that have been selected as outliers after filtering for $q < 0.005$, $He > 0.1$ and highest F_{ST} value in a non-overlapping 50,000 bp sliding window.

Supplementary Figure 5.4. Heatmap of Person correlation coefficients among the 48 environmental variables compiled in the second dataset.

Supplementary Figure 5.5. Correlation plot of the environmental variables after discarding for correlation with mean temperature at mean depth (left), and the final dataset of six lowly correlated variables used in the locus-environment association analysis (right). Colour scale and size of dots correspond to Pearson coefficient value. Dots are plotted only for significant values ($P \leq 0.05$).

Supplementary Figure 5.6. Some of FastQC quality metrics results synthesized by MultiQC for DNA samples of 188 tarakihi and king tarakihi specimens. Top: Per base sequence quality. Middle: Per base sequence base content. Bottom: GC distribution of reads.

Supplementary Figure 5.7. Linkage disequilibrium decay over genetic distance on scaffold 1, calculated on the quality-filtered SNP dataset, including king tarakihi specimens. The horizontal red dashed line shows the threshold of 0.2 that is commonly applied to identify independent degradation of nucleotide sites. The orange line is the background level of linkage disequilibrium (intercept). The blue line is the trend of linkage disequilibrium decay fitted to the plot (minimum and maximum variance in dashed lines for the mean).

Supplementary Figure 5.8. Results of the *K*-means clustering analysis performed on the pruned SNP dataset that includes 183,443 loci from 188 tarakihi and king tarakihi.

Supplementary Figure 5.9. Discriminant analysis of principal components of the pruned SNP dataset that includes 183,443 loci from 188 tarakihi and king tarakihi. Top: Axes 1 and 2. Bottom left: Axes 3 and 4. Bottom right: Eigenvalues. Groups correspond to the inferred groups in Supplementary Figure 5.8.

Supplementary Figure 5.10. Principal component analysis of the neutral SNP dataset that includes 166,022 loci from 165 tarakihi. Ellipses represent the 95% confidence intervals. Left: Axes 1 and 2. Top right: Axes 3 and 4. Bottom right: Eigenvalues. Sampling location codes as referred to in Table 5.1.

Supplementary Figure 5.11. Results of the *K*-means clustering analysis performed on the neutral SNP dataset that includes 166,022 loci from 165 tarakihi.

Supplementary Figure 5.12. Discriminant analysis of principal components of the neutral SNP dataset that includes 166,022 loci from 165 tarakihi. Top: Axes 1 and 2. Bottom left: Axes 3 and 4. Bottom left: Eigenvalues. Groups correspond to the inferred groups in Supplementary Figure 5.11.

5.7 Supplementary Material

Supplementary Figure 5.13. Principal component analysis of the presumably adaptive SNP dataset that includes 389 loci from 165 tarakihi. Ellipses represent the 95% confidence intervals. Top: Axes 1 and 2. Bottom: Axes 3 and 4. Middle: Eigenvalues. Sampling location codes as referred to in Table 5.1.

Supplementary Figure 5.14. Principal component analysis of the second presumably adaptive SNP dataset that includes 61 loci from 140 tarakihi. Ellipses represent the 95% confidence intervals. Top: Axes 1 and 2. Bottom: Axes 3 and 4. Bottom left: Eigenvalues. Sampling location codes as referred to in Table 5.1.

Supplementary Figure 5.15. Heatmap of pairwise weighted F_{ST} estimates (corresponding to the values above and below the diagonal) among sample locations of the neutral SNP dataset. The dendrogram shows the inferred relationship between populations. Significant P -values (≤ 0.05 after false discovery rate correction) are in bold.

Supplementary Figure 5.16. Heatmap of pairwise weighted F_{ST} estimates (corresponding to the values above and below the diagonal) among sample locations of the presumably adaptive SNP dataset. The dendrogram shows the inferred relationship between populations. Significant P -values (≤ 0.01 after false discovery rate correction) are in bold.

Supplementary Figure 5.17. Plots of values of the six environmental variables used in the Redundancy Analysis (RDA).

5.7 Supplementary Material

Supplementary Figure 5.18. Linear regression of the axis 1 coordinates of the RDA in relation to mean temperature (top) and latitude (bottom) of sampling locations.

Supplementary Table 5.1. Parameters used in the K-mean clustering and DAPC analyses.

	find.clusters			dapc	
	max.n.clust	n.pca	n.clust	n.pca	n.da
Pruned	100	150	9	50	8
Neutral	100	150	15	50	14
Adaptive 1	14	150	6	50	5
Adaptive 2	14	150	5	50	4

Supplementary Table 5.2. Groups of environmental variables obtained from Bio-ORACLE and MARSPEC.

Group	<i>n</i> variables
bathymetry	6
calcite	1
carbon phytoplankton biomass	24
chlorophyll	22
cloud fraction	3
current velocity	24
diffuse attenuation	3
dissolved oxygen	25
ice cover	12
iron concentration	24
light penetration	18
nitrate	25
pH (acidity)	1
phosphate	25
photosynthetically available radiation	2
primary production	24
salinity	25
sea surface salinity	17
sea surface temperature	20
sea water temperature	24
silicate	25
topology (MARSPEC "biogeo")	7

Supplementary Table 5.3. Loci candidates for local adaptation according to the genotype-environment association analysis conducted on scaffold number 1.

SNP coordinate	BLASTP label / ID	Most correlated variable	Region	Gene name	Annotation (based on BLATSP best match)
1051143	blastp:XP_027142094.1	mean temperature	gene	LOC104939922	regulation of nuclear pre-mRNA domain-containing protein 2 isoform X2 [<i>Larimichthys crocea</i>]
1507071		mean temperature	repeat		
1586380	blastp:XP_023122240.1	primary production	gene	flad1	FAD synthase [<i>Amphiprion ocellaris</i>]
1626325	blastp:XP_022612259.1	mean temperature	gene	LOC111230007	pre-B-cell leukemia transcription factor-interacting protein 1-like [<i>Seriola dumerili</i>]
2197375	blastp:XP_022621898.1	mean temperature	gene	rad54b	DNA repair and recombination protein RAD54B [<i>Seriola dumerili</i>]
2201535		mean temperature	none		
2238075	blastp:XP_008288143.1	mean temperature	gene	tmem67	PREDICTED: meckelin [<i>Stegastes partitus</i>]
2364205		mean temperature	repeat		
2798928		mean temperature	repeat		
2879291	blastp:XP_023148924.1	mean temperature	gene	mindy3	ubiquitin carboxyl-terminal hydrolase MINDY-3 [<i>Amphiprion ocellaris</i>]
2879525	blastp:XP_023148924.1	mean temperature	gene	mindy3	ubiquitin carboxyl-terminal hydrolase MINDY-3 [<i>Amphiprion ocellaris</i>]
3335367		mean temperature	none		
3577629	blastp:XP_010738915.1	iron concentration	gene	ankrd28	serine/threonine-protein phosphatase 6 regulatory ankyrin repeat subunit A isoform X2 [<i>Larimichthys crocea</i>]
3991506	blastp:XP_023138899.1	mean temperature	gene	LOC111577054	potassium channel subfamily K member 5-like [<i>Amphiprion ocellaris</i>]
4318704	blastp:XP_008283225.1	mean temperature	gene	kmt2b	PREDICTED: histone-lysine N-methyltransferase 2B isoform X1 [<i>Stegastes partitus</i>]
4504812		primary production	repeat		
4560533		mean temperature	none		
4655135	blastp:XP_023275606.1	mean temperature	gene	LOC111664994	mediator of RNA polymerase II transcription subunit 12-like [<i>Seriola lalandi dorsalis</i>]
4655175	blastp:XP_023275606.1	mean temperature	gene	LOC111664994	mediator of RNA polymerase II transcription subunit 12-like [<i>Seriola lalandi dorsalis</i>]
6111582		salinity range	none		
6174709		mean temperature	none		

6217015		mean temperature	none		
6404576		mean temperature	none		
6444298		mean temperature	repeat		
6559859		mean temperature	none		
6640383	blastp:XP_005450598.1	mean temperature	gene	e2f3	transcription factor E2F3 isoform X2 [<i>Oreochromis niloticus</i>]
7018261	TARdn00343	mean temperature	gene		Unannotated gene with multiple isoforms
7132428		primary production	none		
7200389	blastp:AXY87852.1	mean temperature	gene	NA	glyceraldehyde-3-phosphate dehydrogenase [<i>Lateolabrax maculatus</i>]
7333901		mean temperature	none		
7857016	blastp:XP_008298216.1	primary production	gene	dedd2	PREDICTED: DNA-binding death effector domain-containing protein 2 isoform X1 [<i>Stegastes partitus</i>]
8485657	blastp:XP_018527968.1	iron concentration	gene	lipe	PREDICTED: hormone-sensitive lipase isoform X2 [<i>Lates calcarifer</i>]
8559826		mean temperature	repeat		
8720370	PB.386.8 TRPT%2F43556	mean temperature	?		Unannotated single transcript
9111300	blastp:XP_019122739.1	mean temperature	gene	ppp1r8	nuclear inhibitor of protein phosphatase 1 [<i>Larimichthys crocea</i>]
9157049	blastp:XP_018554147.1	mean temperature	gene	maneal	PREDICTED: glycoprotein endo-alpha-1 2-mannosidase-like protein [<i>Lates calcarifer</i>]
9191706	blastp:XP_019122759.2	mean temperature	gene	inpp5b	type II inositol 1 4 5-trisphosphate 5-phosphatase isoform X1 [<i>Larimichthys crocea</i>]
9603410	blastp:XP_022620042.1	mean temperature	gene	oscp1	protein OSCP1 isoform X3 [<i>Seriola dumerili</i>]
9694705	blastp:XP_010729940.3	mean temperature	gene	LOC104919580	thyroid hormone receptor-associated protein 3 isoform X1 [<i>Larimichthys crocea</i>]
9763086		primary production	none		
9879614		mean temperature	none		
10423010		mean temperature	none		
10889404		mean temperature	none		
10912349	blastp:XP_017578417.1	mean temperature	gene	etfb	PREDICTED: electron transfer flavoprotein subunit beta [<i>Pygocentrus nattereri</i>]
11037901	blastp:XP_010743327.2	mean temperature	gene	phldb3	pleckstrin homology-like domain family B member 3 [<i>Larimichthys crocea</i>]
11381521	blastp:XP_010792034.1	mean temperature	gene	LOC104964847	PREDICTED: lysosomal thioesterase PPT2-like [<i>Notothenia coriiceps</i>]

5.7 Supplementary Material

11455617	blastp:XP_027142118.1	mean temperature	gene	LOC104933338	gamma-aminobutyric acid type B receptor subunit 1 isoform X2 [<i>Larimichthys crocea</i>]
11937663	blastp:XP_022602129.1	mean temperature	gene	mtmr11	myotubularin-related protein 11 [<i>Seriola dumerili</i>]
12151987		mean temperature	repeat		
12417477	blastp:XP_008279392.1	mean temperature	gene	LOC103356862	PREDICTED: raftlin-like isoform X1 [<i>Stegastes partitus</i>]
13020685	blastp:XP_014876155.1	mean temperature	gene	LOC106938575	PREDICTED: septin-7 isoform X2 [<i>Poecilia latipinna</i>]
13170301	blastp:XP_022056912.1	mean temperature	gene	thsd7a	thrombospondin type-1 domain-containing protein 7A isoform X3 [<i>Acanthochromis polyacanthus</i>]
13299508	blastp:XP_024241718.1	mean temperature	gene	LOC112223020	von Willebrand factor D and EGF domain-containing protein-like [<i>Oncorhynchus tshawytscha</i>]
13483672		mean temperature	repeat		
13868517	blastp:XP_026229042.1	mean temperature	gene	cspg5a	chondroitin sulfate proteoglycan 5 isoform X1 [<i>Anabas testudineus</i>]

Chapter 6. General Discussion

This thesis provided the first in-depth study of the genome evolution and population structure of the native marine teleost tarakihi (*Nemadactylus macropterus*). By analysing genome-wide genetic variation the main goals of this thesis were addressed, which were to make inferences about the characteristic and evolutionary trends of genomic features, the effect of selection, and the population structure (Figure 1.9). In the present chapter, the main findings of this research are summarised and the limitations are highlighted. The relevance of these findings and associated research perspectives are then discussed. This chapter is concluded with some recommendations for further research to expand the present findings.

6.1 Main findings

6.1.1 There is a general lack of studies on the population genetic structure of fishery species in New Zealand

Chapter 2 provided a summary of the current state of knowledge of genetic stock structure of fishes in New Zealand. The scientific literature was thoroughly reviewed for all cartilaginous and bony fish species up until 2019. It was found that out of the 86 fish species that are managed under the Quota Management System with a total allowable commercial catch limit in New Zealand, only 23 have studies available on their population genetic structure. Of the 23 studies listed several had too few samples with a good enough geographic range to be able to make firm conclusions (especially e.g. alfonsino, blue mackerel, gemfish, and Pacific bluefin tuna, with only two sampled locations). Others applied what are now considered outdated genetic markers techniques, which meant they do not allow for any definitive conclusions to be made about the level of reproductive isolation among stocks (e.g. tarakihi, trevally, barracouta, red cod, jack mackerel, and gurnard, with allozymes that could be subject to selection). At the time of the review, no study had been published on a New Zealand fish species using genome-wide DNA sequencing data. The paucity of comprehensive population genetic studies is most likely the reason why few if any species have an administrative stock structure based on integrative evidence that includes genetics. Moreover, when spatial genetic breaks were found, they rarely matched the administrative

stock boundaries (e.g. probably at least three stocks of orange roughy in ORH3B with Puysegur Bank, Otago, and Chatham Rise, or at least two stocks of ling in LIN6). An exception to this was snapper, for which the inferred genetic stock boundaries have been somewhat consistent among studies, and are also consistent with findings reported using other methods of stock identification. There is a partial match with administrative boundaries (SNA7 vs. SNA8 vs. SNA1).

It was concluded that the lack of strong evidence of genetic stock boundaries found in most fishes could be explained by three reasons. First, the result may simply be the biological reality that stocks are part of a large, genetically panmictic population. Second, the species were made up of structured populations, but the genetic markers used lacked the level of resolution needed to detect genetic differentiation. Third, the sampling design of locations and specimens was limited and therefore has insufficient statistical power to detect any of the differences. The second and third points can be addressed by conducting genome-wide sequencing and using a large spatial and temporal sampling design. Using genome-wide sequencing data from a large number of specimens and a broad mix of locations would also provide enough confidence that the null model of panmixia cannot be rejected and that panmixia is the best population genetic model to apply to the studied species.

6.1.2 Genomics is a promising but under-utilized tool for fisheries management

The second part of Chapter 2 was a brief review of the application of genetics and genomics for fisheries management. It was highlighted that the potential applications of these methods extend far beyond the simple delineation of reproductive stocks. It emphasized that one of the biggest strengths of genomics is to be able to detect selection acting on the genome and to distinguish between presumably adaptive and neutral variation. This has previously been a difficult task using only traditional genetic techniques. In the review, four critical issues related to the sustainability of fisheries that will benefit from the help of genomics were briefly introduced and discussed. (1) The loss of genetic diversity and adaptive potential can make populations more vulnerable to issues related to e.g. inbreeding depression and changes in environmental conditions. (2) Fishing-induced evolution still needs to be properly assessed in commercially exploited wild populations to determine whether

size-selecting fishing methods have a detrimental impact on fish size and maturation time. (3) Climate change could result in adaptive responses and changes to population ranges, which could produce founder events as new suitable habitats develop. This will need to be monitored and the impacts on fisheries managed. (4) Measures of abundance and connectivity within and among populations could be estimated using the non-destructive sampling methods of biopsy hooks to conduct genetic tagging using close-kin mark-recapture analyses. The review chapter ended with recommendations for how genomics could be best incorporated into New Zealand fisheries management decisions. I emphasized the importance of long-term planning of spatial and temporal sampling for each fishery, integration in a tagging context, standardization and archiving of the genetic data and associated data at a national scale, and a possible priority to valuable deep-water species or candidate species for aquaculture.

6.1.3 High heterozygosity and prevalence of intron retention events in the genome of tarakihi

In Chapter 3, the first sequenced and assembled genome of the tarakihi was presented. The most contiguous and complete assembly was obtained using a combination of short Illumina reads and long Nanopore reads, assembling with Flye (the latest algorithm of the MaSuRCA assembler) and polishing with POLCA. The unpolished Flye assembly was only slightly less contiguous and complete, and the CABOG assembly was twofold less contiguous. Moreover, the CABOG assembly contained some unmerged haplotigs and the longest scaffold seemed to have misassembled scaffolds from two different chromosomes. The assembly based on PacBio HiFi reads and assembled with hifiasm had the poorest quality, but this is most probably due to the insufficient amount of coverage provided by this sequencing data. The contiguity and the completeness of the final assembly (1,214 scaffolds, N50 = 3.4 Mb, BUSCO = 97.80%) was particularly good when compared with other reported genome assemblies based on a combination of Illumina and Nanopore reads. Heterozygosity of the tarakihi was estimated separately using *k*-mer frequency and proportion of variants. While these values differed by almost twofold (1.00% and 0.64% respectively), they were both particularly high when compared to the same values reported for other fish species.

The tarakihi transcriptome data was successfully obtained by sequencing full RNA transcripts from four tissue types of a second tarakihi individual. The reads were used to assist in the annotation of the genome and characterize isoforms. Most of the genome features of tarakihi (size, repeat elements (RE), and gene counts) were consistent with other fish genomes. The length of the genome was 568 Mb. This is consistent, albeit a bit on the lower end, with the size range of Percomorpha species. REs and genes were successfully annotated and were used for comparison with other fish genomes in the following chapter (see below). A total of 93,949 full-length, non-redundant Iso-Seq transcripts were characterized from the transcriptome data. Among these, 2,347 were classified as transcripts from REs. The remaining 91,313 transcripts came from 15,515 of the annotated 20,169 protein-coding genes. There was an average of 6 transcripts per gene, with a median of 3 and a maximum of 211. The two genes that produced the most transcripts (211 and 164) were both coding for collagen alpha chains implicated in the structural integrity of the cellular matrix. Finally, a total of 26,644 splicing events were found among the isoforms. The most common splicing event by far was the retention of introns (46% of all events). A prevalence of intron retention events was also reported in the literature for the goldfish (*Carassius auratus auratus*), although the difference in proportion with all other events was even more marked for the tarakihi.

6.1.4 The proportion of repeat elements in percomorph fishes is correlated to the size of their genome

In Chapter 4, Illumina reads were used to assemble the genome of five other Percomorpha species native to New Zealand, the king tarakihi and blue moki (Centrarchiformes: Latridae), the butterfish (Labriformes: Labridae), the barracouta, and the kahawai (Scombriformes). The MaSuRCA assembler produced satisfactory results. Contiguity and completeness were variable among species but consistent with what is expected for assemblies based on Illumina short reads only (58,102–150,595 scaffolds, N50 = 10–30 Kb, BUSCO = 70.20–89.10%). The characteristics of these five genomes were then compared with the genome of the tarakihi reported in Chapter 3. The genome sizes were associated with the taxonomy: the smallest genome was the one of the butterfish (532 Mb), the three Latridae species had similar sizes (568, 576, and 560 Mb for the tarakihi, king tarakihi and blue moki),

and the two longest genomes were the ones of the two Scombriformes species (kahawai = 636 Mb and barracouta = 714 Mb). The total proportion of REs in the six genomes varied from 24.83% (butterfish) to 39.12% (barracouta). The number of annotated genes ranged from 20,169 (tarakihi) to 24,816 (butterfish).

The effects of completeness (as a proxy for genome quality) on the identification of REs and genic features were explored. All genome features metric tested (e.g. proportion of REs and number, total length, and proportion of genes, exons, introns and untranslated regions) were strongly correlated with BUSCO completeness ($|R^2| = 0.79\text{--}0.97$), except notably for the proportion of REs. Conversely, the proportion of REs was the only genome feature significantly correlated with the genome size ($R^2 = 0.97$, $P < 0.01$). Comparison of the proportion of identified RE classes showed that the relative repartition of classes was similar among the six species; DNA transposons were the dominant type. Among these, the hAT-Ac family was the most highly represented family of REs in all species, except the butterfish. By comparison, the butterfish had a higher representation of the Ngaro family of LTR retrotransposons.

6.1.5 Genes in the ATP binding pathway are especially subject to selection among species

A phylogenetic tree of Percomorpha was reported in Chapter 4. This was constructed using 1,481 single-copy orthogroups sequences from the six newly assembled New Zealand species and eight additional well-curated fish genomes. The phylogenomic tree was overall consistent with recently reported molecular phylogenies of Percomorpha and all orders were monophyletic. The radiation of Percomorpha was estimated to have occurred c. 112 million years ago (mid-Cretaceous). Evidence of positive selection in the Latridae and tarakihi lineages was successfully detected using the branch-site model test implemented in PAML. The effect of positive selection was detected in 209 genes of the Latridae lineage. Of the 65 genes under positive selection in tarakihi, only six of them were not also under selection in Latridae as a whole. The most common functions of these genes were related to the integral structure of cellular membranes and the ATP binding pathway. Other common functions of

genes under selection included metal ion binding, zinc ion binding, and regulation of transcription by RNA polymerase II.

6.1.6 Tarakihi in New Zealand are a genetically panmictic population with a distribution that follows a temperature-associated cline

Whole-genome resequencing of 175 tarakihi sampled from around New Zealand and Tasmania, as well as 12 king tarakihi, produced a dataset of 7.5 million high-quality SNPs. King tarakihi displayed high levels of genetic divergence, differences in heterozygosity, and differences in linkage disequilibrium patterns that were all consistent with reproductive isolation from tarakihi from New Zealand and Tasmania. These results were consistent with previous studies that proposed king tarakihi as a different species from tarakihi, but a proper taxonomical description still needs to be conducted. A low but significant level of genetic differentiation ($F_{ST} = 0.0054\text{--}0.0073$, $P \leq 0.05$) and strong evidence of population structure was found between tarakihi populations from Tasmania and New Zealand. These findings suggested that the tarakihi population in Tasmania and New Zealand are partially reproductively isolated from each other. There was no evidence of neutral population structure among New Zealand sampling sites ($\Phi_{ST} < 0.001$, $P = 0.77$), indicating that tarakihi are composed of one single panmictic stock in New Zealand. However, some fine-scale spatial structure was detected around central New Zealand off the east (Wairarapa, Cape Campbell, and Hawke's Bay) and the west coast (Tasman Bay/Golden Bay and Upper West Coast of South Island) when considering the outlier-defined presumably adaptive SNPs. This could indicate the presence of a weak level of local adaptation that is confined to specific areas around New Zealand. This genetic variation might be representative of small groups of locally adapted individuals. Finally, allele frequencies of 47 loci in the New Zealand population were strongly correlated with mean depth temperature. This indicated the existence of a latitudinal temperature cline in New Zealand tarakihi. Of the 47 SNPs associated with temperature, 18 were situated in annotated protein-coding genes. Of these, four of these genes are associated with integral components of the cellular membranes.

6.2 Relevance and prospects

6.2.1 Implication for tarakihi fisheries in New Zealand

Sustainable fisheries require the proper identification of the reproductive boundaries among stocks in order to align them with administrative boundaries (Begg et al., 1999; Cadrin, 2020; Kerr et al., 2017; Waples et al., 2008). This is a fundamental requirement for effective assessment and management. This is particularly the case when the fisheries regulations rely on maximum catch limits to maintain sustainable yield, like in New Zealand: correctly matching reproductive and administrative stocks permits to reduce the risks of depletion that could occur due to differences in abundances among the biological units or an uneven distribution of the harvesting pressure (Benestan, 2019; Reiss et al., 2009). Knowledge of the genomic structure and genomic diversity also enables long-term management strategies to be developed to preserve the genetic variation and evolutionary potential of these fisheries. Genetic diversity is often an overlooked aspect of the capacity for long-term resilience of fisheries (Benestan, 2019; Bernatchez et al., 2017; Cuéllar-Pinzón et al., 2016; Ovenden et al., 2015; Waples et al., 2008; Ward, 2000).

Tarakihi is one of the most heavily harvested fisheries species in New Zealand (Fisheries New Zealand, 2019b). This makes it a fish species that is both commercially valuable but also potentially vulnerable from a conservation and sustainability point of view. According to the results of this study, tarakihi stocks in New Zealand are composed of a single panmictic population that is genetically differentiated from tarakihi in Australia, and both are distinct from king tarakihi. This finding is consistent with what is known of the biology of this species. Due to their particularly long pelagic larval stage and the high dispersal capacity of adults, connectivity and gene flow are expected to be common across their range. The findings reported in this thesis were also concordant with previous studies of the genetic stock structure of tarakihi among sites around New Zealand (Papa, Halliwell, et al., 2021) and Australia (N. G. Elliott & Ward, 1994; Grewe et al., 1994; Richardson, 1982). All of these studies concluded that there was genetic panmixia of tarakihi populations in both (although not between) New Zealand and Australia. The main advances that the present study has made are (1) the informative SNPs used were known to be neutral and therefore allowed the results to be simply interpreted in terms of migration and gene flow (as opposed to selection) and (2) the genome-wide neutral SNP dataset was able to provide the highest possible level of

resolution compared to any other genetic marker set used on this species. This makes the findings reported in this thesis research particularly robust when compared to prior studies.

The results of the present study do not support the hypothesis that tarakihi are constituted of two main stocks in New Zealand, one on the west coast and one on the east coast (Fisheries New Zealand, 2021; Langley, 2018). However, the genomic data cannot conclude that there is also demographic coupling between these areas. It only requires a few migrants per generation between areas to homogenise most genetic loci. Moreover, a large effective population size increases the time that is taken for genetic differentiation to arise through the independent action of genetic drift in each area. If other studies later propose that tarakihi in New Zealand is composed of two demographically uncoupled stocks this would not conflict with the findings of this population genomics study. There is evidence that there is still a level of connectivity between these stocks and that it is enough to result in maintaining genetic homogeneity. A similar study was recently conducted on hoki, the most valuable fisheries of New Zealand, using a similar approach of genome-wide variation data (Koot et al., 2021). Hoki are also thought to be composed of two partially reproductively isolated stocks on the west and the east coast based on non-genetic evidence (see Chapter 2). However, results from the population genomic study resulted in a similar conclusion that for tarakihi: panmixia in New Zealand and genetic differentiation with Australia (Koot et al., 2021).

The presumably adaptive genetic differentiation detected based on outlier loci for some specific locations (Wairarapa, Cape Campbell, Hawke's Bay, Tasman Bay/Golden Bay and Upper West Coast of South Island) might not be directly relevant for the delineation of fisheries management stocks. The genetic variation appears to be highly localized and does not reflect any broader stock boundaries. However, genotype-environment association analysis led to strong evidence for a North-South adaptive cline related to water temperature across the tarakihi range. This type of temperature-associated gradient has been detected in several other marine species (e.g. the American lobster (Benestan et al., 2016; Sandoval-Castillo et al., 2018)) and seems to represent a common pattern in wild fisheries with large populations and/or connectivity. The pattern of genetic variation found at these loci should be monitored through time to test whether changes in water temperatures or other

environmental variables correlated with temperature have influenced the distribution of alleles over time. Gene flow may accelerate an adaptive response to climate change by providing relevant genetic variation to the tarakihi population, but this is still only speculative. The impact of future changes to environmental conditions on the resilience of the tarakihi fishery is still largely unknown.

Tarakihi stocks in New Zealand are currently considered to be in a vulnerable state. The stocks on the east coast appear to be overfished or depleted (Langley, 2018). In response, the total allowable catch limits have been reduced for the first time in 2018 and again in 2019 for these areas, as part of a plan to rebuild the fishery. The finding of the present study, that tarakihi are panmictic in New Zealand, could be beneficial to the implementation of that rebuilding strategy. The high levels of gene flow may mean that the reduction in size of the east coast population has not yet led to a genetic bottleneck. Moreover, given enough time, the eastern stocks could be replenished by the recruitment of individuals from neighbouring stocks through a “rescue effect” mechanism (Benestan, 2019). However, it is crucial to continue to monitor the genetic diversity and structure of tarakihi fisheries and use that information to prevent the detrimental effects of inbreeding depression and loss of adaptability due to the reduction in effective population size caused by overfishing and the loss of habitat due to climate change. Since the quota size of the east coast fishery has been reduced, it can be expected that the harvesting efforts will shift focus on the west coast fishery, which has been less well studied because of historically low catches from that area. It is thus important to begin sampling and monitoring the west coast.

6.2.2 Comparative genomics: towards a better understanding of genomic evolution

A modern understanding of the drivers that shape the diversity of life on Earth started in the mid-nineteenth century, with Darwin’s theory of evolution and Mendel’s classic experiments on the heritability of traits. However, at the time, neither of them had the means to know where the information on trait variation was based and the mechanism that allowed it to be passed on to the next generation. Fast-forward about 170 years, we now know that the genetic information of living beings is stored and transmitted in the DNA and that micro-

evolutionary changes in the genome result in the macro-evolution of individuals, populations, and species. However, there is still a surprisingly big lack of knowledge of the diversity of genomic characteristics among organisms, and how these differences might affect, and be affected by macro-evolutionary processes. For example, the extent of selection on genomes compared to neutral processes is still debated today (R. Nielsen, 2005). Until relatively recently, we thought erroneously that the size of the genome should be correlated with the “apparent complexity” of an organism, but studies showed that this is not the case (the C-value paradox) (T. A. Elliott & Gregory, 2015). This then shifted to the idea that this “apparent complexity” may be correlated to the number of coding genes in the genome. Again, evidence proved that hypothesis wrong (Claverie, 2001; T. A. Elliott & Gregory, 2015). For decades, the numerous repetitive elements found in the genome were thought to be mostly “junk DNA”. Today, we are just beginning to understand the staggering diversity of transposable repeat elements and the direct impact they can have on the evolution of genomes (Biémont, 2010; Biémont & Vieira, 2006). Finally, the emerging field of transcriptomics brings another revolution to advance our understanding of how genomic information is not only stored and transmitted, but also interpreted and translated into molecular functions, and the multiple factors that influence these processes.

Comparative genomic studies are still in a state of relative infancy because the necessary technology to address questions in this field has only recently started to emerge and become affordable (Ellegren, 2008; W. Miller et al., 2004). The genomic resources produced in this study (in the form of genome assemblies) and the resulting findings are additional steps towards this goal. Thanks to the use of full-length transcripts sequencing technology, Chapter 3 is one of the very first studies to report the diversity of splicing events in a fish. Intron retention was thought, until recently, to be the least prevalent alternative splicing form in animals, but our results are part of growing evidence showing that this may not be the case (Q. Gao et al., 2020; X. Wang et al., 2019). Interestingly, intron retention events were more abundant in the tarakihi and the goldfish, while they were much less abundant than alternative 5'/3' splice sites events in the bat (see Chapter 3). More studies of alternative splicing in a diverse range of species will be needed to infer if this finding is linked to a difference between fish and mammals.

The proportion of repetitive elements were found to be positively correlated with the size of the genomes. This pattern has already been observed in a broad number of fish taxa (B. Gao et al., 2016; Shao et al., 2019; Z. Yuan et al., 2018). Integrating these results into the growing amount of repeat element information for other species could allow investigation into the evolution of repeat families in fishes and other vertebrates, exploring e.g. the loss and/or gain of new RE across lineages or if there is a relationship between the amount and the diversity of types of REs in genomes. The metrics for all genic features investigated (genes and their components, notably exons and introns) were significantly correlated with the genome assembly completeness. This result seems to indicate that while scaffold-level assemblies as produced in Chapters 3 and 4 are suitable for population genomics studies, characterization of repeat elements, and identification of selection patterns, they are less suitable for a biologically meaningful comparison of genic features. A lower level of completeness for the genome is usually indicative of a higher amount of fragmentation (Jauhal & Newcomb, 2021), which might lead to a bias towards the detection of smaller genes and gene elements. Subtle changes in the genome assembly pipelines could also have an effect on gene detection. The genome assemblies and annotation pipeline used in this study might thus not be ideal for comparative analyses of genes and their components. If that is the goal for a study, it might be necessary to invest more resources to add a scaffolding method (e.g. Hi-C data) to obtain chromosome-level assemblies to lessen this bias.

Finally, the present study was able to detect patterns of selection in specific regions of the genomes, both at the levels of species (Chapter 4) and populations (Chapter 5). These regions under selection were different depending on the method used, i.e. d_N/d_S ratio, outlier analysis, and locus-environment association analysis. The functions of these regions were retrieved by using the genome annotations done in Chapter 3 and Chapter 4 and looking for the associated GO terms in the UniProt database. Interestingly, the most common functions for these genes were their implication in the structural integrity of cellular membranes and/or the ATP binding pathway. ATP binding is required to provide energy for the active transport of ions across cell membranes (Melkikh & Seleznev, 2012). The ATP-Binding Cassette transporter genes are known to be associated with numerous pathologic disorders (Dean & Annilo, 2005). Interestingly, genes in the ATP binding pathway were also the most common genes found to be under positive selection in a former study based on molecular evolution

rates of four model fish species (Steinke et al., 2006) and were also found to be positively selected in other fish species, like the largemouth bass (Sun et al., 2021). The fact that positive selection was detected could indicate that variation in the ATP binding pathway can have a disproportionate effect on the fitness of individuals compared to other molecular functions. Specific phenotypic variation in these genes might be especially beneficial when faced with new environments, leading to a rapid spread of nonsynonymous mutations when species encounter new selective pressures. As found in Chapter 5, these changes in the ATP binding pathway could be related to temperature, or to other environmental variables (e.g. salinity, oxygen concentration, pH) that are directly related to temperature.

6.3 Recommendations for future research

Since the advent of high-throughput sequencing methods over a decade ago, sequencing technologies and downstream bioinformatics tools have been improving at a fast pace. When this project started, the best and latest standard for genome assemblies of non-model organisms was the hybrid assembly of short accurate reads with longer, less accurate reads to produce scaffold-levels assemblies (e.g. Austin et al., 2017; Dhar et al., 2019; Jiang et al., 2019; Tan et al., 2018; Wiley & Miller, 2020; Zimin, Puiu, et al., 2017; Zimin, Stevens, et al., 2017). Only four years later, following the rapid development of new technologies for scaffolding data (e.g. Hi-C), genome assemblies at the chromosome level can now be more easily achieved (e.g. Q. Gao et al., 2020; Gong et al., 2018; Simison et al., 2020; Sun et al., 2021; Yang et al., 2019; Zhang et al., 2020). While useful for the applications used in this study, the tarakihi genome assembly produced for this thesis could be improved with further Hi-C scaffolding. That would allow the refinement of the annotation of the genetic features and improve the range of possible analyses in comparative genomics studies. The same thing could be done for the five fishes assembled in Chapter 4, with the addition of long-read and Hi-C data, as well as transcriptome data.

It would also be interesting to explore in more detail the functions of the genes under selection detected in Chapters 4 and 5. This has been partially addressed in this study by reporting the frequency of specific gene ontology (GO) terms to aid the interpretation of the results. However, more in-depth GO analyses could be performed, and on more lineages. The

trends in GO patterns could be compared in a multivariate framework to make better inferences of the gene functions that have been particularly influential in shaping the evolution of some species rather than others. For example, a comparative genomics study has found evidence of selection on genes associated with wound healing and the maintenance of genome stability in the white shark and whale shark (Marra et al., 2019). Further methods to detect genomic regions under selection could include selection tests using a sliding window approach to “scan” the genome (Ahrens et al., 2018).

Like all other fisheries species, the tarakihi fisheries should be routinely monitored using genetics/genomics. As already detailed in Chapter 2, this is important to monitor changes in genetic diversity and genetic structure and respond accordingly with management changes. The present population DNA dataset could be improved and built upon by adding more individuals from some locations, especially more remote locations like the Chatham Islands and Fiordland. Routinely adding new data across years could also allow for an analysis of changes in genetic variation and structure through time. In addition, structural variants would likely provide additional insights on selection. Structural variants such as copy number variants and inversions have recently been shown to play an important role in adaptation (Wellenreuther et al., 2019; Wold et al., 2021). For more practical future applications, the fixed DNA differences detected between tarakihi and king tarakihi in Chapter 5 could be used as the basis for an SNP assay to differentiate the two species in the catches. Although king tarakihi are thought to represent only a minor proportion of TAR catches, we have limited knowledge of how harvesting might be impacting these populations due to the lack of information on that species. Although tarakihi is not currently a strong candidate species for aquaculture, the SNPs found to be linked to temperature in Chapter 5 could be useful in that context (Sandoval-Castillo et al., 2018).

6.4 Conclusions

“I believe then that the cod fishery, the herring fishery, the pilchard fishery, the mackerel fishery and probably all the great sea fisheries are inexhaustible: that is to say that nothing we do seriously affects the numbers of fish. And any attempt to regulate these fisheries seems consequently from the nature of the case to be useless.” (Huxley 1884).

Although this became a bit of a cliché, when the topic of fisheries sustainability is addressed, one cannot avoid quoting the infamous citation from evolutionary biologist Thomas H. Huxley, given the sheer irony of this statement to a reader today. It is well established that fish populations subjected to industrial fishing pressures need to be effectively monitored and managed to avoid causing irremediable damages to a natural resource which is of great value commercially, but also (and one can argue, especially) culturally and ecologically. New Zealand, especially, has a diverse and precious marine heritage that should not be compromised. To achieve this, we need to make use of the best of the scientific toolbox available for understanding the biology of harvested species and make evidence-based management decisions. Genomics is a key tool that can be used to support well-informed management. Not only are genomics tools valuable for a single-species stock assessment framework, as thoroughly reviewed in this thesis, but they could also facilitate a shift to a more ecosystem-based approach, with new applications such as environmental DNA and ageing techniques. Coincidentally, it is also an exciting time to work on genomics from a more fundamental research perspective. This study has shown that inferences can be made on evolutionary trends from the nucleotide level (e.g. detection of adaptive SNPs) to individual genomes (e.g. alternative splicing events) to entire populations and species (population structure, phylogenomics). Genomic analyses can address both fundamental research questions about genomic evolution and applications to critical issues related to the future of fishing, conservation and the long-term stability of ecosystems.

References

- Abaunza, P., Murta, A. G., Campbell, N., Cimmaruta, R., Comesaña, A. S., Dahle, G., Gallo, E., García Santamaría, M. T., Gordo, L. S., Iversen, S. A., MacKenzie, K., Magoulas, A., Mattiucci, S., Molloy, J., Nascetti, G., Pinto, A. L., Quinta, R., Ramos, P., Ruggi, A., ... Zimmermann, C. (2008). Considerations on sampling strategies for an holistic approach to stock identification: The example of the HOMSIIR project. *Fisheries Research*, 89(2), 104–113. <https://doi.org/10.1016/j.fishres.2007.09.020>
- Abril, J. F., & Castellano, S. (2018). Genome annotation. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, 1–3, 195–209. <https://doi.org/10.1016/B978-0-12-809633-8.20226-4>
- Ahrens, C. W., Rymer, P. D., Stow, A., Bragg, J., Dillon, S., Umbers, K. D. L., & Dudaniec, R. Y. (2018). The search for loci under selection: trends, biases and progress. *Molecular Ecology*, 27(6), 1–15. <https://doi.org/10.1111/mec.14549>
- Aljanabi, S. M., & Martinez, I. (1997). Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. *Nucleic Acids Research*, 25(22), 4692–4693. <https://doi.org/10.1093/nar/25.22.4692>
- Allendorf, F. W., Hohenlohe, P. A., & Luikart, G. (2010). Genomics and the future of conservation genetics. *Nature Reviews Genetics*, 11(10), 697–709. <https://doi.org/10.1038/nrg2844>
- Altukhov, Y. P. (1981). The stock concept from the viewpoint of population genetics. *Canadian Journal of Fisheries and Aquatic Sciences*, 38(12), 1523–1538. <https://doi.org/10.1139/f81-205>
- An, D., Cao, H. X., Li, C., Humbeck, K., & Wang, W. (2018). Isoform sequencing and State-Of-Art applications for unravelling complexity of plant transcriptomes. *Genes*, 9(1). <https://doi.org/10.3390/genes9010043>
- Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, 17(2), 81–92. <https://doi.org/10.1038/nrg.2015.28>

- Andrews, S. (2018). *FastQC: A quality control tool for high through-put sequence data*. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>
- Annala, J. H. (1987). *The biology and fishery of tarakihi, Nemadactylus macropterus, in New Zealand waters* (S. J. Baird & G. G. Baird (eds.); Vol. 51) [Fisheries Research Division Occasional Publication No. 51]. New Zealand Ministry of Agriculture and Fisheries. <http://docs.niwa.co.nz/library/public/FRDop51.pdf>
- Anne, C. (2006). Choosing the right molecular genetic markers for studying biodiversity: From molecular evolution to practical aspects. *Genetica*, 127(1–3), 101–120. <https://doi.org/10.1007/s10709-005-2485-1>
- Ashton, D. T., Hilario, E., Jaksons, P., Ritchie, P. A., & Wellenreuther, M. (2019). Genetic diversity and heritability of economically important traits in captive Australasian snapper (*Chrysophrys auratus*). *Aquaculture*, 505, 190–198. <https://doi.org/10.1016/j.aquaculture.2019.02.034>
- Ashton, D. T., Ritchie, P. A., & Wellenreuther, M. (2019). High-Density Linkage Map and QTLs for Growth in Snapper (*Chrysophrys auratus*). *G3: Genes, Genomes, Genetics*, 9(4), 1027–1035. <https://doi.org/10.1534/g3.118.200905>
- Assis, J., Tyberghein, L., Bosch, S., Verbruggen, H., Serrão, E. A., De Clerck, O., & Tittensor, D. (2018). Bio-ORACLE v2.0: Extending marine data layers for bioclimatic modelling. *Global Ecology and Biogeography*, 27(3), 277–284. <https://doi.org/10.1111/geb.12693>
- Attard, C. R. M., Beheregaray, L. B., Sandoval-Castillo, J., Jenner, K. C. S., Gill, P. C., Jenner, M.-N. M., Morrice, M. G., & Möller, L. M. (2018). From conservation genetics to conservation genomics: a genome-wide assessment of blue whales (*Balaenoptera musculus*) in Australian feeding aggregations. *Royal Society Open Science*, 5(1), 170925. <https://doi.org/10.1098/rsos.170925>
- Attard, C. R. M., Brauer, C. J., Sandoval-Castillo, J., Faulks, L. K., Unmack, P. J., Gilligan, D. M., & Beheregaray, L. B. (2018). Ecological disturbance influences adaptive divergence despite high gene flow in golden perch (*Macquaria ambigua*): Implications for management and resilience to climate change. *Molecular Ecology*, 27(1), 196–215. <https://doi.org/10.1111/mec.14438>

- Austin, C. M., Tan, M. H., Harrisson, K. A., Lee, Y. P., Croft, L. J., Sunnucks, P., Pavlova, A., & Gan, H. M. (2017). De novo genome assembly and annotation of Australia's largest freshwater fish, the Murray cod (*Maccullochella peelii*), from Illumina and Nanopore sequencing read. *GigaScience*, 6(8), 1–6. <https://doi.org/10.1093/gigascience/gix063>
- Babcock, R. C., Bustamante, R. H., Fulton, E. A., Fulton, D. J., Haywood, M. D. E., Hobday, A. J., Kenyon, R., Matear, R. J., Plagányi, E. E., Richardson, A. J., & Vanderklift, M. A. (2019). Severe continental-scale impacts of climate change are happening now: extreme climate events impact marine habitat forming communities along 45% of Australia's coast. *Frontiers in Marine Science*, 6, 1–14. <https://doi.org/10.3389/fmars.2019.00411>
- Baker, C. S., MacCarthy, M., Smith, P. J., Perry, A. P., & Chambers, G. K. (1992). DNA fingerprints of orange roughy, *Hoplostethus atlanticus*: a population comparison. *Marine Biology*, 113(4), 561–567. <https://doi.org/10.1007/BF00349699>
- Baker, C. S., Perry, A., Chambers, G. K., & Smith, P. J. (1995). Population variation in the mitochondrial cytochrome *b* gene of the orange roughy *Hoplostethus atlanticus* and the hoki *Macruronus novaezelandiae*. *Marine Biology*, 122(4), 503–509. <https://doi.org/10.1007/BF00350673>
- Baldwin, R. E., Banks, M. A., & Jacobson, K. C. (2012). Integrating fish and parasite data as a holistic solution for identifying the elusive stock structure of Pacific sardines (*Sardinops sagax*). *Reviews in Fish Biology and Fisheries*, 22(1), 137–156. <https://doi.org/10.1007/s11160-011-9227-5>
- Ball, A. O., Sedberry, G. R., Zatcoff, M. S., Chapman, R. W., & Carlin, J. L. (2000). Population structure of the wreckfish *Polyprion americanus* determined with microsatellite genetic markers. *Marine Biology*, 137(5–6), 1077–1090. <https://doi.org/10.1007/s002270000392>
- Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, 6(1), 11. <https://doi.org/10.1186/s13100-015-0041-9>
- Barnett, D. W., Garrison, E. K., Quinlan, A. R., Stromberg, M. P., & Marth, G. T. (2011). BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, 27(12), 1691–1692. <https://doi.org/10.1093/bioinformatics/btr174>

- Barrett, R., & Schluter, D. (2008). Adaptation from standing genetic variation. *Trends in Ecology & Evolution*, 23(1), 38–44. <https://doi.org/10.1016/j.tree.2007.09.008>
- Bateman, A. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1), D506–D515. <https://doi.org/10.1093/nar/gky1049>
- Beaumont, M. A., & Nichols, R. A. (1996). Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 263(1377), 1619–1626. <https://doi.org/10.1098/rspb.1996.0237>
- Beddington, J. R., Agnew, D. J., & Clark, C. W. (2007). Current problems in the management of marine fisheries. *Science*, 316(5832), 1713–1716. <https://doi.org/10.1126/science.1137362>
- Beentjes, M. P., Parker, S., & Fu, D. (2012). *Characterisation of TAR 2 & TAR 3 fisheries and age composition of landings in 2010/11*. <http://www.mpi.govt.nz/news-resources/publications.aspx>
- Beer, N. A., & Carbines, G. D. (2012). *Exploratory microchemical analysis of blue cod otoliths from potting surveys in Dusky Sound, Fiordland 2002 and 2008* [New Zealand Fisheries Assessment Report 2012/21]. Ministry for Primary Industries.
- Begg, G. A., Friedland, K. D., & Pearce, J. B. (1999). Stock identification and its role in stock assessment and fisheries management: an overview. *Fisheries Research*, 43(1–3), 1–8. [https://doi.org/10.1016/S0165-7836\(99\)00062-4](https://doi.org/10.1016/S0165-7836(99)00062-4)
- Benestan, L. (2019). Population genomics applied to fishery management and conservation. In M. Oleksiak & O. Rajora (Eds.), *Population Genomics: Marine Organisms* (pp. 399–421). Springer. https://doi.org/10.1007/13836_2019_66
- Benestan, L., Gosselin, T., Perrier, C., Sainte-Marie, B., Rochette, R., & Bernatchez, L. (2015). RAD genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the American lobster (*Homarus americanus*). *Molecular Ecology*, 24(13), 3299–3315. <https://doi.org/10.1111/mec.13245>
- Benestan, L., Quinn, B. K., Maaroufi, H., Laporte, M., Clark, F. K., Greenwood, S. J., Rochette, R., & Bernatchez, L. (2016). Seascape genomics provides evidence for thermal adaptation

- and current-mediated population structure in American lobster (*Homarus americanus*). *Molecular Ecology*, 25(20), 5073–5092. <https://doi.org/10.1111/mec.13811>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Bernal-Ramírez, J. H., Adcock, G. J., Hauser, L., Carvalho, G. R., & Smith, P. J. (2003). Temporal stability of genetic population structure in the New Zealand snapper, *Pagrus auratus*, and relationship to coastal currents. *Marine Biology*, 142(3), 567–574. <https://doi.org/10.1007/s00227-002-0972-9>
- Bernatchez, L., Wellenreuther, M., Araneda, C., Ashton, D. T., Barth, J. M. I., Beacham, T. D., Maes, G. E., Martinsohn, J. T., Miller, K. M., Naish, K. A., Ovenden, J. R., Primmer, C. R., Young Suk, H., Therkildsen, N. O., & Withler, R. E. (2017). Harnessing the power of genomics to secure the future of seafood. *Trends in Ecology & Evolution*, 32(9), 665–680. <https://doi.org/10.1016/j.tree.2017.06.010>
- Bester-van der Merwe, A. E., Bitalo, D., Cuevas, J. M., Ovenden, J., Hernández, S., Da Silva, C., McCord, M., & Roodt-Wilding, R. (2017). Population genetics of Southern Hemisphere tope shark (*Galeorhinus galeus*): Intercontinental divergence and constrained gene flow at different geographical scales. *PLoS ONE*, 12(9), 1–20. <https://doi.org/10.1371/journal.pone.0184481>
- Betancur-R, R., Wiley, E. O., Arratia, G., Acero, A., Bailly, N., Miya, M., Lecointre, G., & Ortí, G. (2017). Phylogenetic classification of bony fishes. *BMC Evolutionary Biology*, 17(1), 162. <https://doi.org/10.1186/s12862-017-0958-3>
- Biémont, C. (2010). A brief history of the status of transposable elements: From junk DNA to major players in evolution. *Genetics*, 186(4), 1085–1093. <https://doi.org/10.1534/genetics.110.124180>
- Biémont, C., & Vieira, C. (2006). Genetics: Junk DNA as an evolutionary force. *Nature*, 443(7111), 521–524. <https://doi.org/10.1038/443521a>
- Bohonak, A. J. (1999). Dispersal, Gene Flow, and Population Structure. *The Quarterly Review of Biology*, 74(1), 21–45. <https://doi.org/10.1086/392950>

- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>
- Boomer, J. J., Harcourt, R. G., Francis, M. P., & Stow, A. J. (2012). Genetic divergence, speciation and biogeography of *Mustelus* (sharks) in the central Indo-Pacific and Australasia. *Molecular Phylogenetics and Evolution*, 64(3), 697–703. <https://doi.org/10.1016/j.ympev.2012.05.024>
- Bosch, S. (2020). *sdmpredictors: Species Distribution Modelling Predictor Datasets*. (R package version 0.2.9). <https://cran.r-project.org/package=sdmpredictors>
- Bradbury, I. R., Hamilton, L. C., Dempson, B., Robertson, M. J., Bourret, V., Bernatchez, L., & Verspoor, E. (2015). Transatlantic secondary contact in Atlantic Salmon, comparing microsatellites, a single nucleotide polymorphism array and restriction-site associated DNA sequencing for the resolution of complex spatial structure. *Molecular Ecology*, 24(20), 5130–5144. <https://doi.org/10.1111/mec.13395>
- Bradbury, I. R., Laurel, B., Snelgrove, P. V. R., Bentzen, P., & Campana, S. E. (2008). Global patterns in marine dispersal estimates: The influence of geography, taxonomic category and life history. *Proceedings of the Royal Society B: Biological Sciences*, 275(1644), 1803–1809. <https://doi.org/10.1098/rspb.2008.0216>
- Branch, T. A. (2001). A review of orange roughy *Hoplostethus atlanticus* fisheries, estimation methods, biology and stock structure. *South African Journal of Marine Science*, 23(1), 181–203. <https://doi.org/10.2989/025776101784529006>
- Braunschweig, U., Barbosa-Morais, N. L., Pan, Q., Nachman, E. N., Alipanahi, B., Gonatopoulos-Pournatzis, T., Frey, B., Irimia, M., & Blencowe, B. J. (2014). Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Research*, 24(11), 1774–1786. <https://doi.org/10.1101/gr.177790.114>
- Bravington, M. V., Grewe, P. M., & Davies, C. R. (2016). Absolute abundance of southern bluefin tuna estimated by close-kin mark-recapture. *Nature Communications*, 7, 1–8. <https://doi.org/10.1038/ncomms13162>
- Bravington, M. V., Skaug, H. J., & Anderson, E. C. (2016). Close-Kin Mark-Recapture. *Statistical*

- Science*, 31(2), 259–274. <https://doi.org/10.1214/16-STS552>
- Brawand, D., Wagner, C. E., Li, Y. I., Malinsky, M., Keller, I., Fan, S., Simakov, O., Ng, A. Y., Lim, Z. W., Bezault, E., Turner-Maier, J., Johnson, J., Alcazar, R., Noh, H. J., Russell, P., Aken, B., Alföldi, J., Amemiya, C., Azzouzi, N., ... Di Palma, F. (2014). The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, 513(7518), 375–381. <https://doi.org/10.1038/nature13726>
- Brian Simison, W., Parham, J. F., Papenfuss, T. J., Lam, A. W., & Henderson, J. B. (2020). An Annotated Chromosome-Level Reference Genome of the Red-Eared Slider Turtle (*Trachemys scripta elegans*). *Genome Biology and Evolution*, 12(4), 456–462. <https://doi.org/10.1093/gbe/evaa063>
- Broad Institute. (2019). *Picard toolkit*. Broad Institute, GitHub Repository. <http://broadinstitute.github.io/picard/>
- Brooks, A. (2019). *Using ecological niche modelling to predict climate change responses of ten key fishery species in Aotearoa New Zealand* [Master's thesis]. Victoria University of Wellington, New Zealand.
- Bruce, B. (2001). Influence of mesoscale oceanographic processes on larval distribution and stock structure in jackass morwong (*Nemadactylus macropterus*: Cheilodactylidae). *ICES Journal of Marine Science*, 58(5), 1072–1080. <https://doi.org/10.1006/jmsc.2001.1099>
- Burridge, C. P. (1999). Molecular phylogeny of *Nemadactylus* and *Acantholatris* (Perciformes: Cirrhitidae: Cheilodactylidae), with implications for taxonomy and biogeography. *Molecular Phylogenetics and Evolution*, 13(1), 93–109. <https://doi.org/10.1006/mpev.1999.0622>
- Burridge, C. P., & Smolenski, A. J. (2003). Lack of genetic divergence found with microsatellite DNA markers in the tarakihi *Nemadactylus macropterus*. *New Zealand Journal of Marine and Freshwater Research*, 37(2), 223–230. <https://doi.org/10.1080/00288330.2003.9517160>
- Burridge, C. P., & Smolenski, A. J. (2004). Molecular phylogeny of the Cheilodactylidae and Latridae (Perciformes: Cirrhitidae) with notes on taxonomy and biogeography. *Molecular Phylogenetics and Evolution*, 30(1), 118–127. <https://doi.org/10.1016/S1055->

7903(03)00157-X

- Burrows, M. T., Schoeman, D. S., Buckley, L. B., Moore, P., Poloczanska, E. S., Brander, K. M., Brown, C., Bruno, J. F., Duarte, C. M., Halpern, B. S., Holding, J., Kappel, C. V., Kiessling, W., O'Connor, M. I., Pandolfi, J. M., Parmesan, C., Schwing, F. B., Sydeman, W. J., & Richardson, A. J. (2011). The pace of shifting climate in marine and terrestrial ecosystems. *Science*, 334(6056), 652–655. <https://doi.org/10.1126/science.1210288>
- Bushnell, B. (2018). *BBMap short read aligner*. Berkeley: University of California. <http://sourceforge.net/projects/bbmap>
- Byrne, A., Cole, C., Volden, R., & Vollmers, C. (2019). Realizing the potential of full-length transcriptome sequencing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1786), 20190097. <https://doi.org/10.1098/rstb.2019.0097>
- Cadrin, S. X. (2020). Defining spatial structure for fishery stock assessment. *Fisheries Research*, 221, 105397. <https://doi.org/10.1016/j.fishres.2019.105397>
- Cadrin, S. X., Bernreuther, M., Danielsdottir, A. K., Hjorleifsson, E., Johansen, T., Kerr, L. A., Kristinsson, K., Mariani, S., Nedreaas, K., Pampoulie, C., Planque, B., Reinert, J., Saborido-Rey, F., Sigurthsson, T., & Stransky, C. (2010). Population structure of beaked redfish, *Sebastes mentella*: evidence of divergence associated with different habitats. *ICES Journal of Marine Science*, 67(8), 1617–1630. <https://doi.org/10.1093/icesjms/fsq046>
- Cadrin, S. X., Kerr, L. A., & Mariani, S. (2014a). Interdisciplinary evaluation of spatial population structure for definition of fishery management units. In S. X. Cadrin, L. A. Kerr, & S. Mariani (Eds.), *Stock Identification Methods : Applications in Fishery Science* (2nd ed., pp. 535–552). Academic Press.
- Cadrin, S. X., Kerr, L. A., & Mariani, S. (2014b). Stock identification methods: an overview. In S. X. Cadrin, L. A. Kerr, & S. Mariani (Eds.), *Stock Identification Methods : Applications in Fishery Science* (2nd ed., pp. 1–5). Academic Press.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1), 1–9. <https://doi.org/10.1186/1471-2105-10-421>
- Capilla, L., Sánchez-Guillén, R. A., Farré, M., Paytuví-Gallart, A., Malinverni, R., Ventura, J.,

- Larkin, D. M., & Ruiz-Herrera, A. (2016). Mammalian comparative genomics reveals genetic and epigenetic features associated with genome reshuffling in rodentia. *Genome Biology and Evolution*, 8(12), 3703–3717. <https://doi.org/10.1093/gbe/evw276>
- Carbines, G., & McKenzie, J. (2004). *Movement patterns and stock mixing of blue cod in Dusky Sound in 2002* [New Zealand Fisheries Assessment Report 2004/36]. Ministry of Fisheries.
- Cárdenas, L., Silva, A. X., Magoulas, A., Cabezas, J., Poulin, E., & Ojeda, F. P. (2009). Genetic population structure in the Chilean jack mackerel, *Trachurus murphyi* (Nichols) across the South-eastern Pacific Ocean. *Fisheries Research*, 100(2), 109–115. <https://doi.org/10.1016/j.fishres.2009.06.013>
- Carvalho, G. R., & Hauser, L. (1994). Molecular genetics and the stock concept in fisheries. *Reviews in Fish Biology and Fisheries*, 4(3), 326–350. <https://doi.org/10.1007/BF00042908>
- Casillas, S., & Barbadilla, A. (2017). Molecular Population Genetics. *Genetics*, 205(3), 1003–1035. <https://doi.org/10.1534/genetics.116.196493>
- Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, 17(4), 540–552. <https://doi.org/10.1093/oxfordjournals.molbev.a026334>
- Catanach, A., Crowhurst, R., Deng, C., David, C., Bernatchez, L., & Wellenreuther, M. (2019). The genomic pool of standing structural variation outnumbers single nucleotide polymorphism by threefold in the marine teleost *Chrysophrys auratus*. *Molecular Ecology*, 28(6), 1210–1223. <https://doi.org/10.1111/mec.15051>
- Chabot, C. L., & Allen, L. G. (2009). Global population structure of the tope (*Galeorhinus galeus*) inferred by mitochondrial control region sequence data. *Molecular Ecology*, 18(3), 545–552. <https://doi.org/10.1111/j.1365-294X.2008.04047.x>
- Challis, R. (2017). *rjchallis/assembly-stats* 17.02. Zenodo. <https://doi.org/https://doi.org/10.5281/zenodo.322347>
- Chalopin, D., Naville, M., Plard, F., Galiana, D., & Volff, J.-N. (2015). Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates.

- Genome Biology and Evolution*, 7(2), 567–580. <https://doi.org/10.1093/gbe/evv005>
- Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1), 7. <https://doi.org/10.1186/s13742-015-0047-8>
- Charlesworth, B. (2010). Molecular population genomics: A short history. *Genetics Research*, 92(5–6), 397–411. <https://doi.org/10.1017/S0016672310000522>
- Chen, Y., Wan, S., Li, Q., Dong, X., Diao, J., Liao, Q., Wang, G.-Y., & Gao, Z.-X. (2021). Genome-Wide Integrated Analysis revealed functions of lncRNA–miRNA–mRNA interaction in growth of intermuscular bones in *Megalobrama amblycephala*. *Frontiers in Cell and Developmental Biology*, 8(603815), 1–15. <https://doi.org/10.3389/fcell.2020.603815>
- Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, 18(2), 170–175. <https://doi.org/10.1038/s41592-020-01056-5>
- Cheng, S. H., Gold, M., Rodriguez, N., & Barber, P. H. (2021). Genome-wide SNPs reveal complex fine scale population structure in the California market squid fishery (*Doryteuthis opalescens*). *Conservation Genetics*, 22(1), 97–110. <https://doi.org/10.1007/s10592-020-01321-2>
- Chow, S., Okamoto, H., Uozumi, Y., Takeuchi, Y., & Takeyama, H. (1997). Genetic stock structure of the swordfish (*Xiphias gladius*) inferred by PCR-RFLP analysis of the mitochondrial DNA control region. *Marine Biology*, 127(3), 359–367. <https://doi.org/10.1007/s002270050022>
- Chow, S., & Takeyama, H. (2000). Nuclear and mitochondrial DNA analyses reveal four genetically separated breeding units of the swordfish. *Journal of Fish Biology*, 56(5), 1087–1098. <https://doi.org/10.1006/jfbi.2000.1229>
- Chuong, E. B., Elde, N. C., & Feschotte, C. (2017). Regulatory activities of transposable elements: From conflicts to benefits. *Nature Reviews Genetics*, 18(2), 71–86. <https://doi.org/10.1038/nrg.2016.139>
- Clark, M. R., McMillan, P. J., Anderson, O. F., & Roux, M.-J. (2016). *Stock management areas for orange roughy (Hoplostethus atlanticus) in the Tasman Sea and western South Pacific*

- Ocean* [New Zealand Fisheries Assessment Report 2016/19]. Ministry for Primary Industries.
- Clarke, C. R., Karl, S. A., Horn, R. L., Bernard, A. M., Lea, J. S., Hazin, F. H., Prodöhl, P. A., & Shivji, M. S. (2015). Global mitochondrial DNA phylogeography and population structure of the silky shark, *Carcharhinus falciformis*. *Marine Biology*, 162(5), 945–955. <https://doi.org/10.1007/s00227-015-2636-6>
- Claverie, J.-M. (2001). What If There Are Only 30,000 Human Genes? *Science*, 291(5507), 1255–1257. <https://doi.org/10.1126/science.1058969>
- Colgan, D. J., & Paxton, J. R. (1997). Biochemical genetics and recognition of a western stock of the common gemfish, *Rexea solandri* (Scombroidea : Gempylidae), in Australia. *Marine and Freshwater Research*, 48(2), 103. <https://doi.org/10.1071/MF96048>
- Collins, F. S., Morgan, M., & Patrinos, A. (2003). The Human Genome Project: Lessons from large-scale biology. *Science*, 300(5617), 286–290. <https://doi.org/10.1126/science.1084564>
- Conover, D. O., & Munch, S. B. (2002). Sustaining fisheries yields over evolutionary time scales. *Science*, 297(5578), 94–96. <https://doi.org/10.1126/science.1074085>
- Corrigan, S., Lowther, A. D., Beheregaray, L. B., Bruce, B. D., Cliff, G., Duffy, C. A., Foulis, A., Francis, M. P., Goldsworthy, S. D., Hyde, J. R., Jabado, R. W., Kacev, D., Marshall, L., Mucientes, G. R., Naylor, G. J. P., Pepperell, J. G., Queiroz, N., White, W. T., Wintner, S. P., & Rogers, P. J. (2018). Population connectivity of the highly migratory shortfin mako (*Isurus oxyrinchus* Rafinesque 1810) and implications for management in the Southern Hemisphere. *Frontiers in Ecology and Evolution*, 6, 1–15. <https://doi.org/10.3389/fevo.2018.00187>
- Coyne, J. A., & Orr, H. A. (1998). The evolutionary genetics of speciation. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 353(1366), 287–305. <https://doi.org/10.1098/rstb.1998.0210>
- Cuéllar-Pinzón, J., Presa, P., Hawkins, S. J., & Pita, A. (2016). Genetic markers in marine fisheries: Types, tasks and trends. *Fisheries Research*, 173, 194–205. <https://doi.org/10.1016/j.fishres.2015.10.019>

- Dahle, G., Johansen, T., Westgaard, J.-I., Aglen, A., & Glover, K. A. (2018). Genetic management of mixed-stock fisheries “real-time”: The case of the largest remaining cod fishery operating in the Atlantic in 2007–2017. *Fisheries Research*, 205, 77–85. <https://doi.org/10.1016/j.fishres.2018.04.006>
- Dainat, J. (2021). *AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF format.* (Version v0.6.0). Zenodo. <https://doi.org/https://www.doi.org/10.5281/zenodo.3552717>
- Daley, R. K., Ward, R. D., Last, P. R., Reilly, A., Appleyard, S. A., & Gledhill, D. C. (2000). *Stock delineation of the pink ling (Genypterus blacodes) in Australian waters using genetic and morphometric techniques* [FRDC Final Report]. CSIRO Marine Research.
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156–2158. <https://doi.org/10.1093/bioinformatics/btr330>
- Darriba, Di., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B., & Flouri, T. (2020). ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Molecular Biology and Evolution*, 37(1), 291–294. <https://doi.org/10.1093/molbev/msz189>
- Dayan, D. D. (2020). Clinal Adaptation in the Marine Environment. In M. F. Oleksiak & O. P. Rajora (Eds.), *Population Genomics: Marine Organisms*. Springer International Publishing. <https://doi.org/10.1007/978-3-030-37936-0>
- De Bie, T., Cristianini, N., Demuth, J. P., & Hahn, M. W. (2006). CAFE: a computational tool for the study of gene family evolution. *Bioinformatics*, 22(10), 1269–1271. <https://doi.org/10.1093/bioinformatics/btl097>
- De Coster, W. (2017). *Per base sequence content and quality (new basecaller)*. <https://gigabaseorgigabyte.wordpress.com/2017/05/10/per-base-sequence-content-and-quality-new-basecaller/>
- De Coster, W., D’Hert, S., Schultz, D. T., Cruts, M., & Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*, 34(15), 2666–

2669. <https://doi.org/10.1093/bioinformatics/bty149>
- de Lannoy, C., de Ridder, D., & Risse, J. (2017). The long reads ahead: de novo genome assembly using the MinION. *F1000Research*, 6(1), 1083. <https://doi.org/10.12688/f1000research.12012.2>
- Dean, M., & Annilo, T. (2005). Evolution of the ATP-binding cassette (ABC) transporter superfamily in vertebrates. *Annual Review of Genomics and Human Genetics*, 6, 123–142. <https://doi.org/10.1146/annurev.genom.6.080604.162122>
- Deli Antoni, M. Y., González-Castro, M., & Díaz de Astarloa, J. M. (2015). New tools (DNA barcoding), old hypothesis: the case of the taxonomic identity of the Argentine hakes (Actinopterygii: Merluccius). *Journal of Fish Biology*, 87(3), 783–793. <https://doi.org/10.1111/jfb.12745>
- Devloo-Delva, F., Maes, G. E., Hernández, S. I., Mcallister, J. D., Gunasekera, R. M., Grewe, P. M., Thomson, R. B., & Feutry, P. (2019). Accounting for kin sampling reveals genetic connectivity in Tasmanian and New Zealand school sharks, *Galeorhinus galeus*. *Ecology and Evolution*, 9(8), 4465–4472. <https://doi.org/10.1002/ece3.5012>
- Dhar, R., Seethy, A., Pethusamy, K., Singh, S., Rohil, V., Purkayastha, K., Mukherjee, I., Goswami, S., Singh, R., Raj, A., Srivastava, T., Acharya, S., Rajashekhar, B., & Karmakar, S. (2019). De novo assembly of the Indian blue peacock (*Pavo cristatus*) genome using Oxford Nanopore technology and Illumina sequencing. *GigaScience*, 8(5), 1–13. <https://doi.org/10.1093/gigascience/giz038>
- Dijkstra, L. H., & Jellyman, D. J. (1999). Is the subspecies classification of the freshwater eels *Anguilla australis australis* Richardson and *A. a. schmidtii* Phillipps still valid? *Marine and Freshwater Research*, 50(3), 261–263. <https://doi.org/10.1071/MF98144>
- Diopere, E., Vandamme, S. G., Hablützel, P. I., Cariani, A., Van Houdt, J., Rijnsdorp, A., Tinti, F., Volckaert, F. A. M., & Maes, G. E. (2018). Seascape genetics of a flatfish reveals local selection under high levels of gene flow. *ICES Journal of Marine Science*, 75(2), 675–689. <https://doi.org/10.1093/icesjms/fsx160>
- Domingues, R. R., Bruels, C. C., Gadig, O. B. F., Chapman, D. D., Hilsdorf, A. W. S., & Shivji, M. S. (2019). Genetic connectivity and phylogeography of the night shark (*Carcharhinus*

- signatus*) in the western Atlantic Ocean: Implications for conservation management. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 29(1), 102–114. <https://doi.org/10.1002/aqc.2961>
- Domingues, R. R., Hilsdorf, A. W. S., & Gadig, O. B. F. (2018). The importance of considering genetic diversity in shark and ray conservation policies. *Conservation Genetics*, 19(3), 501–525. <https://doi.org/10.1007/s10592-017-1038-3>
- Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1), 27–46. <https://doi.org/10.1111/j.1600-0587.2012.07348.x>
- Dray, S., & Dufour, A.-B. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software*, 22(4), 1–20. <https://doi.org/10.18637/jss.v022.i04>
- Drummond, D. A., Silberg, J. J., Meyer, M. M., Wilke, C. O., & Arnold, F. H. (2005). On the conservative nature of intragenic recombination. *Proceedings of the National Academy of Sciences*, 102(15), 5380–5385. <https://doi.org/10.1073/pnas.0500729102>
- Duhamel, G. (1989). Ichtyofaune des Iles Saint-Paul et Amsterdam (Océan Indien Sud). *Mésogée*, 49, 21–47.
- Dunn, M. R., & Forman, J. S. (2011). Hypotheses of spatial stock structure in orange roughy *Hoplostethus atlanticus* inferred from diet, feeding, condition, and reproductive activity. *PLoS ONE*, 6(11). <https://doi.org/10.1371/journal.pone.0026704>
- Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, 7(10), e1002195. <https://doi.org/10.1371/journal.pcbi.1002195>
- Ellegren, H. (2008). Comparative genomics and the study of evolution by natural selection. *Molecular Ecology*, 17(21), 4586–4596. <https://doi.org/10.1111/j.1365-294X.2008.03954.x>
- Ellegren, H., & Galtier, N. (2016). Determinants of genetic diversity. *Nature Reviews Genetics*, 17(7), 422–433. <https://doi.org/10.1038/nrg.2016.58>

- Elliott, N. G., Smolenski, A. J., & Ward, R. D. (1994). Allozyme and mitochondrial DNA variation in orange roughy, *Hoplostethus atlanticus* (Teleostei: Trachichthyidae): little differentiation between Australian and North Atlantic populations. *Marine Biology*, 119(4), 621–627. <https://doi.org/10.1007/BF00354326>
- Elliott, N. G., & Ward, R. D. (1992). Enzyme variation in orange roughy, *Hoplostethus atlanticus* (Teleostei: Trachichthyidae), from southern Australian and New Zealand waters. *Marine and Freshwater Research*, 43(6), 1561–1571. <https://doi.org/10.1071/MF9921561>
- Elliott, N. G., & Ward, R. D. (1994). Enzyme variation in jackass morwong, *Nemadactylus macropterus* (Schneider, 1801) (Teleostei: Cheilodactylidae), from Australian and New Zealand waters. *Marine and Freshwater Research*, 45(1), 51–67. <https://doi.org/10.1071/MF9940051>
- Elliott, T. A., & Gregory, T. R. (2015). What’s in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1678). <https://doi.org/10.1098/rstb.2014.0331>
- Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20(1), 238. <https://doi.org/10.1186/s13059-019-1832-y>
- Espeland, S. H., Olsen, E. M., Knutsen, H., Gjøsæter, J., Danielssen, D., & Stenseth, N. C. (2008). New perspectives on fish movement: Kernel and GAM smoothers applied to a century of tagging data on coastal Atlantic cod. *Marine Ecology Progress Series*, 372, 231–241. <https://doi.org/10.3354/meps07721>
- Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19), 3047–3048. <https://doi.org/10.1093/bioinformatics/btw354>
- Fan, G., Song, Y., Yang, L., Huang, X., Zhang, S., Zhang, M., Yang, X., Chang, Y., Zhang, H., Li, Y., Liu, S., Yu, L., Chu, J., Seim, I., Feng, C., Near, T. J., Wing, R. A., Wang, W., Wang, K., ... He, S. (2020). Initial data release and announcement of the 10,000 Fish Genomes Project (Fish10K). *GigaScience*, 9(8), 1–7. <https://doi.org/10.1093/gigascience/giaa080>
- FAO. (2020). The State of World Fisheries and Aquaculture 2020. In *Fao*. FAO.

<https://doi.org/10.4060/ca9229en>

- Feng, S., Stiller, J., Deng, Y., Armstrong, J., Fang, Q., Reeve, A. H., Xie, D., Chen, G., Guo, C., Faircloth, B. C., Petersen, B., Wang, Z., Zhou, Q., Diekhans, M., Chen, W., Andreu-Sánchez, S., Margaryan, A., Howard, J. T., Parent, C., ... Zhang, G. (2020). Dense sampling of bird diversity increases power of comparative genomics. *Nature*, 587(7833), 252–257. <https://doi.org/10.1038/s41586-020-2873-9>
- Fenton, G. E., Short, S. A., & Ritz, D. A. (1991). Age determination of orange roughy, *Hoplostethus atlanticus* (Pisces: Trachichthyidae) using ^{210}Pb : ^{226}Ra disequilibria. *Marine Biology*, 109(2), 197–202. <https://doi.org/10.1007/BF01319387>
- Feron, R., Zahm, M., Cabau, C., Klopp, C., Roques, C., Bouchez, O., Ech, C., Valire, S., Donnadieu, C., Haffray, P., Bestin, A., Morvezen, R., Acloque, H., Euclide, P. T., Wen, M., Jouano, E., Scharl, M., Postlethwait, J. H., Schraidt, C., ... Guiguen, Y. (2020). Characterization of a Y-specific duplication/insertion of the anti-Mullerian hormone type II receptor gene based on a chromosome-scale genome assembly of yellow perch, *Perca flavescens*. *Molecular Ecology Resources*, 20(2), 531–543. <https://doi.org/10.1111/1755-0998.13133>
- Finnegan, D. J. (2012). Retrotransposons. *Current Biology*, 22(11), R432–R437. <https://doi.org/10.1016/j.cub.2012.04.025>
- Fisheries New Zealand. (2018a). *Fisheries Assessment Plenary, May 2018: stock assessments and stock status* (Issue May). Compiled by the Fisheries Science and Information Group, Fisheries New Zealand.
- Fisheries New Zealand. (2018b). *Fisheries Assessment Plenary: Stock Assessment and Stock Status Volume 3: Pipi to Yellow-eyed Mullet*. Ministry for Primary Industries.
- Fisheries New Zealand. (2019a). *Quota Management System*. <https://www.fisheries.govt.nz/law-and-policy/legal-overviews/fisheries/quota-management-system/>
- Fisheries New Zealand. (2019b). *Species*. <https://fs.fish.govt.nz/Page.aspx?pk=6&tk=97&ey=2019>
- Fisheries New Zealand. (2021). *Fisheries Assessment Plenary: Stock Assessment and Stock*

- Status Volume 3: Pipi to Yellow-eyed Mullet*. Ministry for Primary Industries.
- Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, 117(17), 9451–9457. <https://doi.org/10.1073/pnas.1921046117>
- Forester, B. R., Jones, M. R., Joost, S., Landguth, E. L., & Lasky, J. R. (2016). Detecting spatial genetic signatures of local adaptation in heterogeneous landscapes. *Molecular Ecology*, 25(1), 104–120. <https://doi.org/10.1111/mec.13476>
- Forester, B. R., Lasky, J. R., Wagner, H. H., & Urban, D. L. (2018). Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations. *Molecular Ecology*, 27(9), 2215–2233. <https://doi.org/10.1111/mec.14584>
- Francis, R. I. C. C., & Clark, M. R. (1998). Inferring spawning migrations of orange roughy (*Hoplostethus atlanticus*) from spawning oöives. *Marine and Freshwater Research*, 49(2), 103. <https://doi.org/10.1071/MF97253>
- Francis, R. I. C. C., Neil, H. L., Horn, P. L., Gillanders, B., Marriott, P., & Vorster, J. (2011). *A pilot study to evaluate the utility of otolith microchemistry for determining natal fidelity in New Zealand hoki* [Final Research Report for Ministry of Fisheries Research Project HOK2006/05 Objective 1]. Unpublished report held by Ministry for Primary Industries, Wellington, New Zealand.
- Fricke, R., Eschmeyer, W. N., & Van Der Laan, R. (2020). *Eschmeyer's Catalog of Fishes: Genera, Species, References*. <http://researcharchive.calacademy.org/research/ichthyology/catalog/fishcatmain.asp>
- Funk, W. C., McKay, J. K., Hohenlohe, P. A., & Allendorf, F. W. (2012). Harnessing genomics for delineating conservation units. *Trends in Ecology & Evolution*, 27(9), 489–496. <https://doi.org/10.1016/j.tree.2012.05.012>
- Gan, H. M., Lee, Y. P., & Austin, C. M. (2017). Nanopore long-read guided complete genome assembly of *Hydrogenophaga* intermedia, and genomic insights into 4-aminobenzenesulfonate, p-aminobenzoic acid and hydrogen metabolism in the genus *Hydrogenophaga*. *Frontiers in Microbiology*, 8(OCT), 1–9.

- <https://doi.org/10.3389/fmicb.2017.01880>
- Gan, W., Chung-Davidson, Y. W., Chen, Z., Song, S., Cui, W., He, W., Zhang, Q., Li, W., Li, M., & Ren, J. (2021). Global tissue transcriptomic analysis to improve genome annotation and unravel skin pigmentation in goldfish. *Scientific Reports*, 11(1), 1–14. <https://doi.org/10.1038/s41598-020-80168-6>
- Gao, B., Shen, D., Xue, S., Chen, C., Cui, H., & Song, C. (2016). The contribution of transposable elements to size variations between four teleost genomes. *Mobile DNA*, 7(1), 1–16. <https://doi.org/10.1186/s13100-016-0059-7>
- Gao, Q., Xiong, Z., Larsen, R. S., Zhou, L., Zhao, J., Ding, G., Zhao, R., Liu, C., Ran, H., & Zhang, G. (2020). High-quality chromosome-level genome assembly and full-length transcriptome analysis of the pharaoh ant *Monomorium pharaonis*. *GigaScience*, 9(12), 1–14. <https://doi.org/10.1093/gigascience/giaa143>
- Gao, Y., Xi, F., Liu, X., Wang, H., Reddy, A. S., & Gu, L. (2019). Single-molecule Real-time (SMRT) Isoform Sequencing (Iso-Seq) in Plants: The status of the bioinformatics tools to unravel the transcriptome complexity. *Current Bioinformatics*, 14(7), 566–573.
- Garcia-Mayoral, E., Olsen, M., Hedeholm, R., Post, S., Nielsen, E. E., & Bekkevold, D. (2016). Genetic structure of West Greenland populations of lumpfish *Cyclopterus lumpus*. *Journal of Fish Biology*, 89(6), 2625–2642. <https://doi.org/10.1111/jfb.13167>
- Garcia, S. M., Kolding, J., Rice, J., Rochet, M. J., Zhou, S., Arimoto, T., Beyer, J. E., Borges, L., Bundy, A., Dunn, D., Fulton, E. A., Hall, M., Heino, M., Law, R., Makino, M., Rijnsdorp, A. D., Simard, F., & Smith, A. D. M. (2012). Reconsidering the Consequences of Selective Fisheries. *Science*, 335(6072), 1045–1047. <https://doi.org/10.1126/science.1214594>
- Gardner, M. G., & Ward, R. D. (2002). Taxonomic affinities within Australian and New Zealand *Mustelus* sharks (Chondrichthyes: Triakidae) inferred from allozymes, mitochondrial DNA and precaudal vertebrae counts. *Copeia*, 2002(2), 356–363. [https://doi.org/https://doi.org/10.1643/0045-8511\(2002\)002\[0356:TAWAAN\]2.0.CO;2](https://doi.org/https://doi.org/10.1643/0045-8511(2002)002[0356:TAWAAN]2.0.CO;2)
- Gauldie, R. W., & Johnston, A. J. (1980). The geographical distribution of phosphoglucomutase and glucose phosphate isomerase alleles of some New Zealand fishes. *Comparative Biochemistry and Physiology Part B: Comparative Biochemistry*, 66(2), 171–183.

- [https://doi.org/10.1016/0305-0491\(80\)90051-6](https://doi.org/10.1016/0305-0491(80)90051-6)
- Gauldie, R. W., & Nathan, A. (1977). Iron content of the otoliths of tarakihi (Teleostei: Cheilodactylidae). *New Zealand Journal of Marine and Freshwater Research*, 11(2), 179–191. <https://doi.org/10.1080/00288330.1977.9515672>
- Ge, H., Lin, K., Shen, M., Wu, S., Wang, Y., Zhang, Z., Wang, Z., Zhang, Y., Huang, Z., Zhou, C., Lin, Q., Wu, J., Liu, L., Hu, J., Huang, Z., & Zheng, L. (2019). De novo assembly of a chromosome-level reference genome of red-spotted grouper (*Epinephelus akaara*) using nanopore sequencing and Hi-C. *Molecular Ecology Resources*, 19(6), 1461–1469. <https://doi.org/10.1111/1755-0998.13064>
- Gebbie, C. L. (2014). *Population genetic structure of New Zealand blue cod (Parapercis colias) based on mitochondrial and microsatellite DNA markers* [Master's thesis, Victoria University of Wellington, New Zealand]. <http://researcharchive.vuw.ac.nz/handle/10063/4245>
- Gibbs, M. T. (2008). The historical development of fisheries in New Zealand with respect to sustainable development principles. *The Electronic Journal of Sustainable Development*, 1(2), 23–33.
- Gong, G., Dan, C., Xiao, S., Guo, W., Huang, P., Xiong, Y., Wu, J., He, Y., Zhang, J., Li, X., Chen, N., Gui, J. F., & Mei, J. (2018). Chromosomal-level assembly of yellow catfish genome using third-generation DNA sequencing and Hi-C analysis. *GigaScience*, 7(11), 1–9. <https://doi.org/10.1093/gigascience/giy120>
- Gonzalez, A., Ronce, O., Ferriere, R., & Hochberg, M. E. (2012). Evolutionary rescue : an emerging focus at the intersection between ecology and evolution. *Philosophical Transactions of the Royal Society B*, 368. <https://doi.org/10.1098/rstb.2012.0404>
- Grewe, P. M., Smolenski, A. J., & Ward, R. D. (1994). Mitochondrial DNA Diversity in Jackass Morwong (*Nemadactylus macropterus* : Teleostei) from Australian and New Zealand Waters. *Canadian Journal of Fisheries and Aquatic Sciences*, 51(5), 1101–1109. <https://doi.org/10.1139/f94-109>
- Gruber, B., Georges, A., Mijangos, J. L., Unmack, P. J., Berry, O., Clark, L. V., & Devloo-Delva, F. (2021). *Package 'dartR': Importing and analysing SNP and silicodart data generated*

- by genome-wide restriction fragment analysis*. The Comprehensive R Archive Network. <https://cran.r-project.org/web/packages/dartR/dartR.pdf>
- Gruber, B., Unmack, P. J., Berry, O. F., & Georges, A. (2018). DARTR: An R package to facilitate analysis of SNP data generated from reduced representation genome sequencing. *Molecular Ecology Resources*, 18(3), 691–699. <https://doi.org/10.1111/1755-0998.12745>
- Halliwell, A. G. (2019). *Population genetics of New Zealand Nemadactylus macropterus (tarakihi) and characterisation of their mitochondrial genome* [Master's thesis]. Victoria University of Wellington, New Zealand.
- Halvorsen, K. A. T., Árnason, E., Smith, P. J., & Mork, J. (2012). Mitochondrial DNA differentiation between the antitropical blue whiting species *Micromesistius poutassou* and *Micromesistius australis*. *Journal of Fish Biology*, 81(1), 253–269. <https://doi.org/10.1111/j.1095-8649.2012.03334.x>
- Hamilton, M. (2009). Genetic drift and effective population size. In *Population genetics* (pp. 53–78). John Wiley & Sons, Incorporated.
- Hanchet, S. M., & Field, K. (2001). *Review of current and historical data for tarakihi (Nemadactylus macropterus) fishstocks TAR 1,2,3, and 7, and recommendations for future monitoring* [New Zealand Fisheries Assessment Report 2001/59]. Ministry of Fisheries. http://docs.niwa.co.nz/library/public/FAR2001_59.pdf
- Hansen, K. D., Brenner, S. E., & Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, 38(12), e131–e131. <https://doi.org/10.1093/nar/gkq224>
- Hardison, R. C. (2003). Comparative Genomics. *PLoS Biology*, 1(2), e58. <https://doi.org/10.1371/journal.pbio.0000058>
- Hardwick, S. A., Joglekar, A., Flicek, P., Frankish, A., & Tilgner, H. U. (2019). Getting the Entire Message: Progress in Isoform Sequencing. *Frontiers in Genetics*, 10(709), 1–10. <https://doi.org/10.3389/fgene.2019.00709>
- Harrell, F. E. (2021). *Hmisc: Harrell Miscellaneous* (R package version 4.5-0). <https://cran.r-project.org/package=Hmisc>

- Hastings, P. J., Lupski, J. R., Rosenberg, S. M., & Ira, G. (2009). Mechanisms of change in gene copy number. *Nature Reviews Genetics*, 10(8), 551–564. <https://doi.org/10.1038/nrg2593>
- Hauser, L., Adcock, G. J., Smith, P. J., Bernal Ramírez, J. H., & Carvalho, G. R. (2002). Loss of microsatellite diversity and low effective population size in an overexploited population of New Zealand snapper (*Pagrus auratus*). *Proceedings of the National Academy of Sciences*, 99(18), 11742–11747. <https://doi.org/10.1073/pnas.172242899>
- Hauser, L., & Carvalho, G. R. (2008). Paradigm shifts in marine fisheries genetics: Ugly hypotheses slain by beautiful facts. *Fish and Fisheries*, 9(4), 333–362. <https://doi.org/10.1111/j.1467-2979.2008.00299.x>
- Heino, M., Díaz Pauli, B., & Dieckmann, U. (2015). Fisheries-Induced Evolution. *Annual Review of Ecology, Evolution, and Systematics*, 46, 461–480. <https://doi.org/10.1146/annurev-ecolsys-112414-054339>
- Hemmer-Hansen, J., Therkildsen, N. O., & Pujolar, J. M. (2014). Population Genomics of Marine Fishes: Next-Generation Prospects and Challenges. *The Biological Bulletin*, 227(2), 117–132. <https://doi.org/10.1086/BBLv227n2p117>
- Henderson, I. M. (2009). Optimising recreational harvests of blue cod: The effects of catch-and-release mortality and size selectivity. *Fisheries Research*, 99(3), 184–195. <https://doi.org/10.1016/j.fishres.2009.06.003>
- Hendry, R. T. (2004). *An assessment of the spatial extent and relative importance of nurseries, and of the genetic structure among nurseries of rig (Mustelus lenticulatus), an endemic New Zealand shark* [Master’s thesis, Victoria University of Wellington, New Zealand]. <http://library.victoria.ac.nz/library/>
- Hernández, S., Daley, R., Walker, T., Braccini, M., Varela, A., Francis, M. P., & Ritchie, P. A. (2015). Demographic history and the South Pacific dispersal barrier for school shark (*Galeorhinus galeus*) inferred by mitochondrial DNA and microsatellite DNA mark. *Fisheries Research*, 167, 132–142. <https://doi.org/10.1016/j.fishres.2015.02.010>
- Hewitt, G. M. (2000). The genetic legacy of the quaternary ice ages. *Nature*, 405(6789), 907–913. <https://doi.org/10.1038/35016000>

- Hewitt, G. M. (2004). Genetic consequences of climatic oscillations in the Quaternary. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 359(1442), 183–195. <https://doi.org/10.1098/rstb.2003.1388>
- Hicks, A. C., Smith, P. J., Horn, P. L., & Gilbert, D. J. (2003). *Differences in otolith measurements and gill raker counts between the two major spawning stocks of hoki (Macruronus novaezelandiae) in New Zealand* (Issue March) [New Zealand Fisheries Assessment Report 2003/7]. Ministry of Fisheries.
- Hijmans, R. J. (2019). *raster: Geographic Data Analysis and Modeling* (R package version 2.9-5). <https://cran.r-project.org/package=raster>
- Hillary, R. M., Bravington, M. V., Patterson, T. A., Grewe, P., Bradford, R., Feutry, P., Gunasekera, R., Peddemors, V., Werry, J., Francis, M. P., Duffy, C. A. J., & Bruce, B. D. (2018). Genetic relatedness reveals total population size of white sharks in eastern Australia and New Zealand. *Scientific Reports*, 8, 1–9. <https://doi.org/10.1038/s41598-018-20593-w>
- Hoang, N. V., & Henry, R. J. (2021). Iso-Seq Long Read Transcriptome Sequencing. In A. Cifuentes (Ed.), *Comprehensive Foodomics* (pp. 486–500). Elsevier. <https://doi.org/10.1016/b978-0-08-100596-5.22729-7>
- Hoarau, G., Boon, E., Jongma, D. N., Ferber, S., Palsson, J., Van der Veer, H. W., Rijnsdorp, A. D., Stam, W. T., & Olsen, J. L. (2005). Low effective population size and evidence for inbreeding in an overexploited flatfish, plaice (*Pleuronectes platessa* L.). *Proceedings of the Royal Society B: Biological Sciences*, 272(1562), 497–503. <https://doi.org/10.1098/rspb.2004.2963>
- Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., Poss, M. L., Reed, L. K., Storfer, A., & Whitlock, M. C. (2016). Finding the Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future Directions. *The American Naturalist*, 188(4), 379–397. <https://doi.org/10.1086/688018>
- Hoey, J. A., & Pinsky, M. L. (2018). Genomic signatures of environmental selection despite near-panmixia in summer flounder. *Evolutionary Applications*, 11(9), 1732–1747. <https://doi.org/10.1111/eva.12676>

- Holt, C., & Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, 12(491), 1–14. <https://doi.org/10.1186/1471-2105-12-491>
- Horn, P. L. (2005). *A review of the stock structure of ling (Genypterus blacodes) in New Zealand waters* [New Zealand Fisheries Assessment Report 2005/59]. Ministry of Fisheries.
- Horn, P. L., & Sullivan, K. J. (1996). Validated aging methodology using otoliths, and growth parameters for hoki (*Macruronus novaezelandiae*) in New Zealand waters. *New Zealand Journal of Marine and Freshwater Research*, 30(2), 161–174. <https://doi.org/10.1080/00288330.1996.9516705>
- Hurst, R. J., Stevenson, M. L., Bagley, N. W., Griggs, L. H., Morrison, M. A., Francis, M. P., & Duffy, C. A. (2000). *Areas of importance for spawning, pupping or egg-laying, and juveniles of New Zealand coastal fish* (Issue December) [NIWA Technical Report]. National Institute of Water and Atmospheric Research.
- Iwasaki, W., Fukunaga, T., Isagozawa, R., Yamada, K., Maeda, Y., Satoh, T. P., Sado, T., Mabuchi, K., Takeshima, H., Miya, M., & Nishida, M. (2013). MitoFish and MitoAnnotator: A mitochondrial genome database of fish with an accurate and automatic annotation pipeline. *Molecular Biology and Evolution*, 30(11), 2531–2540. <https://doi.org/10.1093/molbev/mst141>
- Izzo, C., Ward, T. M., Ivey, A. R., Suthers, I. M., Stewart, J., Sexton, S. C., & Gillanders, B. M. (2017). Integrated approach to determining stock structure: implications for fisheries management of sardine, *Sardinops sagax*, in Australian waters. *Reviews in Fish Biology and Fisheries*, 27(1), 267–284. <https://doi.org/10.1007/s11160-017-9468-z>
- Jain, C., Koren, S., Dilthey, A., Phillippy, A. M., & Aluru, S. (2018). A fast adaptive algorithm for computing whole-genome homology maps. *Bioinformatics*, 34(17), i748–i756. <https://doi.org/10.1093/bioinformatics/bty597>
- Jansen, H. J., Liem, M., Jong-Raadsen, S. A., Dufour, S., Weltzien, F.-A., Swinkels, W., Koelewijn, A., Palstra, A. P., Pelster, B., Spaink, H. P., Thillart, G. E. van den, Dirks, R. P., & Henkel, C. V. (2017). Rapid de novo assembly of the European eel genome from nanopore sequencing reads. *Scientific Reports*, 7(1), 7213. <https://doi.org/10.1038/s41598-017-07650-6>

- Jauhal, A. A., & Newcomb, R. D. (2021). Assessing genome assembly quality prior to downstream analysis: N50 versus BUSCO. *Molecular Ecology Resources*, February, 1416–1421. <https://doi.org/10.1111/1755-0998.13364>
- Jeffares, D. C., Tomiczek, B., Sojo, V., & dos Reis, M. (2015). A Beginners Guide to Estimating the Non-synonymous to Synonymous Rate Ratio of all Protein-Coding Genes in a Genome. In C. Peacock (Ed.), *Parasite Genomics Protocols* (pp. 65–90). Humana Press. https://doi.org/10.1007/978-1-4939-1438-8_4
- Jeffries, D. L., Copp, G. H., Lawson Handley, L., Olsén, K. H., Sayer, C. D., & Hänfling, B. (2016). Comparing RADseq and microsatellites to infer complex phylogeographic patterns, an empirical perspective in the Crucian carp, *Carassius carassius*, L. *Molecular Ecology*, 25(13), 2997–3018. <https://doi.org/10.1111/mec.13613>
- Jenkins, T. L., Ellis, C. D., Triantafyllidis, A., & Stevens, J. R. (2019). Single nucleotide polymorphisms reveal a genetic cline across the north-east Atlantic and enable powerful population assignment in the European lobster. *Evolutionary Applications*, 12(10), 1881–1899. <https://doi.org/10.1111/eva.12849>
- Jiang, J. B., Quattrini, A. M., Francis, W. R., Ryan, J. F., Rodríguez, E., & McFadden, C. S. (2019). A hybrid de novo assembly of the sea pansy (*Renilla muelleri*) genome. *GigaScience*, 8(4), 1–7. <https://doi.org/10.1093/gigascience/giz026>
- Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, 24(11), 1403–1405. <https://doi.org/10.1093/bioinformatics/btn129>
- Jones, M. R., Forester, B. R., Teufel, A. I., Adams, R. V., Anstett, D. N., Goodrich, B. A., Landguth, E. L., Joost, S., & Manel, S. (2013). Integrating landscape genomics and spatially explicit approaches to detect loci under selection in clinal populations. *Evolution*, 67(12), 3455–3468. <https://doi.org/10.1111/evo.12237>
- Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9), 1236–1240. <https://doi.org/10.1093/bioinformatics/btu031>

- Joost, S., Vuilleumier, S., Jensen, J. D., Schoville, S., Leempoel, K., Stucki, S., Widmer, I., Melodelima, C., Rolland, J., & Manel, S. (2013). Uncovering the genetic basis of adaptive change: On the intersection of landscape genomics and theoretical population genetics. *Molecular Ecology*, 22(14), 3659–3665. <https://doi.org/10.1111/mec.12352>
- Kadobianskyi, M., Schulze, L., Schuelke, M., & Judkewitz, B. (2019). Hybrid genome assembly and annotation of *Danionella translucida*. *Scientific Data*, 6(156), 1:7. <https://doi.org/10.1038/s41597-019-0161-z>
- Kamvar, Z. N., Tabima, J. F., & Grünwald, N. J. (2014). Poppr : an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, 2(1), e281. <https://doi.org/10.7717/peerj.281>
- Kaschner, K., Kesner-Reyes, K., Garilao, C., Segschneider, J., Rius-Barile, J., Rees, T., & Froese, R. (2019). *AquaMaps: Predicted range maps for aquatic species version 10/2019*. www.aquamaps.org
- Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4), 772–780. <https://doi.org/10.1093/molbev/mst010>
- Kawecki, T. J., & Ebert, D. (2004). Conceptual issues in local adaptation. *Ecology Letters*, 7(12), 1225–1241. <https://doi.org/10.1111/j.1461-0248.2004.00684.x>
- Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., & Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12), 1647–1649. <https://doi.org/10.1093/bioinformatics/bts199>
- Keel, B. N., & Snelling, W. M. (2018). Comparison of Burrows-Wheeler transform-based mapping algorithms used in high-throughput whole-genome sequencing: Application to illumina data for livestock genomes 1. *Frontiers in Genetics*, 9(35), 1–6. <https://doi.org/10.3389/fgene.2018.00035>
- Kerr, L. A., Hintzen, N. T., Cadrin, S. X., Clausen, L. W., Dickey-Collas, M., Goethel, D. R., Hatfield, E. M. C., Kritzer, J. P., & Nash, R. D. M. (2017). Lessons learned from practical

- approaches to reconcile mismatches between biological population structure and stock units of marine fish. *ICES Journal of Marine Science*, 74(6), 1708–1722. <https://doi.org/10.1093/icesjms/fsw188>
- Kersey, P. J., Allen, J. E., Armean, I., Boddu, S., Bolt, B. J., Carvalho-Silva, D., Christensen, M., Davis, P., Falin, L. J., Grabmueller, C., Humphrey, J., Kerhornou, A., Khobova, J., Aranganathan, N. K., Langridge, N., Lowy, E., McDowall, M. D., Maheswari, U., Nuhn, M., ... Staines, D. M. (2016). Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Research*, 44(D1), D574–D580. <https://doi.org/10.1093/nar/gkv1209>
- Kimura, K., Imamura, H., & Kawai, T. (2018). Comparative morphology and phylogenetic systematics of the families Cheilodactylidae and Latridae (Perciformes: Cirrhitidae), and proposal of a new classification. *Zootaxa*, 4536(1), 1–72. <https://doi.org/10.11646/zootaxa.4536.1.1>
- Kimura, M. (1968). Evolutionary Rate at the Molecular Level. *Nature*, 217(5129), 624–626. <https://doi.org/10.1038/217624a0>
- Klingström, T., Bongcam-Rudloff, E., & Pettersson, O. V. (2018). A comprehensive model of DNA fragmentation for the preservation of High Molecular Weight DNA. *BioRxiv*. <https://doi.org/10.1101/254276>
- Knaus, B. J., & Grünwald, N. J. (2017). VCFR: a package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources*, 17(1), 44–53. <https://doi.org/10.1111/1755-0998.12549>
- Koepfli, K. P., Paten, B., O'brien, S. J., Antunes, A., Belov, K., Bustamante, C., Castoe, T. A., Clawson, H., Crawford, A. J., Diekhans, M., Distel, D., Durbin, R., Earl, D., Fujita, M. K., Gamble, T., Georges, A., Gemmell, N., Gilbert, M. T. P., Graves, J. M., ... Ryder, O. (2015). The genome 10K project: A way forward. *Annual Review of Animal Biosciences*, 3, 57–111. <https://doi.org/10.1146/annurev-animal-090414-014900>
- Kolde, R. (2019). *pheatmap: Pretty Heatmaps*. R package version 1.0.12. <https://cran.r-project.org/package=pheatmap>
- Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, 37(5), 540–546.

- <https://doi.org/10.1038/s41587-019-0072-8>
- Koot, E., Wu, C., Ruza, I., Hilario, E., Storey, R., Wells, R., Chagné, D., & Wellenreuther, M. (2021). Genome-wide analysis reveals the genetic stock structure of hoki (*Macruronus novaezelandiae*). *Evolutionary Applications*. <https://doi.org/10.1111/eva.13317>
- Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. A., McCombie, W. R., Jarvis, E. D., & Phillippy, A. M. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology*, 30(7), 693–700. <https://doi.org/10.1038/nbt.2280>
- Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, 5(59). <https://doi.org/10.1186/1471-2105-5-59>
- Kovach, A. I., Breton, T. S., Berlinsky, D. L., Maceda, L., & Wirgin, I. (2010). Fine-scale spatial and temporal genetic structure of Atlantic cod off the Atlantic coast of the USA. *Marine Ecology Progress Series*, 410, 177–195. <https://doi.org/10.3354/meps08612>
- Kraft, D. W., Conklin, E. E., Barba, E. W., Hutchinson, M., Toonen, R. J., Forsman, Z. H., & Bowen, B. W. (2020). Genomics versus mtDNA for resolving stock structure in the silky shark (*Carcharhinus falciformis*). *PeerJ*, 8(October). <https://doi.org/10.7717/peerj.10186>
- Kryazhimskiy, S., & Plotkin, J. B. (2008). The Population Genetics of dN/dS. *PLoS Genetics*, 4(12), e1000304. <https://doi.org/10.1371/journal.pgen.1000304>
- Kumar, G., & Kocour, M. (2017). Applications of next-generation sequencing in fisheries research: A review. *Fisheries Research*, 186, 11–22. <https://doi.org/10.1016/j.fishres.2016.07.021>
- Kuo, R. I., Cheng, Y., Zhang, R., Brown, J. W. S., Smith, J., Archibald, A. L., & Burt, D. W. (2020). Illuminating the dark side of the human transcriptome with long read transcript sequencing. *BMC Genomics*, 21(751), 1:22. <https://doi.org/10.1186/s12864-020-07123-7>
- Kuo, R. I., Tseng, E., Eory, L., Paton, I. R., Archibald, A. L., & Burt, D. W. (2017). Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. *BMC Genomics*, 18(323), 1–19. <https://doi.org/10.1186/s12864-017-3691-9>
- Laikre, L., Palm, S., & Ryman, N. (2005). Genetic population structure of fishes: implications

- for coastal zone management. *AMBIO: A Journal of the Human Environment*, 34(2), 111–119. <https://doi.org/10.1579/0044-7447-34.2.111>
- Lal, M. M., Southgate, P. C., Jerry, D. R., & Zenger, K. R. (2016). Fishing for divergence in a sea of connectivity: The utility of ddRADseq genotyping in a marine invertebrate, the black-lip pearl oyster *Pinctada margaritifera*. *Marine Genomics*, 25, 57–68. <https://doi.org/10.1016/j.margen.2015.10.010>
- Lane, H. S., Symonds, J. E., & Ritchie, P. A. (2016). The phylogeography and population genetics of *Polyprion oxygeneios* based on mitochondrial DNA sequences and microsatellite DNA markers. *Fisheries Research*, 174, 19–29. <https://doi.org/10.1016/j.fishres.2015.08.009>
- Langley, A. D. (2017). *Fishery characterisation and Catch-Per-Unit-Effort indices for tarakihi in TAR 1, TAR 2 and TAR 3* [New Zealand Fisheries Assessment Report 2017/44]. Ministry for Primary Industries. <http://fs.fish.govt.nz/Doc/24486/FAR-2017-44-tarakihi-CPUE.pdf.ashx>
- Langley, A. D. (2018). *Stock assessment of tarakihi off the east coast of mainland New Zealand* [New Zealand Fisheries Assessment Report 2018/05]. Ministry for Primary Industries.
- Lantz, H., Dominguez Del Angel, V., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Vinnere Pettersson, O., Amselem, J., Bouri, L., Bocs, S., Klopp, C., Gibrat, J. F., Vlasova, A., Leskosek, B. L., Soler, L., & Binzer-Panchal, M. (2018). Ten steps to get started in Genome Assembly and Annotation. *F1000Research*, 7. <https://doi.org/10.12688/f1000research.13598.1>
- Lemopoulos, A., Prokkola, J. M., Uusi-Heikkilä, S., Vasemägi, A., Huusko, A., Hyvärinen, P., Koljonen, M.-L., Koskiniemi, J., & Vainikka, A. (2019). Comparing RADseq and microsatellites for estimating genetic diversity and relatedness — Implications for brown trout conservation. *Ecology and Evolution*, 9(4), 2106–2120. <https://doi.org/10.1002/ece3.4905>
- Lenormand, T. (2002). Gene flow and the limits to natural selection. *Trends in Ecology & Evolution*, 17(4), 183–189. [https://doi.org/10.1016/S0169-5347\(02\)02497-7](https://doi.org/10.1016/S0169-5347(02)02497-7)
- Lerat, E. (2018). Repeat in genomes: How and why you should consider them in genome

- analyses? In *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics* (Vols. 1–3, Issue 1950). Elsevier Ltd. <https://doi.org/10.1016/B978-0-12-809633-8.20227-6>
- Leslie, R. W., Gon, O., & Gouws, G. (2018). The taxonomic status of the South African straptail, *Macruronus capensis* Davies, 1950 (Pisces, Gadiformes, Macruronidae). *Zootaxa*, 4374(1), 91–98. <https://doi.org/10.11646/zootaxa.4374.1.5>
- Lévy-Hartmann, L., Roussel, V., Letourneur, Y., & Sellos, D. Y. (2011). Global and New Caledonian patterns of population genetic variation in the deep-sea splendid alfonsino, *Beryx splendens*, inferred from mtDNA. *Genetica*, 139(11–12), 1349–1365. <https://doi.org/10.1007/s10709-012-9628-y>
- Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21), 2987–2993. <https://doi.org/10.1093/bioinformatics/btr509>
- Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. <https://arxiv.org/abs/1303.3997v2>
- Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18), 3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Li, Y.-H., & Wang, H.-P. (2017). Advances of genotyping-by-sequencing in fisheries and aquaculture. *Reviews in Fish Biology and Fisheries*, 27(3), 535–559. <https://doi.org/10.1007/s11160-017-9473-2>
- Liu, K., Xu, D., Li, J., Bian, C., Duan, J., Zhou, Y., Zhang, M., You, X., You, Y., Chen, J., Yu, H., Xu, G., Fang, D. an, Qiang, J., Jiang, S., He, J., Xu, J., Shi, Q., Zhang, Z., & Xu, P. (2017). Whole genome sequencing of Chinese clearhead icefish, *Protosalanx hyalocranius*. *GigaScience*,

- 6(4), 1–6. <https://doi.org/10.1093/gigascience/giw012>
- Livingston, M. E., & Schofield, K. A. (1996). Stock discrimination of hoki (*Macruronus novaezelandiae*, Merlucciidae) in New Zealand waters using morphometrics. *New Zealand Journal of Marine and Freshwater Research*, 30(2), 197–208. <https://doi.org/10.1080/00288330.1996.9516708>
- Long, M., Betrán, E., Thornton, K., & Wang, W. (2003). The origin of new genes: glimpses from the young and old. *Nature Reviews Genetics*, 4(11), 865–875. <https://doi.org/10.1038/nrg1204>
- Lu, L., Zhao, J., & Li, C. (2020). High-Quality Genome Assembly and Annotation of the Big-Eye Mandarin Fish (*Siniperca kneri*). *G3: Genes, Genomes, Genetics*, 10(3), 877–880. <https://doi.org/10.1534/g3.119.400930>
- Ludt, W. B., BurrIDGE, C. P., & Chakrabarty, P. (2019). A taxonomic revision of Cheilodactylidae and Latridae (Centrarchiformes: Cirrhitidae) using morphological and genomic characters. *Zootaxa*, 4585(1), 121–141. <https://doi.org/10.11646/zootaxa.4585.1.7>
- Mace, P., Ritchie, P., Wellenreuther, M., McKenzie, J., Hupman, K., Hillary, R., Francis, M., & Bunce, M. (2020). *Report of the Workshop on the Utility of Genetic Analyses for Addressing New Zealand Fisheries Questions*.
- Machado-Schiaffino, G., Campo, D., & Garcia-Vazquez, E. (2009). Strong genetic differentiation of the Austral hake (*Merluccius australis*) across the species range. *Molecular Phylogenetics and Evolution*, 53(1), 351–356. <https://doi.org/10.1016/j.ympev.2009.02.017>
- Malmstrøm, M., Matschiner, M., Tørresen, O. K., Jakobsen, K. S., & Jentoft, S. (2017). Data descriptor: Whole genome sequencing data and de novo draft assemblies for 66 teleost species. *Scientific Data*, 4, 1–13. <https://doi.org/10.1038/sdata.2016.132>
- Malmstrøm, M., Matschiner, M., Tørresen, O. K., Star, B., Snipen, L. G., Hansen, T. F., Baalsrud, H. T., Nederbragt, A. J., Hanel, R., Salzburger, W., Stenseth, N. C., Jakobsen, K. S., & Jentoft, S. (2016). Evolution of the immune system influences speciation rates in teleost fishes. *Nature Genetics*, 48(10), 1204–1210. <https://doi.org/10.1038/ng.3645>
- Marandel, F., Lorange, P., Berthel, O., Trenkel, V. M., Waples, R. S., & Lamy, J.-B. (2019).

- Estimating effective population size of large marine populations, is it feasible? *Fish and Fisheries*, 20(1), 189–198. <https://doi.org/10.1111/faf.12338>
- Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics*, 27(6), 764–770. <https://doi.org/10.1093/bioinformatics/btr011>
- Marques, D. A., Jones, F. C., Di Palma, F., Kingsley, D. M., & Reimchen, T. E. (2018). Experimental evidence for rapid genomic adaptation to a new niche in an adaptive radiation. *Nature Ecology & Evolution*, 2(7), 1128–1138. <https://doi.org/10.1038/s41559-018-0581-8>
- Marra, N. J., Stanhope, M. J., Jue, N. K., Wang, M., Sun, Q., Pavinski Bitar, P., Richards, V. P., Komissarov, A., Rayko, M., Kliver, S., Stanhope, B. J., Winkler, C., O'Brien, S. J., Antunes, A., Jorgensen, S., & Shivji, M. S. (2019). White shark genome reveals ancient elasmobranch adaptations associated with wound healing and the maintenance of genome stability. *Proceedings of the National Academy of Sciences*, 116(10), 4446–4455. <https://doi.org/10.1073/pnas.1819778116>
- Masel, J. (2011). Genetic drift. *Current Biology*, 21(20), R837–R838. <https://doi.org/10.1016/j.cub.2011.08.007>
- Maunder, M. N., & Piner, K. R. (2015). Contemporary fisheries stock assessment: many issues still remain. *ICES Journal of Marine Science*, 72(1), 7–18. <https://doi.org/10.1093/icesjms/fsu015>
- McCauley, D. J., Pinsky, M. L., Palumbi, S. R., Estes, J. A., Joyce, F. H., & Warner, R. R. (2015). Marine defaunation: Animal loss in the global ocean. *Science*, 347(6219). <https://doi.org/10.1126/science.1255641>
- McKenzie, J. R., Beentjes, M., Parker, S., Parsons, D. M., Armiger, H., Wilson, O., Middleton, D., Langley, A., Buckthought, D., Walsh, C., Bian, R., Maolagáin, C. Ó., Stevenson, M., Sutton, C., Spong, K., Rush, N., & Smith, M. (2017). *Fishery characterisation and age composition of tarakihi in TAR 1, 2 and 3 for 2013/14 and 2014/15* [New Zealand Fisheries Assessment Report 2017/36]. Ministry for Primary Industries. <http://www.mpi.govt.nz/news-and-resources/publications>

- McKenzie, M. K. (1961). Fisheries research in the East Cape area. *Fisheries Technical Report*, 3, 1–9.
- Mejía-Ruíz, P., Perez-Enriquez, R., Mares-Mayagoitia, J. A., & Valenzuela-Quiñonez, F. (2020). Population genomics reveals a mismatch between management and biological units in green abalone (*Haliotis fulgens*). *PeerJ*, 8, e9722. <https://doi.org/10.7717/peerj.9722>
- Melkikh, A. V., & Seleznev, V. D. (2012). Mechanisms and models of the active transport of ions and the transformation of energy in intracellular compartments. *Progress in Biophysics and Molecular Biology*, 109(1–2), 33–57. <https://doi.org/10.1016/j.pbiomolbio.2012.04.008>
- Migiro, G. (2018). *Countries with the Largest Exclusive Economic Zones*. <https://www.worldatlas.com/articles/countries-with-the-largest-exclusive-economic-zones.html>
- Miller, J. R., Delcher, A. L., Koren, S., Venter, E., Walenz, B. P., Brownley, A., Johnson, J., Li, K., Mobarry, C., & Sutton, G. (2008). Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics*, 24(24), 2818–2824. <https://doi.org/10.1093/bioinformatics/btn548>
- Miller, L. M., Ward, M. C., & Schultz, D. W. (2015). Using Genetic Markers as Individual Tags: a Case Study of a Mark–Recapture Estimate of Adult Muskellunge Population Size. *North American Journal of Fisheries Management*, 35(2), 210–215. <https://doi.org/10.1080/02755947.2014.996688>
- Miller, P. A., Fitch, A. J., Gardner, M., Hutson, K. S., & Mair, G. (2011). Genetic population structure of Yellowtail Kingfish (*Seriola lalandi*) in temperate Australasian waters inferred from microsatellite markers and mitochondrial DNA. *Aquaculture*, 319(3–4), 328–336. <https://doi.org/10.1016/j.aquaculture.2011.05.036>
- Miller, W., Makova, K. D., Nekrutenko, A., & Hardison, R. C. (2004). Comparative genomics. *Annual Review of Genomics and Human Genetics*, 5(1), 15–56. <https://doi.org/10.1146/annurev.genom.5.061903.180057>
- Minegishi, Y., Aoyama, J., Inoue, J. G., Azanza, R. V., & Tsukamoto, K. (2009). Inter-specific and subspecific genetic divergences of freshwater eels, genus *Anguilla* including a recently

- described species , *A. luzonensis* , based on whole mitochondrial genome sequences. *Coastal Marine Science*, 33(1), 1–14.
- Minegishi, Y., Aoyama, J., Inoue, J. G., Miya, M., Nishida, M., & Tsukamoto, K. (2005). Molecular phylogeny and evolution of the freshwater eels genus *Anguilla* based on the whole mitochondrial genome sequences. *Molecular Phylogenetics and Evolution*, 34(1), 134–146. <https://doi.org/10.1016/j.ympev.2004.09.003>
- Ministry for Primary Industries. (2017). *Fisheries Assessment Plenary May 2017: Stock Assessments and Stock Status*.
- Morgan, J. A. T., Sumpton, W. D., Jones, A. T., Campbell, A. B., Stewart, J., Hamer, P., & Ovenden, J. R. (2019). Assessment of genetic structure among Australian east coast populations of snapper *Chrysophrys auratus* (Sparidae). *Marine and Freshwater Research*, 70(7), 964–976. <https://doi.org/10.1071/MF18146>
- Morjan, C. L., & Rieseberg, L. H. (2004). How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles. *Molecular Ecology*, 13(6), 1341–1356. <https://doi.org/10.1111/j.1365-294X.2004.02164.x>
- Morley, J. W., Selden, R. L., Latour, R. J., Frölicher, T. L., Seagraves, R. J., & Pinsky, M. L. (2018). Projecting shifts in thermal habitat for 686 species on the North American continental shelf. *PLoS ONE*, 13(5), 1–28. <https://doi.org/10.1371/journal.pone.0196127>
- Morrison, M. A., Jones, E. G., Parsons, D. P., & Grant, C. M. (2014). *Habitats and areas of particular significance for coastal finfish fisheries management in New Zealand: A review of concepts and life history knowledge, and suggestions for future research* [New Zealand Aquatic Environment and Biodiversity Report No. 125]. Ministry for Primary Industries. <http://www.mpi.govt.nz/news-resources/publications.aspx>
- Mullon, C., Freon, P., & Cury, P. (2005). The dynamics of collapse in world fisheries. *Fish and Fisheries*, 6(2), 111–120. <https://doi.org/10.1111/j.1467-2979.2005.00181.x>
- Munday, P. L., Donelson, J. M., & Domingos, J. A. (2017). Potential for adaptation to climate change in a coral reef fish. *Global Change Biology*, 23(1), 307–317. <https://doi.org/10.1111/gcb.13419>
- Munday, P. L., Schunter, C., Allan, B. J. M., Nicol, S., Parsons, D. M., Pether, S. M. J., Pope, S.,

- Ravasi, T., Setiawan, A. N., Smith, N., & Domingos, J. A. (2019). Testing the adaptive potential of yellowtail kingfish to ocean warming and acidification. *Frontiers in Ecology and Evolution*, 7, 1–10. <https://doi.org/10.3389/fevo.2019.00253>
- Muñoz, N. J., Farrell, A. P., Heath, J. W., & Neff, B. D. (2015). Adaptive potential of a Pacific salmon challenged by climate change. *Nature Climate Change*, 5, 163–166. <https://doi.org/10.1038/nclimate2473>
- Myers, R. A., Hutchings, J. A., & Barrowman, N. J. (1997). Why do fish stocks collapse? The example of cod in Atlantic Canada. *Ecological Applications*, 7(1), 91–106. [https://doi.org/10.1890/1051-0761\(1997\)007\[0091:WDFSCT\]2.0.CO;2](https://doi.org/10.1890/1051-0761(1997)007[0091:WDFSCT]2.0.CO;2)
- Myers, R. A., & Worm, B. (2003). Rapid worldwide depletion of predatory fish communities. *Nature*, 423(6937), 280–283. <https://doi.org/10.1038/nature01610>
- Nazareno, A. G., Bemmels, J. B., Dick, C. W., & Lohmann, L. G. (2017). Minimum sample sizes for population genomics: an empirical study from an Amazonian plant species. *Molecular Ecology Resources*, 17(6), 1136–1147. <https://doi.org/10.1111/1755-0998.12654>
- Nguinkal, J. A., Brunner, R. M., Verleih, M., Rebl, A., de los Ríos-Pérez, L., Schäfer, N., Hadlich, F., Stüeken, M., Wittenburg, D., & Goldammer, T. (2019). The first highly contiguous genome assembly of pikeperch (*Sander lucioperca*), an emerging aquaculture species in Europe. *Genes*, 10(9), 708. <https://doi.org/10.3390/genes10090708>
- Nielsen, E. E., Hemmer-Hansen, J., Larsen, P. F., & Bekkevold, D. (2009). Population genomics of marine fishes: Identifying adaptive variation in space and time. *Molecular Ecology*, 18(15), 3128–3150. <https://doi.org/10.1111/j.1365-294X.2009.04272.x>
- Nielsen, R. (2005). Molecular Signatures of Natural Selection. *Annual Review of Genetics*, 39(1), 197–218. <https://doi.org/10.1146/annurev.genet.39.073003.112420>
- Nosil, P., & Feder, J. L. (2013). GENOME EVOLUTION AND SPECIATION: TOWARD QUANTITATIVE DESCRIPTIONS OF PATTERN AND PROCESS. *Evolution*, 67(9), 2461–2467. <https://doi.org/10.1111/evo.12191>
- Nudelman, G., Frasca, A., Kent, B., Sadler, K. C., Sealfon, S. C., Walsh, M. J., & Zaslavsky, E. (2018). High resolution annotation of zebrafish transcriptome using long-read sequencing. *Genome Research*, 28(9), 1415–1425.

- <https://doi.org/10.1101/gr.223586.117>
- Nugroho, E., Ferrell, D. J., Smith, P., & Taniguchi, N. (2001). Genetic divergence of kingfish from Japan, Australia and New Zealand inferred by microsatellite DNA and mitochondrial DNA control region markers. *Fisheries Science*, 67(5), 843–850. <https://doi.org/10.1046/j.1444-2906.2001.00331.x>
- Nurk, S., Walenz, B. P., Rhie, A., Vollger, M. R., Logsdon, G. A., Grothe, R., Miga, K. H., Eichler, E. E., Phillippy, A. M., & Koren, S. (2020). HiCanu: Accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Research*, 30(9), 1291–1305. <https://doi.org/10.1101/GR.263566.120>
- Ohta, T. (1992). The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics*, 23(1), 263–286. <https://doi.org/10.1146/annurev.es.23.110192.001403>
- Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., & Wagner, H. (2020). *vegan: Community Ecology Package* (R package version 2.5-7). <https://cran.r-project.org/package=vegan>
- Olavarria, C., Balbontin, F., Bernal, R., & Baker, C. S. (2006). Lack of divergence in the mitochondrial cytochrome *b* gene between *Macruronus species* (Pisces: Merlucciidae) in the Southern Hemisphere. *New Zealand Journal of Marine and Freshwater Research*, 40(2), 299–304. <https://doi.org/10.1080/00288330.2006.9517422>
- Oleksiak, M. F., & Rajora, O. P. (Eds.). (2020). *Population Genomics: Marine Organisms*. Springer International Publishing. <https://doi.org/10.1007/978-3-030-37936-0>
- Oosting, T. (2021). *Connecting the past, present and future: A population genomic study of Australasian snapper (Chrysophrys auratus) in New Zealand* [Doctoral thesis]. Victoria University of Wellington, New Zealand.
- Oosting, T., Hilario, E., Wellenreuther, M., & Ritchie, P. A. (2020). DNA degradation in fish: Practical solutions and guidelines to improve DNA preservation for genomic research. *Ecology and Evolution*, 10(16), 8643–8651. <https://doi.org/10.1002/ece3.6558>
- Oosting, T., Star, B., Barrett, J. H., Wellenreuther, M., Ritchie, P. A., & Rawlence, N. J. (2019).

- Unlocking the potential of ancient fish DNA in the genomic era. *Evolutionary Applications*, 12(8), 1513–1522. <https://doi.org/10.1111/eva.12811>
- Orensanz, J. M., Armstrong, J., Armstrong, D., & Hilborn, R. (1998). Crustacean resources are vulnerable to serial depletion: the multifaceted decline of crab and shrimp fisheries in the Greater Gulf of Alaska. *Reviews in Fish Biology and Fisheries*, 8(September), 117–176. <https://doi.org/10.1023/A:1008891412756>
- Ovenden, J. R. (2013). Crinkles in connectivity: combining genetics and other types of biological data to estimate movement and interbreeding between populations. *Marine and Freshwater Research*, 64(3), 201. <https://doi.org/10.1071/mf12314>
- Ovenden, J. R., Berry, O., Welch, D. J., Buckworth, R. C., & Dichmont, C. M. (2015). Ocean's eleven: a critical evaluation of the role of population, evolutionary and molecular genetics in the management of wild fisheries. *Fish and Fisheries*, 16(1), 125–159. <https://doi.org/10.1111/faf.12052>
- Ovenden, J. R., Smolenski, A. J., & Whit, R. W. G. (1989). Mitochondrial DNA restriction site variation in Tasmanian populations of orange roughy (*Hoplostethus atlanticus*), a deep-water marine teleost. *Marine and Freshwater Research*, 40(1), 1–9. <https://doi.org/10.1071/MF9890001>
- PacBio. (2020). *SMRT Link v9.0*. <https://www.pacb.com/support/software-downloads/>
- Palsbøll, P. J., Bérubé, M., & Allendorf, F. W. (2007). Identification of management units using population genetic data. *Trends in Ecology & Evolution*, 22(1), 11–16. <https://doi.org/10.1016/j.tree.2006.09.003>
- Papa, Y., Halliwell, A. G., Morrison, M. A., Wellenreuther, M., & Ritchie, P. A. (2021). Phylogeographic structure and historical demography of tarakihi (*Nemadactylus macropterus*) and king tarakihi (*Nemadactylus n.sp.*) in New Zealand. *New Zealand Journal of Marine and Freshwater Research*, 1–25. <https://doi.org/10.1080/00288330.2021.1912119>
- Papa, Y., Oosting, T., Valenza-Troubat, N., Wellenreuther, M., & Ritchie, P. A. (2021). Genetic stock structure of New Zealand fish and the use of genomics in fisheries management: an overview and outlook. *New Zealand Journal of Zoology*, 48(1), 1–31.

- <https://doi.org/10.1080/03014223.2020.1788612>
- Parsons, D., Sim-Smith, C., Cryer, M., Francis, M., Hartill, B., Jones, E., Le Port, A., Lowe, M., McKenzie, J., Morrison, M., Paul, L., Radford, C., Ross, P., Spong, K., Trnski, T., Usmar, N., Walsh, C., & Zeldis, J. (2014). Snapper (*Chrysophrys auratus*): a review of life history and key vulnerabilities in New Zealand. *New Zealand Journal of Marine and Freshwater Research*, 48(2), 256–283. <https://doi.org/10.1080/00288330.2014.892013>
- Paul, L. J. (1986). *New Zealand fishes: an identification guide*. Reed Methuen.
- Pauly, D., Christensen, V., Gunette, S., Pitcher, T. J., Sumaila, U. R., Walters, C. J., Watson, R., & Zeller, D. (2002). Towards sustainability in world fisheries. *Nature*, 418(6898), 689–695. <https://doi.org/10.1038/nature01017>
- Pauly, D., & Zeller, D. (2016). Catch reconstructions reveal that global marine fisheries catches are higher than reported and declining. *Nature Communications*, 7. <https://doi.org/10.1038/ncomms10244>
- Pecoraro, C., Babbucci, M., Franch, R., Rico, C., Papetti, C., Chassot, E., Bodin, N., Cariani, A., Bargelloni, L., & Tinti, F. (2018). The population genomics of yellowfin tuna (*Thunnus albacares*) at global geographic scale challenges current stock delineation. *Scientific Reports*, 8(1), 1–10. <https://doi.org/10.1038/s41598-018-32331-3>
- Pembleton, L. W., Cogan, N. O. I., & Forster, J. W. (2013). StAMPP: an R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. *Molecular Ecology Resources*, 13(5), 946–952. <https://doi.org/10.1111/1755-0998.12129>
- Piccoli, G. R. (2021). *grpicoli/assemblies-stats: (Version 1.1.1)*. Zenodo. <https://doi.org/10.5281/zenodo.4703697>
- Pinsky, M. L., & Palumbi, S. R. (2014). Meta-analysis reveals lower genetic diversity in overfished populations. *Molecular Ecology*, 23(1), 29–39. <https://doi.org/10.1111/mec.12509>
- Plank, M. J., Kolding, J., Law, R., Gerritsen, H. D., & Reid, D. (2017). Balanced harvesting can emerge from fishing decisions by individual fishers in a small-scale fishery. *Fish and Fisheries*, 18(2), 212–225. <https://doi.org/10.1111/faf.12172>
- Pootakham, W., Sonthirod, C., Naktang, C., Nawae, W., Yoocha, T., Kongkachana, W.,

- Sangsrakru, D., Jomchai, N., Uthoomporn, S., Sheedy, J. R., Buaboocha, J., Mekiyanon, S., & Tangphatsornruang, S. (2020). *De novo* assemblies of *Luffa acutangula* and *Luffa cylindrica* genomes reveal an expansion associated with substantial accumulation of transposable elements. *Molecular Ecology Resources*, 1–14. <https://doi.org/10.1111/1755-0998.13240>
- Portnoy, D. S., & Heist, E. J. (2012). Molecular markers: Progress and prospects for understanding reproductive ecology in elasmobranchs. *Journal of Fish Biology*, 80(5), 1120–1140. <https://doi.org/10.1111/j.1095-8649.2011.03206.x>
- Potts, W. M., Henriques, R., Santos, C. V., Munnik, K., Ansorge, I., Dufois, F., Booth, A. J., Kirchner, C., Sauer, W. H. H., & Shaw, P. W. (2014). Ocean warming, a rapid distributional shift, and the hybridization of a coastal fish species. *Global Change Biology*, 20(9), 2765–2777. <https://doi.org/10.1111/gcb.12612>
- Pullin, R. S. V. (2000). Management of Aquatic Biodiversity and Genetic Resources. *Reviews in Fisheries Science*, 8(4), 379–393. <https://doi.org/10.1080/10408340308951117>
- Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841–842. <https://doi.org/10.1093/bioinformatics/btq033>
- R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <http://www.r-project.org/>
- Raj, A., Stephens, M., & Pritchard, J. K. (2014). FastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics*, 197(2), 573–589. <https://doi.org/10.1534/genetics.114.164350>
- Ramos, J. E., Pecl, G. T., Moltschaniwskyj, N. A., Semmens, J. M., Souza, C. A., & Strugnell, J. M. (2018). Population genetic signatures of a climate change driven marine range extension. *Scientific Reports*, 8, 1–12. <https://doi.org/10.1038/s41598-018-27351-y>
- Randhawa, S. S., & Pawar, R. (2021). Fish genomes: Sequencing trends, taxonomy and influence of taxonomy on genome attributes. *Journal of Applied Ichthyology*, 37(4), 553–562. <https://doi.org/10.1111/jai.14227>
- Reiss, H., Hoarau, G., Dickey-Collas, M., & Wolff, W. J. (2009). Genetic population structure of

- marine fish: Mismatch between biological and fisheries management units. *Fish and Fisheries*, 10(4), 361–395. <https://doi.org/10.1111/j.1467-2979.2008.00324.x>
- Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., & Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology*, 24(17), 4348–4370. <https://doi.org/10.1111/mec.13322>
- Revelle, W. (2021). *psych: Procedures for Psychological, Psychometric, and Personality Research* (R package version 2.1.3). <https://cran.r-project.org/package=psych>
- Rice, E. S., & Green, R. E. (2019). New approaches for genome assembly and scaffolding. *Annual Review of Animal Biosciences*, 7(1), 17–40. <https://doi.org/10.1146/annurev-animal-020518-115344>
- Richard, G.-F., Kerrest, A., & Dujon, B. (2008). Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiology and Molecular Biology Reviews*, 72(4), 686–727. <https://doi.org/10.1128/mnbr.00011-08>
- Richardson, B. J. (1982). Geographical Distribution of Electrophoretically Detected Protein Variation in Australian Commercial Fishes. II.* Jackass Morwong, *Cheilodactylus macropterus* Bloch & Schneider. *Australian Journal of Marine and Freshwater Research*, 33, 927–931. <http://www.publish.csiro.au/mf/pdf/MF9820927>
- Roach, M. J., Schmidt, S. A., & Borneman, A. R. (2018). Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, 19(460), 1–10. <https://doi.org/10.1186/s12859-018-2485-7>
- Roberts, C. D., Stewart, A. L., & Struthers, C. D. (2015). *The Fishes of New Zealand* (C. D. Roberts, A. L. Stewart, & C. D. Struthers (Eds.)). Te Papa Press.
- 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*. Museum of New Zealand Te Papa Tongarewa, Wellington. <https://collections.tepapa.govt.nz/document/10564>
- Robinson, N., Skinner, A., Sethuraman, L., McPartlan, H., Murray, N., Knuckey, I., Smith, D. C., Hindell, J., & Talman, S. (2008). Genetic stock structure of blue-eye trevalla (*Hyperoglyphe antarctica*) and warehou (*Seriola lalandi* and *Seriola punctata*) in south-eastern Australian waters. *Marine and Freshwater Research*, 59(6), 502–514.

<https://doi.org/10.1071/MF07175>

- Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., & Huelsenbeck, J. P. (2012). MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3), 539–542. <https://doi.org/10.1093/sysbio/sys029>
- RStudio Team. (2020). *RStudio: Integrated development environment for R*. RStudio, PBC. <http://www.rstudio.com/>
- Russello, M. A., Kirk, S. L., Frazer, K. K., & Askey, P. J. (2011). *Detection of outlier loci and their utility for fisheries management*. <https://doi.org/10.1111/j.1752-4571.2011.00206.x>
- Ryan, A. W., Smith, P. J., & Mork, J. (2002). Genetic differentiation between the New Zealand and Falkland Islands populations of southern blue whiting *Micromesistius australis*. *New Zealand Journal of Marine and Freshwater Research*, 36(3), 637–643. <https://doi.org/10.1080/00288330.2002.9517119>
- Sanciango, M. D., Carpenter, K. E., & Betancur-R., R. (2016). Phylogenetic placement of enigmatic percomorph families (Teleostei: Percomorphaceae). *Molecular Phylogenetics and Evolution*, 94, 565–576. <https://doi.org/10.1016/j.ympev.2015.10.006>
- Sandoval-Castillo, J., Robinson, N. A., Hart, A. M., Strain, L. W. S., & Beheregaray, L. B. (2018). Seascape genomics reveals adaptive divergence in a connected and commercially important mollusc, the greenlip abalone (*Haliotis laevis*), along a longitudinal environmental gradient. *Molecular Ecology*, 27(7), 1603–1620. <https://doi.org/10.1111/mec.14526>
- Sarvas, T. H., & Fevolden, S. E. (2005). The scnDNA locus Pan I reveals concurrent presence of different populations of Atlantic cod (*Gadus morhua* L.) within a single fjord. *Fisheries Research*, 76(3), 307–316. <https://doi.org/10.1016/j.fishres.2005.07.013>
- Sbrocco, E. J., & Barber, P. H. (2013). MARSPEC: ocean climate layers for marine spatial ecology. *Ecology*, 94(4), 979–979. <https://doi.org/10.1890/12-1358.1>
- Schmarr, D. W., Whittington, I. D., Ovenden, J. R., & Ward, T. M. (2011). Discriminating Stocks of Scomber australis Using a Holistic Approach: a Pilot Study. *American Fisheries Society Symposium*, 76(January), 1–22. <https://doi.org/10.13140/2.1.1326.2724>

- Scoles, D. R., Collette, B. B., & Graves, J. E. (1998). Global phylogeography of mackerels of the genus *Scomber*. In *Fishery Bulletin* (Vol. 96, Issue 4, pp. 823–842).
- Sedberry, G. R., Carlin, J. L., Chapman, R. W., & Eleby, B. (1996). Population structure in the pan-oceanic wreckfish, *Polyprion americanus* (Teleostei: Polyprionidae), as indicated by mtDNA variation. *Journal of Fish Biology*, 49(sa), 318–329. <https://doi.org/10.1111/j.1095-8649.1996.tb06085.x>
- Sela, N., Kim, E., & Ast, G. (2010). The role of transposable elements in the evolution of non-mammalian vertebrates and invertebrates. *Genome Biology*, 11(6). <https://doi.org/10.1186/gb-2010-11-6-r59>
- Sgrò, C. M., Lowe, A. J., & Hoffmann, A. A. (2011). Building evolutionary resilience for conserving biodiversity under climate change. *Evolutionary Applications*, 4(2), 326–337. <https://doi.org/10.1111/j.1752-4571.2010.00157.x>
- Shao, F., Han, M., & Peng, Z. (2019). Evolution and diversity of transposable elements in fish genomes. *Scientific Reports*, 9(1), 1–8. <https://doi.org/10.1038/s41598-019-51888-1>
- Shen, K. N., & Tzeng, W. N. (2007). Genetic differentiation among populations of the shortfinned eel *Anguilla australis* from East Australia and New Zealand. *Journal of Fish Biology*, 70(sb), 177–190. <https://doi.org/10.1111/j.1095-8649.2007.01399.x>
- Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLOS ONE*, 11(10), e0163962. <https://doi.org/10.1371/journal.pone.0163962>
- Shou, C., & Han, Z. (2021). Genome-wide phylogenetic study of Percomorpha providing robust support for previous molecular classification. *Marine and Freshwater Research*, 11, 1387–1396. <https://doi.org/10.1071/MF20167>
- Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19), 3210–3212. <https://doi.org/10.1093/bioinformatics/btv351>
- Simison, W. B., Parham, J. F., Papenfuss, T. J., Lam, A. W., Henderson, J. B., Brian Simison, W., Parham, J. F., Papenfuss, T. J., Lam, A. W., & Henderson, J. B. (2020). An annotated

- chromosome-level reference genome of the red-eared slider turtle (*Trachemys scripta elegans*). *Genome Biology and Evolution*, 12(4), 456–462. <https://doi.org/10.1093/gbe/evaa063>
- Slatkin, M. (1987). Gene flow and the geographic structure of natural populations. *Science*, 236(4803), 787–792. <https://doi.org/10.1126/science.3576198>
- Smedbol, R. K., & Stephenson, R. (2001). The importance of managing within-species diversity in cod and herring fisheries of the north-western Atlantic. *Journal of Fish Biology*, 59(SUPPL. A), 109–128. <https://doi.org/10.1006/jfbi.2001.1765>
- Smit, A., Hubley, R., & Green, P. (2013). *RepeatMasker Open-4.0*. <http://www.repeatmasker.org>
- Smith, H. M. (2012). *Characterisation of the mitochondrial genome and the phylogeographic structure of blue cod (Parapercis colias)* [Master's thesis, Victoria University of Wellington, New Zealand]. <http://researcharchive.vuw.ac.nz/handle/10063/2139>
- Smith, P. J. (1979). Glucosephosphate isomerase and phosphoglucomutase polymorphisms in the New Zealand ling *Genypterus blacodes*. *Comparative Biochemistry and Physiology Part B: Comparative Biochemistry*, 62(4), 573–577. [https://doi.org/10.1016/0305-0491\(79\)90136-6](https://doi.org/10.1016/0305-0491(79)90136-6)
- Smith, P. J. (1986). Genetic similarity between samples of the orange roughy *Hoplostethus atlanticus* from the Tasman Sea, South-west Pacific Ocean and North-east Atlantic Ocean. *Marine Biology*, 91(2), 173–180. <https://doi.org/10.1007/BF00569433>
- Smith, P. J., & Benson, P. G. (1997). Genetic diversity in orange roughy from the east of New Zealand. *Fisheries Research*, 31(3), 197–213. [https://doi.org/10.1016/S0165-7836\(97\)00034-9](https://doi.org/10.1016/S0165-7836(97)00034-9)
- Smith, P. J., Benson, P. G., & McVeagh, S. M. (1997). A comparison of three genetic methods used for stock discrimination of orange roughy, *Hoplostethus atlanticus*: allozymes, mitochondrial DNA, and random amplified polymorphic DNA. *Fishery Bulletin*, 95(4), 800–811. <https://spo.nmfs.noaa.gov/fishery-bulletin-journal/954>
- Smith, P. J., Benson, P. G., Stanger, C., Chisnall, B. L., & Jellyman, D. J. (2001). Genetic structure of New Zealand eels *Anguilla dieffenbachii* and *A. australis* with allozyme markers.

- Ecology of Freshwater Fish*, 10(3), 132–137. <https://doi.org/10.1034/j.1600-0633.2001.100302.x>
- Smith, P. J., Diggles, B., & Kim, S. (2005). *Stock structure of blue mackerel, Scomber australasicus* (Issue August) [New Zealand Fisheries Assessment Report 2005/43]. Ministry of Fisheries.
- Smith, P. J., & Francis, R. I. C. C. (1982). A glucosephosphate isomerase polymorphism in New Zealand ling *Genypterus blacodes*. *Comparative Biochemistry and Physiology Part B: Comparative Biochemistry*, 73(3), 451–455. [https://doi.org/10.1016/0305-0491\(82\)90057-8](https://doi.org/10.1016/0305-0491(82)90057-8)
- Smith, P. J., Francis, R. I. C. C. C., & Paul, L. J. (1978). Genetic variation and population structure in the New Zealand snapper. *New Zealand Journal of Marine and Freshwater Research*, 12(4), 343–350. <https://doi.org/10.1080/00288330.1978.9515761>
- Smith, P. J., Gaffney, P. M., & Roberts, C. D. (2003). Phylogenetic relationships of the silver trumpeter *Latris pacifica* (Teleostei, Percomorpha, Latridae) based on allozymes and mitochondrial cytochrome b sequences. *Journal of the Royal Society of New Zealand*, 33(4), 755–767. <https://doi.org/10.1080/03014223.2003.9517757>
- Smith, P. J., & Johnston, A. D. (1985). Glucosephosphate isomerase and α -glycerophosphate dehydrogenase electromorph frequencies in groper *Polyprion oxygeneiosus* from central New Zealand. *New Zealand Journal of Marine and Freshwater Research*, 19(2), 173–177. <https://doi.org/10.1080/00288330.1985.9516084>
- Smith, P. J., McMillan, P. J., Bull, B., McVeagh, S. M., Gaffney, P. M., & Chow, S. (2002). Genetic and meristic variation in black and smooth oreos in the New Zealand exclusive economic zone. *New Zealand Journal of Marine and Freshwater Research*, 36(4), 737–750. <https://doi.org/10.1080/00288330.2002.9517127>
- Smith, P. J., McPhee, R. P., & Roberts, C. D. (2006). DNA and meristic evidence for two species of giant stargazer (Teleostei: Uranoscopidae: Kathetostoma) in New Zealand waters. *New Zealand Journal of Marine and Freshwater Research*, 40(3), 379–387. <https://doi.org/10.1080/00288330.2006.9517429>
- Smith, P. J., McVeagh, S. M., & Ede, A. (1996). Genetically isolated stocks of orange roughy

- (*Hoplostethus atlanticus*), but not of hoki (*Macruronus novaezelandiae*), in the Tasman Sea and Southwest Pacific Ocean around New Zealand. *Marine Biology*, 125(4), 783–793. <https://doi.org/10.1007/BF00349261>
- Smith, P. J., Patchell, G. J., & Benson, P. G. (1979). Glucosephosphate isomerase and isocitrate dehydrogenase polymorphisms in the hake, *Merluccius australis*. *New Zealand Journal of Marine and Freshwater Research*, 13(4), 545–547. <https://doi.org/10.1080/00288330.1979.9515831>
- Smith, P. J., Patchell, G. J., & Benson, P. G. (1981). Genetic tags in the New Zealand hoki *Macruronus novaezelandiae*. *Animal Blood Groups and Biochemical Genetics*, 12(1), 37–45. <https://doi.org/10.1111/j.1365-2052.1981.tb01529.x>
- Smith, P. J., & Paulin, C. D. (2003). Genetic and morphological evidence for a single species of pink ling (*Genypterus blacodes*) in New Zealand waters. *New Zealand Journal of Marine and Freshwater Research*, 37(1), 183–194. <https://doi.org/10.1080/00288330.2003.9517156>
- Smith, P. J., Roberts, C. D., McVeagh, S. M., & Benson, P. G. (1996). Genetic evidence for two species of tarakihi (Teleostei: Cheilodactylidae: *Nemadactylus*) in New Zealand waters. *New Zealand Journal of Marine and Freshwater Research*, 30(2), 209–220. <https://doi.org/10.1080/00288330.1996.9516709>
- Smith, P. J., Robertson, S. G., Horn, P. L., Bull, B., Anderson, O. F., Stanton, B. R., & Oke, C. S. (2002). Multiple techniques for determining stock relationships between orange roughy, *Hoplostethus atlanticus*, fisheries in the eastern Tasman Sea. *Fisheries Research*, 58(2), 119–140. [https://doi.org/10.1016/S0165-7836\(01\)00389-7](https://doi.org/10.1016/S0165-7836(01)00389-7)
- Smith, P. J., Steinke, D., Smith, P. J., Steinke, D., Mcmillan, P. J., Mcveagh, S. M., Carl, D., & Struthers, C. D. (2008). *DNA database for commercial marine fish* (Issue 22) [New Zealand Aquatic Environment and Biodiversity Report]. National Institute of Water and Atmospheric Research.
- Smith, P. J., Struthers, C. D., Paulin, C. D., McVeagh, S. M., & Daley, R. K. (2009). Shallow genetic and morphological divergence among seaperches in the South Pacific (family Scorpaenidae; Genus *Helicolenus*). *Journal of Fish Biology*, 74(5), 1104–1128. <https://doi.org/10.1111/j.1095-8649.2008.02172.x>

- Smolenski, A. J., Ovenden, J. R., & White, R. W. G. G. (1993). Evidence of stock separation in southern hemisphere orange roughy (*Hoplostethus atlanticus*, Trachichthyidae) from restriction-enzyme analysis of mitochondrial DNA. *Marine Biology: International Journal on Life in Oceans and Coastal Waters*, 116(2), 219–230. <https://doi.org/10.1007/BF00350011>
- Sotero-Caio, C. G., Platt, R. N., Suh, A., & Ray, D. A. (2017). Evolution and diversity of transposable elements in vertebrate genomes. *Genome Biology and Evolution*, 9(1), 161–177. <https://doi.org/10.1093/gbe/evw264>
- Spielman, D., Brook, B. W., & Frankham, R. (2004). Most species are not driven to extinction before genetic factors impact them. *Proceedings of the National Academy of Sciences*, 101(42), 15261–15264. <https://doi.org/10.1073/pnas.0403809101>
- Stanke, M., Steinkamp, R., Waack, S., & Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Research*, 32(Web Server), W309–W312. <https://doi.org/10.1093/nar/gkh379>
- Stanley, R. R. E., DiBacco, C., Lowen, B., Beiko, R. G., Jeffery, N. W., Van Wyngaarden, M., Bentzen, P., Brickman, D., Benestan, L., Bernatchez, L., Johnson, C., Snelgrove, P. V. R., Wang, Z., Wringe, B. F., & Bradbury, I. R. (2018). A climate-associated multispecies cryptic cline in the northwest Atlantic. *Science Advances*, 4(3). <https://doi.org/10.1126/sciadv.aag0929>
- Steinke, D., Salzburger, W., Braasch, I., & Meyer, A. (2006). Many genes in fish have species-specific asymmetric rates of molecular evolution. *BMC Genomics*, 7, 1–18. <https://doi.org/10.1186/1471-2164-7-20>
- Storer, J., Hubley, R., Rosen, J., Wheeler, T. J., & Smit, A. F. (2021). The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA*, 12(2), 1–14. <https://doi.org/10.1186/s13100-020-00230-y>
- Sun, C., Li, J., Dong, J., Niu, Y., Hu, J., Lian, J., Li, W., Li, J., Tian, Y., Shi, Q., & Ye, X. (2021). Chromosome-level genome assembly for the largemouth bass *Micropterus salmoides* provides insights into adaptation to fresh and brackish water. *Molecular Ecology Resources*, 21(1), 301–315. <https://doi.org/10.1111/1755-0998.13256>

- Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, 34(Web Server), W609–W612. <https://doi.org/10.1093/nar/gkl315>
- Tabata, K., & Taniguchi, N. (2000). Differences between *Pagrus major* and *Pagrus auratus* through mainly mtDNA control region analysis. *Fisheries Science*, 66(1), 9–18. <https://doi.org/10.1046/j.1444-2906.2000.00032.x>
- Takehana, Y., Zahm, M., Cabau, C., Klopp, C., Roques, C., Bouchez, O., Donnadieu, C., Barrachina, C., Journot, L., Kawaguchi, M., Yasumasu, S., Ansai, S., Naruse, K., Inoue, K., Shinzato, C., Scharl, M., Guiguen, Y., & Herpin, A. (2020). Genome sequence of the euryhaline javafish medaka, *Oryzias javanicus* : A small aquarium fish model for studies on adaptation to salinity. *G3: Genes, Genomes, Genetics*, 10(3), 907–915. <https://doi.org/10.1534/g3.119.400725>
- Takeshima, H., Hatanaka, A., Yamada, S., Yamazaki, Y., Kimura, I., & Nishida, M. (2011). Low genetic differentiation between two geographically separated populations of demersal gadiform fishes in the Southern Hemisphere. *Genes & Genetic Systems*, 86(5), 339–349. <https://doi.org/10.1266/ggs.86.339>
- Tan, M. H., Austin, C. M., Hammer, M. P., Lee, Y. P., Croft, L. J., & Gan, H. M. (2018). Finding Nemo: hybrid assembly with Oxford Nanopore and Illumina reads greatly improves the clownfish (*Amphiprion ocellaris*) genome assembly. *GigaScience*, 7(3), 1–6. <https://doi.org/10.1093/gigascience/gix137>
- Tanner, S. E., Pérez, M., Presa, P., Thorrold, S. R., & Cabral, H. N. (2014). Integrating microsatellite DNA markers and otolith geochemistry to assess population structure of European hake (*Merluccius merluccius*). *Estuarine, Coastal and Shelf Science*, 142, 68–75. <https://doi.org/10.1016/j.ecss.2014.03.010>
- Tanner, S. E., Reis-Santos, P., & Cabral, H. N. (2016). Otolith chemistry in stock delineation: A brief overview, current challenges and future prospects. *Fisheries Research*, 173, 206–213. <https://doi.org/10.1016/j.fishres.2015.07.019>
- Teske, P. R., Golla, T. R., Sandoval-Castillo, J., Emami-Khoyi, A., van der Lingen, C. D., von der Heyden, S., Chiazzari, B., Jansen van Vuuren, B., & Beheregaray, L. B. (2018). Mitochondrial DNA is unsuitable to test for isolation by distance. *Scientific Reports*, 8, 1–

9. <https://doi.org/10.1038/s41598-018-25138-9>
- Thai, B. T., Lee, Y. P., Gan, H. M., Austin, C. M., Croft, L. J., Trieu, T. A., & Tan, M. H. (2019). Whole genome assembly of the snout otter clam, *Lutraria rhynchaena*, using Nanopore and Illumina data, benchmarked against bivalve genome assemblies. *Frontiers in Genetics*, 10(1158), 1–8. <https://doi.org/10.3389/fgene.2019.01158>
- Therkildsen, N. O., Hemmer-Hansen, J., Als, T. D., Swain, D. P., Morgan, M. J., Trippel, E. A., Palumbi, S. R., Meldrup, D., & Nielsen, E. E. (2013). Microevolution in time and space: SNP analysis of historical DNA reveals dynamic signatures of selection in Atlantic cod. *Molecular Ecology*, 22(9), 2424–2440. <https://doi.org/10.1111/mec.12260>
- Therkildsen, N. O., Wilder, A. P., Conover, D. O., Munch, S. B., Baumann, H., & Palumbi, S. R. (2019). Contrasting genomic shifts underlie parallel phenotypic evolution in response to fishing. *Science*, 365(6452), 487–490. <https://doi.org/10.1126/science.aaw7271>
- Thomson, A. I., Archer, F. I., Coleman, M. A., Gajardo, G., Goodall-Copestake, W. P., Hoban, S., Laikre, L., Miller, A. D., O'Brien, D., Pérez-Espona, S., Segelbacher, G., Serrão, E. A., Sjøtun, K., & Stanley, M. S. (2021). Charting a course for genetic diversity in the UN Decade of Ocean Science. *Evolutionary Applications*, November 2020, 1–22. <https://doi.org/10.1111/eva.13224>
- Thrasher, D. J., Butcher, B. G., Campagna, L., Webster, M. S., & Lovette, I. J. (2018). Double-digest RAD sequencing outperforms microsatellite loci at assigning paternity and estimating relatedness: A proof of concept in a highly promiscuous bird. *Molecular Ecology Resources*, 18(5), 953–965. <https://doi.org/10.1111/1755-0998.12771>
- Thresher, R. E., Proctor, C. H., & Gunn, J. S. (1994). An evaluation of electron-probe microanalysis of otoliths for stock delineation and identification of nursery areas in a southern temperate groundfish, *Nemadactylus macropterus* (Cheilodactylidae). *Fishery Bulletin*, 92, 817–840. <https://spo.nmfs.noaa.gov/sites/default/files/pdf-content/1994/924/thresher.pdf>
- Tong, L. J., & Vooren, C. M. (1972). *The biology of the New Zealand tarakihi, Cheilodactylus macropterus (Bloch and Schneider)*. New Zealand Ministry of Agriculture and Fisheries.
- Tracey, S. R., Smolenski, A., & Lyle, J. M. (2007). Genetic structuring of *Latris lineata* at

- localized and transoceanic scales. *Marine Biology*, 152(1), 119–128. <https://doi.org/10.1007/s00227-007-0666-4>
- Trincado, J. L., Entizne, J. C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D. J., & Eyras, E. (2018). SUPPA2: Fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biology*, 19(1), 1–11. <https://doi.org/10.1186/s13059-018-1417-1>
- Tseng, M. C., & Smith, P. J. (2012). Lack of genetic differentiation observed in Pacific bluefin tuna (*Thunnus orientalis*) from Taiwanese and New Zealand waters using mitochondrial and nuclear DNA markers. *Marine and Freshwater Research*, 63(3), 198–209. <https://doi.org/10.1071/MF11126>
- Tyberghein, L., Verbruggen, H., Pauly, K., Troupin, C., Mineur, F., & De Clerck, O. (2012). Bio-ORACLE: a global environmental dataset for marine species distribution modelling. *Global Ecology and Biogeography*, 21(2), 272–281. <https://doi.org/10.1111/j.1466-8238.2011.00656.x>
- Valenzuela-Quiñonez, F. (2016). How fisheries management can benefit from genomics? *Briefings in Functional Genomics*, 15(5), 352–357. <https://doi.org/10.1093/bfgp/elw006>
- van den Enden, T., White, R. W. G., & Elliott, N. G. (2000). Genetic variation in the greenback flounder *Rhombosolea tapirina* Günther (Teleostei, Pleuronectidae) and the implications for aquaculture. *Marine and Freshwater Research*, 51(1), 23. <https://doi.org/10.1071/MF99022>
- van Etten, J. (2017). R package gdistance: Distances and routes on geographical grids. *Journal of Statistical Software*, 76(13), 1–21. <https://doi.org/10.18637/jss.v076.i13>
- van Wijk, S. J., Taylor, M. I., Creer, S., Dreyer, C., Rodrigues, F. M., Ramnarine, I. W., van Oosterhout, C., & Carvalho, G. R. (2013). Experimental harvesting of fish populations drives genetically based shifts in body size and maturation. *Frontiers in Ecology and the Environment*, 11(4), 181–187. <https://doi.org/10.1890/120229>
- Varela, A. I., Ritchie, P. A., & Smith, P. J. (2012). Low levels of global genetic differentiation and population expansion in the deep-sea teleost *Hoplostethus atlanticus* revealed by mitochondrial DNA sequences. *Marine Biology*, 159(5), 1049–1060.

- <https://doi.org/10.1007/s00227-012-1885-x>
- Varela, A. I., Ritchie, P. A., & Smith, P. J. (2013). Global genetic population structure in the commercially exploited deep-sea teleost orange roughy (*Hoplostethus atlanticus*) based on microsatellite DNA analyses. *Fisheries Research*, 140, 83–90. <https://doi.org/10.1016/j.fishres.2012.12.011>
- Vaux, F., Bohn, S., Hyde, J. R., & O'Malley, K. G. (2021). Adaptive markers distinguish North and South Pacific Albacore amid low population differentiation. *Evolutionary Applications*, October 2020, 1–22. <https://doi.org/10.1111/eva.13202>
- Vendrami, D. L. J., Telesca, L., Weigand, H., Weiss, M., Fawcett, K., Lehman, K., Clark, M. S., Leese, F., McMinn, C., Moore, H., & Hoffman, J. I. (2017). RAD sequencing resolves fine-scale population structure in a benthic invertebrate: implications for understanding phenotypic plasticity. *Royal Society Open Science*, 4(2), 160548. <https://doi.org/10.1098/rsos.160548>
- Veríssimo, A., McDowell, J. R., & Graves, J. E. (2010). Global population structure of the spiny dogfish *Squalus acanthias*, a temperate shark with an antitropical distribution. *Molecular Ecology*, 19(8), 1651–1662. <https://doi.org/10.1111/j.1365-294X.2010.04598.x>
- Vezzi, F., Narzisi, G., & Mishra, B. (2012a). Reevaluating assembly evaluations with Feature Response Curves: GAGE and assemblathons. *PLoS ONE*, 7(12), e52210. <https://doi.org/10.1371/journal.pone.0052210>
- Vezzi, F., Narzisi, G., & Mishra, B. (2012b). Feature-by-Feature – Evaluating De Novo Sequence Assembly. *PLoS ONE*, 7(2), e31002. <https://doi.org/10.1371/journal.pone.0031002>
- Vij, S., Kuhl, H., Kuznetsova, I. S., Komissarov, A., Yurchenko, A. A., Van Heusden, P., Singh, S., Thevasagayam, N. M., Prakki, S. R. S., Purushothaman, K., Saju, J. M., Jiang, J., Mbandi, S. K., Jonas, M., Hin Yan Tong, A., Mwangi, S., Lau, D., Ngoh, S. Y., Liew, W. C., ... Orbán, L. (2016). Chromosomal-level assembly of the Asian seabass genome using long sequence reads and multi-layered scaffolding. *PLoS Genetics*, 12(4), 1–35. <https://doi.org/10.1371/journal.pgen.1005954>
- Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting natural selection in genomic data. *Annual Review of Genetics*, 47(1), 97–120. <https://doi.org/10.1146/annurev-genet->

111212-133526

- Vooren, C. M. (1972). Postlarvae and juveniles of the tarakihi (teleostei: Cheilodactylidae) in New Zealand. *New Zealand Journal of Marine and Freshwater Research*, 6(4), 602–618. <https://doi.org/10.1080/00288330.1972.9515448>
- Vooren, C. M. (1975). Nursery grounds of Tarakihi (Teleostei: Cheilodactylidae) around New Zealand. *New Zealand Journal of Marine and Freshwater Research*, 9(2), 121–158. <https://doi.org/10.1080/00288330.1975.9515555>
- Vrijenhoek, R. C. (1997). Gene Flow and Genetic Diversity in Naturally Fragmented Metapopulations of Deep-Sea Hydrothermal Vent Animals. *Journal of Heredity*, 88(4), 285–293. <https://doi.org/10.1093/oxfordjournals.jhered.a023106>
- Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*, 33(14), 2202–2204. <https://doi.org/10.1093/bioinformatics/btx153>
- Wallis, G. P., & Trewick, S. A. (2009). New Zealand phylogeography: evolution on a small continent. *Molecular Ecology*, 18(17), 3548–3580. <https://doi.org/10.1111/j.1365-294X.2009.04294.x>
- Walsh, C., McKenzie, J., Bian, R., Armiger, H., Buckthought, D., Smith, M., Ferguson, H., & Miller, A. (2012). *Snapper catch-at-length and catch-at-age heterogeneity between spatial strata in SNA 2 bottom trawl landings, 2007-08 and 2008-09* [New Zealand Fisheries Assessment Report 2012/40]. Ministry for Primary Industries. http://fs.fish.govt.nz/Doc/23071/12_40_FAR_TN2535.pdf.ashx
- Wang, X., You, X., Langer, J. D., Hou, J., Rupprecht, F., Vlatkovic, I., Quedenau, C., Tushev, G., Epstein, I., Schaefke, B., Sun, W., Fang, L., Li, G., Hu, Y., Schuman, E. M., & Chen, W. (2019). Full-length transcriptome reconstruction reveals a large diversity of RNA and protein isoforms in rat hippocampus. *Nature Communications*, 10(5009), 1–15. <https://doi.org/10.1038/s41467-019-13037-0>
- Wang, Y., Lu, Y., Zhang, Y., Ning, Z., Li, Y., Zhao, Q., Lu, H., Huang, R., Xia, X., Feng, Q., Liang, X., Liu, K., Zhang, L., Lu, T., Huang, T., Fan, D., Weng, Q., Zhu, C., Lu, Y., ... Zhu, Z. (2015).

- The draft genome of the grass carp (*Ctenopharyngodon idellus*) provides insights into its evolution and vegetarian adaptation. *Nature Genetics*, 47(6), 625–631. <https://doi.org/10.1038/ng.3280>
- Waples, R. S. (2002). Effective size of fluctuating salmon populations. *Genetics*, 161(2), 783–791.
- Waples, R. S., & Gaggiotti, O. (2006). What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology*, 15(6), 1419–1439. <https://doi.org/10.1111/j.1365-294X.2006.02890.x>
- Waples, R. S., Grewe, P. M., Bravington, M. W., Hillary, R., & Feutry, P. (2018). Robust estimates of a high N_e/N ratio in a top marine predator, southern bluefin tuna. *Science Advances*, 4, 1–9. <https://doi.org/10.1126/sciadv.aar7759>
- Waples, R. S., Punt, A. E., & Cope, J. M. (2008). Integrating genetic data into management of marine resources: How can we do it better? *Fish and Fisheries*, 9(4), 423–449. <https://doi.org/10.1111/j.1467-2979.2008.00303.x>
- Ward, R. D. (2000). Genetics in fisheries management. *Hydrobiologia*, 420(1), 191–201. <https://doi.org/10.1023/A:1003928327503>
- Ward, R. D., Costa, F., Holmes, B., & Steinke, D. (2008). DNA barcoding of shared fish species from the North Atlantic and Australasia: minimal divergence for most taxa, but Zeus faber and Lepidopus caudatus each probably constitute two species. *Aquatic Biology*, 3(1), 71–78. <https://doi.org/10.3354/ab00068>
- Ward, R. D., & Elliott, N. G. (2001). Genetic population structure of species in the South East Fishery of Australia. *Marine and Freshwater Research*, 52(4), 563–573. <https://doi.org/10.1071/MF99184>
- Ward, R. D., Elliott, N. G., Grewe, P. M., Last, P. R., Lowry, P. S., Innes, B. H., & Yearsley, G. K. (1998). Allozyme and mitochondrial DNA variation in three species of oreos (Teleostei: Oreosomatidae) from Australasian waters. *New Zealand Journal of Marine and Freshwater Research*, 32(2), 233–245. <https://doi.org/10.1080/00288330.1998.9516822>
- Ward, R. D., & Gardner, M. G. (1997). *Stock structure and species identification of school and*

- gummy sharks in Australasian waters* (Issue February) [Report of FRRF 93/11 & FRDC 93/64]. CSIRO Marine Research.
- Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, 38(6), 1358–1370.
- Wellenreuther, M., Mérot, C., Berdan, E., & Bernatchez, L. (2019). Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. *Molecular Ecology*, 28(6), 1203–1209. <https://doi.org/10.1111/mec.15066>
- Wen, M., Ng, J. H. J., Zhu, F., Chionh, Y. T., Chia, W. N., Mendenhall, I. H., Lee, B. P. Y. H., Irving, A. T., & Wang, L. F. (2018). Exploring the genome and transcriptome of the cave nectar bat *Eonycteris spelaea* with PacBio long-read sequencing. *GigaScience*, 7(10), 1–8. <https://doi.org/10.1093/gigascience/giy116>
- Westgaard, J.-I., Saha, A., Kent, M., Hansen, H. H., Knutsen, H., Hauser, L., Cadrin, S. X., Albert, O. T., & Johansen, T. (2017). Genetic population structure in Greenland halibut (*Reinhardtius hippoglossoides*) and its relevance to fishery management. *Canadian Journal of Fisheries and Aquatic Sciences*, 74(4), 475–485. <https://doi.org/10.1139/cjfas-2015-0430>
- Whibley, A., Kelley, J. L., & Narum, S. R. (2021). The changing face of genome assemblies: Guidance on achieving high-quality reference genomes. *Molecular Ecology Resources*, 21(3), 641–652. <https://doi.org/10.1111/1755-0998.13312>
- Whibley, A., Kelley, J., & Narum, S. (2020). The changing face of genome assemblies: guidance on achieving high-quality reference genomes. *Molecular Ecology Resources*, September 2020, 1–12. <https://doi.org/10.1111/1755-0998.13312>
- Whitlock, M. C., & Lotterhos, K. E. (2015). Reliable Detection of Loci Responsible for Local Adaptation: Inference of a Null Model through Trimming the Distribution of F ST. *The American Naturalist*, 186(S1), S24–S36. <https://doi.org/10.1086/682949>
- Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., & Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, 8(12), 973–982. <https://doi.org/10.1038/nrg2165>

- Wickham, H. (2009). *ggplot2: elegant graphics for data analysis* (1st ed.). Springer.
<https://doi.org/10.1007/978-0-387-98141-3>
- Wiley, G., & Miller, M. J. (2020). A highly contiguous genome for the golden-fronted woodpecker (*Melanerpes aurifrons*) via hybrid Oxford Nanopore and short read assembly. *G3: Genes, Genomes, Genetics*, 10(6), 1829–1836.
<https://doi.org/10.1534/g3.120.401059>
- Williams, G. C. (1966). *Adaptation and Natural Selection*. Princeton University Press.
- Willing, E., Dreyer, C., & van Oosterhout, C. (2012). Estimates of genetic differentiation measured by F_{ST} do not necessarily require large sample sizes when using many SNP markers. *PLoS ONE*, 7(8), e42649. <https://doi.org/10.1371/journal.pone.0042649>
- Wold, J., Koepfli, K., Galla, S. J., Eccles, D., Hogg, C. J., Le Lec, M. F., Guhlin, J., Santure, A. W., & Steeves, T. E. (2021). Expanding the conservation genomics toolbox: incorporating structural variants to enhance genomic studies for species of conservation concern. *Molecular Ecology*, mec.16141. <https://doi.org/10.1111/mec.16141>
- Wood, D. E. (2019). *MiniKraken2 v2 8GB database*. Johns Hopkins University.
ftp://ftp.ccb.jhu.edu/pub/data/kraken2_dbs/old/minikraken2_v2_8GB_201904.tgz
- Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20(257), 1–13. <https://doi.org/10.1186/s13059-019-1891-0>
- Worley, K. C., Richards, S., & Rogers, J. (2017). The value of new genome references. *Experimental Cell Research*, 358(2), 433–438.
<https://doi.org/10.1016/j.yexcr.2016.12.014>
- Worm, B., Hilborn, R., Baum, J. K., Branch, T. A., Collie, J. S., Costello, C., Fogarty, M. J., Fulton, E. A., Hutchings, J. A., Jennings, S., Jensen, O. P., Lotze, H. K., Mace, P. M., Mcclanahan, T. R., Palumbi, S. R., Parma, A. M., Ricard, D., Rosenberg, A. A., Watson, R., & Zeller, D. (2009). *Rebuilding Global Fisheries*. 325(July), 578–586.
- Wright, S. (1943). Isolation by distance. *Genetics*, 28(2), 114–138.
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1209196/>
- Wu, C., Zhang, D., Kan, M., Lv, Z., Zhu, A., Su, Y., Zhou, D., Zhang, J., Zhang, Z., Xu, M., Jiang, L., Guo, B., Wang, T., Chi, C., Mao, Y., Zhou, J., Yu, X., Wang, H., Weng, X., ... Liu, Y. (2014).

- The draft genome of the large yellow croaker reveals well-developed innate immunity. *Nature Communications*, 5(5227), 1–7. <https://doi.org/10.1038/ncomms6227>
- Xu, J., Bian, C., Chen, K., Liu, G., Jiang, Y., Luo, Q., You, X., Peng, W., Li, J., Huang, Y., Yi, Y., Dong, C., Deng, H., Zhang, S., Zhang, H., Shi, Q., & Xu, P. (2017). Draft genome of the Northern snakehead, *Channa argus*. *GigaScience*, 6(4), 1–5. <https://doi.org/10.1093/gigascience/gix011>
- Xuereb, A., Benestan, L., Normandeau, É., Daigle, R. M., Curtis, J. M. R., Bernatchez, L., & Fortin, M. J. (2018). Asymmetric oceanographic processes mediate connectivity and population genetic structure, as revealed by RADseq, in a highly dispersive marine invertebrate (*Parastichopus californicus*). *Molecular Ecology*, 27(10), 2347–2364. <https://doi.org/10.1111/mec.14589>
- Yang, X., Liu, H., Ma, Z., Zou, Y., Zou, M., Mao, Y., Li, X., Wang, H., Chen, T., Wang, W., & Yang, R. (2019). Chromosome-level genome assembly of *Triplophysa tibetana*, a fish adapted to the harsh high-altitude environment of the Tibetan Plateau. *Molecular Ecology Resources*, 19(4), 1027–1036. <https://doi.org/10.1111/1755-0998.13021>
- Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24(8), 1586–1591. <https://doi.org/10.1093/molbev/msm088>
- Yang, Z., & Bielawski, J. P. (2000). Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*, 15(12), 496–503. [https://doi.org/10.1016/S0169-5347\(00\)01994-7](https://doi.org/10.1016/S0169-5347(00)01994-7)
- Yang, Z., & Rannala, B. (2006). Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Molecular Biology and Evolution*, 23(1), 212–226. <https://doi.org/10.1093/molbev/msj024>
- Yin, R., Kwoh, C. K., & Zheng, J. (2018). Whole genome sequencing analysis. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, 1–3, 176–183. <https://doi.org/10.1016/B978-0-12-809633-8.20095-2>
- Ying, Y., Chen, Y., Lin, L., & Gao, T. (2011). Risks of ignoring fish population spatial structure in fisheries management. *Canadian Journal of Fisheries and Aquatic Sciences*, 68(12), 2101–2120. <https://doi.org/10.1139/f2011-116>

- Yuan, Y. W., & Wessler, S. R. (2011). The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proceedings of the National Academy of Sciences of the United States of America*, 108(19), 7884–7889. <https://doi.org/10.1073/pnas.1104208108>
- Yuan, Z., Liu, S., Zhou, T., Tian, C., Bao, L., Dunham, R., & Liu, Z. (2018). Comparative genome analysis of 52 fish species suggests differential associations of repetitive elements with their living aquatic environments. *BMC Genomics*, 19(141), 1–10. <https://doi.org/10.1186/s12864-018-4516-1>
- Zemeckis, D. R., Martins, D., Kerr, L. A., & Cadrin, S. X. (2014). Stock identification of Atlantic cod (*Gadus morhua*) in US waters: an interdisciplinary approach. *ICES Journal of Marine Science*, 71(6), 1490–1506. <https://doi.org/10.1093/icesjms/fsu032>
- Zhang, H. H., Xu, M. R. X., Wang, P. L., Zhu, Z. G., Nie, C. F., Xiong, X. M., Wang, L., Xie, Z. Z., Wen, X., Zeng, Q. X., Zhang, X. G., & Dai, F. Y. (2020). High-quality genome assembly and transcriptome of *Ancherythroculter nigrocauda*, an endemic Chinese cyprinid species. *Molecular Ecology Resources*, 20(4), 882–891. <https://doi.org/10.1111/1755-0998.13158>
- Zhang, J., Nielsen, R., & Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular Biology and Evolution*, 22(12), 2472–2479. <https://doi.org/10.1093/molbev/msi237>
- Zhang, X., Li, G., Jiang, H., Li, L., Ma, J., Li, H., & Chen, J. (2019). Full-length transcriptome analysis of *Litopenaeus vannamei* reveals transcript variants involved in the innate immune system. *Fish & Shellfish Immunology*, 87, 346–359. <https://doi.org/10.1016/j.fsi.2019.01.023>
- Zheng, S., Shao, F., Tao, W., Liu, Z., Long, J., Wang, X., Zhang, S., Zhao, Q., Carleton, K. L., Kocher, T. D., Jin, L., Wang, Z., Peng, Z., Wang, D., & Zhang, Y. (2021). Chromosome-level assembly of Southern catfish (*Silurus meridionalis*) provides insights into visual adaptation to the nocturnal and benthic lifestyles. *Molecular Ecology Resources*. <https://doi.org/10.1111/1755-0998.13338>
- Zhou, S., Kolding, J., Garcia, S. M., Plank, M. J., Bundy, A., Charles, A., Hansen, C., Heino, M., Howell, D., Jacobsen, N. S., Reid, D. G., Rice, J. C., & van Zwieten, P. A. M. (2019). Balanced

- harvest: concept, policies, evidence, and management implications. *Reviews in Fish Biology and Fisheries*, 29(3), 711–733. <https://doi.org/10.1007/s11160-019-09568-w>
- Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., & Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics*, 29(21), 2669–2677. <https://doi.org/10.1093/bioinformatics/btt476>
- Zimin, A. V., Puiu, D., Hall, R., Kingan, S., Clavijo, B. J., & Salzberg, S. L. (2017). The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. *GigaScience*, 6(11), 1–7. <https://doi.org/10.1093/gigascience/gix097>
- Zimin, A. V., Puiu, D., Luo, M.-C., Zhu, T., Koren, S., Marçais, G., Yorke, J. A., Dvořák, J., & Salzberg, S. L. (2017). Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Research*, 27(5), 787–792. <https://doi.org/10.1101/gr.213405.116>
- Zimin, A. V., & Salzberg, S. L. (2020). The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLoS Computational Biology*, 16(6), 1–8. <https://doi.org/10.1371/journal.pcbi.1007981>
- Zimin, A. V., Stevens, K. A., Crepeau, M. W., Puiu, D., Wegrzyn, J. L., Yorke, J. A., Langley, C. H., Neale, D. B., & Salzberg, S. L. (2017). An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. *GigaScience*, 6(1), 1–4. <https://doi.org/10.1093/gigascience/giw016>