

Assessment of genomic diversity and population sub-structuring of Kingklip (*Genypterus capensis*) off southern Africa

By

Melissa Jane Schulze



Thesis presented in fulfilment of the requirements of the degree of Master of Science
in the Department of Botany and Zoology at Stellenbosch University

Supervisor: Prof. Sophie von der Heyden

Co-supervisor: Dr Romina Henriques

April 2019

DECLARATION

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualifications.

Date: December 2019

ABSTRACT

Kingklip (*Genypterus capensis*) represents a valuable marine resource for both South Africa and Namibia. Historical exploitation levels led to substantial declines in abundance, resulting in the species being considered over-exploited in the past. Currently, there is a lack of consensus regarding kingklip stock structure, with previous studies providing evidence for both multiple and single stocks. Understanding stock structure is vital for the appropriate assessment and management of marine resources. Taking into account both the commercial importance and trans-boundary nature of this species, it is therefore evident that a consensus regarding the fine-scale genetic structure is needed in order to best inform future management decisions. Next Generation Sequencing (NGS) has revolutionised population genetics allowing for the sequencing and identification of thousands of loci at reduced costs, thereby helping to identify weak genetic differentiation and adaptive divergence even in species with high gene flow levels. By employing a pooled ezRAD sequencing technique, the first chapter of this thesis isolated and identified a novel set of genome-wide molecular markers (Single Nucleotide Polymorphisms – SNPs). Over 40 000 SNP loci were identified in chapter 1, both neutral as well as putative outlier loci, potentially under selection. The second chapter of this thesis subsequently employed the SNP database developed in chapter 1 to investigate i) the relation of previous genetic versus genomic divergence levels and patterns of sub-structuring along the South African coastline, as well as ii) genome-wide patterns of fine-scale sub-structuring along Kingklip's southern African distribution, thereby providing novel insight into the genetic relation of Namibian and South African Kingklip. Overall, the results of chapter 2 provided evidence for a three-stock hypothesis with significant levels of adaptive divergence identified between “Northern Benguela” (North of Lüderitz), “Southern Benguela” (South of Lüderitz to Cape Agulhas) and “Eastern Cape (Cape Agulhas to Algoa Bay) populations. However, adaptive divergence appears to be occurring in the face of high levels of gene flow, thereby creating a dynamic system across the southern African distribution. Based on the findings of chapter 2, the third chapter addresses management recommendations and the potential for the use of the newly developed marker panel for future Kingklip fisheries management.

OPSOMMING

Kingklip (*Genypterus capensis*) verteenwoordig 'n waardevolle mariene hulpbron vir beide Suid-Afrika en Namibië. Historiese uitbuitingsvlakke het gelei tot aansienlike afname, wat veroorsaak het dat die spesies in die verlede uitgenuit is. Tans is daar 'n gebrek aan konsensus oor Kingklip visbevolkings, met vorige studies wat bewyse lewer vir beide veelvuldige en enkele bevolkings. Om visbevolking struktuur te verstaan is noodsaaklik vir die toepaslike assessering en bestuur van mariene hulpbronne. Met inagneming van beide die kommersiële belang en die grensvlak van hierdie spesie, is dit dus duidelik dat 'n konsensus aangaande die fynskaalse genetiese struktuur nodig is om toekomstige bestuursbesluite die beste in te lig. Next Generation Sequencing (NGS) het populasiegenetika gewoloseer, wat die sequencing en identifikasie van duisende loci teen verlaagde koste moontlik gemaak het, en sodoende help om swak genetiese differensiasie en adaptiewe divergensie te identifiseer selfs in spesies met hoë genevloei vlakke. Deur die gebruik van 'n saamgevoegde ezRAD-sequencingtegniek, het die eerste hoofstuk van hierdie proefskrif 'n nuwe stel genom-wye molekulêre merkers (enkel-nukleotied-polimorfismes - SNP's) geïsoleer en geïdentifiseer. Meer as 40 000 SNP loci is geïdentifiseer in hoofstuk 1, beide neutrale sowel as potensiële outlier loci, moontlik onder seleksie. Die tweede hoofstuk van hierdie proefskrif het daarna die SNP-databasis wat in hoofstuk 1 ontwikkel is, aangewend om i) die verband tussen vorige genetiese versus genomiese divergensievlakke en patrone van substrukturering langs die Suid-Afrikaanse kus te ondersoek, asook ii) genomwye patrone van fynskaalse substruktuur van Kingklip se suidelike Afrika-verspreiding, en bied dus nuwe insig in die genetiese verhouding van Namibiese en Suid-Afrikaanse Kingklip. Die algehele resultate van hoofstuk 2 het getuienis gelewer vir 'n drie-visbevolking hipotese met betekenisvolle vlakke van adaptiewe divergensie wat geïdentifiseer is tussen "Northern Benguela" (Noord van Lüderitz), "Southern Benguela" (suid van Lüderitz tot Kaap Agulhas) en "Oos-Kaap" (Kaap Agulhas tot Algoabaai) bevolkings. Adaptiewe afwykings blyk egter voor te kom in die lig van hoë vlakke van geenvloei, en skep daardeur 'n dinamiese stelsel oor die suidelike Afrika-verspreiding. Op grond van die bevindings van hoofstuk 2 word die aanbevelings in die derde hoofstuk aangespreek

en die potensiaal vir die gebruik van die nuut ontwikkelde merkpaneel vir toekomstige Kingklip-visserybestuur.

ACKNOWLEDGEMENTS

I would like to thank my supervisors Prof Sophie von der Heyden and Dr Romina Henriques for allowing me this opportunity as well as for their guidance and assistance throughout my masters. I would also like to give thanks to the National Research Foundation (NRF) for their funding of this project, as well as the University and Stellenbosch and Department of Botany and Zoology for their personal sponsorship. Thank you to the Department of Agriculture Forestry and Fisheries (DAFF), CapFish and the University of Namibia for the collection of samples, without your help this project would not have been possible. Further I would like to thank the Hawaii Institute of Marine Biology for their assistance in library preparation and sequencing, as well as the Central Analytical Facility (CAF) of Stellenbosch University for their help with DNA quality and quantity control. I would also like to take this opportunity to thank my lab associates in the Evolutionary Genomics Group, especially Fawzia Gordon and Henry for their technical assistance and making sure everything runs smoothly. A special thanks to the members of the von der Heyden Lab, particularly Lisa Mertens, Nikki Phair and Erica Nielsen for their assistance with NGS, as well as Molly Czachur for the daily chats by the kettle. I would additionally like to thank my friends and “Stellenbosch family” for making the journey so enjoyable, with a special thanks to the members of 11 Piet Retief Straat for the endless stoep kuiers and support throughout the last two years. Finally, I would like to thank my mom and dad for their unwavering support and for allowing me the opportunity to continue my studies.

TABLE OF CONTENTS

DECLARATION.....	1
ABSTRACT.....	2
OPSOMMING.....	3
ACKNOWLEDGEMENTS.....	5
TABLE OF CONTENTS.....	6
LIST OF FIGURES.....	9
LIST OF TABLES.....	11
LIST OF SUPPLEMENTARY MATERIALS.....	13
GENERAL INTRODUCTION.....	14
Southern African Kingklip, <i>Genypterus capensis</i>	14
Southern African Kingklip fisheries.....	15
Current state of knowledge and management of southern African Kingklip resources.....	16
Fisheries Management.....	19
Population structure and fisheries management.....	19
Molecular technologies and their use in fisheries management.....	21
Aims and objectives.....	24
CHAPTER 1: SNP development and identification of local adaptation in southern African Kingklip, <i>Genypterus capensis</i>	
INTRODUCTION.....	25
Restriction Site Associated Sequencing.....	28
Single Nucleotide Polymorphisms – SNPs.....	30
Chapter aims.....	31
METHODS.....	31
Sample collection.....	31
DNA extraction and pooling of samples.....	32
ezRAD Sequencing.....	33

SNP development pipeline.....	33
Quality control.....	33
Assembly and mapping of the mitochondrial DNA dataset.....	34
Assembly and mapping of the nuclear DNA dataset.....	36
Regional diversity measures.....	37
Outlier detection.....	39
Additional/Exploratory outlier detection.....	40
RESULTS	41
Pooling, ezRAD sequencing and quality control.....	41
Mitochondrial dataset: assembly and mapping.....	42
Nuclear dataset: assembly and mapping.....	42
Regional diversity measures.....	43
mtDNA dataset	43
nDNA dataset.....	44
Outlier detection.....	44
mtDNA dataset	44
nDNA dataset	45
Additional/Exploratory outlier detection.....	48
DISCUSSION.....	49
Genome-wide levels of diversity.....	51
Detection of putative outlier loci.....	54
Methodological considerations.....	58
CHAPTER 2: Genetic sub-structuring of southern African Kingklip within and between South Africa and Namibia	
INTRODUCTION.....	59
Kingklip distribution and the genetic population sub-structuring	59
Chapter aims and objectives.....	62
METHODS.....	63
Genome-wide population differentiation: fixation index.....	63
Genetic versus genomic patterns of differentiation.....	64
Pop 1 versus Pop 2 differentiation	64

Mapping of microsatellite primer sequences to reference genome.....	65
South African population sub-structuring.....	66
Population sub-structuring and genome-wide differentiation.....	66
Population sub-structuring and genome-wide differentiation – top 500 loci.....	67
Southern African population sub-structuring of Kingklip.....	68
RESULTS	69
Genetic versus genomic patterns of differentiation.....	69
Genome-wide differentiation: Pop 1 versus Pop 2.....	69
Microsatellite primer sequence mapping.....	73
South African population sub-structuring.....	75
Population sub-structuring and genome-wide differentiation.....	75
Population sub-structuring and genome-wide differentiation – top 500 loci.....	78
South African versus Namibian population sub-structuring.....	80
DISCUSSION.....	82
Pop 1 versus Pop 2: genetic versus genomic differentiation.....	84
Pop 1 and Pop 2 versus 2017 South African regions – relation of past clusters to contemporary sampling sites.....	86
Contemporary South African genomic sub-structuring.....	88
South African versus Namibian genomic sub-structuring.....	91
Conclusion.....	97
CHAPTER 3: Molecular tools in action: Conservation recommendations and implications, as well as development towards a genomic tool for post-harvest control of Kingklip	
Conservation recommendations and implications.....	99
Post-harvest control	104
BIBLIOGRAPHY.....	108
SUPPLEMENTARY MATERIAL.....	139

LIST OF FIGURES

Figure 1: Map of the southern African coastline and Benguela system showing depth contour (200m) as well as oceanographic features including the Agulhas Current, Benguela Current, Angola Current and approximate location of the Angola-Benguela frontal zone, the Lüderitz upwelling cell and the Agulhas Bank.

Figure 2: Sampling locations for Kingklip (2014 & 2017): CB - Child's Bank, TB – Table Bay, SC - South Coast, EC - Eastern Cape and NAM - Namibia. Kingklip distribution indicated in orange.

Figure 3: Bioinformatic pipeline followed for the identification and development of Chapter 1 Single Nucleotide Polymorphism (SNP) databases.

Figure 4: Venn diagram illustrating overlap of outlier loci detected by three methodologies for the nuclear dataset: pcadapt, PoPoolation2 and Bayescan 2.1.

Figure 5: Frequency of candidate nuclear outlier loci (outlier loci identified by two or more outlier detection approaches) across sampling sites. Outliers identified by Node number and SNP position. Pool names as per Table 2.

Figure 6: Principal component analysis (PCA) of variation in allele frequencies per pool, based on neutral and outlier loci within the **A.** simulated and **B.** complete nuclear datasets. Pool names as per Table 2.

Figure 7: Manhattan plot of pairwise genomic estimates (F_{ST}) per SNP loci for P1 versus P2 (P1 – Pop 1, P2 – Pop 2), against SNP loci position within nuclear reference sequence (nDNA_ref). Plotted for neutral and outlier loci contained within the simulated, full dataset.

Figure 8: Manhattan plot of pairwise genomic estimates (F_{ST}) per SNP loci for P1 versus P2 (P1 – Pop 1, P2 – Pop 2), against SNP loci position within nuclear reference sequence (nDNA_ref). Plotted for neutral and outlier loci contained within the complete dataset.

Figure 9: BAPS Bayesian clustering analysis of P1, P2 and 2017 South African sampling sites/pools for simulated **A.** full (neutral and outlier loci) and **B.** outlier (outlier loci only) datasets. Pool names as per Table 2.

Figure 10: Principal component analysis (PCA) of variation in allele frequencies per pool, based on outlier and neutral loci contained within the **A.** simulated and **B.** complete nuclear datasets. Pool names as per Table 2.

Figure 11: Principal component analysis (PCA) of variation in allele frequencies per pool, based the top 500 loci (top 500 dataset), identified based on the loci loadings of P1 versus P2 PCA. Pool names as per Table 2

Figure 12: BAPS Bayesian clustering analysis of Namibian (NAM1 –and NAM 2) and 2017 South African sampling sites/pools for simulated **A.** full (neutral and outlier loci) and **B.** outlier (outlier loci only) datasets. Pool names as per Table 2.

Figure 13: Principal component analysis (PCA) of variation in allele frequencies per pool, based on outlier and neutral loci contained within the **A.** simulated and **B.** complete nuclear datasets. Pool names as per Table 2.

Figure 14: Development of molecular tools for stock identification and individual assignment of Kingklip (*Genypterus capensis*).

LIST OF TABLES

Table 1: Marine species with available genome-wide datasets, as well as associated Next-Generation Sequencing (NGS) approaches and references. Note, this is not an exhaustive list, but a short representation of different species and NGS approaches employed.

Table 2: Sampling location, code, year and number of individuals sampled per pool.

Table 3: Sequencing results per pool for Kingklip. Number of raw reads sequenced, number of quality-controlled reads (post initial quality control) and percentage of raw reads remaining after initial quality control (% high quality reads) per pool. Pool names as per Table 2.

Table 4: Mapping statistics for the mitochondrial dataset of Kingklip. Number of properly-paired, unique reads mapped to *Gadus morhua* mitochondrial genome (*G. morhua* reference) per pool. Number of reads mapped to Kingklip mitochondrial reference sequence (KK_mtDNA_ref) per pool. Pool names as per Table 2.

Table 5: Filtering and mapping statistics for nuclear dataset of Kingklip. Number of reads prior to filtering of mtDNA reads (quality-controlled reads) and remaining following the removal of potential mitochondrial reads (nDNA reads), per pool. Number and percentage of filtered reads mapped to nDNA reference per pool. Pool names as per Table 2.

Table 6: Regional diversity measures based on mitochondrial dataset per pool for Kingklip: nucleotide diversity (Tajima's π), population mutation rate (Watterson's θ_w) and Tajima's D, total biallelic SNPs, private biallelic SNPs and percentage private, biallelic SNPs. Pool names as per Table 2.

Table 7: Regional diversity measures based on nuclear dataset per pool for Kingklip: nucleotide diversity (Tajima's π), population mutation rate (Watterson's θ_w) and Tajima's D. Total number biallelic SNPs, private biallelic SNPs and percentage private, biallelic SNPs. Pool names as per Table 2.

Table 8: Total number of outlier SNPs identified per pool for mitochondrial (mtDNA) and nuclear (nDNA) datasets. *Outlier SNPs refer to outlier loci identified by

PoPoolation 2 **Outlier SNPs refer to outlier loci identified by two or more outlier detection approaches. Pool names as per Table 2.

Table 9: Top BLASTX search results, with corresponding E-value scores and percentage identity, for nodes containing candidate outlier loci identified by two or more outlier detection approaches, for the nuclear dataset.

Table 10: Top BLASTX search results, with corresponding E-value scores and percentage identity, for nodes containing outlier loci identified by pcadapt and PoPoolation2, based on 41 369 loci.

Table 11: Estimates of pairwise genomic differentiation (F_{ST} , below diagonal) and 95% confidence intervals (above diagonal) for all sampling sites/pools, based on simulated **A.** full (outlier & neutral loci), **B.** neutral (neutral loci only) and **C.** outlier (outlier loci only) datasets, and **D.** complete dataset. Statistically significant results in bold. Pool names as per Table 2.

Table 12: Results of mapping of 10 microsatellite primers (Ward & Reilly, 2010), forward (F) and reverse (R), to *de novo* reference sequence (nDNA_ref). Refer to Supplementary Table S2 and Ward and Reilly (2001) for primer notes.

Table 13: Estimates of pairwise genomic differentiation (F_{ST}) and 95% confidence intervals (95% CI) for South African sampling sites/pools, based top 500 loci dataset. Statistically significant results indicated in bold. Pool names as per Table 2.

Table 14: Top BLASTX search results, with corresponding E-value scores and percentage identity, for candidate outlier loci identified for the top 500 dataset.

LIST OF SUPPLEMENTARY MATERIALS

Supplementary Material 1: Scripts used for bioinformatic analyses and pipeline.

Supplementary Table S1: Major allele frequencies of shared outlier SNPs, identified for nuclear dataset, per pool. SNPs not found within pools indicated by -. Pool names as per Table 2.

Supplementary Table S2: Repeat motifs, primer sequences (Forward – F & Reverse – R) and GenBank accession number of 10 microsatellite primers employed. Refer to Ward and Reilly (2001) for primer notes.

GENERAL INTRODUCTION

Southern African Kingklip, *Genypterus capensis*

Kingklip, *Genypterus capensis* (Smith, 1847), is a deep-water, slow-growing, long-lived benthic fish endemic to the southern African coastline (Japp, 1990; Punt & Japp, 1994). Being one of six species belonging to the genus *Genypterus*, which are found only within the temperate waters of the southern hemisphere, *G. capensis* is most closely related to the Pacific Pink ling (*Genypterus blacodes* – Punt & Japp, 1994; Bisby et al., 2012; Santaclara et al., 2014). With a geographical distribution extending from the north of Walvis Bay, on the Namibian coast, to Algoa Bay, on the South-east coast of South Africa (Figure 1; Smith, 1847; Olivar & Sabatés, 1989), Kingklip occurs within a unique region comprising three oceanographic currents: the warm Angola Current to the north, the cold nutrient rich Benguela Current off the west coast, and the warm Agulhas Current on the south coast. This creates a dynamic system composed of three distinct, but overlapping oceanographic regimes (Figure 1; Hutchings et al., 2009).

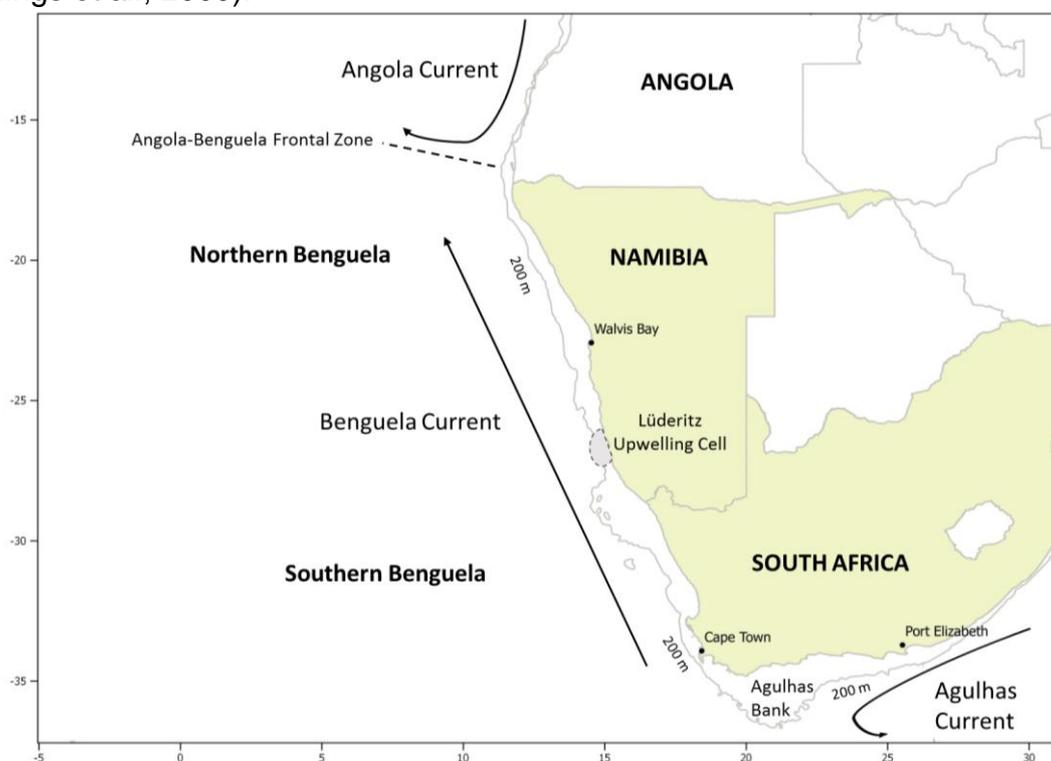


Figure 1: Map of southern African coastline and Benguela system showing depth contour (200m) as well as oceanographic features including the Agulhas Current, Benguela Current, Angola Current and approximate location of the Angola-Benguela frontal zone, the Lüderitz upwelling cell and the Agulhas Bank.

Genypterus capensis occurs at depths ranging from shallow (+/- 42 meters) to 485 meters, exhibiting a general trend of increased age with depth (Badenhorst & Smale, 1991). Adults are relatively sedentary, despite seasonal spawning migrations, with recruitment found to take place in shallower waters (<200 meters; Badenhorst & Smale, 1991; Payne & Badenhorst, 1995). Kingklip females are larger than males, with adults aged to 25 years and recorded to have a maximum length of 150 cm (Payne, 1977; Japp, 1990). While the morphology of Kingklip larvae and eggs has been described (Brownell, 1979; Olivar & Sabatés, 1989), little is known with regards to their early life stages. Spawning is found to occur between March and November/December, with maximum intensity between June and September (Olivar & Sabatés, 1989). Currently only one spawning aggregation has been recorded and reported for Kingklip, east of the Agulhas Bank in the vicinity of Port Elizabeth (Japp, 1989). However, the possibility of a separate spawning aggregation off the West coast cannot be excluded, as mature and fertile females (Durbholtz, pers. comms), as well as two spawning aggregations (vicinity of Dassen Island and Cape Point – Japp, pers. comms), have been reported off the West coast.

Southern African Kingklip fisheries

Kingklip represents a commercially important resource for both Namibia and South Africa (Olivar & Sabatés, 1989; Pecquerie et al., 2004; de Moor et al., 2015; Henriques et al., 2017). The demersal fishing industry contributes substantially to the Gross Domestic Product (GDP) of these countries, and in South Africa supports roughly 27 000 jobs (DAFF, 2014), contributing more than R 5 billion per year in sales (SADSTIA, 2017). In Namibia, the fishing industry represents the third largest economic sector, with a projected export value of N\$ 2 900 million in 2000 (Boyer & Hampton, 2001). The ocean economy has thus been identified as a priority field in the National Development Plan of South Africa, with Operation Phakisa specifically aiming to unlock the potential of the blue economy whilst ensuring the sustainable exploitation of marine resources (Operation Phakisa, 2014). Being a valuable by-catch of hake-directed trawls and long-line fishing, and one of the most commercially important marine resources within South African and Namibian waters, the effective management of Kingklip is considered a national priority (Shannon et al., 1992; Punt & Japp, 1994; de Moor et al., 2015).

Kingklip has supported intense fishing pressures throughout decades, with the earliest records of trawl catches dating back to the 1930s (Punt & Japp, 1994). The introduction of a Kingklip directed longline fishery in 1983 resulted in a rapid increase in catches, with the combined longline and trawl catches of 1986 being more than double that of 1973, totalling 11 370 tons (Badenhorst, 1988; Shannon et al., 1992; Punt & Japp, 1994). This rapid increase in intense exploitation resulted in considerable declines in abundance levels. Within four years, Kingklip catches had decreased from 11 370 tons in 1986 to 2 533 tons in 1990, with the spawning biomass estimated to be less than 50% of its pre-exploited level (Punt & Japp, 1994; Brandão & Butterworth, 2013). Subsequently, the species was considered overexploited (Punt & Japp, 1994), and the direct fishery terminated. Following the closure of Kingklip directed long-line fishing in 1990, and the implementation of regulatory policies, Kingklip by-catch has gradually increased (Brandão & Butterworth, 2013; de Moor et al., 2015), and the species is now considered to be optimally exploited, with evidence for a recent recovery in abundance on both the South and West coast of South Africa (Brandão & Butterworth, 2013; de Moor et al., 2015). Despite this, the effects of past exploitation on the effective population size (*sensus* Wright, 1931: the number of breeding individuals within an idealized population that would experience similar levels of genetic loss, via genetic drift and/or inbreeding, as the population, census, being studied – N_e) and contemporary diversity levels are evident, with a recent study by Henriques et al. (2017) reporting relatively lower estimates of contemporary genetic diversity as compared to historical levels, as well as low estimates of contemporary effective population sizes (CN_e). This, in conjunction with the continued by-catch of Kingklip in the hake-directed fisheries, remains a concern for Kingklip resources (Brandão & Butterworth, 2008; Henriques et al., 2017).

Current state of knowledge and management of southern African Kingklip resources

Increasing evidence for population structure and genetic differentiation among marine species suggests that few species are truly panmictic. Instead, increasing evidence suggests that marine species are composed of several discrete populations (Hauser & Carvalho, 2008; Reiss et al., 2009; Gaither et al., 2016). In fisheries management, identification of biological stocks is central to support sustainable fishing practices,

with stocks generally referring to groups of individuals, of the same species, with similar characteristics (demographic and/or genetic – Ovenden et al., 2015). Stocks should react more or less independently to harvesting pressures and the effects of exploitation (Hauser & Carvalho, 2008; Benestan et al., 2015; Ovenden et al., 2015; Spies et al., 2015). Knowledge regarding stock structure is thus key for fisheries management (Carvalho & Hauser, 1994; Ovenden et al., 2015; Spies et al., 2015), with stock delimitation providing an accurate basis for the assessment of marine resources, as well as for the spatial-temporal delineation of harvest quotas and management units (Grant & Leslie, 2005; Ovenden et al., 2015; Spies et al., 2015; Henriques et al., 2017). In addition, managing a resource as a single unit if it is in fact a mixture of several stocks (or vice-versa) has inherent risks such as under- or over-exploitation, with potential loss of genetic diversity associated with the latter (Carvalho & Hauser, 1994; Henriques et al., 2017; Pinsky et al., 2018). These risks are of greater concern for long-lived, slow growing, sedentary species such as Kingklip, since these life-history traits make them generally more vulnerable to the effects of over-exploitation (Grant & Leslie, 2005; FAO 2009; Henriques et al., 2017). However, despite their economic value and transboundary nature, a consensus regarding the population sub-structuring of southern African Kingklip is currently lacking, with previous studies revealing contrasting results.

The earliest demographic studies based on differences in otolith morphology, vertebral count and growth rate, identified three distinct stocks along the southern African coastline: the “Walvis” stock extending from Walvis Bay northwards, the “Cape” stock extending from Lüderitz to Cape Point and the “South-East” stock found off the South-east coast of South Africa (Payne 1977, 1985). However, morphology, growth rates and number of vertebrae are environmentally influenced (Payne, 1985), therefore differentiation may be due to the relatively sedentary nature of Kingklip adults, allowing for phenotypic differentiation to develop as a result of plasticity and environmental variation between areas (Grant & Leslie, 2005). A later study based on larval distribution patterns provided further support for the existence of two discrete South African stocks (Olivar & Sabatés, 1989), with spatio-temporal variation in spawning strategies observed between the West and South-east coast. Multiple spawning grounds and/or periods have been shown to influence the genetic sub-structuring of marine species in the Benguela (Henriques et al., 2012, 2015). The existence of two

different spawning strategies observed between the West and South-east coasts therefore suggests the existence of two stocks, as previously proposed by Payne (1985) and Olivar and Sabatés (1989). These results must however be interpreted with caution as biological and behavioural differentiation may not necessarily reflect genetic divergence (Carvalho & Hauser, 1994).

To date, only two molecular studies have been conducted on Kingklip stock structure, each revealing contrasting results. The earliest study, based on allozymes, did not detect significant genetic differentiation along the West and South coast of South Africa, suggesting the existence of one single South African population (Grant & Leslie, 2005). In contrast, a more recent study by Henriques et al. (2017), based on mitochondrial DNA (mtDNA) and nuclear microsatellite markers, provided evidence for two sub-populations along the South African coastline, with disruption in gene flow detected between the West and South-east coasts. Although patterns of differentiation were not temporally stable, potentially due to the effects of reproductive sweepstakes, observed genetic differentiation appeared to be associated with the two oceanographic regimes within the region (Henriques et al., 2017). The observed discrepancy between Henriques et al. (2017) and Grant and Leslie (2005) may thus be a result of the different markers used, as allozymes may not be adequate to detect low levels of genetic differentiation (Grant & Leslie, 2005). Furthermore, no genetic studies have been conducted to investigate the population structure of Kingklip along its entire southern African distribution, as neither of the previous studies included samples from the Namibian coastline. Understanding stock structure of Kingklip across the political border is therefore vital for the accurate assessment of current management strategies between the two countries.

Due to a lack of evidence for the existence of a transboundary stock, as well as for practical and political simplicity, South African and Namibian Kingklip resources are currently managed separately (DAFF, 2016). The potential existence of a single stock across the political border is however of commercial importance, as excessive and increased fishing in one area may possibly influence population dynamics across the entire distribution, resulting in a misalignment in management strategies between the two countries (Duncan et al., 2015). Subsequently if a single transboundary stock is detected, joint management may be advisable (von der Heyden et al., 2007; Pinsky et al., 2018), as is currently under debate for the Deep-water Cape hake (*Merluccius*

paradoxus) and Slinger (*Chrysoblephus puniceus*) following the work of Henriques et al. (2016) and Duncan et al. (2015).

In addition, despite possible evidence for the existence of two stocks along the coastline, South African Kingklip management currently follows a one stock approach with an overall Precautionary Upper Catch Limit (PUCL; de Moor et al., 2015; DAFF, 2016). Assessments conducted in 2008 highlighted the importance of stock structure assumptions in influencing estimates of resource status, with a single stock evaluation finding South African Kingklip to be fully exploited. When assessed separately, however, the West coast stock was found to have a greater abundance (replacement yield = 4 102 tons) as compared to the South coast stock (replacement yield = 1 614 tons - Brandão & Butterworth, 2013). Recent Replacement Yield (RY) models found the South coast biomass to be at 40% of its pre-exploitation level (DAFF, 2016). Managing South African Kingklip resources as one stock thus poses the risk of overexploitation, as the precautionary catch limit of the South coast is 1 553 tons compared to 4 302 tons along the West coast (Brandão & Butterworth, 2013). It is therefore generally recommended that a more conservative two-stock management approach be employed within South Africa (Japp, 1990; Punt & Japp, 1994; Grant & Leslie, 2005), as “oversplitting” (i.e. managing a single stock as multiple stocks) is harmless when compared to managing separate stocks as one (Laikre et al., 2005).

Fisheries Management

Population structure and fisheries management

With an estimated \$102 billion USD in marine resources traded globally (estimates based on 2008), fisheries represent a valuable economic and food resource, with capture fisheries exploiting one of the last remaining wild sources of protein (WWF, 2011; Bernatchez et al., 2017). As a result, many world fisheries are near collapse, with 63% of fisheries stocks requiring rebuilding (WWF, 2011). In South Africa, 19 of 35 commercially exploited species are deemed as ‘collapsed’, i.e. they are below 40% of original spawner biomass, six species are over-exploited, and only ten of 35 species are optimally exploited (Bruce Mann, ORI, pers. comm). With predicted increases in anthropogenic pressures, habitat degradation and climate change, the status of global fisheries is cause for concern (Bernatchez et al., 2017). Effectively managing and

assessing marine resources is thus of global importance, with marine management largely focused on commercially valuable species at risk of population declines and overexploitation due to overharvesting (Lundy et al., 2000).

Fundamental to the regional management of fisheries is the identification and incorporation of biologically meaningful population sub-structuring in policy making. The stock concept is thus central to effective management, representing the basic management unit for harvested marine species (Reiss et al., 2009; Ovenden et al., 2015; Lal et al., 2017). Numerous definitions of stock can be found within the literature, including genetic, phenotypic, environmental, fishery and harvest stock (Coyle et al., 1998). In the context of fisheries management, a stock may refer to a semi-discrete, intraspecific group with definable attributes that occur in the same geographical area (Begg et al., 1999). A genetic stock, on the contrary, may refer to an interbreeding group of individuals with a shared/common gene pool, where separate genetic stocks are reproductively isolated and genetically different from one another (Ward et al., 1994; Coyle et al., 1998). Therefore, stock definitions employed to identify management units may differ, depending on management aims, time-scales and interpretations (Coyle et al., 1998). Regardless of the definition employed, demographic differences are central to the stock concept, with stocks reacting more or less independently to harvesting pressures and external influences (Hauser & Carvalho, 2008; Benestan et al., 2015; Ovenden et al., 2015; Spies et al., 2015).

A range of direct and indirect methods are available to infer migration/gene flow and delineate stock structure, with the techniques employed differing over time (Hawkins et al., 2016; Izzo et al., 2017). “Traditional” stock definition methods include differences in life-history parameters (e.g. spawning period and time), morphometrics (e.g. scales and otoliths), meristics (repeated morphological features) as well as tagging data, and have provided extensive evidence for population sub-structuring (Pawson & Jennings, 1996; Cadrin et al., 2005; Campana, 2005; Hawkins et al., 2016). While these methods have proven successful at stock delineation for some species, they may fail to reflect underlying genetic differentiation, as many phenotypic traits/differences arise as a result of population plasticity and environmental variation (Payne, 1985; Carvalho & Hauser, 1994). By employing molecular markers to determine levels of genetic or genomic differentiation, molecular-based methods have provided a universally comparative means for the identification of genetic stock boundaries (Carvalho &

Hauser, 1994; Cadrin et al., 2005), and have proven to be a valuable tool for elucidating stock structure and connectivity at different spatio-temporal scales, with measures of genetic/genomic differentiation allowing for the identification of genetic stock boundaries.

Molecular technologies and their use in fisheries management

Molecular techniques have proven invaluable for fisheries management, offering a range of versatile and useful tools that provide insights into N_e and population dynamics, genetic variation, gene flow and connectivity (Carvalho & Hauser, 1994; Reiss et al., 2009; Seeb et al., 2011; Henriques et al., 2017). Casey et al. (2016) highlights the application of molecular technologies to address three main themes critical to fisheries management: i) resolving stock structure, ii) assessing mixed stock fisheries, and iii) estimate harvest quotas and abundance. Indeed, several studies have argued for the routine integration of genetic and genomic data into fisheries management (Laikre et al., 2005; Ovenden et al., 2015; Casey et al., 2016; Hawkins et al., 2016; Valenzuela-Quinonez, 2016).

The majority of previous studies on genetic population structure of marine species have employed few neutral markers (Hauser & Carvalho, 2008; Nielsen et al., 2009a; Milano et al., 2014). These led to the general observation of weak genetic structure, and low levels of genetic differentiation, which is generally argued to be a result of historically high N_e , high levels of gene flow and/or the lack of effective dispersal barriers found within marine systems (Carvalho & Hauser, 1994; Ward et al., 1994; Hauser & Carvalho, 2008; Nielsen et al., 2009a; Milano et al., 2014). Accordingly, signals of adaptive divergence/local adaptation, which arise as a result of the homogenizing influences of gene flow and diversifying effects of selection (Garant et al., 2007), are predicted to be rare for marine species given the observed high levels of gene flow, which homogenize allele frequencies among populations and limit the effects of natural selection (Hauser & Carvalho, 2008; Nielsen et al., 2009a; Limborg et al., 2012; Milano et al., 2014).

The concepts of no local adaptation and lack of genetic differentiation for marine species have however, been challenged with increasing evidence for genetic structure and local adaptation found for several marine species, despite high levels of gene flow

(Hauser & Carvalho, 2008; Reiss et al., 2009; Helyar et al., 2012; Lamichhaney et al., 2012; Milano et al., 2014; Benestan et al., 2015; Guo et al., 2016). Signals of fine-scale population sub-structuring, as well as adaptive diversity, in species such as the highly mobile Atlantic herring (*Clupea harengus* - Limborg et al., 2012), Atlantic cod (*Gadus morhua* - Nielsen et al., 2009a; Bradbury et al., 2012) and European hake (*Merluccius merluccius* – Milano et al., 2012, 2014), suggest that local adaptation and population divergence may in fact be more common than previously realised (Hauser & Carvalho, 2008; Nielsen et al., 2009a; Bradbury et al., 2012; Di Battista et al., 2017). It has been suggested that previous inability to detect structure may result from the marker type used, with the usefulness of molecular markers depending largely on the sensitivity of the marker employed as well as the number of variable loci analysed (Carvalho & Hauser, 1994; Hauser & Carvalho, 2008). In fact, evidence suggests that previously used neutral markers may not be sensitive enough to reveal population differentiation for species with large N_e (Narum et al., 2013; Milano et al., 2014). Sampling strategies as well as the statistical algorithms employed must additionally be taken into consideration, with continuous advances in analytical methods as well as sampling design influencing population structure analyses and potentially resulting in discrepancies between past and current studies (Guillot et al., 2009; Tucker et al., 2014).

Increasing the number of markers, such as with Single Nucleotide Polymorphism (SNPs) studies that can include thousands of loci, provides opportunities for detecting population differentiation even in high gene flow systems, or between populations shaped by recent divergence (Waples & Gaggiotti, 2006; Reiss et al., 2009; Benestan et al., 2015). Furthermore, putatively adaptive (outlier) loci have been shown to improve the resolution of population structure and assignment success, revealing additional barriers to gene flow and providing insight into the occurrence of adaptive divergence (Reiss et al., 2009; Milano et al., 2011; Nielsen et al., 2012; Bradbury et al., 2012), and the influence of environment in shaping the genetic structure of natural populations (Nielsen et al., 2009a; Limborg et al., 2012; Selkoe et al. 2016). As a result, genome-wide polymorphisms (e.g. SNPs) are argued to be better in the context of fisheries management, with hundreds to thousands of markers improving the resolution of population structure (Nielsen et al., 2009b; Funk et al., 2012; Hess et al.,

2013; Hawkins et al., 2016; Rodríguez-Ezpeleta et al., 2016), thereby ensuring the accurate delineation of stock boundaries vital for effective fisheries management.

Despite these advantages the identification and sequencing of hundreds to thousands of genome-wide molecular markers is associated with high monetary, bioinformatic and computational costs, requiring large infrastructure for library preparation, sequencing and bioinformatic analyses (Narum et al., 2013; Hess et al., 2015; Li & Wang, 2017). Subsequently, genome-wide marker panels have yet to be developed for several marine species (Helyar et al., 2012), with the majority of datasets and studies available focusing on commercially important, Northern hemisphere species (Table 1). Given the commercial value of, and the continued anthropogenic pressures experienced by, Kingklip, the development and employment of genomic molecular tools for its management and conservation should be considered a priority (Helyar et al., 2012). Within this context, genomic marker discovery is a vital first step in the development of such molecular resources (Hubert et al., 2010).

Table 1: Marine species with available genome-wide datasets as well as associated NGS approaches and references. Note, this is not an exhaustive list, but a short representation of different species and NGS approaches employed.

Species	Sequencing approach	Reference
Atlantic cod (<i>Gadus morhua</i>)	WGS cDNA	Star et al., 2011 Hubert et al., 2010
Atlantic herring (<i>Clupea harengus</i>)	RADSeq cDNA cDNA & gDNA	Corander et al., 2013; Guo et al., 2016 Helyar et al., 2012 Lamichhaney et al., 2012
Atlantic mackerel (<i>Scomber scaombrus</i>)	RADSeq	Rodríguez-Ezpeleta, 2016
European hake (<i>Merluccius merluccius</i>)	cDNA	Milano et al., 2011
Spotted sea bass (<i>Lateolabrax maculatus</i>)	RAD-PE	Wang et al., 2016
Striped snapper (<i>Lutjanus carponotatus</i>)	RRSeq	DiBattista et al., 2017
Pacific blue fin tuna (<i>Thunnus orientalis</i>)	WGS	Nakamura et al., 2013

WGS: Whole Genome Sequencing; cDNA: complementary DNA; gDNA: genomic DNA; RADSeq: Restriction Site Associated Sequencing; RRSeq: Reduced Representation Sequencing; RAD-PE: Paired-end sequencing of restriction site associated DNA

Aims and objectives

Considering all of the above, specifically the lack of consensus regarding population structuring, lack of genetic studies including Namibian samples and risks of over-exploitation, the need for a transboundary genetic study of Kingklip population sub-structuring is self-evident. This study therefore aims to develop a novel set of molecular markers (SNPs) in order to assess Kingklip population sub-structuring, with the results intended to contribute towards the establishment of effective and sustainable fisheries management policies. In addition, this project forms part of a collaboration between Stellenbosch University, the University of Namibia and the Department of Agriculture, Forestry and Fisheries (DAFF) aimed at generating high-throughput data for commercially important southern African fishes, thus contributing towards a more comprehensive understanding of population structuring for offshore, demersal species that underpin future management decisions in the region.

The present thesis is split into three inter-connected chapters, outlined below:

CHAPTER 1 - SNP development and identification of local adaptation in southern African Kingklip, *Genypterus capensis*

CHAPTER 2 - Genetic sub-structuring of southern African Kingklip within and between South Africa and Namibia

CHAPTER 3 - Molecular tools in action: Conservation recommendations and implications, as well as development towards a genomic tool for post-harvest control of Kingklip

CHAPTER 1: SNP development and identification of local adaptation in southern African Kingklip, *Genypterus capensis*

INTRODUCTION

The distribution of southern African Kingklip falls within a globally unique region, the cold and productive Benguela Large Marine Ecosystem, bordering Namibia as well as the West and South-west coasts of South Africa (Figure 1). This region is bounded by two warm water systems, the Angola Current to the north and the Agulhas Current to the south (Figure 1 – Shillington et al., 2006; Hutchings et al., 2009). Such variable oceanic conditions translate into environmental heterogeneity, with oxygen availability, salinity and temperature varying longitudinally across the Benguela system (Hutchings et al., 2009). In particular, the year-round Lüderitz upwelling cell (26 °S), characterised by strong winds, turbulence and offshore transport, acts to partially divide the system into two sub-systems: the northern and southern Benguela (Hutchings et al., 2009). The northern sub-system is characterised by Low Oxygen Waters (LOW) and a seasonally shifting Angola-Benguela Frontal Zone, while the southern sub-system is characterised by seasonal-driven upwelling events, and is strongly influenced by the Agulhas current flowing along the Agulhas bank (Figure 1 - Shillington et al., 2006; Hutchings et al., 2009). Freshwater outflow from the Orange River at the border between South Africa and Namibia, as well as the temperature transition zone between Cape Point and Cape Agulhas, are additional features found within this region (Stephenson & Stephenson, 1972; Emanuel et al., 1992; Turpie et al., 2000). These features in conjunction with variation in bathymetry, oxygen availability and upwelling patterns, act to create environmental and seascape heterogeneity throughout the region (Shillington et al., 2006; Hutchings et al., 2009; Teske et al., 2011; Henriques et al., 2016).

Evidence for the potential influence of such oceanographic and environmental features on population genetic structure of pelagic and demersal species has been provided for fishes such as the Shallow-water Cape hake, *Merluccius capensis* (Henriques et al., 2016), Geelbek (*Atractoscion aequidens* – Henriques et al., 2014), Leervis (*Lichia amia* – Henriques et al., 2012), Silver kob (*Argyrosomus inodorus* – Henriques et al., 2015; Mirimin et al., 2016), Bluefish (*Pomatomus saltatrix* – Reid et al., 2016) and

sardines (*Sardinops sagax* – van der Lingen, 2015). In particular, for the Shallow-water Cape hake, oceanographic features, including oxygen availability (LOW conditions), Sea Surface Temperature (SST), depth and chlorophyll a concentration (chl a), were found to significantly influence the genetic differentiation observed between Namibian and South African populations, thereby suggesting that adaptation to local environmental conditions may have contributed towards differentiating populations (Henriques et al., 2016). It must be noted however that these studies were mainly based on surface measurements, with seascape studies of deep-sea species being largely hampered by a lack of available abiotic data.

By providing large sets of genomic data at increasing speeds and decreasing costs, as compared to conventional sequencing techniques, Next Generation Sequencing (NGS) has greatly facilitated genome-wide analyses of genetic variation (Milano et al., 2011; Helyar et al., 2012; Toonen et al., 2013; Hess et al., 2015). This has provided researchers with the ability to investigate a range of evolutionary questions on non-model taxa (Helyar et al., 2012; Narum et al., 2013; Toonen et al., 2013; Hess et al., 2015; Ovenden et al., 2015). More specifically, by increasing the number of variable markers analysed, allowing for hundreds to thousands of genome-wide polymorphisms (largely SNPs), NGS has revolutionized the field of population genomics by providing increased statistical power, accuracy and precision of population genetic estimates, as well as making it possible to detect interspecific differentiation and cryptic population structure despite high levels of gene flow (Reiss et al., 2009; Allendorf et al., 2010; Corander et al., 2013; Shafer et al., 2014; Benestan et al., 2015; Hawkins et al., 2016; Rodríguez-Ezpeleta et al., 2016; Di Battista et al., 2017). Furthermore, by favouring genome scans and increasing genomic coverage, NGS approaches are able to simultaneously identify both neutral and potentially adaptive variation in natural populations, through the detection of outlier loci (Nielsen et al., 2009a; Seeb et al., 2011; Limborg et al., 2012). Such putative outlier loci (i.e. loci with high F_{ST} values that are significantly different to other loci, or that are associated with known environmental features such as SST, salinity, etc.) are assumed to be subject to selective pressures, representing genomic regions which may be under selection, subsequently providing insight into the potential occurrence of adaptive differentiation and/or local adaptation (Milano et al., 2011; Nielsen et al., 2011, but see also Hoban et al. 2016; Lowry et al. 2017). While signals of potential

local adaptation are expected to be rare within high gene flow biological systems, large N_e in conjunction with environmental heterogeneity may increase selective pressures within marine systems, resulting in adaptation to local environmental conditions (Helyar et al., 2012; Limborg et al., 2012). In fact, there is mounting evidence for local adaptation even in the face of high gene flow, including the maintenance of adaptive polymorphisms despite high gene flow, as seen for the Purple sea urchin (*Strongylocentrotus purpuratus* – Pespeni et al., 2010; Pespeni & Palumbi, 2013) and Rainbow trout (*Oncorhynchus mykiss* – Baerwald et al., 2016; but see also the review by Tigano & Friesen, 2016). Furthermore, increasing evidence suggests that gene flow may also act to promote local adaptation, as shown for the Three-spined stickleback (*Gasterosteus aculeatus* – Jones et al., 2012a). Although it is difficult to accurately disentangle the biotic and abiotic variables/drivers that shape marine populations, variations in salinity and SST are generally identified as two of the main environmental factors resulting in selection and local adaptation (Nielsen et al., 2009a; Bradbury et al., 2012; Limborg et al., 2012; Milano et al., 2012; Lal et al., 2017). Other drivers, such as dissolved oxygen, depth and precipitation have also been linked to population divergence (Selkoe et al., 2016).

For many studies that focus on marine species with large N_e and wide geographic distributions, the inclusion of outlier loci has improved the resolution of population structure and individual assignment success, revealing additional barriers to gene flow (Reiss et al., 2009; Ackerman et al., 2011; Helyar et al., 2011; Bradbury et al., 2012; Limborg et al., 2012; Benestan et al., 2015; Ovenden et al., 2015). In some cases, outlier loci and adaptive differences may present the only discriminating factor between populations, uncovering differentiation previously undetected by neutral loci alone (Hawkins et al., 2016; Li & Wang, 2017). Although the adaptive significance of outlier loci is often elusive, putatively adaptive markers are useful for detecting locally adapted populations and can therefore help delineate conservation management units (Nielsen et al., 2009b; Limborg et al., 2012; Funk et al., 2012; Shafer et al., 2014). Further, by improving our understanding of how locally adapted populations may respond to environmental change, putatively adaptive markers can aid in conservation efforts, targeting adaptive and intraspecific diversity as well as evolutionary processes (Nielsen et al., 2009b; Milano et al., 2011; Funk et al., 2012; Limborg et al., 2012; Shafer et al., 2014). In the face of continued fishing pressures and climate change, it

is vital that fisheries management should include the conservation and protection of intra-specific adaptive variation (von der Heyden 2007; Reiss et al., 2009; Ovenden et al., 2015). This is of particular importance for Kingklip, which have been found to display reduced levels of contemporary genetic diversity, potentially as a result of past over-exploitation (Henriques et al., 2017).

Restriction Site Associated Sequencing

Despite reduced costs compared to traditional sequencing methods, sequencing the entire genome of hundreds of individuals still remains prohibitively expensive, particularly for countries with low investment in research and development that may not have the capacity and infrastructure required for library preparation, sequencing and downstream analyses. In addition, sequencing the entire genome is often unnecessary for the purpose of most population and phylogeographic studies, with whole genome sequencing (WGS) inflating computational and bioinformatics costs (Narum et al., 2013; Hess et al., 2015; Li & Wang, 2017). The development of Genotyping-by-Sequencing (GBS) methods presents a solution to this problem. By combining the power of high-throughput sequencing and large-scale genotyping, GBS approaches target a fraction of the genome whilst still providing large sets of data and allowing for genomic regions potentially affected by selection to be identified (Helyar et al., 2011; Narum et al., 2013).

This approach is used in techniques such as Restriction-Site Associated DNA sequencing (RADseq - Baird et al., 2008; Hohenlohe et al., 2011), one of the most popular and widely used GBS methodologies (Baird et al., 2008; Davey et al., 2013; Rodríguez-Ezpeleta et al., 2016). By implementing several filters, quality control steps and employing restriction enzymes, RADseq reduces genome complexity whilst still producing thousands of short sequence reads spread throughout the genome (Davey & Blaxter, 2010; Hohenlohe et al., 2011; Toonen et al., 2013). As a result, RADseq approaches are able to genotype thousands of genome-wide polymorphisms, regardless of species genome size or state of prior genomic knowledge available (Baird et al., 2008; Davey & Blaxter, 2010; Seeb et al., 2011; Davey et al., 2013; Narum et al., 2013). As such, RADseq approaches have been employed for several population structure studies with a sufficient number of unbiased SNPs accurately

reflecting genome-wide diversity (Corander et al., 2013; Larson et al., 2014; Rodríguez-Ezpeleta et al., 2016; Catchen et al., 2017; Fischer et al., 2017) (Table 1). This methodology is thus highly advantageous for genomic studies in non-model organisms, such as Kingklip.

Despite the reduction in NGS costs provided by GBS methodologies, sequencing a large number of individuals still remains expensive (Huang et al., 2015). As an alternative, sequencing pools of DNA samples (Pool-Seq) provides a more cost-effective approach for SNP discovery and genome-wide sequencing, allowing for increased samples to be analysed at a fraction of the cost (Futschik & Schlötterer, 2010; Schlötterer et al., 2014; Fu et al., 2016). As a result, Pool-Seq has been shown to increase the probability of SNP detection as well as accuracy of allele frequency and population genetic parameter estimates, as it increases the number of individuals analysed (Futschik & Schlotterer, 2010; Toonen et al., 2013).

There are, however, several limitations of Pool-Seq that need to be considered. These include less accurate base calling and the effects of unequal representation and contamination of pools (Schlötterer et al., 2014). Pool size is an important consideration, as small pools risk unequal individual representation. This can be overcome by increasing the pool size, thereby reducing the impact of differential individual representation (Schlötterer et al., 2014). A further limitation is the difficulty of distinguishing sequencing errors from low-frequency alleles (Schlötterer et al., 2014). However, SNP calling software has greatly improved, including features allowing for the identification of sequencing errors, such as false positives and misalignments (Schlötterer et al., 2014). Furthermore, by employing strict quality filtering steps, sequencing errors can be reduced helping to improve the reliability of SNP detection (Futschik & Schlotterer, 2010; Henriques et al. in review). Pool-Seq has thus been identified as a valuable tool for population genomic analyses (Fu et al., 2016), being previously employed in the study of local adaptation and patterns of population genomic differentiation of the Three-spined stickleback (*G. aculeatus* – Guo et al., 2015, 2016), Great scallop (*Pecten maximus* – Vendrami et al., 2017), Cape urchin (*Paranichius angulosus* – Nielsen et al. 2018), Granular limpet (*Scutellastra granulata* – Nielsen et al., 2018) and Prickly sculpin (*Cottus asper* – Dennenmoser et al., 2017).

Single Nucleotide Polymorphisms - SNPs

The advancement of high-throughput genotyping approaches, such as RADseq, aided in overcoming the challenges associated with the development of large genomic data sets, enabling the production of highly diagnostic marker panels. Single Nucleotide Polymorphisms represent the most abundant and widespread DNA sequence polymorphisms within the eukaryote genome, making them well suited for high-throughput genotyping (Glover et al., 2010; Hubert et al., 2010; Milano et al., 2011). Compared to widely used microsatellite markers traditionally employed in fisheries management, SNPs are found to have lower sequencing error rates whilst providing higher quality data and better fine-scale population structure resolution (Martinsohn & Ogden, 2009; Clemento et al., 2014; Anderson et al., 2017). In addition, SNPs do not require inter-laboratorial calibration, allowing them to be compared across different laboratories (Martinsohn & Ogden, 2009; Milano et al., 2011; Clemento et al., 2014; Anderson et al., 2017). Due to their genome-wide distribution and abundance, information can be obtained from several regions, capturing both neutral variation and loci potentially under selection. Furthermore, the use of multiple polymorphic markers often enables researchers to assign a sample to a single source (Martinsohn & Ogden, 2009), making genome-wide SNPs ideal for collaborative traceability and genetic stock identification efforts (Ackerman et al., 2011). The improved probability of assignment success can aid in the accurate identification of harvested individuals and improved product traceability thereby helping to deter Illegal, Unregulated and Unreported (IUU) fishing and aiding in eco-certification efforts (Martinsohn & Ogden, 2009; Benestan et al., 2015). For example, this was the aim of FishPopTrace, a Pan-European initiative aimed at developing SNP marker panels and DNA techniques to facilitate product traceability and monitoring of commercially valuable species (Martinsohn & Ogden, 2009; FishPopTrace, 2017). By identifying hundreds of novel genetic markers (SNPs), FishPopTrace provided new, cost-effective and fast traceability tools and baseline data, for several European fish species including Atlantic herring (*C. harengus*), Atlantic cod (*G. morhua*) and European hake (*M. merluccius*), allowing for fish and fish products to be traced back to their population/area of origin (FishPopTrace, 2017).

Chapter aims

Within the context provided above, the identification of hundreds to thousands of genome-wide SNPs is ideal for assessing the population sub-structuring and adaptive diversity of southern African Kingklip, by isolating both neutral and putatively adaptive loci. As such, the aim of this chapter is to identify a novel set of variable SNP markers that captures both putatively adaptive and neutral regions, that can be utilised in the management and conservation policies of southern African Kingklip resources.

METHODS

Sample collection

Samples of mature individuals (total length > 30 cm) were collected from commercial fishing operations (through fisheries observers), as well as research surveys, spanning the distributional range of Kingklip in South Africa and Namibia (Figure 2 and Table 2).

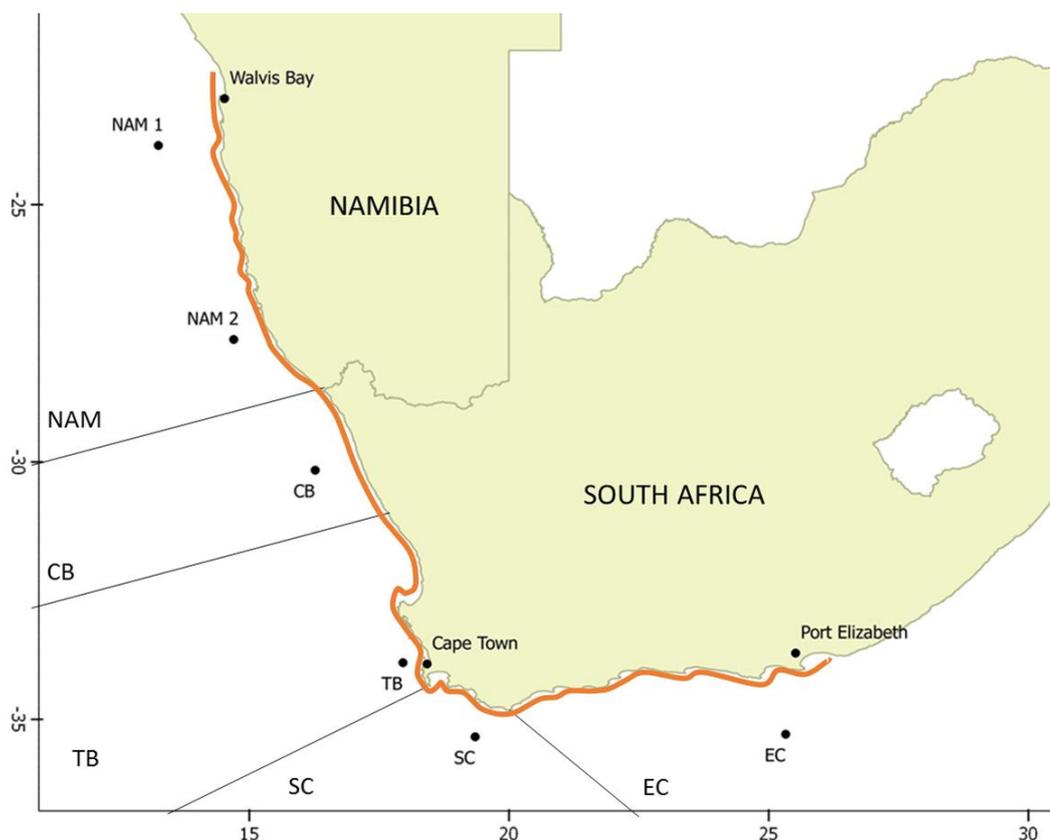


Figure 2: Sampling locations for Kingklip (2014 & 2017). CB - Child's Bank, TB – Table Bay, SC - South Coast, EC – Eastern Cape and NAM – Namibia. Kingklip distribution indicated in orange.

Muscle tissue from each individual was stored in 95% ethanol. Sampling in South Africa took place in 2012, 2014, 2015 and 2017, with samples collected off the West coast (Child's Bank; CB & Table Bay; TB) and South-east coast (South Coast; SC & East Coast; EC), east of Cape Point. Namibian samples were collected from two main sampling areas in 2017 (Figure 2 and Table 2). As such, 2017 was the only year with samples from all regions and was subsequently used for population sub-structuring analyses in Chapter 2

Table 2: Sampling location, code, year and number of individuals sampled per pool.

Country	Site	Pool ID	Year	Nb. Samples per pool
South Africa	CB, TB, SC, EC	P1	2014,2015,2016	20
	CB, TB, SC, EC	P2	2014,2015,2016	20
Namibia	northern Namibia	NAM 1	2017	39
	southern Namibia	NAM 2	2017	44
South Africa	Child's Bank	CB	2017	28
	Table Bay	TB	2017	28
	South Coast	SC	2017	20
	East Coast	EC	2017	31

In addition, two pools containing 20 individuals identified by Henriques et al. (2017) as belonging to two separate sub-populations/groupings were also included. Due to spatial and temporal variation, these two proposed sub-populations comprised of a mixture of individuals collected from different years and sampling locations along the South African coastline (Table 2 & Figure 2).

DNA extraction and pooling of samples

Total genomic DNA was extracted from tissue samples following the CTAB protocol (Winnepenninckx et al., 1993) and stored at -20°C. A 1% Agarose gel with 1 Kb DNA ladder (®Promega) was run to assess DNA quality and degradation for each sample. DNA concentration was quantified using the Qubit Quanti It dsDNA HS Assay system at the Central Analytical Facility (CAF), Stellenbosch, with samples with a concentration below 5 ng/ul excluded. Following DNA extraction, between 20 to 50 individuals were pooled by sampling location, year and depth. DNA concentrations were standardized based on Qubit results to ensure equal representation of individuals

within each pool. Each pool comprised a final concentration of 3000 ng/ul. Pooled samples were flash frozen and sent to the Hawaii Institute of Marine Biology for library construction and Mi-Seq Illumina sequencing.

ezRAD Sequencing

For the purpose of this study, library preparation and sequencing was conducted at the Hawaii Institute of Marine Biology using the ezRAD sequencing protocol developed by Toonen et al. (2013). Unlike conventional RADseq methods, ezRAD (Toonen et al., 2013) is a RADseq strategy that can make use of any restriction enzyme, or combination of enzymes, to double digest DNA to produce suitably sized sequencing fragments. Digested DNA is then inserted into a TruSeqDNA kit, following the sample preparation guide. Standard Illumina TruSeq library preparation with agarose gel size selection is used to select sequencing fragments. By simply altering the restriction enzyme and/or size of selection used, ezRAD provides researchers with the ability to optimize the number of fragments sequenced (Toonen et al., 2013).

SNP development pipeline

Quality control

A number of bioinformatic steps were completed to produce the final SNP panels (Figure 3, see also Supplementary Material 1 for scripts used). Base calling was completed by the sequencing facility, with ezRAD incorporating a standard quality control filter providing Illumina reads in FastQ format (Toonen et al., 2013). Read data was then analysed using FASTQC and FASTQ toolkits available on the Basespace Illumina platform (Andrews, 2010). All raw reads were trimmed for over-represented sequences, adapter sequences, and reads with a Phred quality score below 25 ($Q > 25$) in TrimGalore! V 0.4.4 (Babraham Bioinformatics, 2017), producing quality-controlled reads for each pool. Quality controlled reads were then assessed in FASTQC as before, and used for further assembly, mapping and bioinformatic analyses.

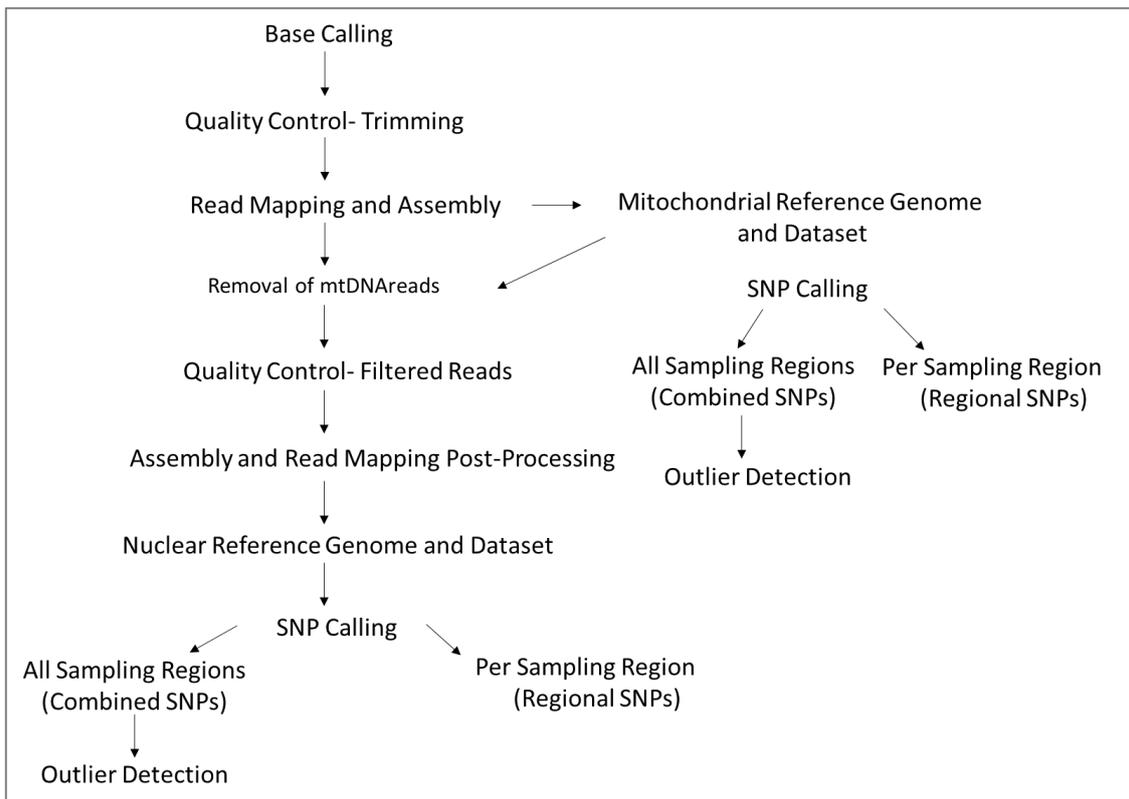


Figure 3: Bioinformatic pipeline followed for the identification and development of Chapter 1 Single Nucleotide Polymorphism (SNP) databases.

Assembly and mapping of the mitochondrial DNA dataset

Whole genome and NGS datasets contain both mitochondrial and nuclear DNA (Al-Nakeeb et al., 2017). Therefore, in addition to developing highly informative SNP panels, NGS and high-throughput sequencing represents a valuable resource for extracting and assembling mitochondrial genomes (Hahn et al., 2013; Al-Nakeeb et al., 2017). ‘Traditional’ mitochondrial genome assembly and sequencing is both labour intensive and resource demanding, with mtDNA reads needing to be isolated beforehand (Al-Nakeeb et al., 2017). However, NGS datasets already contain mtDNA, thereby removing the need to isolate it prior to sequencing. As a result, the increased accessibility and use of NGS and WGS has led to an increase in mitochondrial genome assembly and sequencing (Coulson et al., 2006; Hahn et al., 2013; Ding et al., 2015). By employing an available mitochondrial genome of a well-studied species as a reference, mtDNA reads can be extracted and subsequently assembled *de novo*, with the effectiveness of such approaches being previously demonstrated (Hahn et al., 2013; Al-Nakeeb et al., 2017). As such the data generated within this study provides

the opportunity to employ the above methodology to generate and assemble the first partial Kingklip mitochondrial genome/mitogenome, as this will allow to generate a predominantly nuclear dataset for downstream analyses (please refer to Assembly and mapping of the nuclear DNA dataset).

In order to identify mtDNA reads, trimmed reads were mapped and aligned to the mitochondrial genome of Atlantic cod, *Gadus morhua* (Genbank HG514359.1), in BWA 0.7.13 (Li & Durbin, 2009). Due to the lack of a complete mitochondrial reference genome for *Genypterus* species, the mitochondrial genome of *G. morhua* was selected as a “backbone” for the identification and assembly of mitochondrial reads. BWA makes use of the Burrow-Wheeler Transform (BWT) base algorithm for read mapping (Li & Durbin, 2009; Nielsen et al., 2011; Altman et al., 2012). This is a fast, memory efficient approach that is well suited for the alignment of repetitive reads (Nielsen et al., 2011; Altman et al., 2012). Mapping was done using the MEM sub-command and default mapping parameters, with reads with a mapping quality score of less than 20 ($T < 20$) being filtered out.

Downstream processing and filtering made use of sorted SAM and BAM file formats, with alignments needing to be sorted with regards to their positions prior to filtering (Altman et al., 2012). This was done by converting SAM files into BAM file format and then sorting by position in SAMtools 1.3 (Li et al., 2009). The resulting sorted BAM files (mtDNA_BAM) were subsequently filtered to include properly-paired, unique reads that mapped to the Atlantic cod mtDNA genome only (Ding et al., 2015). These properly-paired, uniquely mapped reads were then converted from BAM to FastQ file format in Picard (Galaxy platform; Blankenberg et al., 2010), and subsampled (SAMtools 1.3; Li et al., 2009) to ensure uniform coverage from each pool. In order to create a Kingklip-specific mitochondrial reference sequence (KK_mtDNA_ref), the properly-paired, uniquely mapped reads were *de novo* assembled in SPAdes (Bankevich et al., 2012). SPAdes employs multiple k-mer values for assembly (Bankevich et al., 2012), with the quality of assemblies largely influenced by the selected k-mer values. A k-mer refers to all possible sequences (of length k) found within a read, with multi-kmer assemblers largely outperforming approaches based on a single k-mer value, as a result of a single value rarely fitting all genes (Durai & Schulz, 2016). KmerGenie (Chikhi & Medvedev, 2013) was used to select optimal k-mer lengths for *de novo* assembly, and k-values of 19 (lower), 21 (optimal) and 31

(higher) were used for assembly. Assembly statistics, including number of contigs, total contig length, largest contig, N50 and L50, were estimated for the resulting mtDNA reference sequence using Quast (Gurevich et al., 2013).

In order to produce a predominantly mitochondrial dataset (KK_mtDNA), previously mapped mtDNA reads were mapped to the *de novo* Kingklip reference (KK_mtDNA_ref). Mapping was completed using BWA (Li & Durbin, 2009) as outlined above. The resulting SAM files were then sorted, subsampled to the median coverage across all pools and converted into BAM file format (KK_mtDNA_BAM), with “ambiguous” reads with a mapping quality below 20 (MAPQ <20) removed (SAMtools 1.3; Li et al., 2009). Mapping results were assessed using the ‘samstats’ command in SAMtools (Li et al., 2009). Sorted files (KK_mtDNA_BAM) were then used to call mitochondrial specific variants/SNPs. SNP calling, also referred to as variant calling, aims to determine in which position there are polymorphisms and variation from the reference genome (Nielsen et al., 2011). SNP calling was completed using the SAMtools ‘mpileup’ command (Li et al., 2009). Filter parameters of a minimum base quality of 20 (Q>20), mapping quality of 10 (MAPQ>10) and maximum of 10,000 reads per position were used for the initial SNP calling. Calling was completed per sampling site/pool (Per region SNPs; pileup) as well as for all sampling sites/pools combined (combined SNPs; mpileup). The mpileup file (combined SNPs) was converted to a sync file format, using the ‘mpileup2sync.pl’ command in PoPoolation2 (Kofler et al., 2011a), and subsequently filtered by minimum count, minimum and maximum coverage (see below), to ensure accurate SNP calling. The resulting pileup files (per region SNPs) and sync file (combined SNPs) were used for subsequent analyses.

Assembly and mapping of the nuclear DNA dataset

In order to develop a predominantly nuclear DNA (nDNA) dataset, potential mtDNA reads found to map to the cod mitochondrial genome were filtered from the original data. The previously obtained mtDNA BAM files from all pools were merged and converted to SAM file format using BAMTools (Barnett et al., 2011). The resulting merged SAM file was then filtered from the original quality-controlled reads using the ‘filterbyname.sh’ command of BBMAP (Barnett, 2011), producing a set of ‘filtered’ FastQ files per pool.

Following the removal of mtDNA reads, filtered reads were assembled *de novo* in SPAdes (Bankevich et al., 2012), creating a Kingklip-specific filtered nDNA reference sequence (nDNA_ref), used for subsequent mapping and SNP calling. K-mer lengths of 19, 21 and 31 were selected for the *de novo* assembly based on the outputs of KmerGenie (Chikhi & Medvedev, 2013). Assembly statistics for the resulting reference sequence were estimated in Quast (Gurevich et al., 2013). In addition, to ensure the successful removal of mtDNA reads, the largest contig of the resulting reference sequence (nDNA_ref) underwent a BLASTN search. Upon the successful removal of mtDNA reads, filtered reads were mapped to the reference sequence (nDNA_ref) using BWA 0.7.13 (Li & Durbin, 2009), as previously outlined. The resulting SAM files were then sorted by their position, converted into BAM file format and filtered for ambiguous reads with a mapping quality score of less than 20 (MAPQ<20) (SAMTools; Li et al., 2009). Sorted BAM files (nDNA_BAM) were subsequently used for SNP/variant calling.

SNP calling was completed using the SAMtools 'mpileup' command (Li et al., 2009). Filter parameters of a minimum base quality of 20 (Q>20), mapping quality of 10 (MAPQ>10) and maximum of 10,000 reads per position was used for SNP calling. Calling was completed per sampling site/pool (Per region SNPs; pileup) as well as for all sampling sites/pools combined (combined SNPs; mpileup). As before, the mpileup file (combined SNPs) was converted to a sync file format, using the 'mpileup2sync.pl' command in PoPoolation2 (Kofler et al., 2011a), and subsequently filtered by minimum count, minimum and maximum coverage (see below), to ensure accurate SNP calling. The resulting pileup files (per region SNPs) and sync file (combined SNPs) were used for subsequent analyses.

Regional diversity measures

Summary measures of genomic diversity were calculated from individual pileup files (per region SNPs) in PoPoolation 1.2.2 (Kofler et al., 2011b), using the 'variance-sliding.pl' command (Supplementary Material 1). Nucleotide diversity (Tajima's π), population mutation rate (Watterson's θ_w) and Tajima's D were calculated for the mtDNA and nDNA dataset, per sampling site/pool, to obtain regional diversity measures. Diversity measures were calculated using a window- and step-size of 100

bp, with SNPs subject to analysis required to have a minimum coverage of 10, maximum coverage of 500 and Phred quality score of 10. Based on the median and standard deviation of the number of mapped reads across all pools, for each dataset, a minimum allele count of 2 and 3 was set for calculations of Tajima's π and Watterson's θ_W for the mtDNA and nDNA datasets respectively. Tajima's D requires a minimum allele count of 2 and was subsequently calculated for SNPs with a minimum allele count of 2 for both datasets.

For the overall mpileup file, SNPs were identified based on allele counts estimated from sync files (combined SNPs), using the 'snp-frequency-diff.pl' command in PoPoolation 2 (Kofler et al., 2011a; Supplementary Material 1). A non-overlapping 100 bp window and minimum Phred quality score of 10 was used to identify SNPs for each dataset. Allele counts were subsequently estimated for a subset of identified SNPs based on a stringent criterion. SNP calling in the mtDNA dataset required a minimum coverage of 10 reads, maximum coverage of 500 reads and a minimum allele count of 3, whereas nDNA SNPs were identified based on a minimum coverage of 20 reads, maximum coverage of 500 reads and a minimum allele count of 4. These parameters were set to prevent the possible loss of rare alleles and/or inclusion of sequencing errors, with too low a minimum count and coverage potentially resulting in the inclusion of sequencing errors, whilst too strict a minimum count and coverage potentially resulting in a loss of rare alleles. Employed parameters differed between the two datasets due to differences in the median number and standard deviation of mapped reads across pools, with stricter parameters employed for the nDNA dataset due to increased coverage. In addition, 'Reference call' (rc) SNPs were filtered out in order to only retain SNPs that are present among sampling sites (pop), and not against the reference file. The resulting allele counts were used to identify the number of biallelic SNPs as well as private (SNPs specific to a certain pool), biallelic SNPs per pool for each dataset (mtDNA and nDNA), using a final filtering step. With SNPs found to be largely biallelic in nature, multiallelic or non-biallelic SNPs are believed to be more likely to represent sequencing errors (Kumar et al., 2012) and as such, only biallelic SNPs were used in all subsequent analyses.

Outlier detection

Outlier loci can be identified using one, or several, F_{ST} and genome scan outlier detection methods (Seed et al., 2011). Based on stringent filtering criteria, a subset of previously identified biallelic, 'pop' SNPs were used for outlier detection, with mtDNA biallelic SNPs required to have a minimum allele count of 2, minimum coverage of 28 reads and maximum coverage of 500 reads. In contrast, due to the larger number of SNPs and coverage found for the nDNA dataset, "stricter" parameters were employed for outlier detection, with biallelic SNPs requiring a minimum allele count of 4, minimum coverage of 28 reads and maximum coverage of 100 reads.

For the purpose of this study, outlier loci were identified using three methodologies. The first method made use of a Bayesian approach to directly estimate the posterior probability of a locus being under selection, as implemented in Bayescan 2.1 (Foll & Gaggiotti, 2008). Here, outlier loci were identified by estimating the posterior probability for two alternative models, one including the effects of selection and the other excluding it, using a reversible-jump Markov Chain Monte Carlo (MCMC) approach. Sync files and biallelic SNP lists were used to create simulated Genepop files (Rousset, 2008) using the 'subsample-sync2GenePop.pl' command in PoPoolation2 (Kofler et al., 2011a; Supplementary Material 1), as that is the required input for Bayescan. In addition to converting sync files into Genepop format (Rousset, 2008) this command simulates a set number of individuals per pool (target coverage), performing an internal random subsampling step in order to ensure uniform coverage. A target coverage of 28 was selected for each pool based on the median number of samples per pool. In addition to simulating a set number of individuals per pool, the defined target coverage acts as the minimum coverage threshold, discarding entries with a coverage below the defined target. Resulting Genepop files were edited and converted into Bayescan file format using PGDSpider2 v2.1.03 (Lischer & Excoffier, 2012). A total of 20 pilot runs of 5,000 iterations each, and a burn-in period of 50,000 reversible jump chains with a thinning interval of 10, was employed for the mtDNA dataset. Prior odds were set to 10, suitable for fewer than a hundred loci, with a target False Discovery Rate (FDR) of 0.05 used to identify candidate outlier loci. For the nDNA dataset, a total of 40 pilot runs of 5,000 iterations each and a burn-in period of 50,000 reversible jump chains with a thinning interval of 20 were used. Due to the

nDNA dataset containing thousands of SNPs, prior odds were set to 100, with a FDR of 0.05 used to identify candidate outlier loci.

The second method employed an empirical outlier detection approach in PoPoolation2 (Kofler et al., 2011b). Pairwise F_{ST} values were calculated for each SNP using the 'fst-sliding.pl' command of PoPoolation2 (Kofler et al., 2011a). SNPs/loci falling into the upper 95th percentile of the empirical distribution of pairwise F_{ST} values were subsequently identified as outlier loci.

Finally, outlier loci were identified using the R package pcadapt (Luu et al., 2017). Pcadapt identifies outliers from pooled datasets based on a matrix of allele frequencies calculated for each pool. By performing genome scans and ascertaining population structure through PCAs, correlations between SNPs and principal components are used to calculate tests statistics, Mahalanobis distance and p-values. Outlier loci are subsequently identified based on their relation to population structure, with loci excessively related to structure (q-value < 0.05) identified as candidates for selection (Luu et al., 2017).

Potential outliers identified for the nDNA dataset by either of the three approaches were subsequently removed or extracted from simulated Genepop files as well as biallelic SNP lists, producing both "neutral" (putative outlier loci removed) and "outlier" (outlier loci only) files to be used for population sub-structuring analyses in Chapter 2. In addition, loci pinpointed by two of the three methodologies for either dataset, were identified as candidate outliers. In order to understand the potential functional role of these candidate outliers, nodes containing one or more SNPs, underwent a BLASTX search. BLASTX searches were done using default parameters and the non-redundant protein sequence database, available online (NCBI; National Centre for Biotechnology Information, 2018).

Additional/Exploratory outlier detection

As a result of previously identified candidate outliers failing to map to known areas of interest an additional, exploratory outlier detection was performed on the complete, nDNA dataset (>40K loci). By increasing the input dataset employed for outlier detection, additional outliers may be identified, highlighting potential selective forces

previously overlooked. As before, biallelic SNP loci were required to have a minimum allele count of 4 and coverage of 20 to 500 reads. Pcadapt (Luu et al., 2017) and empirical outlier detection via PoPoolation2 (Kofler et al., 2011a) were performed using the same parameters as previously employed (Chapter 1 methods; outlier detection). The top 1% of outliers were subsequently identified with those detected by both pcadapt and PoPoolation2 being subject to BLASTX searches. BLASTX searches were done using default parameters and the non-redundant protein sequence database in order to understand the potential functional role of the identified outliers.

RESULTS

Pooling, ezRAD sequencing and quality control

A total of 230 samples were pooled with a median of 28.75 samples per pool. Due to difficulties in obtaining high-quality DNA for the older samples (i.e. 2012), a total of 20 samples were pooled for P1, P2, and a maximum of 44 samples pooled for NAM 1 (Table 2). A total of 42.7 million paired-end reads were received following ezRAD sequencing, with an average of 5.3 million paired-end reads per pool (Table 3). A total of 41.8 million reads remained following initial quality control (Phred quality score > 20) and the removal of adapter and overrepresented sequences, with an average of 5.2 million reads per pool (Table 3). These quality-controlled reads were used for further bioinformatic analyses.

Table 3: Sequencing results per pool for Kingklip. Number of raw reads sequenced, number of quality-controlled reads (after initial quality control) and percentage of raw reads remaining after initial quality control (% high quality reads) per pool. Pool names as per Table 2.

Pool ID	Nb. Raw reads	Nb. Quality controlled reads (after initial quality control)	% High quality reads (after quality control)
P1	4 391 730	4 327 938	98.55
P2	4 661 820	4 553 164	97.67
NAM 1	5 881 884	5 741 648	97.62
NAM 2	5 841 342	5 569 726	95.35
CB	5 842 392	5 751 270	98.44
TB	5 821 848	5 726 396	98.36
SC	5 939 718	5 825 752	98.08
EC	4 417 046	4 327 898	97.98

Mitochondrial dataset: assembly and mapping

A total of 15 246 quality-controlled reads mapped to the mtDNA reference genome of *G. morhua*, with an average of 1 906 reads per pool (Table 4). Of these mapped reads, and following subsampling, a total of 9 968 were properly-paired, unique reads with an average of 1 246 reads per pool (Table 4). *De novo* assembly of properly-paired, uniquely mapped reads produced a Kingklip-specific mtDNA reference sequence (KK_mtDNA_ref), with 11 contigs and a total length of 9 110 base-pairs (bp). A total of 8 902 reads (89.31 %) of the properly paired, uniquely mapped reads, mapped to KK_mtDNA_ref, with an average of 1 112 reads per pool ranging from 646 (P2) to 1 364 (NAM 1) reads (Table 4).

Table 4: Mapping statistics for the mitochondrial dataset of Kingklip. Number of properly-paired, unique reads mapped to *Gadus morhua* mitochondrial genome (*G. morhua* reference) per pool. Number of reads mapped to Kingklip mitochondrial reference sequence (KK_mtDNA_ref) per pool. Pool names as per Table 2.

Pool ID	Nb. Mapped reads (<i>G. morhua</i> reference)	Nb. Mapped reads (KK_mtDNA_ref)
P1	1 454	1 308
P2	740	646
NAM 1	1 520	1 364
NAM 2	1 438	1 234
CB	1 450	1 308
TB	1 112	976
SC	1 446	1 372
EC	808	694

Nuclear dataset: assembly and mapping

A total of 41.8 million quality-controlled reads remained following the removal of reads that mapped to the mitochondrial genome, with an average of 5.2 million reads per pool remaining (Table 5). *De novo* assembly of filtered reads produced a nuclear reference sequence (nDNA_ref) of 269 771 contigs and a total length of 240 421 810 bp, with the longest contig being 7 228 bp and N50 and L50 equalling 918 bp and 91 520 bp respectively. BLASTN results showed the successful removal of mtDNA reads. A total of 32.5 million (77.75 %) filtered reads mapped to the nDNA_ref, with an average of 4.1 million reads per pool, ranging from 3 390 428 (P1) to 4 533 875 (CB) reads (Table 5).

Table 5: Filtering and mapping statistics for nuclear dataset of Kingklip. Number of reads prior to filtering of mtDNA reads (quality-controlled reads) and remaining following the removal of potential mitochondrial reads (nDNA reads), per pool. Number and percentage of filtered reads mapped to nDNA reference per pool. Pool names as per Table 2.

Pool ID	Nb. Of reads processed (Prior to filtering)	Nb. nDNA reads (Post filtering)	Nb. Mapped reads	% Reads mapped
P1	4 327 938	4 326 460	3 390 428	78.36
P2	4 553 164	4 552 420	3 434 610	75.45
NAM 1	5 741 648	5 737 974	4 449 554	77.55
NAM 2	5 569 726	5 566 762	4 334 571	77.87
CB	5 751 270	5 749 792	4 533 875	78.85
TB	5 726 396	5 725 284	4 426 393	77.31
SC	5 825 752	5 824 196	4 603 744	79.05
EC	4 327 898	4 327 076	3 393 515	78.43

Regional diversity measures

mtDNA dataset

A total of 12 biallelic SNPs were identified across all pools, with 3 to 7 biallelic SNPs found per pool. Only two pools, namely NAM 2 and TB, contained private SNPs representing 33.33% and 20% of the total number of biallelic SNPs identified for each pool respectively (Table 6). Genome-wide regional diversity measures ranged from 0.0017 to 0.0037 for Tajima's π and 0.0037 to 0.0053 for Watterson's θ_w . Tajima's D was negative across all pools (overall D = -1.12), possibly indicating the occurrence of positive selection and/or a recent population bottleneck followed by expansion (Table 6).

Table 6: Regional diversity measures based on mitochondrial dataset per pool for Kingklip: nucleotide diversity (Tajima's π), population mutation rate (Watterson's θ_w) and Tajima's D, total biallelic SNPs, private biallelic SNPs and percentage private, biallelic SNPs. Pool names as per Table 2.

Pool ID	Tajima's π	Watterson's θ_w	Tajima's D	Nb. Biallelic SNPs	Nb. Private biallelic SNPs	% Private biallelic SNPs
P1	0.0021	0.0050	-1.4135	5	0	00.00
P2	0.0029	0.0053	-1.1651	4	0	00.00
NAM 1	0.0022	0.0047	-1.3184	7	0	00.00
NAM 2	0.0021	0.0037	-0.8912	3	1	33.33
CB	0.0018	0.0039	-1.1555	3	0	00.00
TB	0.0022	0.0045	-1.1571	5	1	20.00
SC	0.0017	0.0038	-1.1755	5	0	00.00
EC	0.0037	0.0053	-0.7128	4	0	00.00

nDNA dataset

A total of 41 369 biallelic SNPs were identified across all pools with the number of biallelic SNPs per pool ranging from 23 564 (P2) to 27 838 (CB) SNPs (Table 7). Private biallelic SNPs ranged from 79 (EC) to 167 (CB) SNPs per pool, representing less than one percent of the total number of biallelic SNPs identified per pool (Table 7). Genome-wide diversity measures ranged from 0.0139 to 0.0199 for Tajima's π and 0.0138 to 0.0207 for Watterson's θ_w , with an average of $\pi = 0.0173$ and $\theta_w = 0.0177$ across all pools respectively (Table 7). All pools displayed a slightly negative Tajima's D value (average D = -0.47).

Table 7: Regional diversity measures based on nuclear dataset per pool for Kingklip: nucleotide diversity (Tajima's π), population mutation rate (Watterson's θ_w) and Tajima's D, total number biallelic SNPs, private biallelic SNPs and percentage private, biallelic SNPs. Pool names as per Table 2.

Pool ID	Tajima's π	Watterson's θ_w	Tajima's D	Nb. Biallelic SNPs	Nb. Private biallelic SNPs	% Private biallelic SNPs
P1	0.0199	0.0207	-0.5503	24 434	95	00.39
P2	0.0178	0.0182	-0.4896	23 564	90	00.38
NAM 1	0.0139	0.0138	-0.4376	26 604	160	00.60
NAM 2	0.0173	0.0176	-0.4303	26 949	109	00.40
CB	0.0165	0.0168	-0.4483	27 838	167	00.60
TB	0.0171	0.0175	-0.4492	27 418	143	00.52
SC	0.0172	0.0175	-0.4521	27 274	146	00.54
EC	0.0185	0.0191	-0.4695	25 432	79	00.31

Outlier detection*mtDNA dataset*

Outlier loci were identified from a total of nine biallelic loci. Bayescan found no outlier loci (FDR = 0.05), while two outlier loci (22.23%) were detected through PoPoolation2 (Kofler et al., 2011a). Pcadapt (Luu et al., 2017) required a minimum allele frequency of 0.00, and therefore could not be used for outlier detection. The two identified outliers were not found within all pools, occurring at different frequencies across pools (Table 8). The BLASTX search identified the corresponding nodes, containing the identified outlier loci, as NADH dehydrogenase subunit 1 (percentage identical: 81%) and an uncharacterised protein (percentage identical: 79%).

Table 8: Total number of outlier SNPs identified per pool for mitochondrial (mtDNA) and nuclear (nDNA) datasets. *Outlier SNPs refer to outlier loci identified by PoPoolation 2 **Outlier SNPs refer to outlier loci identified by two or more outlier detection approaches. Pool names as per Table 2.

Pool ID	Nb. mtDNA outlier SNPs *	Nb. nDNA outlier SNPs **
P1	1	23
P2	0	23
NAM 1	2	25
NAM 2	2	23
CB	0	23
TB	1	24
SC	1	24
EC	0	21

nDNA dataset

Outlier loci were identified from a total of 10 068 loci (simulated full dataset; outlier and neutral loci), due to the sub-sampling method involved in creating a GenePop file. Bayescan analyses (Foll & Gaggiotti, 2008) identified a single outlier locus (Figure 4). Empirical outlier detection through PoPoolation2 (Kofler et al., 2011a) identified a total of 678 (6.73%) outlier loci, the highest of all three methodologies. A total of 179 (1.78%) outlier loci were identified by pcadapt (Luu et al., 2017). The single outlier identified through Bayescan analysis was also identified by both pcadapt and the empirical outlier approach. In addition, 24 outlier loci, across 20 contigs, were identified by both pcadapt and PoPoolation2 analyses (Figure 4).

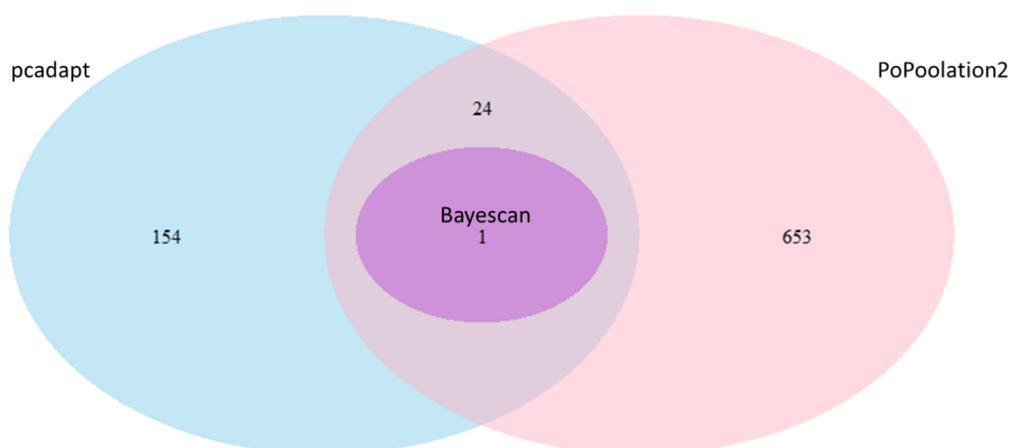


Figure 4: Venn diagram illustrating overlap of outlier loci detected by three methodologies for the nuclear dataset: pcadapt, PoPoolation2 and Bayescan 2.1.

The number, major allele frequency and combination of shared outliers varied across pools, ranging from 21 (EC) to 25 (NAM 1) outliers per pool (Table 8, Figure 5 and Supplementary Table S1).

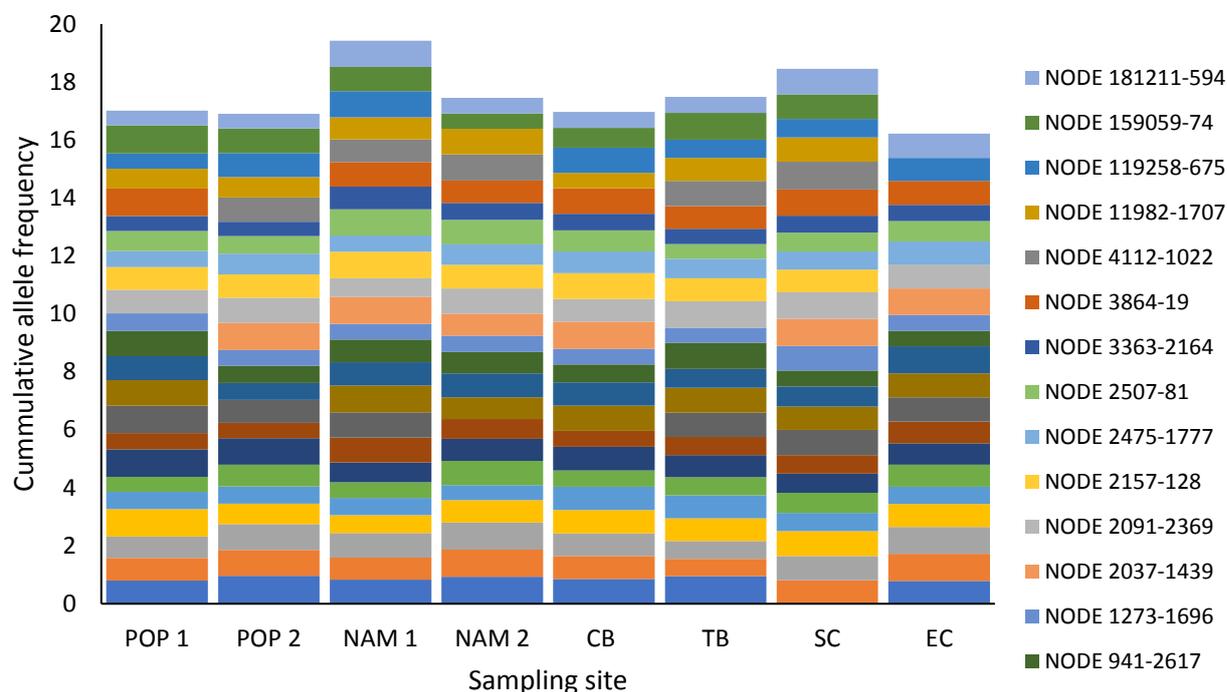


Figure 5: Frequency of candidate nuclear outlier loci (outlier loci identified by two or more outlier detection approaches) across sampling sites. Outliers identified by Node number and SNP position. Pool names as per Table 2.

Furthermore, BLASTX searches of the 20 identified contigs containing one or more candidate outlier loci, matched largely to hypothetical and uncharacterized proteins (Table 9). Several of the identified proteins were involved in cellular transport, regulation and functioning. For example, “Predicted: vacuolar protein sorting-associated protein 13B isoform X1 [*Larimichthys crocea*]” is largely associated with protein transport and sorting, while “Predicted: protein NYNRIN-like [*Nothobranchius furzeri*]” and “Predicted: nucleolar protein 8 isoform X3 [*Paralichthys olivaceus*]” are involved with nucleic acid binding (Anantharaman & Aravind, 2006; Gu et al., 2018). Furthermore, “GTPase IMAP family member 4-like [*Monopterus albus*]” is associated with GTP binding proteins, playing a role in signal transduction pathways for both hormones and neurotransmitters. A more specific function is suggested/predicted for “Predicted: von Willebrand factor A domain-containing protein 7-like [*Stegastes partitus*]”, which is a large protein involved in hemostasis (Carrillo et al., 2010), while “Predicted: myosin-IIIa-like [*Sinocyclocheilus rhinoceros*]” suggests a protein in the myosin superfamily, which has a potential role in photoreception, hearing and sensory

transduction (Montell & Rubin, 1988; Lin-Jones et al., 2004). More specifically, myosin proteins are cytoskeletal motors needed to generate cell structures and are found in the retinas of several fish species (Lin-Jones, 2009). The final neutral dataset contained 9 326 loci (91.74%) of the initial reads, whereas the final outlier dataset contained 832 loci (8.26%).

Table 9: Top BLASTX search results, with corresponding E-value scores and percentage identity, for nodes containing candidate outlier loci identified by two or more outlier detection approaches, for the nuclear dataset.

Node	BLASTX Results	E-Score	% Identity
NODE 2	Uncharacterized protein LOC111192030 [<i>Astyanax mexicanus</i>]	0.00	59
NODE 6	Predicted: protein NYNRIN-like [<i>Nothobranchius furzeri</i>]	0.00	69
NODE 72	No significant similarity	-	-
NODE 171	Predicted: uncharacterized protein LOC106911977 [<i>Poecilia mexicana</i>]	0.00	46
NODE 610	Uncharacterized protein K02A2.6-like [<i>Cynoglossus semilaevis</i>]	0.00	70
NODE 929	Predicted: nucleolar protein 8 isoform X3 [<i>Paralichthys olivaceus</i>]	0.01	62
NODE 941	Unnamed protein product, partial [<i>Oncorhynchus mykiss</i>]	0.00	28
NODE 1273	GTPase IMAP family member 4-like [<i>Monopterus albus</i>]	0.00	70
NODE 2037	Predicted: vacuolar protein sorting-associated protein 13B isoform X1 [<i>Larimichthys crocea</i>]	0.00	81
NODE 2091	Unnamed protein product, partial [<i>Tetraodon nigroviridis</i>]	0.00	97
NODE 2157	Hypothetical protein DD884_13045, partial [<i>Staphylococcus pseudintermedius</i>]	4.00	30
NODE 2475	Predicted: myosin-IIIa-like [<i>Sinocyclocheilus rhinoceros</i>]	0.00	72
NODE 2507	Predicted: von Willebrand factor A domain-containing protein 7-like [<i>Stegastes partitus</i>]	0.00	84
NODE 3364	No significant similarity	-	-
NODE 4112	Predicted: protein FAM160A1-like, partial [<i>Paralichthys olivaceus</i>]	0.00	52
NODE 11982	Predicted: tripartite motif-containing protein 7-like [<i>Lates calcarifer</i>]	0.00	56
NODE 119258	RNA-directed DNA polymerase from mobile element jockey [<i>Larimichthys crocea</i>]	0.00	91
NODE 159059	Capsule biosynthesis protein CapD [<i>Verrucomicrobia bacterium</i>]	8.40	40
NODE 181211	General transcription factor II-I repeat domain- containing protein 2A-like [<i>Camponotus floridanus</i>]	0.00	49

Additional/Exploratory outlier detection

Outlier loci were identified from a total of 41 369 loci. PoPoolation2 (Kofler et al., 2011a) identified 2 593 outlier loci (6.27%), while 519 outlier loci were identified by pcadapt (Luu et al., 2017). Of the top 1% of outlier loci, 17 outlier loci across 12 nodes were identified by both pcadapt and PoPoolation2 analyses. BLASTX searches of the 12 identified nodes mapped largely to predicted as well as hypothetical proteins, with four outlier loci found to have no significant similarities (Table 10). Successfully identified nodes mapped to genes or proteins involved in nucleic acid binding (“Predicted: zinc finger BED domain-containing protein 5-like [*Larimichthys crocea*]” and “Predicted: RNA-binding protein 45 [*Notothenia coriiceps*]”) as well as catalytic activity (“Predicted: RNA-binding protein 45 [*Notothenia coriiceps*]”).

Table 10: Top BLASTX search results, with corresponding E-value scores and percentage identity, for nodes containing outlier loci identified by pcadapt and PoPoolation2, based on 41 369 loci.

Node	BLASTX Results	E-Score	% Identity
NODE 11925	No significant similarity	-	-
NODE 12194	Uncharacterized protein LOC 112140660 [<i>Colletotrichum simmondsii</i>]	7.20	63
NODE 23553	Predicted: zinc finger BED domain-containing protein 5-like [<i>Larimichthys crocea</i>]	0.00	45
NODE 65959	Putative RNA-directed DNA polymerase from mobile element jockey-like [<i>Apostichopus japonicus</i>]		
NODE 119258	RNA-directed DNA polymerase from mobile element jockey [<i>Larimichthys crocea</i>]	0.00	91
NODE 146611	Predicted: RNA-binding protein 45 [<i>Notothenia coriiceps</i>]	0.029	76
NODE 162494	Predicted: uncharacterized protein LOC106930616 [<i>Poecilia Mexicana</i>]	0.00	44
NODE 208114	No significant similarity	-	-
NODE 427108	Hypothetical protein B5Z25 19610 [<i>Bacillus velezensis</i>]	0.00	86
NODE 436833	No significant similarity	-	-
NODE 436834	3.oxo-5-alpha-steroid 4-dehydrogenase [<i>Colletotrichum simmondsii</i>]	3.00	38
NODE 465029	No significant similarity	-	-

DISCUSSION

Advances in sequencing technologies have allowed for genome-wide studies to be conducted on a range of organisms, including non-model species such as Kingklip (Helyar et al., 2012; Hess et al., 2015; Carreras et al., 2017). More specifically, by genotyping hundreds to thousands of genome-wide polymorphisms, NGS has been shown to improve the ability to accurately estimate levels of genomic diversity and differentiation, whilst simultaneously providing insights into the potential influence of selection (Allendorf et al., 2010; Limborg et al., 2012; Carreras et al., 2017; Fischer et al., 2017). By employing a pooled RADseq approach, this chapter identified genome-wide molecular markers (SNPs) for southern African Kingklip, thereby contributing to improve the current lack of genome-wide based studies conducted on commercially important Southern Hemisphere marine species.

In order to distinguish between mtDNA and nDNA SNPS, I used the Atlantic cod (*G. morhua*) mitogenome to identify, extract and assemble all mtDNA reads, thus providing the first rough draft of the Kingklip mitogenome (mtDNA_ref). The Kingklip mitogenome was found to be smaller (total length = 9 110 bp) as compared to previously sequenced and available mitogenomes. For example, the complete mitogenome of the Atlantic cod (*G. morhua* – Karlsen et al., 2013), the East Siberian cod (*Arctogadus glacialis* – Coulson et al., 2006), the Pacific cod (*G. microcephalus* – Coulson et al., 2006), and the Polar cod (*Boreogadus saida* – Coulson et al., 2006) are found to be 16 696 bp, 14 875 bp, 15 511 bp and 15 564 bp in length, respectively, whilst a recent study by Li et al. (2018) found the mitogenomes of ten fish belonging to the sub-family Gobioninae to range between 16 605 bp to 16 617 bp in length (Coulson et al., 2006; Karlsen et al., 2013; Li et al., 2018). The smaller size of the Kingklip mitogenome may be as a result of the methods employed. Particularly, by identifying reads for *de novo* assembly based on the mapping of reads to an available reference genome as well as employing a RADseq approach, the resulting mtDNA_ref represents a partial mitogenome, thereby being smaller as compared to the complete mitogenome examples provided above.

The generated nuclear reference sequence (nDNA_ref) similarly provides a first draft of the Kingklip nuclear genome. In general, three different strategies can be applied for SNP discovery in non-model species such as Kingklip, namely WGS and assembly,

reduced genome complexity sequencing (RADseq & Reduced Representation Library; RRL) and cDNA sequencing (RNA-seq) methods (Helyar et al., 2012). Compared to recent studies by Gou et al. (2016) and Nielsen et al. (2018), which employed a similar pooled-RADseq approach, the *de novo* generated Kingklip nDNA_ref was found have a greater total length (>240 Mb) than that of the Granular limpet (*S. granularis*: 180Mb) and Cape urchin (*P. angulosus*: 200Mb), whilst variation in the mean (891 bp) as well as longest (7 223 bp) contig length was found when compared to that of the Atlantic herring (*C. harengus*; mean contig length = 316 bp, longest contig length = 16 434 bp). Additionally, the N50 of the Kingklip reference sequence (N50 = 918 bp) was found to be larger than the limpet (N50 = 717 bp) and urchin (N50 = 719 bp), whilst the accompanying L50 of the Kingklip nDNA_ref (L50 = 91 520 bp) was comparable (limpet L50 = 87 790 bp and urchin L50 = 94 187 bp). Although the observed differences may reflect biological variation in genomic sizes between these species, the most likely explanation for the observed variation in test statistics (total length and number of contigs) for the assembled Kingklip mitogenome and nuclear genome, as compared to one another as well as available genomes and *de novo* reference sequences, may result from the influence of the employed approaches, namely the different assemblers as well as the assembly parameters used (Zhang et al., 2011; Mastretta-Yanes et al., 2015; Smolka et al., 2015). As such, direct comparisons between species and studies must be carried out with caution and take into consideration not only biological differences between species, but also analytical processes from sampling regime to downstream genomic analyses.

As expected, more biallelic SNPs were identified for the nDNA than the mtDNA. This may be due to potential differences in genome size, as explained above, as well as the possible influence of the number of raw reads, reference contigs as well as mapped reads on the identified number of SNPs (Shafer et al., 2017; Nielsen et al., 2018). While the percentage of mapped reads were found to be relatively similar between the two datasets, the increased size and contig number of the nDNA_ref may have resulted in the increased number of identified SNPs. Furthermore, evidence for a relationship between the number of mapped reads and identified SNPs has previously been reported (Guo et al., 2015; Nielsen et al., 2018), with the number of mapped reads found to be positively correlated to the number of identified SNPs for the Three-spined stickleback (*G. aculeatus* - Guo et al., 2015), Cape urchin (*P.*

angulosus) and Granular limpet (*S. granularis*; Nielsen et al., 2018). Accordingly, the increased number of mapped nuclear reads, as compared to mitochondrial, may explain observed SNP differences.

Furthermore, observed differences in the number of SNPs may be a result of the methods employed in generating the mtDNA genome and dataset. While the effectiveness of employing an available reference genome for the identification and subsequent assembly of mitochondrial reads has been widely demonstrated, this approach requires strict mapping parameters (Hahn et al., 2013; Al-Nakeeb et al., 2017). As a result, the stringent mapping approach followed by this study, with only properly-paired, uniquely mapped reads being extracted and subsequently assembled, may have resulted in the potential loss of highly variable regions. Furthermore, by mapping to the genome of a distantly related species, reads are most likely found to map to highly conserved regions, potentially explaining the lack of polymorphisms (SNPs) observed for the mitochondrial dataset (Hahn et al., 2013).

Genome-wide levels of diversity

With RADseq targeting a random sample of the genome, including regulatory, coding and non-coding regions, estimates of genome-wide SNP diversity are argued to reflect more accurately genomic levels of diversity, whilst additionally providing information regarding potentially adaptive and functionally important variation (Catchen et al., 2017; Fischer et al., 2017). Based on the identified SNPs, genome-wide levels of diversity were characterised for both the nuclear and mitochondrial datasets. Although levels of diversity were largely comparable between sampling sites/pools, within each dataset respectively, observed levels of nucleotide diversity (Tajima's π) were considerably lower for the mitochondrial as compared to nuclear dataset (average mtDNA $\pi = 0.0023$ & nDNA $\pi = 0.0173$). Two hypotheses may explain the observed discordance. Firstly, reduced nucleotide diversity may be a result of the lower number of mtDNA versus nDNA SNPs, as previously discussed, with a reduced number of SNPs potentially resulting in reduced diversity levels. Alternatively, mtDNA diversity estimates may not reflect patterns of nDNA diversity as a result of biological differences between the two markers, namely mutation rates, patterns of inheritance,

N_e and “time-scale” (historical versus contemporary) (Zhan et al., 2003; Ballard & Whitlock, 2004; Bensch et al., 2006).

Average genome-wide SNP nucleotide diversity values (Tajima's π) were different when compared to previous values obtained for other Pool-Seq based marine studies, with associated π diversity reported at $\pi = 0.003$ for Three-spined sticklebacks (*G. aculeatus* – Guo et al., 2015), $\pi = 0.006$ to $\pi = 0.009$ for the Atlantic herring (*C. harengus* – Gou et al., 2016) and $\pi = 0.25$ to $\pi = 0.37$ for the Prickly sculpin (*C. asper* – Dennenmoser et al., 2017). Similarly, the observed average π values differed from those previously reported by studies employing individual-level RADseq and/or WGS approaches: e.g. $\pi = 0.0053$ for the European eel (*Anguilla Anguilla* – Pujolar et al., 2013), $\pi = 0.0300$ for the white-beaked dolphin (*Lagenorhynchus albirostris* – Fernández et al., 2016), $\pi = 0.0492$ for the white-sided dolphin (*L. acutus* - Fernández et al., 2016), and $\pi = 0.32$ for the Atlantic herring (*C. harengus* - Barrio et al., 2016). This observed variation may be as a result of differences in the respective bioinformatic pipelines employed by each study (Shafer et al., 2017), as well as species differences. Specifically, variation in filtering, assembly and mapping parameters may affect both the number and diversity of identified SNPs, subsequently influencing observed levels of diversity (Mullins, 2017; Shafer et al., 2017). Furthermore, none of the previous studies employed the ezRAD (Toonen et al., 2013) sequencing approach. As a result, observed variation may relate to the regions in which the respective restriction enzymes splice the genome (Davey et al., 2013). Therefore, with genetic summary statistics and levels of diversity found to be influenced by the bioinformatic pipeline employed, direct comparisons should not be made across studies. Similarly, with several evolutionary and demographic factors/features, such as population size, gene flow, reproductive strategy as well as mutation and migration rate, influencing observed genetic/genomic diversity direct comparisons between species must be made with caution (Gaggiotti et al., 2009; Pinsky & Palumbi, 2014; Henriques et al., 2017).

Compared to the previous estimates of Kingklip nucleotide diversity (Henriques et al., 2017 - based on a fragment of the mtDNA Control Region), the average genome-wide nucleotide diversity based on the nDNA dataset was found to be greater, whilst the mtDNA diversity estimates were once again lower by comparison. Such variation in diversity estimates based on traditional mtDNA versus SNP markers has been found

by previous studies (Fernández et al., 2016; Nielsen et al., 2018). Broadly, the observed discordance may be due to previous estimates being based on the hyper-variable CR, under the general assumption that variation within/of that region represents the entire genome. This assumption may however be violated, with loci often found to display varying patterns of diversity across the genome (Ballard et al., 2002; Ballard & Whitlock, 2004; Bensch et al., 2006; Guo et al., 2015). More specifically, the observed discordance between the previous mtDNA CR estimates and those based on the nDNA dataset may be as a result of incongruence between the two marker types (mtDNA versus nDNA), as previously discussed (Zhan et al., 2003; Ballard & Whitlock, 2004; Bensch et al., 2006). The lower nucleotide diversity levels reported for the mtDNA database, as compared to those based on the mtDNA CR, may be due to the methods employed in identifying the mtDNA SNPs, with the mapping of reads to a distantly related species as well as stringent filtering parameters potentially leading to a loss of highly variable regions and thereby associated lower diversity levels.

Overall, Kingklip genomic diversity was found to be largely similar across its entire distribution with no specific regions found to display significantly increased or decreased levels of genomic diversity. It is noted however that each sampling sites/pool has a certain level of uniqueness with private, biallelic SNPs found across all pools within the nuclear dataset. In addition, the generally negative Tajima D values suggested the occurrence of demographic fluctuation, pointing to a recent population expansion (Tajima, 1989; Fischer et al., 2017; Henriques et al., 2017). This is in line with the previous findings of Henriques et al. (2017), where significantly negative Tajima D and Fu's F_s values were reported for South African Kingklip across all sampling regions and years. Such evidence for demographic change (population expansion or bottleneck) is generally associated with environmental change(s) (Ruzzante et al., 2008; Henriques et al., 2014, 2017). In the case of Kingklip, past exploitation, linked to direct longline fishing, resulted in considerable declines in Kingklip population abundance and spawning biomass (Punt & Japp, 1994). As such, the observed negative Tajima D values may reflect the recorded recovery of Kingklip abundance, observed along the West and East coast of South Africa, following the closure of direct longline fisheries (Brandão & Butterworth, 2013).

Detection of putative outlier loci

The detection of highly differentiated loci (referred to for ease as 'outlier loci') across the Kingklip genome provides evidence for the possible influence of selection in different populations. While acknowledging the potential impact of genetic drift in driving and maintaining patterns of outlier loci, the effects of genetic drift are argued to be negligible as a result of large N_e that generally characterise marine species, including Kingklip (Tigano & Friesen, 2016; Dennenmoser et al., 2017). As such, these findings are believed to support the view, and provide further evidence that, local adaptive diversity and selection may be more prevalent within high gene flow marine environments than previously believed (Hauser & Carvalho, 2008; Bradbury et al., 2012; Benestan et al., 2015; Guo et al., 2015; Dennenmoser et al., 2017). Future selective experiments are required to confirm the possible influence of selection in driving the occurrence of these outlier loci, although carrying out such experiments for deep-water species will not be without difficulties.

Overall, the inclusion of outlier loci has been found to increase the resolution of population sub-structuring in several high gene flow species (Freamo et al., 2011; Helyar et al., 2011; Bradbury et al., 2012; Nielsen et al., 2012; Benestan et al., 2015). Therefore, the identification and subsequent isolation of outlier loci may prove highly valuable within the context of this study, providing valuable information regarding potential selective forces and adaptive divergence within the system.

The joint use of three separate outlier detection methods, each employing a different analytical approach, provides increased confidence that a wide range of potential outlier loci were identified. However, as a result of the different analytical approaches employed and the discordance between methods, different outliers were identified using the different detection methods. The total percentage of surveyed SNPs identified as outliers, 11% and 8% for the mtDNA and nDNA datasets respectively, are largely consistent with patterns of past studies, where between 5-10% of the genome was considered to be under selection (Nosil et al. 2009; Galindo et al., 2010; Bradbury et al., 2012; Strasburg et al., 2012; Guo et al., 2015, 2016). Candidate outlier loci, identified by more than one method, provide insights into the possible selective pressures acting within Kingklip populations. Interestingly, candidate nDNA outlier loci were shared between the majority of pools, with no private outliers identified, despite

considerable environmental variation across sampled sites. This shared adaptive variation may result from high levels of gene flow across the system, as previously reported for Kingklip (Grant & Leslie, 2005; Henriques et al., 2017). Furthermore, observed patterns of shared putative adaptive variation may indicate the influence of common selective pressures experienced by Kingklip, across both space and time. Similar evidence for high levels of shared adaptive variation between sites has been reported for several other species, including the Atlantic Salmon (*Salmon salar* – Freamo et al., 2011), southern Africa seagrass (*Zostera capensis* – Phair et al., in review) as well as Atlantic cod (*G. morhua* – Nielsen et al., 2009a). Additionally, observed shared adaptive variation may be due to the reduced genomic cover obtained through the RADseq approach employed within this study, with the lack of genome-wide coverage potentially resulting in several loci under selection being missed (Lowry et al., 2017).

Shared outlier loci are often employed for assignment analyses, as differences in allele frequencies can be used to differentiate regions, stocks, populations and/or ecotypes (Nielsen et al., 2009a; Freamo et al., 2011; Jones et al., 2012b; Phair et al., in review). Such an example is provided in the case of the Atlantic salmon, where individuals were accurately assigned (85%) to their metapopulation of origin based on differences in outlier allele frequencies (Freamo et al., 2011). Therefore, the frequency of the identified outlier loci were used in Chapter 2 to test their ability to identify fine-scale patterns of population sub-structuring in southern African Kingklip.

With regards to the potential functional roles of the identified outlier loci, the majority of candidate nDNA outliers were found to map to genes and/or proteins involved in signal transduction, nucleic-acid binding and catalytic activity. Additionally, certain outlier loci were found to be within genomic regions related to genes and/or proteins involved in vision (myosin IIIA protein), haemostasis (von Willebrand Factor; vWF protein) and adipogenesis (VPS13 gene). The role and/or occurrence of these genes and proteins, as well as associated processes/functions, has previously been investigated within fish. For example, the myosin IIIA protein is proposed to play a functional role in calycal processes (Montell & Rubin, 1988; Lin-Jones et al., 2004). Specifically, a previous study by Lin-Jones et al. (2009) investigating the expression and localization of myosins within fish retina, identified myosin IIIA within the retina of several fish species thereby proposing its potential role in fish vision. Likewise, studies

on Zebrafish (*Dania rerio*) have identified the vWF protein as being involved in primary hemostasis, a physiological process involved in the first stage of wound healing (Carrillo et al., 2010; Gale, 2011), while the VPS13B gene has been shown to be involved in adipogenesis, influencing the storage and distribution of fats within the body (Limoge et al., 2015). In particular, with regards to fish, adipogenesis is found to be associated with adipose tissue (AT), a main form of lipid storage with a functional role in energy homeostasis (Salmerón, 2018). The identification of outlier loci found to be associated with these genes and proteins may therefore indicate potential selective forces acting on different physiological processes in southern African Kingklip.

Taking into consideration the large number of SNPs needed to identify outlier loci via genome scans, as highlighted by the high percentage of neutral loci, increasing the number of initial SNPs used for outlier detection may identify additional, putative outliers previously overlooked. Interestingly while resulting in an increase in the number of detected outlier loci, as expected, outlier loci identified based on the full set of biallelic nDNA SNPs (exploratory/additional outlier detection) were once again associated with genes and/or proteins involved in nucleic-acid binding, molecular pathways and catalytic activity, thereby providing little additional insight into the potential selective forces/pressures experienced by Kingklip. With the initial SNPs and approaches employed for outlier detection found to influence results, researchers are constantly faced with a trade-off between including false positives (as a result of too lenient parameters) or the loss of rare alleles (Helyar et al., 2011; Nielsen et al., 2012; Rodríguez-Ezpeleta et al., 2016). As such, only the initial outliers identified from the simulated dataset were used in all further analyses. This was done so as to ensure continuity and comparability across and between all three outlier detection approaches, thereby increasing statistical confidence in the obtained findings. Furthermore, by identifying a subset of biallelic SNPs based on more stringent filtering parameters, the potential influence of false positives was reduced.

In contrast to the candidate nDNA outlier SNPs, the two mtDNA outliers identified were not shared across pools/sampling sites. Both outliers were found for the Namibain sampling sites (NAM 1 and NAM 2), whilst neither were found for EC and/or CB. It is noted, however, that considerations regarding these two outliers must be made with caution as a result of them only being identified by one approach. Nevertheless, the identified candidate outliers were found to BLAST to NADH dehydrogenase subunit 1

and an uncharacterised protein. Broadly, while a function cannot be assigned to the uncharacterised protein, NADH dehydrogenase is found to be involved in energy metabolism, with the various subunits encoded by the mtDNA genome (Galindo et al., 2010; Porcelli et al., 2015). Similar evidence for the potential selection of NADH dehydrogenase has been reported for the marine gastropod *Littorina saxatilis*, with variation in energetic metabolism proposed to result from environmental factors and stressors, relating to wave action, anoxia and temperature (Galindo et al., 2010). Moreover, additional studies regarding marine molluscs have found potential evidence for the selection of NADH dehydrogenase genes in response to temperature variation (Gleason & Burton, 2016; Pante et al., 2013). While direct deductions regarding the potential selective pressures acting on the identified outlier loci cannot be made, observed selection of NADH dehydrogenase subunit 1 suggests the likely influence of environmental variation, between sites, in driving the observed patterns of selection.

Ultimately, while genome scans prove highly valuable in identifying loci that may be under selection (Nielsen et al., 2009a; Seeb et al., 2011; Limborg et al., 2012; Milano et al., 2014), demonstrating the adaptive significance and function of these loci requires selection experiments and functional tests, a considerable challenge within marine systems. Moreover, evaluations of RADseq-based genome scan studies and/or approaches have been found to miss potential loci under selection, owing to the reduced genome complexity (Lowry et al., 2017). Accordingly, it is acknowledged that genomic regions potentially under selection may have been altogether missed within this thesis. As such, the outlier detection results presently included should be considered as an exploratory exercise, providing preliminary insights into the possibility of local adaptation in Kingklip. However, given the number of loci as well as multiple analytical approaches employed, there is strong evidence for differential selective pressures acting on Kingklip, with the identified outliers also providing increased discriminatory power for further population structure analyses. Future studies incorporating candidate gene and functional genetic approaches, as well as whole genome coverage, may aid to identify additional outlier loci (Lowry et al., 2017).

Methodological considerations

With several genomic techniques currently available, each with a separate set of advantages and limitations, the chosen genomic approach is largely influenced by the study question, trade-offs of each approach as well as available resources and funding. As this Chapter aimed to produce a genome-wide marker set to investigate genomic diversity and population sub-structuring of southern African Kingklip, the use of a pooled RADseq approach was highly advantageous and successful. While it is noted that only a fraction of the genome was sequenced as a result of the RADseq approach, information was gained from throughout the genome, including coding, non-coding and regulatory regions (Catchen et al., 2017).

Additionally, SNP based analyses are prone to ascertainment bias, where deviation in allele frequencies from the expected distribution can occur due to the sampling and SNP discovery strategies employed. Ascertainment bias may result in a bias in nucleotide diversity estimates (increase or decrease in levels of polymorphism), while similarly influencing population structure inferences (Akey et al., 2003; Helyar et al., 2011; Lachance & Tishkoff, 2013). However, by employing a RADseq approach, which has been shown to reduce the influence of ascertainment bias, as well as stringent filtering and large pool sizes, the potential influence of ascertainment bias was greatly reduced. The Pool-Seq approach has been shown to provide more robust and consistent allele frequency estimates as compared to individual sequencing, further supporting the accuracy of the population-based estimates of genetic diversity (Futschik & Schlötterer et al., 2010; Guo et al., 2015; Anand et al., 2016).

Therefore, on the basis of the above arguments, the genome-wide datasets produced in this chapter, both nuclear and mitochondrial, are reasoned to be both robust and reliable. Furthermore, these datasets represent a valuable first step towards the development and sequencing of a complete Kingklip reference genome, with continuous improvements and advances in sequencing technology holding great promise for the future.

CHAPTER 2: Genetic sub-structuring of southern African Kingklip within and between South Africa and Namibia

INTRODUCTION

Kingklip distribution and genetic population sub-structuring

The genetic structure of a species is shaped by a complex interplay of historical and contemporary environmental factors, demographic change dynamics and physiological traits (Selkoe et al., 2008; Gaggiotti et al., 2009; White et al., 2010a; Henriques et al., 2012; 2016). Marine species tend to demonstrate low levels of genetic population differentiation, as compared to freshwater or terrestrial species, most likely as a result of their historically high N_e , high dispersal abilities and lack of effective dispersal barriers within marine systems (Ward et al., 1994; Hauser & Carvalho, 2008; Reiss et al., 2009; Henriques et al., 2012). However, the view that marine species exhibit low or no population structure has increasingly been disputed, as numerous studies have provided evidence for genetically structured populations and spatially discrete units across the marine tree of life (Hauser & Carvalho, 2008; Reiss et al., 2009; Vendrami et al., 2017). Although species population structuring, as well as stock boundaries, are most often spatially defined, they may have an important temporal component as well (Banks et al., 2007). Importantly, discrete stocks may react differently to harvesting and exploitation pressures (Pawson & Jennings, 1996; Hauser & Carvalho, 2008; Reiss et al., 2009; Ovenden et al., 2015; Spies et al., 2015), which makes understanding stock structure in relation to effective fisheries management an important requirement (Carvalho & Hauser, 1994; Ovenden et al., 2015; Spies et al., 2015).

Gene flow and dispersal within marine systems is mostly influenced by underlying seascape and environmental features (Banks et al., 2007, Gaggiotti et al., 2009; Saha et al., 2015; Selkoe et al., 2010, 2016), with an increasing number of seascape studies relating population genetic structure to oceanographic and environmental features, such as continuity of habitat and oceanographic features (D'Aloia et al., 2014; Johansson et al., 2015), bathymetry (Glazier & Etter, 2014; Saenz-Agudelo et al., 2015), currents (Banks et al., 2007; White et al. 2010a; Dambach et al., 2015; Johansson et al., 2015), and SSTs (Banks et al., 2007; Eberl et al., 2013), in addition

to numerous other factors (Liggins et al. 2016; Riginos et al. 2016; Selkoe et al. 2016). Such abiotic features can either act to promote dispersal and gene flow, or serve as effective dispersal barriers, thereby influencing species population and stock structure (Selkoe et al., 2010; Benestan et al., 2015). In particular, coastal oceanographic features such as upwelling cells, currents, riverine outflows and coastal geography, have been identified to potentially influence the genetic structure of inshore and coastal species (Banks et al., 2007; von der Heyden et al., 2007; Henriques et al., 2016; Selkoe et al., 2016). Population genetic structuring and gene flow of deep-water benthic species, such as Kingklip, may therefore be influenced by deep water features such as bathymetric gradients and hypoxic conditions at depth (White et al., 2010b; Glazier & Etter, 2014; Henriques et al., 2016). Bathymetric barriers (e.g. deep ocean basins and depth gradients) have been shown to influence the genetic structure of marine species, as reported for the deep-water Fish tusk (*Brosme brosme* – Knutsen et al., 2009) and the Roundnose grenadier (*Coryphaenoides rupestris* – Gaither et al., 2018). The overall influence of abiotic and environmental barriers on the genetic sub-structuring of marine species depends greatly on the life history traits of the species, with features such as Pelagic Larval Dispersal (PLD), larval duration, adult size as well as mobility having been hypothesized to influence the extent of environmental barriers on dispersal (Bradbury et al., 2008; Pascual et al., 2017). In addition, spawning and feeding aggregations can also affect the signal of population genetic dynamics. Equally, biotic features/processes relating to homing (Lundy et al., 2000) and migration (Gaggiotti et al., 2009; Saha et al., 2015) behaviour, as well as the presence of multiple spawning grounds and/or periods (Henriques et al., 2012, 2015, 2017), have been shown to influence the genetic sub-structure of both deep water and coastal species. Therefore, a growing body of evidence suggests that the genetic structuring of species is influenced by both abiotic and biotic features.

Kingklip is endemic to the southern African coastline with a geographic distribution spanning from the north of Walvis Bay, off central Namibia, to Algoa Bay, off the South-east coast of South Africa (Figure 2). This region has unique oceanographic features, with the warm Angola Current to the north, the warm Agulhas Current on the South coast and the cold, nutrient rich Benguela Current off the West coast, creating a dynamic oceanographic system (Figure 1; Hutchings et al., 2009). While the cold northward moving and warm southward flowing currents influence the shallower

regions of the coastline, deep water movement (>350 meters deep) is mainly poleward within the Benguela region (Shillington et al., 2006).

Through their influence on larval supply and recruitment, prevailing ocean currents are believed to play an important role in shaping the observed genetic structure of several marine species (Banks et al., 2007; White et al. 2010a; Coscia et al., 2013; Benestan et al., 2015; Johansson et al., 2015). As such, the prevailing currents in the Benguela system may similarly act to mediate the movement of larvae and eggs, thereby potentially influencing the genetic sub-structure of Kingklip populations (Grant & Leslie, 2005; Henriques et al., 2017). In a previous study investigating the possible role of drift routes on the transport of Cape hake (*M. capensis* and *M. paradoxus*) larvae and eggs, the coastal jet flowing along the West coast of southern Africa was identified as the principal transport mechanism (Stenevik et al., 2008). More specifically, the coastal Agulhas and Benguela Currents, as well as inshore counter currents, have been hypothesized to facilitate the passive movement of Kingklip larvae and eggs, potentially influencing patterns of Kingklip connectivity along its range (Grant & Leslie, 2005).

Several studies have investigated regional signals of population genetic structure and gene flow, specifically within the context of oceanographic features of the Benguela Current. One major barrier appears to be the Lüderitz upwelling cell located off central Namibia. This oceanographic feature has been suggested to limit gene flow for several fish species, including sardine *S. sagax* (Beckley & van der Lingen, 1999), Leervis (*L. amia* – Henriques et al., 2012), Silver kob (*A. inodorus* – Henriques et al., 2015; Mirimin et al., 2016), Geelbek (*A. aequidens* – Henriques et al., 2014), as well as Kingklip (Shannon et al., 1985). Additionally, the Agulhas-Benguela transition zone has been described to influence population sub-structuring for a number of marine species (reviewed Teske et al., 2011), with the central Agulhas bank found to separate western and eastern sardine (*S. sagax*) populations (Miller et al., 2006). Contrasting results have, however, been reported for the Bluefish (*P. saltatrix* - Reid et al., 2016) and Silver kob (*A. inodurus* - Mirimin et al., 2016), with a single genetic stock found to occur across this boundary, thus providing evidence for the potential transport of larvae and/or movement of adults across this biogeographic barrier (Reid et al., 2016).

Furthermore, variation in bathymetry, upwelling patterns and oxygen availability across the Benguela region creates further environmental and seascape heterogeneity (Shillington et al., 2006; Hutchings et al., 2009; Henriques et al., 2016; Selkoe et al., 2016). For example, insight into the potential impact of environmental heterogeneity on population structure across the Benguela region was provided for the Shallow-water hake, *M. capensis* (Henriques et al., 2016). By combining seascape and genetic analyses of microsatellites, an association was found between observed population differentiation between the northern and southern Benguela regions and local oceanographic and environmental features, specifically bathymetry, SST and upwelling events, suggesting that observed differentiation may result from local adaptation to environmental conditions. Importantly, the location of genetic differentiation was found to vary between years, with environmental stochasticity related to Low Oxygen Water (LOW) events influencing the spatial population structure of this species (Henriques et al. 2016).

With the increasing use of genome-wide marker sets, such as SNPs, additional evidence for local adaptation and adaptive divergence has been presented for numerous marine species (Nielsen et al., 2009a; White et al., 2010b; Milano et al., 2014). Specifically, signals of local adaptation and adaptive divergence have been demonstrated for European hake (*M. merluccius*), with outlier loci revealing increased levels of divergence and fine-scale population structure as compared to neutral-based estimates. These outlier loci were found to be correlated with SST and salinity, thereby supporting the hypothesis of adaptation to local conditions (Milano et al., 2014). Several additional studies have detected similar trends, with the inclusion of outlier loci revealing adaptive divergence and population sub-structuring, previously missed when based on neutral markers alone, illustrating the potential influence of selection on observed population sub-structuring (Nielsen et al., 2009a; White et al., 2010b; Limborg et al., 2012).

Chapter aims and objectives

The stock structure of southern African Kingklip is currently poorly understood and hindered by the lack of agreement between available studies. Furthermore, a lack of knowledge persists regarding the distribution of spatial genetic variation within, and

between, Namibia and South Africa, as previous genetic studies did not include Namibian samples. This study represents the first attempt to plug this knowledge gap, by incorporating samples collected along the Namibian coastline, as well as along the South African coastline. From a regional perspective, Kingklip resources are currently managed separately between the two countries. However, if a panmictic population is detected, joint management may be advisable (von der Heyden et al., 2007), as is currently being debated for *M. paradoxus* after the work of Henriques et al. (2016).

In light of the lack of agreement regarding Kingklip stock structure, and the differences in abundance yield between West and East coast South African stocks (Brandão & Butterworth, 2013), it is necessary to have a more informative and comprehensive approach to stock identification. The aim of this chapter is thus to identify the level of population genetic structure of Kingklip along the southern African coastline, spanning Namibia to the Eastern Cape of South Africa. To do this, I utilised the SNP panels (both neutral and outlier loci) generated in Chapter 1. I hypothesize that three sub-populations/stocks will be identified, consisting of one in Namibia and two within South African waters (Western Cape; TB, CB, SC and Eastern Cape; EC).

METHODS

Population sub-structuring analyses were performed on the simulated nDNA datasets: “full” (neutral and outlier loci), “neutral” (neutral loci only) and “outlier” (outlier loci only), with biallelic SNPs having a minimum allele count of 4 and coverage of 28 to 100 reads, as well as the complete nDNA dataset (>40K SNPs), with biallelic SNPs having a minimum allele count of 4 and coverage of 20 to 500 reads (as outlined in Chapter 1).

Genome-wide population differentiation: fixation index

In order to investigate genome-wide patterns of differentiation and population sub-structuring, pairwise differentiation, measured by Weir and Cockerham’s fixation index (F_{ST}) (1984), was estimated between all pools using the ‘diffCalc’ command of the R package *diveR*sity (Keenan et al., 2013). F_{ST} values were estimated for all the simulated, nDNA datasets (full, neutral and outlier) produced in Chapter 1, with 95%

Confidence intervals (95 % CI) calculated based on 1 000 bootstrap replicates to assess statistical significance. Significant pairwise F_{ST} values were considered to represent discrete genetic stocks. In addition, pairwise F_{ST} values (*sensus* Hartl & Clark, 2007) were estimated using the 'fst-sliding.pl' command of PoPoolation2 (Kofler et al., 2011a). Unlike *diveRsity* (Keenan et al., 2013), PoPoolation2 does not require input files to be in Genepop file format. As such pairwise F_{ST} values were estimated between all pools for the complete, biallelic dataset (>40K SNPs), identified in Chapter 1, so to allow for a comparison between estimates based on the complete versus simulated datasets.

Genetic versus genomic patterns of differentiation

Henriques et al. (2017) proposed the occurrence of two sub-populations off the coast of South Africa, based on microsatellite sequencing data. As a result, samples from the two proposed sub-populations were pooled and sequenced (P1 and P2 respectively), as detailed in Chapter 1. In order to determine if the observed clustering found by Henriques et al. (2017) could still be identified using the SNP nDNA dataset, pairwise comparisons as well as multivariate analyses were performed for the P1 and P2 pools, as well as the 2017 South African pools. In addition, in order to assess the possibility of a marker-effect, I mapped the previously employed microsatellite primer sequences to the *de novo* nDNA reference (nDNA_ref).

Pop 1 versus Pop 2 differentiation

Fixation index (F_{ST}) for the pairwise comparison of P1 and P2 was estimated for all simulated nDNA datasets (full, neutral and outlier) using the 'diffCalc' command of *diveRsity* (Keenan et al., 2013), with statistical significance estimated based on 95% CI calculated following 1 000 bootstrap replicates, as well as for the complete dataset (>40K SNPs) using the 'fst-sliding.pl' command of PoPoolation2 (Kofler et al., 2011a), as outlined above.

In order to investigate the relative contribution of SNP loci to the total differentiation observed between P1 and P2, a Manhattan plot was generated in R using the 'manhattan' function available in the qqman package (Turner, 2018). A Manhattan plot

is a specific form of scatterplot largely used for Genome Wide Association Studies (GWAS) to visualize the relationship of loci to a given variable, as well as to better understand the distribution of loci across chromosomes or genomes (Turner, 2018). For the purpose of this analysis, I plotted per locus F_{ST} values estimated based on the full simulated dataset (10 068 SNPs) in *diveRsim* (Keenan et al., 2013), as well as per locus F_{ST} values estimated based on the complete dataset (>40 K SNPs) in *PoPoolation2* (Kofler et al., 2011a).

In addition, to visualize the relationship between the two proposed groupings, a Principal Component Analysis (PCA) based on SNP allele frequencies per pool was generated in R using the 'prcomp' function (R Core Development Team 2008). PCAs do not require prior information regarding Hardy-Weinberg Equilibrium (HWE) or Linkage disequilibrium (LD), and are thus suitable for pooled data sets, providing a simplified picture of the genetic relationship between pools/sampling regions and potentially identifying any form of weak structuring within the dataset (Jombart, 2008). The PCA was performed by calculating allele frequencies per pool for the complete (>40K SNPs) and the full simulated (10 068 SNPs) datasets. In both cases, PCAs were based on allele counts estimated from sync files using the 'snp-freq-diff.pl' command of *PoPoolation2* (Kofler et al., 2011a). Allele frequency data for each SNP loci was then transformed into Principal Components (PCs) along which variation was maximized.

Mapping of microsatellite primer sequences to reference genome

To further investigate the potential relationship between the microsatellite and SNP datasets, sequences for the ten microsatellite loci developed by Ward and Reilly (2001) and employed by Henriques et al. (2017) (Supplementary Table S2), were mapped to the *de novo* nDNA_ref sequence generated in Chapter 1. This was done to determine to what extent the previously employed microsatellite sequences are represented in the RADseq reference sequence and, consequently, SNP dataset.

Primer sequences were first converted into FASTA file format. The resulting files, two per primer (forward and reverse sequences), were then mapped to the *de novo* reference (nDNA_ref) in Bowtie2 (Langmead & Salzberg, 2012). Bowtie2 is a memory efficient tool which employs full-text minute indexing to achieve sensitive and accurate

alignment and mapping across a range of read lengths (Langmead & Salzberg, 2012). Mapping was done using default parameters and end-to-end alignment. Due to the primers being developed for the sister-species *G. blacodes* (Ward & Reilly, 2001), a mismatch rate of 0.05% was set, based on previously reported interspecific genetic distances (Daley et al., 2000; Smith & Pauline, 2003; Santaclara et al., 2014). Primer sequences found to map to the reference genome were subsequently viewed in the Integrative Genomics Viewer (IGV; Robinson et al., 2011).

South African population sub-structuring

Population sub-structuring and genome-wide differentiation

In order to determine how the previously proposed South African sub-populations (P1 and P2) relate to the 2017 South African sampling sites (CB, TB, SC & EC), as well as to investigate population sub-structuring within the 2017 samples, genome-wide pairwise differentiation was estimated for the simulated datasets (full, neutral and outlier), as well as the complete dataset, as outlined above.

Population clustering was investigated using the software Bayesian Analysis of Population Structure (BAPS; Corander & Marttinen, 2006). Clustering analyses were performed on both the simulated full and outlier datasets with $K = 1 - 10$ tested (K being the most likely number of clusters). BAPS employs Bayesian mixture models in order to cluster individuals into genetically or genomically divergent groups, to best explain underlying population structure (Corander & Marttinen, 2006). The mixture analysis 'Clustering of groups of individuals' was performed in BAPS (Corander & Marttinen, 2006), with the resulting binary files subsequently used for admixture analyses, to evaluate the statistical significance of clusters ($p < 0.05$). Admixture coefficients were estimated after 100 iterations.

Furthermore, patterns of differentiation were visualised using a PCA based on SNP allele frequencies per pool, generated using the 'prcomp' function in R (R Core Development Team 2008). Allele frequencies were calculated per pool, for the simulated full dataset (10 068 SNPs) as well as the complete dataset (>40K SNPs), based on allele counts estimated from sync files using the 'snp-freq-diff.pl' command of PoPoolation2 (Kofler et al., 2011a).

Population sub-structuring and genome-wide differentiation – top 500 loci

The possible influence of the number and the methods for calling SNPs on the patterns of genetic differentiation observed between P1, P2 and the 2017 South Africa sampling sites was investigated. In order to do so, the most informative loci contributing to the differentiation of P1 and P2 were identified. The contributions, also referred to as loadings, of variables (loci) to the principal components can be used to identify the most informative loci; those that contribute to the greatest proportion of variance between pools (Kassambara, 2017; Mullins, 2017). Based on the PCA of P1 versus P2 generated for the simulated full dataset, the contribution of each loci to the differentiation of P1 versus P2 was determined using the 'get_pca_var', 'var\$contrib' and 'fviz_contrib' commands of the FactoExtra 1.0.5 package (Kassambara & Mundt, 2017), available in R. Relative contributions were then ranked to identify the top 500 loci (+/- 5%) contributing to the differentiation of the two pools, and the loci were extracted from the simulated full Genepop file, produced in Chapter 1, resulting in a "top 500" dataset for P1, P2 and the 2017 South African sampling sites/pools. Population sub-structuring analyses were subsequently performed on these top 500 loci to compare results to the full dataset. This dataset was then used to estimate pairwise genetic differentiation among 2017 South African pools, P1 and P2 in *diveRsity* (Keenan et al., 2013), with 95% CI calculated after 1 000 bootstrap replicates, and to generate a PCA based on the SNP allele frequencies per pool, using the 'prcomp' function in R (R Core Development Team 2008).

Outlier detection was performed on the top 500 loci for all South African pools (P1, P2, CB, TB, SC and EC), using the same three methodologies described in Chapter 1: Bayescan 2.1 (Foll & Gaggiotti, 2008), empirical (PoPoolation2; Kofler et al., 2011a) and *pcadapt* (Luu et al., 2017), in order to investigate the potential role of local adaptation on the observed patterns of genetic differentiation. The top 500 loci file was edited and converted into the Bayescan format in PGDSpider2 v2.1.03 (Lischer & Excoffier, 2012). BayeScan analyses were performed using a total of 20 pilot runs of 5 000 iterations each and a burn-in period of 50 000 reversible jump chains with a thinning interval of 10 steps. Prior odds were set to 10, as this is the most suitable approach for a few hundred loci, with a target FDR of 0.05 used to identify candidate outlier loci. Empirical outlier detection in PoPoolation2 (Kofler et al., 2011a) used the calculated pairwise F_{ST} values for each SNP, obtained with the 'fst-sliding.pl' command

(Kofler et al., 2011b). SNPs/loci falling into the upper 0.5% tail of the empirical distribution of pairwise F_{ST} values were subsequently identified as outlier loci, as recommended by Guo et al. (2015 & 2016). Finally, outlier loci were identified using the R package pcadapt (Luu et al., 2017), with loci excessively related to structure (q -value < 0.05) identified as candidates for selection (Luu et al., 2017). Loci identified by two of the three outlier detection methods were subsequently recognized as candidate outlier loci. The functional role of candidate outliers was evaluated by subjecting 1 000 bp upstream and downstream of the candidate outlier SNP to BLASTX searches. BLASTX searches were completed online (NCBI; National Centre for Biotechnology Information, 2018) using default parameters and the non-redundant protein sequence database.

Southern African population sub-structuring of Kingklip

Similarly, to the methodology described above, I investigated the possibility of population sub-structuring across the Benguela Current region, by comparing Namibian versus South African samples. Pairwise F_{ST} comparisons were performed between the Namibian (Nam 1 & Nam 2) and the 2017 South African sampling sites (CB, TB, SC & EC), based on all simulated datasets (full, neutral and outlier) using the 'diffCalc' command of diveRsity (Keenan et al., 2013). Statistical significance of F_{ST} estimates was assessed based on 95% CI calculated after 1 000 bootstrap replicates. Pairwise F_{ST} was additionally estimated based on the complete dataset using the 'sliding-fst.pl' command of PoPoolation2 (Kofler et al., 2011a). Population clustering, based on the simulated full and outlier datasets, was assessed using BAPS (Corander & Marttinen, 2006), as explained above.

To visualize the multi-loci patterns and genetic relationships among sampling regions, a Principal Component Analysis (PCA) was generated using the 'prcomp' function in R (R Core Development Team 2008), based on SNP allele frequencies per pool, calculated for the full simulated dataset (10 068 SNPs) as well as the complete dataset (>40K SNPs). Allele frequencies were calculated based on allele counts estimated from sync files using the 'snp-freq-diff.pl' command of PoPoolation2 (Kofler et al., 2011a), as described above.

RESULTS

Genetic versus genomic patterns of differentiation

Genome-wide differentiation: Pop 1 versus Pop 2

The pairwise F_{ST} values estimated for the comparison between P1 and P2 were not significantly different from 0, based on the simulated neutral and full datasets (Table 11). Significance of pairwise F_{ST} , based on the complete dataset is not provided by PoPoolation 2, and thus could not be assessed. Overall, genomic differentiation was higher when based on the complete dataset ($F_{ST} = 0.0189$; Table 11), than when using the reduced simulated full (outlier and neutral) dataset ($F_{ST} = 0.0150$, 95% CI = -0.0089, 0.0732), or the neutral only dataset ($F_{ST} = 0.0072$, 95% CI = -0.0178, 0.0552). However, genetic differentiation was found to be statistically significant ($F_{ST} = 0.0486$, 95% CI = 0.0156, 0.1091) when using the simulated outlier dataset (Table 11C).

Table 11: Estimates of pairwise genomic differentiation (F_{ST} , below diagonal) and 95% confidence intervals (above diagonal) for all sampling sites/pools, based on simulated **A.** full (outlier & neutral loci), **B.** neutral (neutral loci only) and **C.** outlier (outlier loci only) datasets, and **D.** complete dataset. Statistically significant results in bold. Pool names as per Table 2.

A.	P1	P2	NAM1	NAM2	CB	TB	SC	EC
P1	-	-0.0089, 0.0732	-0.0129, 0.0654	-0.0164, 0.0614	-0.0177, 0.0627	-0.0178, 0.0599	-0.0182, 0.0609	-0.0131, 0.0628
P2	0.0150	-	-0.0126, 0.0626	-0.0154, 0.0580	-0.0141, 0.0569	-0.0166, 0.0570	-0.0110, 0.0640	-0.0129, 0.0755
NAM1	0.0131	0.0130	-	-0.0180, 0.0609	-0.0202, 0.0610	-0.0212, 0.0627	-0.0188, 0.0584	-0.0184, 0.0615
NAM2	0.0074	0.0092	0.0063	-	-0.0219, 0.0557	-0.0229, 0.0536	-0.0205, 0.0548	-0.0180, 0.0569
CB	0.0074	0.0097	0.0058	0.0027	-	-0.0238, 0.0514	-0.0220, 0.0560	-0.0184, 0.0647
TB	0.0069	0.0079	0.0044	0.0027	0.0001	-	-0.0236, 0.0568	-0.0188, 0.0566
SC	0.0075	0.0135	0.0062	0.0046	0.0034	0.0026	-	-0.0195, 0.0584
EC	0.0115	0.0130	0.0078	0.0066	0.0066	0.0063	0.0059	-

B.	P1	P2	NAM1	NAM2	CB	TB	SC	EC
P1	-	-0.0178, 0.0552	-0.0215, 0.0555	-0.0251, 0.0541	-0.0255, 0.0563	-0.0254, 0.0581	-0.0249, 0.0589	-0.0229, 0.0606
P2	0.0072	-	-0.0214, 0.0529	-0.0238, 0.0503	-0.0215, 0.0517	-0.0229, 0.0524	-0.0202, 0.0556	-0.0217, 0.0608
NAM1	0.0042	0.0042	-	-0.0276, 0.0551	-0.0271, 0.0495	-0.0279, 0.0530	-0.0278, 0.0517	-0.0248, 0.0518
NAM2	0.0009	0.0022	-0.0018	-	-0.0279, 0.0444	-0.0282, 0.0463	-0.0267, 0.0525	-0.0270, 0.0525
CB	-0.0002	0.0023	-0.0010	-0.0033	-	-0.0295, 0.0457	-0.0286, 0.0553	-0.0257, 0.0527
TB	0.0003	0.0013	-0.0026	-0.0035	-0.0054	-	-0.0281, 0.0457	-0.0258, 0.0529
SC	0.0008	0.0054	-0.0012	-0.0021	-0.0027	-0.0032	-	-0.0260, 0.0510
EC	0.0038	0.0044	0.0006	-0.0006	0.0000	-0.0009	-0.0007	-
C.	P1	P2	NAM1	NAM2	CB	TB	SC	EC
P1	-	0.0156, 0.1091	0.0181, 0.1100	0.0067, 0.0963	0.0090, 0.1001	0.0044, 0.1050	0.0059, 0.0975	0.0121, 0.1106
P2	0.0486	-	0.0147, 0.1111	0.0063, 0.1014	0.0082, 0.1019	0.0046, 0.1051	0.0151, 0.1099	0.0163, 0.1140
NAM1	0.0497	0.0491	-	0.0048, 0.1048	-0.0010, 0.0973	0.0012, 0.1052	0.0053, 0.1053	0.0073, 0.1054
NAM2	0.0361	0.0396	0.0400	-	-0.0025, 0.0873	-0.0013, 0.0915	0.0029, 0.1004	0.0075, 0.1070
CB	0.0403	0.0416	0.0339	0.0291	-	-0.0059, 0.0882	-0.0016, 0.0931	0.0036, 0.1028
TB	0.0356	0.0366	0.0338	0.0300	0.0241	-	-0.0024, 0.0924	0.0083, 0.0948
SC	0.0365	0.0486	0.0374	0.0337	0.0299	0.0282	-	0.0035, 0.1026
EC	0.0447	0.0498	0.0377	0.0380	0.0352	0.0374	0.0345	-
D.	P1	P2	NAM1	NAM2	CB	TB	SC	EC
P1	-							
P2	0.0189	-						
NAM1	0.0183	0.0178	-					
NAM2	0.0166	0.0169	0.0142	-				
CB	0.0167	0.0176	0.0141	0.0134	-			
TB	0.0166	0.0166	0.0143	0.0135	0.0136	-		
SC	0.0165	0.0184	0.0148	0.0140	0.0136	0.0142	-	
EC	0.0168	0.0176	0.0161	0.0152	0.0153	0.0152	0.0155	-

Furthermore, although levels of differentiation were non-significant, the PCAs revealed separation between the two pools (Figure 6), with the primary Principal Component (Dim 1) explaining 100 % variance.

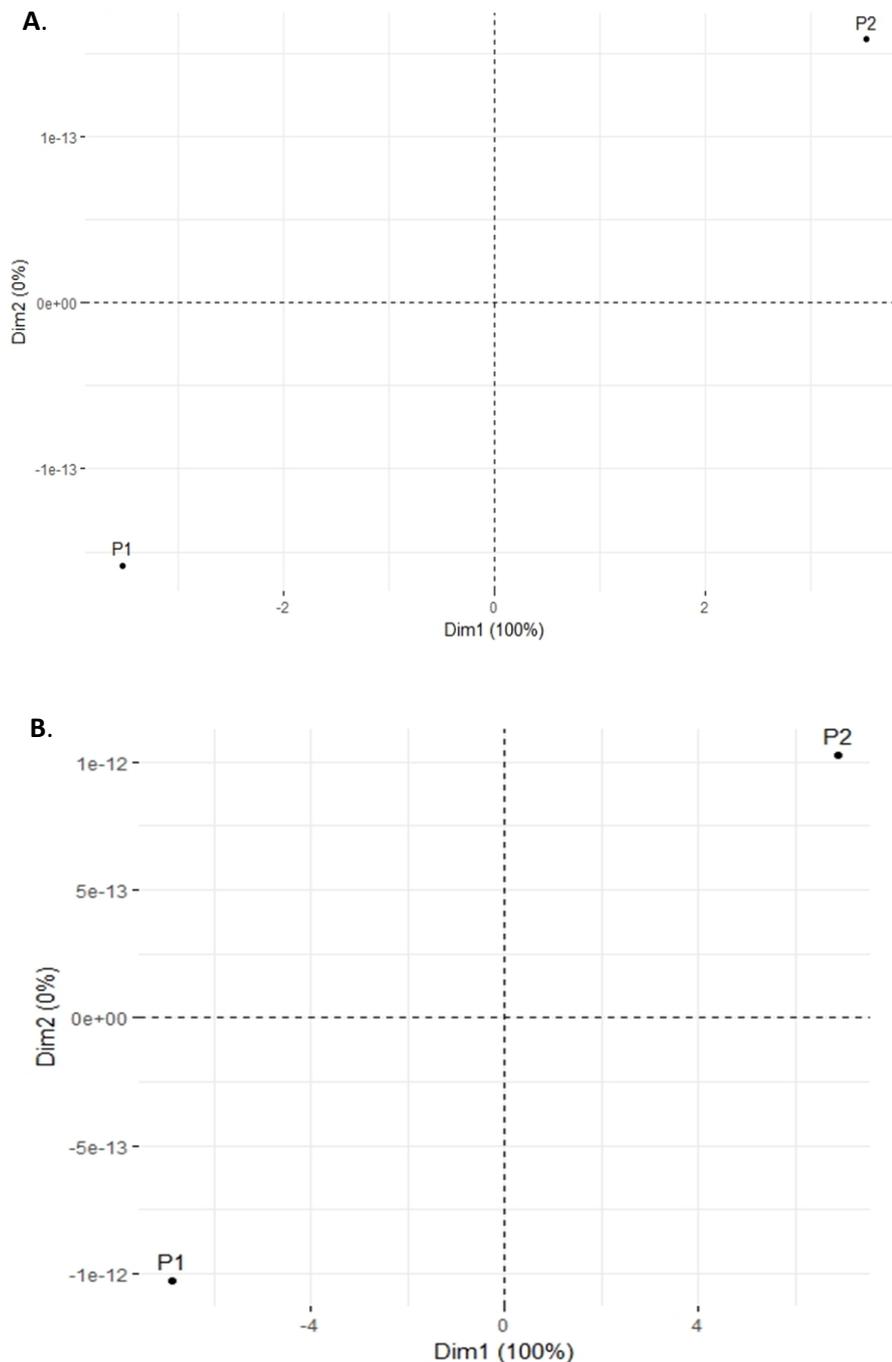


Figure 6: Principal component analysis (PCA) of variation in allele frequencies per pool, based on neutral and outlier loci within the **A.** simulated and **B.** complete nuclear datasets. Pool names as per Table 2.

The generated Manhattan plots of pairwise F_{ST} values illustrates the unequal contribution of loci to the total observed differentiation, with certain loci having larger F_{ST} values (more differentiated) than others (Figure 7 and 8).

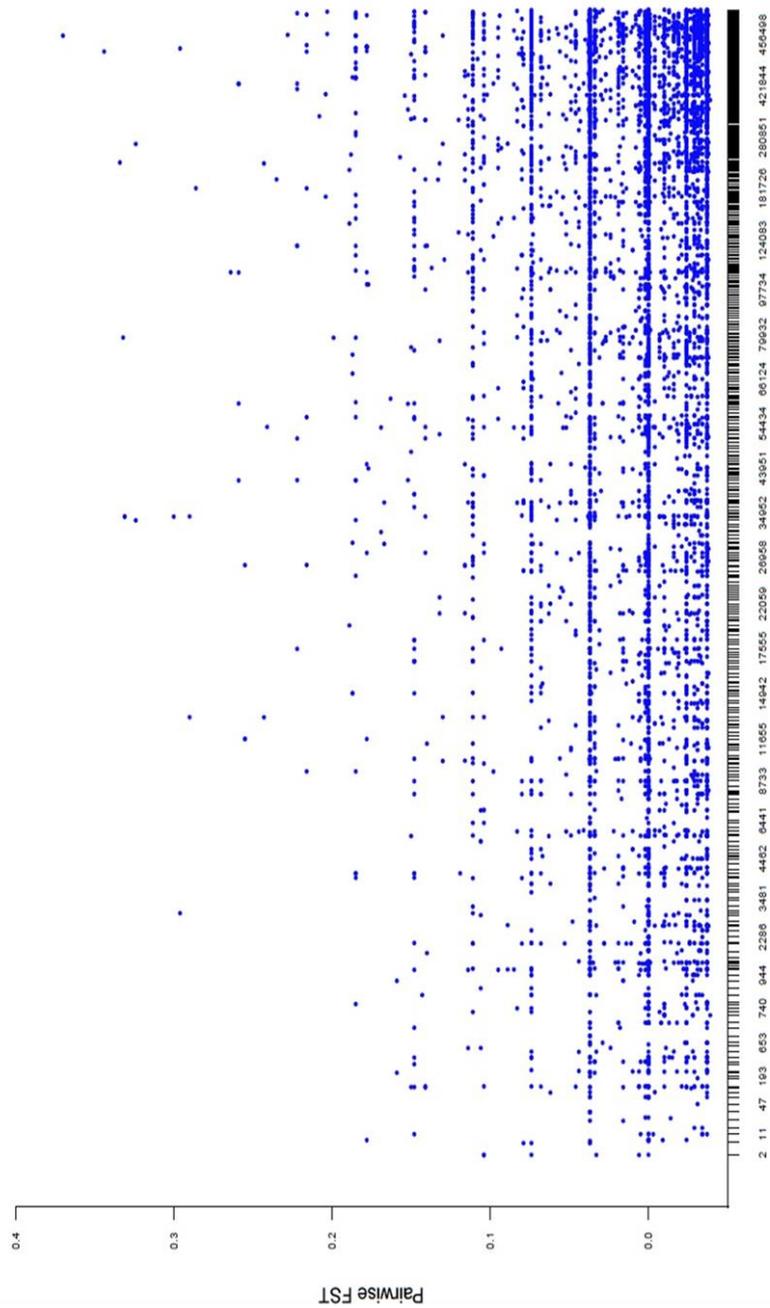


Figure 7: Manhattan plot of pairwise genomic estimates (F_{ST}) per SNP loci for P1 versus P2 (P1 – Pop 1, P2 – Pop 2), against SNP loci position within nuclear reference sequence (nDNA_ref). Plotted for neutral and outlier loci contained within the simulated, full dataset.

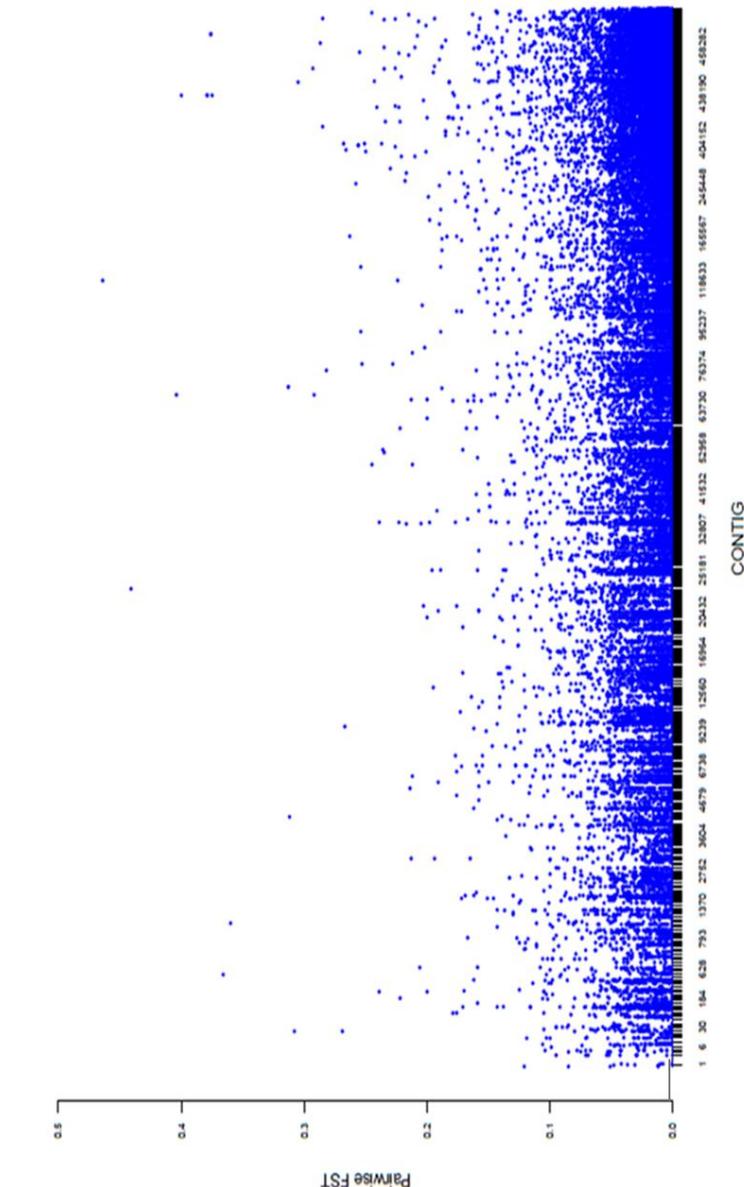


Figure 8: Manhattan plot of pairwise genomic estimates (F_{ST}) per SNP loci for P1 versus P2 (P1 – Pop 1, P2 – Pop 2), against SNP loci position within nuclear reference sequence (nDNA_ref). Plotted for neutral and outlier loci contained within complete, biallelic dataset.

Microsatellite primer sequence mapping

A total of six of 10 previously described microsatellite primer sequences mapped to the *de novo* reference sequence (nDNA_ref), with at least one read mapping per primer pair (Table 12). Of these six primers, only two had both the forward and reverse sequences mapped: cmrGb2.6.1 and cmrGb5.8B. It must be noted, however, that

some primer sequences were found to be split between nodes, with forward and reverse sequences mapping to different nodes along the reference sequence. When viewed in IGV, the associated trinucleotide repeat motifs “GTT” and “CAA” (Supplementary Table S2) were found to be flanked by the forward and reverse primer sequences of cmrGb2.6.1, respectively. Similarly, the forward primer sequence of cmrGb58.B was found to flank the dinucleotide repeat motif “CA”. However, the reverse repeat motif “GT” was not found to be within the vicinity.

The RADseq *de novo* generated reference sequence was thus found to have captured a total of six out of ten microsatellite primer sequences. Interestingly, none of the contigs to which primer sequences mapped to contained either neutral or outlier SNPs. Therefore, although microsatellite primer sequences were found within the generated reference sequence, no SNP variation was captured within these repeat regions.

Table 12: Results of mapping of 10 microsatellite primers (Ward & Reilly, 2010), forward (F) and reverse (R), to *de novo* reference sequence (nDNA_ref). Refer to Supplementary Table S2 and Ward and Reilly (2001) for primer notes.

Locus	Primer Sequences	Mapped to Reference Sequence (Yes/No)
cmrGb4.11B	F - CCTGAGTGCTTAAAGAGGA	No
	R - GAGGAGGAGACGATGAAA	Yes
cmrGb5.5	F - ACTCCTGGACTGGATCTAA	No
	R - TGCAAATTTTCATGTAAATG	Yes
cmrGb5.9	F - AGGGTCACTTTTCAGTTTTA	No
	R - TGCAGAACACACTCCAC	Yes
cmrGb4.2A	F - ATCGGGCAGTTCCTTGCTAT	No
	R - GGGAAGCTTTTGTGAGCATC	No
cmrGb5.2B	F - CGGTCTGAGCAATGATACGA	No
	R - TACAGAGGGGAGGTAAATCAAGTC	No
cmrGb2.6.1	F - AGAACTAAACCAGCAGAATC	Yes
	R - CACAACAAGAGGGAAGTC	Yes
cmrGb5.8B	F - CACTTTGGGGCTTCTCCTC	Yes
	R - CCCGATTCATTCATCCATC	Yes
cmrGb4.2B	F - AGTTGGTGTTTGCCTGA	No
	R - GTCTGGAGTGTTTTGGATCATT	No
cmrGb5.8A	F - AACCTCTGGCATCCATTTTC	No
	R - CCCAAAGTGCTGCTACTG	Yes
cmrGb5.2A	F - AAACAGTGTTTCGCGTTACT	No
	R - CCTGACATGTGTCGTTGA	No

South African population sub-structuring

Population sub-structuring and genome-wide differentiation

Overall, pairwise genomic differentiation estimates based on the full simulated dataset were not significantly different from 0 (average $F_{ST} = 0.0073$) and ranged from $F_{ST} = 0.0034$ for the comparisons between SC and CB, to $F_{ST} = 0.0150$ for P1 versus P2 (Table 11A). Pairwise genomic differentiation based on neutral loci alone (average $F_{ST} = 0.0022$), revealed smaller, non-significant differences between 2017 South African, P1 and P2 pools, with pairwise F_{ST} ranging from $F_{ST} = -0.0002$ for the comparison of CB versus P1, to $F_{ST} = 0.0054$ for the comparison SC versus P2 (Table 11B). The observed difference between the full and neutral datasets thus suggests the possible influence of outlier loci in driving genetic differentiation across pools, as the full dataset had generally higher, yet still non-significant, F_{ST} values by comparison (average $F_{ST} = 0.0073$). In fact, F_{ST} estimates based on the outlier loci dataset was found to yield both higher and statistically significant F_{ST} estimates, with significant differentiation found for 12 out of 15 pairwise comparisons (average $F_{ST} = 0.0373$), resolving four distinguishable sub-populations among the six sampled pools, namely; P1, P2, West coast (CB, TB, SC) and EC (Table 11C).

Within South African populations, sub-structuring revealed contrasting patterns. While all comparisons across the Western and South-western coasts were not statistically significant (average $F_{ST} = 0.0124$ across all datasets; average F_{ST} full and outlier dataset = 0.0147; Table 11A, 11C and 11D), all comparisons containing the Eastern coast were largely different from 0 (average $F_{ST} = 0.0136$ across all datasets; average full and outlier dataset = 0.0179; Table 11A, 11C and 11D). In fact, pairwise comparisons between the 2017 South African sampling sites CB, TB and SC, retrieved a spatially consistent pattern for the complete as well as simulated full and outlier datasets (although not significantly different from 0 for the former), with neighbouring pools found to be less differentiated than the pool from further down the coast (Table 11B, 11C and 11 D). The eastern Agulhas Bank region (EC) was found to be largely differentiated from all 2017 South African sampling sites, yielding the largest pairwise F_{ST} values (average F_{ST} across all datasets = 0.0148; average full and outlier = 0.0210; Table 11A, 11C and 11D), across all four datasets, and displaying statistically significant levels of differentiation based on the outlier dataset (Table 11C).

Bayesian clustering analysis conducted in BAPS (Corander & Marttinen, 2006), on the simulated full dataset failed to reveal sub-structuring, with all sampling sites falling within a single cluster ($K=1$, $p < 0.05$; Figure 9A). However, when analyses were conducted on the outlier dataset, three clusters ($K=3$, $p < 0.05$; Figure 9B) were identified, with cluster one consisting of 2017 sampling sites found west of the Agulhas Bank (CB, TB and SC) as well as P1, and cluster two and three represented by sampling sites P2 and EC respectively (Figure 9). Notably, in contrast to what was observed for outlier F_{ST} values, the pool P1 was found to cluster with CB, SC and TB, while the separate clustering of P2 and EC was in line with F_{ST} patterns of differentiation.

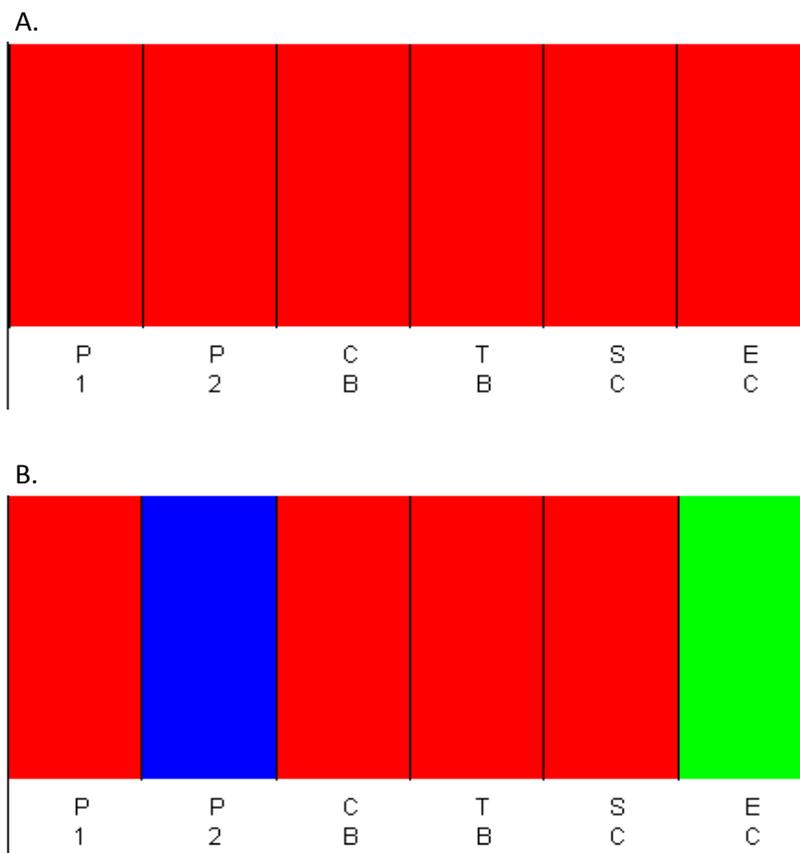


Figure 9: BAPS Bayesian clustering analysis of P1, P2 and 2017 South African sampling sites/pools for simulated **A.** full (neutral and outlier loci) and **B.** outlier (outlier loci only) datasets. Pool names as per Table 2.

The generated PCA plot based on allele frequencies of the simulated full dataset explained 24.1% and 22% of the total variation respectively (Figure 10A), while the PCA generated for the complete dataset explained 24.6% and 22% of the total variation (Figure 10B). Although F_{ST} values based on the simulated full dataset suggest the existence of a single population, the obtained PCAs revealed a geographically ordered pattern, with neighbouring West coast sites CB and TB clustering together and differentiating from southern and eastern coast sampling sites SC and EC, along DIM 1. Previously identified clusters P1 and P2 appear to differentiate from 2017 South African sampling sites along the Dim 2 axis, with P1 and P2 failing to group with any 2017 South African sampling sites.

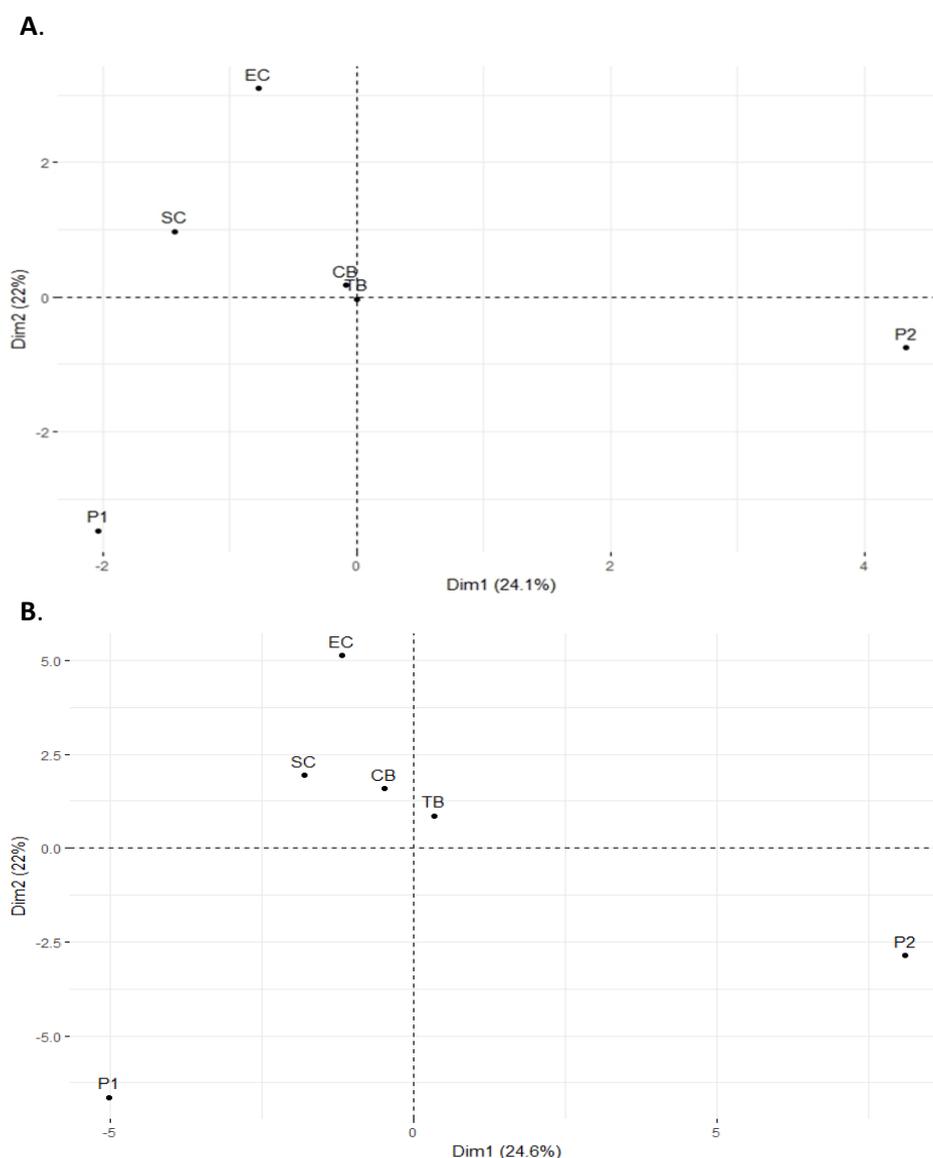


Figure 10: Principal component analysis (PCA) of variation in allele frequencies per pool, based on outlier and neutral loci contained within the **A.** simulated and **B.** complete nuclear datasets. Pool names as per Table 2.

Population sub-structuring and genome-wide differentiation – top 500 loci

As expected, pairwise differentiation of P1 versus P2 based on the top 500 loci was higher and statistically different from 0 ($F_{ST} = 0.1105$; 95% CI = 0.0683 – 0.1729; Table 13). In addition, patterns of differentiation and sub-structuring among P1, P2, and contemporary 2017 South African sampling sites were also different from the previously obtained results, regardless of the dataset used (Table 11). Namely, pairwise F_{ST} estimates of P1 versus CB, TB and SC were found to be non-significant, suggesting the possible relation of these pools with P1 and contrasting the previous F_{ST} estimates based on the outlier dataset (Table 11C). Furthermore, P2 was found to be significantly different from TB ($F_{ST} = 0.0403$, 95% CI = 0.0015 – 0.1109) and SC ($F_{ST} = 0.0435$, 95% CI = 0.0074 – 0.1212), but not from CB ($F_{ST} = 0.0360$, 95% CI = -0.0013 – 0.1122; Table 13). As with F_{ST} estimates based on outlier loci alone, EC was found to be significantly differentiated from P1 and P2 (Table 13).

Table 13: Estimates of pairwise genomic differentiation (F_{ST}) and 95% confidence intervals (95% CI) for South African sampling sites/pools, based top 500 loci dataset. Statistically significant results indicated in bold. Pool names as per Table 2.

Pool ID	P1		P2	
	F_{ST}	95 % CI	F_{ST}	95 % CI
P1	-	-	0.1105	0.0683, 0.1729
P2	0.1105	0.0683, 0.1729	-	-
CB	0.0314	-0.0024, 0.1059	0.0360	-0.0013, 0.1122
TB	0.0346	-0.0004, 0.1065	0.0403	0.0015, 0.1109
SC	0.0289	-0.0042, 0.1063	0.0435	0.0074, 0.1212
EC	0.0365	0.0039, 0.1172	0.0443	0.0098, 0.1148

Interestingly, the generated PCA revealed similar patterns of differentiation as found by the PCA based on the full dataset: P1 and P2 remained differentiated from all 2017 South African sampling sites (Figure 11). The observed patterns of differentiation potentially reflected pairwise F_{ST} estimates, with EC found to be largely differentiated from all sampling sites and 2017 South African sampling sites falling largely between P1 and P2 along the principal axis (Dim 1 = 52.6 % variance).

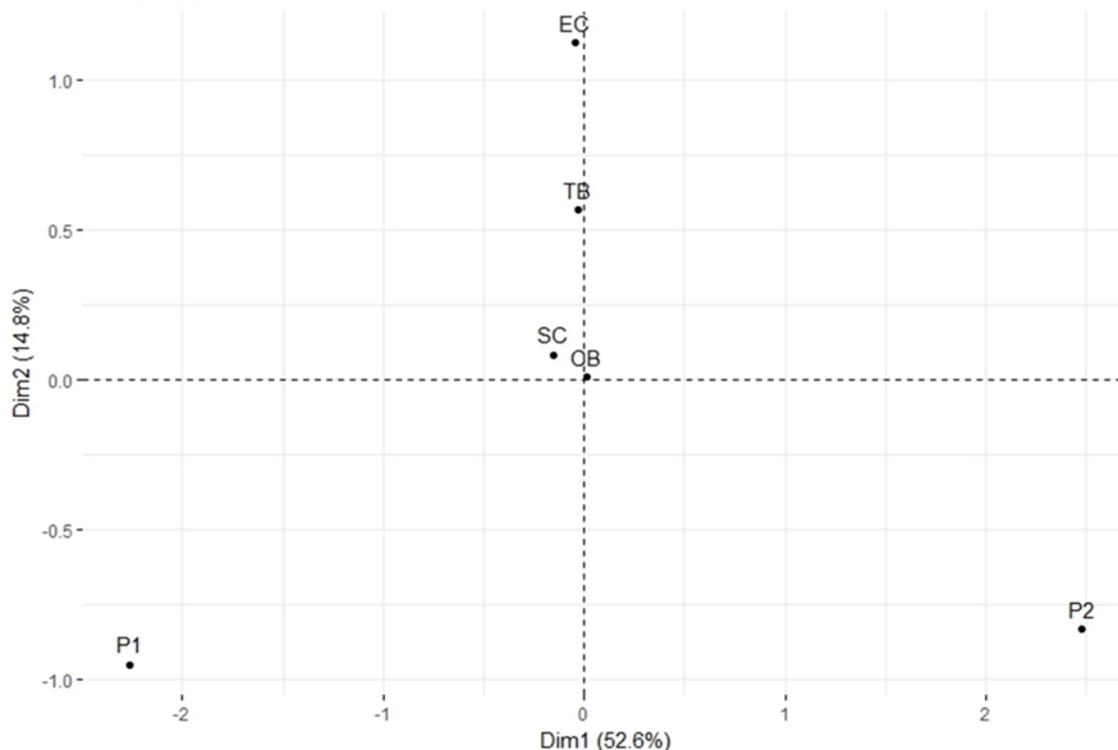


Figure 11: Principal component analysis (PCA) of variation in allele frequencies per pool, based on the top 500 loci (top 500 dataset), identified based on the loci loadings of P1 versus P2 PCA. Pool names as per Table 2.

Outlier analysis of the top 500 loci identified 32 outlier loci. Empirical outlier detection through PoPoolation2 (Kofler et al., 2011a) identified a total of 32 (6.45%) outlier loci, whilst pcadapt (Luu et al., 2017) identified a total of 4 (0.8%) outlier loci. In contrast, Bayescan (Foll & Gaggiotti, 2008) failed to identify potential outlier loci, based on a FDR of 0.05. Four outlier loci were identified by both pcadapt and PoPoolation2 analyses, with two of the four outlier loci matching to protein sequences of known function (Table 14).

Table 14: Top BLASTX search results, with corresponding E-value scores and percentage identity, for candidate outlier loci identified for the top 500 dataset.

Outlier Loci	BLASTX Results	E-Score	% Identity
941-2617	Retinoic acid receptor responder protein 3-like [<i>Anabas restudineus</i>]	0.00	30
38174-239	Uncharacterized protein LOC 111651664 [<i>Seriaa lalandi dorsalis</i>]	0.00	61
256003-469	Uncharacterized protein LOC 10915546 [<i>Esox lucius</i>]	0.00	40
414656-271	Predicted: signal transducer and activator of transcription 1-like [<i>Clupea harengus</i>]	0.00	92

South African versus Namibian population sub-structuring

Genome-wide sub-structuring analyses across the South African (CB, TB, SC and EC) and Namibian (Nam 1 and Nam 2) sampling sites generated non-significant pairwise F_{ST} estimates (average $F_{ST} = 0.0035$), for both the simulated full and neutral datasets (Table 11A and 11B). As previously found, the inclusion of potentially outlier loci in the complete (average $F_{ST} = 0.0144$; Table 11D) and simulated full (average $F_{ST} = 0.0052$; Table 11A) datasets led to higher F_{ST} values, as compared to those based on neutral loci alone (average $F_{ST} = 0.0019$; Table 11B). Pairwise comparisons based on outlier loci revealed increased pairwise differentiation levels, with statistically significant comparisons observed for multiple tests (Table 11C). Specifically, Nam 1 was found to be significantly different from all regions except CB ($F_{ST} = 0.0339$, 95% CI = -0.0010 – 0.0973), whilst Nam 2 was found to be significantly different from sampling sites found along the South-east (SC) ($F_{ST} = 0.0029$, 95% CI = 0.0029 – 0.1004) and East coast (EC) ($F_{ST} = 0.0380$, 95% CI = 0.0075 – 0.1070) only. For all datasets, the Eastern Cape (EC) was found to be most differentiated, with the largest F_{ST} values (average $F_{ST} = 0.0150$; Table 11A – 11D), potentially indicating a break in gene flow.

Clustering analysis in BAPS using the simulated full dataset detected a single cluster, suggesting no population sub-structuring as previously found for pairwise F_{ST} estimates (Figure 12). However, when based on outlier loci only, BAPS (Corander & Marttinen, 2006) clustering analysis revealed evidence for population sub-structuring between sites with a total of two clusters detected (Figure 12). Cluster one included sampling sites falling along the West and South coast of southern Africa, West of the Agulhas Bank, whilst cluster two was represented by the EC sampling site alone (Figure 12).

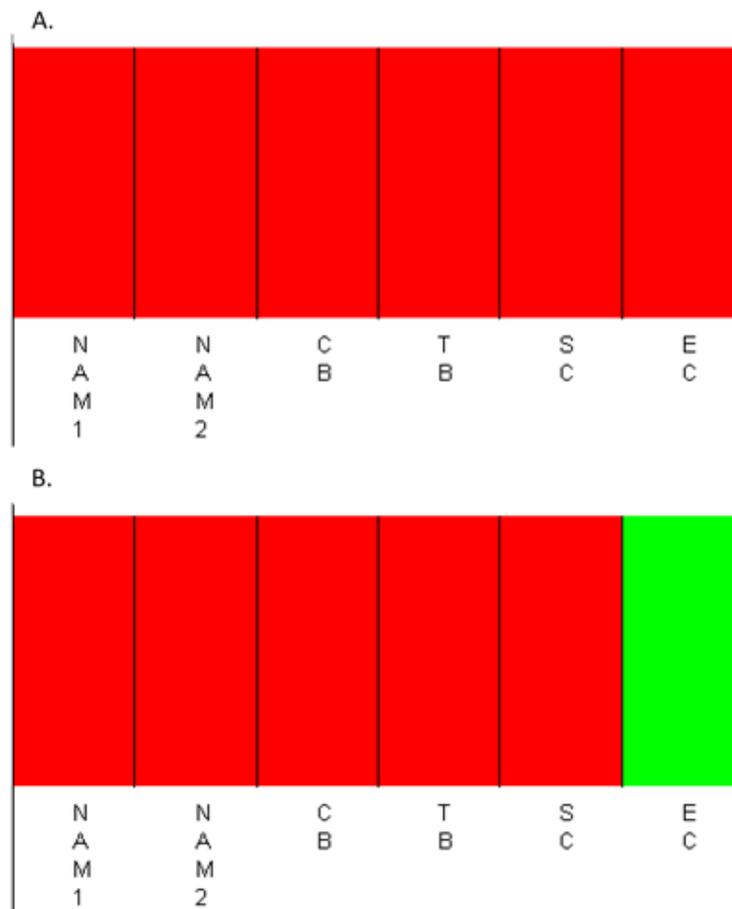


Figure 12: BAPS Bayesian clustering analysis of Namibian (NAM1 –and NAM 2) and 2017 South African sampling sites/pools for simulated **A.** full (neutral and outlier loci) and **B.** outlier (outlier loci only) datasets. Pool names as per Table 2.

The generated PCA, based on the simulated full dataset, where the first two axes explained a total of 42.9% of variance, revealed the grouping of Nam 2, CB, TB and SC sampling sites and the separation of Nam 1 and EC (Figure 13A). Sites corresponding to cluster 1 appear to differentiate from cluster 2 (EC) along the primary axis (Dim 1), with Nam 1 largely differentiating along the secondary axis (Dim 2; Figure 13A). Overall, patterns of variation reflect pairwise F_{ST} estimates, based on outlier loci alone, suggesting the potential occurrence of three distinct sub-populations across southern Africa. Interestingly, in contrast, the generated PCA based on the complete dataset, was found to group Nam 1 with Nam 2, CB and TB whilst SC and EC were found to differ along Dim 2 (20.2%) and Dim 1 (22.4%) respectively (Figure 13B).

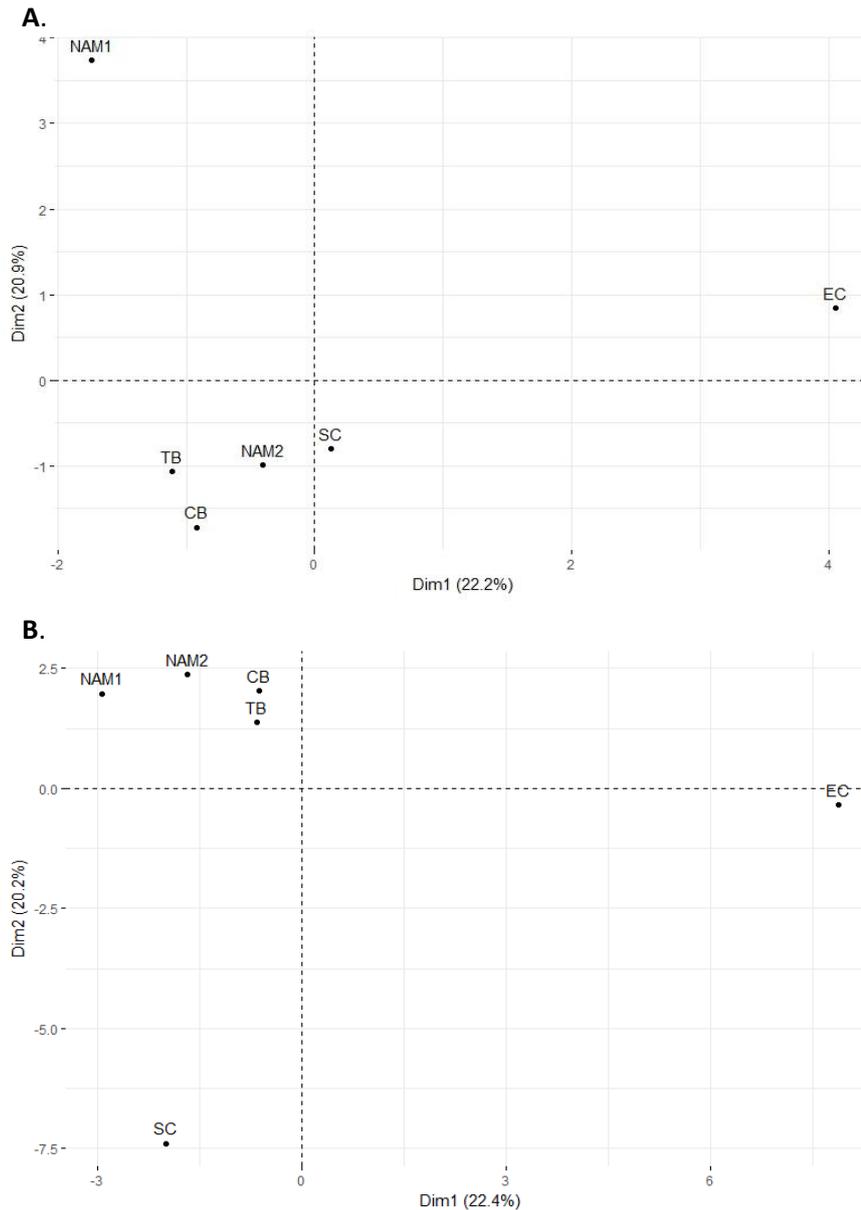


Figure 13: Principal component analysis (PCA) of variation in allele frequencies per pool, based on outlier and neutral loci contained within the **A.** simulated and **B.** complete nuclear datasets. Pool names as per Table 2.

DISCUSSION

With the number of alleles assessed found to influence the power to detect genetic structure more so than the number of alleles per marker (Helyar et al., 2011), the use of hundreds to thousands of SNPs has been shown to provide increased discriminatory power and precision with regards to population structure analyses (Helyar et al., 2011; Benestan et al., 2015; Rodríguez-Ezpeleta et al., 2016; Mullins,

2017). Additionally, with the discriminatory power of 100 SNPs estimated to be equivalent to that of between 10 to 20 nuclear microsatellites, SNP-based analyses have been shown to outperform nuclear microsatellites, by detecting weakly structured populations in the face of high levels of gene flow (Helyar et al., 2011; Carreras et al., 2017; Vendrami et al., 2017). By employing over 10 000 nuclear SNPs, this chapter therefore investigated the fine-scale population sub-structuring of southern African Kingklip, from northern Namibia to southern South Africa, covering most of its known distribution. Moreover, by employing nuclear SNPs only, information regarding contemporary levels of population sub-structuring was provided, which can be used to support future fisheries management policies (Helyar et al., 2012).

The genetic structure of a species is shaped by the interaction of different micro-evolutionary forces, namely genetic drift, gene flow, selection and mutation (Gaggiotti et al., 2009; Carreras et al., 2017). However, as a result of large N_e found for marine species, such as Kingklip, the effects of drift are argued to be minimal (Allendorf et al., 2010; Tigano & Friesen, 2016; Carreras et al., 2017; Dennenmoser et al., 2017; Mullins, 2017). Instead, under large population scenarios, the effects of selection and gene flow are expected to be greater than that of genetic drift in influencing allele frequency differences between populations (Allendorf et al., 2010; Carreras et al., 2017; Mullins, 2017). As such, the observed genetic structure found within this study is reasoned to reflect these two separate components, gene flow and selection. By separating neutral (gene flow) and outlier (putatively under selection) loci into two datasets, inferences regarding the relative role of these two components in shaping the observed patterns of differentiation could be made. Moreover, analysing outlier loci alone provided novel insights into the potential occurrence of local adaptive differentiation, whilst providing information regarding structure at an ecological time-scale (Helyar et al., 2011; Milano et al., 2014; Rodríguez-Ezpeleta et al., 2016).

The full datasets on the other hand, containing both neutral and adaptive loci, provided an overall, yet reduced, genome-wide understanding of Kingklip population sub-structuring. More specifically, analyses based on the simulated full (10 068 SNPs) and complete (41 369 SNPs) datasets were found to be largely congruent. While pairwise F_{ST} estimates were higher based on the complete dataset, the general patterns of genomic sub-structuring and differentiation were similar across the two datasets,

thereby providing confidence that the results obtained based on the simulated datasets are not statistical artefacts.

In light of this, as well as the fact that the simulated dataset could be used for all population analyses, allowing for direct comparisons, inferences and conclusions regarding Kingklip population sub-structuring will be drawn from the results relating to the reduced, simulated datasets (“neutral”, “outlier” and “full”). Furthermore, unlike PoPoolation2 analyses of the complete dataset, 95% confidence intervals of the pairwise F_{ST} estimates of the simulated datasets, based on diveRcity analyses, could be calculated; thereby allowing for the significance of pairwise differentiation levels to be assessed. Overall, assessment of Kingklip sub-structuring revealed complex patterns, with different analyses and marker sets (neutral vs. outlier vs. full) producing different results.

Pop 1 versus Pop 2: genetic versus genomic differentiation

Based on the generated PCA (based on the full dataset), as well as significant levels of differentiation observed for the outlier dataset, the identified SNPs appear to be effective in confirming the existence of the previously identified/hypothesized sub-populations, P1 and P2. It is noted that in contrast to the outlier-based F_{ST} values, non-significant levels of pairwise differentiation were detected for the full and neutral datasets, thereby failing to differentiate between the two pools. However due to differences in marker characteristics including, for example, allelic state and mutation rate, direct comparisons with regards to pairwise F_{ST} values between previous microsatellite-based and reported SNP-based estimates cannot be taken at face-value (DeFaveri et al., 2013; Fischer et al., 2017; Vendrami et al., 2017). Nevertheless, previous microsatellite-based pairwise F_{ST} values were found to be lower (overall $F_{ST} = 0.004$, $p < 0.001$) as compared to those based on the SNP datasets (Full $F_{ST} = 0.0150$, Neutral $F_{ST} = 0.0072$, Outlier $F_{ST} = 0.0486$, overall $F_{ST} = 0.0236$). As such, based on the PCA results, which are comparable between studies and similarly reflect the separation of the two pools, confidence in the discriminatory ability of the newly developed SNP panel is provided. It is noted that whilst the full dataset-based F_{ST} values failed to significantly differentiate the two pools, the generated PCA, similarly based on the full dataset, clearly differentiated P1 and P2. This observed variation in

results may be due to multivariate analyses (e.g. PCAs) not relying on underlying models of differentiation, modes of inheritance or assumptions of HWE or LD, as with F_{ST} -based analyses. As such, PCAs are argued to provide a less biased and simplified image of population sub-structuring with the principal components maximizing variation and thereby potentially detecting subtle structure “overlooked” by F_{ST} values alone (Jombart, 2008; Sham et al., 2009; Allendorf et al., 2010; Helyar et al., 2011; Mullins, 2017).

With microsatellite markers representing differences in repeat units and not necessarily single mutations, discordance between the two marker types may be expected. Relating the newly developed SNP dataset to the previously employed microsatellite primers revealed that, while the microsatellite sequences were represented within the *de novo* reference genome, no SNPs were found within the vicinity of the repeat units. As such, the previous variation contributing to the observed differentiation of P1 and P2 is not accurately reflected within the SNP dataset, thereby explaining the observed differences seen between genetic (microsatellite markers) versus genomic (SNPs) relationships of P1 and P2, whilst simultaneously emphasizing the difficulty of directly comparing separate marker sets, as is widely debated (DeFaveri et al., 2013; Fischer et al., 2017; Vendrami et al., 2017). Ultimately, the detection of differentiation between these two pools, based on the genomic dataset (SNPs), corroborates the results of Henriques et al. (2017), providing further support for the existence of two separate South African sub-populations.

The relative contribution of loci to the observed differentiation of P1 and P2 was unequal across the genome, with certain loci found to have higher levels of differentiation as compared to others. Similar observations of highly variable patterns of differentiation have been reported for several marine species, mainly represented by either few large “genomic islands” of divergence, or many small divergent loci and/or regions spread across the genome (Bradbury et al., 2012; Hemmer-Hansen et al., 2013, 2014; Pujolar et al., 2014; Guo et al., 2015; Reid et al., 2016; Dennenmoser et al., 2017). Furthermore, with divergence proposed to occur in the face of high gene flow, as suggested by neutral F_{ST} values, the distribution of differentiation is predicted to be unequal across the genome, with only a select few loci displaying increased divergence (Hemmer-Hansen et al., 2014). As such, observed heterogeneity of differentiation aligns with that of past studies, illustrating the occurrence of several

highly differentiated loci spread across the genome, potentially as a result of divergence in the face of gene flow. Due to the lack of a complete reference genome, this study was however unable to identify the specific chromosomal locations of increased differentiation, thus suffering from incomplete genomic resolution and preventing inferences with regards to the potential adaptive significance of divergent regions from being made (Hemmer-Hansen et al., 2013).

Pop 1 and Pop 2 versus 2017 South African regions – relation of past clusters to contemporary sampling sites

With P1 and P2 representing two potential sub-populations found along the South African coastline, contemporary 2017 South African sampling sites may be expected to group with either of the two pools. Yet, based on the full and neutral datasets, no clear pattern of grouping was detected between 2017 South African sampling sites and the previously identified clusters, with pairwise comparisons being non-significant and largely similar across pools. Furthermore, outlier based F_{ST} values, as well as the generated PCA, showed clear differentiation between 2017 sampling sites and the previously identified clusters.

This observed inability to assign 2017 samples to the previously identified clusters may be a result of spatial and temporal variation in Kingklip population sub-structuring, as previously described by Henriques et al. (2017). The previously identified clusters consisted of samples collected from different sampling years (2012, 2014 and 2015) and various sampling regions, and thus the failure to group with contemporary (2017) sampling sites may be a result of stochastic temporal and/or spatial variation. Furthermore, the pooling of individuals may have equally contributed to the observed findings. It is thus possible that sequencing of individual fish in the future may allow for a more comprehensive comparison between P1, P2 and contemporary samples, with the expectation that a mix of individuals from various sampling sites will group more clearly with either of the previously hypothesized clusters, as compared to pools of individuals.

Interestingly, BAPS clustering analysis based on the outlier dataset, clustered P1 (representing the most western cluster) with 2017 western and southwestern sampling regions (CB, TB and SC), contrasting with previous outlier F_{ST} results, whilst P2

remained a separate cluster. This observed, discordance may result due to the different assumptions and statistical algorithms employed by the two methodologies (Sham et al., 2009; Allendorf et al., 2010; Wilson & Rannala et al., 2013). Overall, while BAPS clustering analyses are considered appropriate for Pool-Seq datasets, assessing allele frequencies per pool, future individual-based clustering analyses may prove highly valuable, revealing the possible admixture of individuals within/between identified clusters (Pritchard et al., 2000).

Overall, evidence for the potential influence of marker selection, and number, on the observed patterns of sub-structuring between P1, P2 and 2017 South African sampling sites was found. SNP loci are generally found to be unequally informative (Hess et al., 2015), as demonstrated by the differential levels of divergence observed between/across loci as well as datasets in the present study. Marker selection, and number, have been shown to potentially influence observed patterns of population sub-structuring as well as assignment success (Ackerman et al., 2011; Freamo et al., 2011; Benestan et al., 2015; Hess et al., 2015). Accordingly, observed variation between the full and top 500 loci results was observed. By selecting the top 500 loci contributing to the differentiation of P1 and P2, increased levels of divergence were detected as compared to past results. Furthermore, marker selection and number were found to influence the relations of P1 and P2 to 2017 South African sites, with F_{ST} values found to be intermediate as compared to past estimates based on the full and outlier datasets. Notably, P1 was found to be non-significantly different from CB, TB and SC, largely reflecting BAPS outlier results. Correspondingly, as with the BAPS outlier results, P2 was found to be significantly different from the majority of 2017 South African regions, with the exception of CB.

Interestingly, SNP selection was not found to influence PCA results, as the general patterns observed between the full and top 500 datasets were largely congruent, differentiating P1 and P2 from the 2017 South African sampling sites. Similar results were found for the Atlantic mackerel (*Scomber scombrus*), where multivariate analyses found the same population structure patterns regardless of the combinations of individuals and SNPs employed (Rodríguez-Ezpeleta et al., 2016). Furthermore, by not relying on assumptions of underlying population structure, PCA analyses are argued to provide a less biased visualization of genetic relationships among individuals and/or populations, potentially detecting subtle genetic differentiation

previously overlooked or missed (Sham et al., 2009; Allendorf et al., 2010; Helyar et al., 2011; Mullins, 2017).

In summary, while 2017 South African sampling sites failed to clearly group with previously identified clusters when based on the simulated dataset, the selection of highly informative loci was found to influence observed patterns of population sub-structuring. Specifically, 2017 Western and South-western coast regions were found to cluster with P1, when based on the top 500 loci and outlier datasets, thereby indicating the potential occurrence of a single sub-population off western South Africa. While these results are highly valuable, it is noted that future approaches should include the sequencing of individual fishes, which may provide a more comprehensive understanding of the genomic relationship of contemporary samples to previously identified clusters.

Contemporary South African genomic sub-structuring

Sub-structuring analyses of 2017 South African sampling sites revealed complex patterns, with different analyses and datasets producing different results. The distribution of South African Kingklip extends across two distinct biogeographic provinces, the warm temperate region found along the East coast, and the cool temperate region occurring off the West coast (Hutchings et al., 2009), with Cape Point and Cape Agulhas identified as phylogeographic barriers for several marine species (although all available evidence comes from coastal species; Teske et al., 2011). Pairwise F_{ST} and Bayesian clustering analyses of the full and neutral datasets could not detect genetic sub-structure across these provinces suggesting a single South African population of Kingklip. This is in agreement with the earliest Kingklip molecular study by Grant & Leslie (2005) based on allozymes, yet in contrast with the more recent study by Henriques et al. (2017), based on microsatellite and mtDNA markers, which indicated the possibility of a break in gene flow between the Eastern and Western Cape regions. This observed contrast in results between Henriques et al. (2017) and some of the results in the present study is not unexpected, however, as summary statistics of microsatellites and SNPs are generally found to differ (Fischer et al., 2017; Vendrami et al., 2017). While previous comparative studies have found microsatellite and SNP-based F_{ST} estimates to be largely correlated, the relationship

between these markers is highly complex. Previous studies found microsatellite-derived F_{ST} estimates to be greater than SNPs-based estimates, and vice-versa (DeFaveri et al., 2013; Fischer et al., 2017; Vendrami et al., 2017).

In the case of Kingklip, F_{ST} estimates based on the simulated full SNP dataset were found to be higher than microsatellite-based estimates of Henriques et al. (2017). This observed discrepancy may be due to variation between the two types of markers, as higher mutation rates in microsatellites lead to higher genetic diversity levels, which may result in reduced F_{ST} estimates, as compared to biallelic SNP-based estimates (Whitlock et al., 1999). However, in contrast to previous results of Henriques et al. (2017), SNP-derived F_{ST} values for the full and neutral datasets were found to be non-significant for all pairwise comparisons. This observed lack of significant differentiation, despite increased F_{ST} values, may be due to only a small amount of gene flow being needed to overshadow any potential signals of sub-structuring (Hauser & Carvalho, 2008). Interestingly, closer examination of SNP-based F_{ST} values found EC, east of the Agulhas Bank, to have the highest levels of differentiation for all pairwise comparisons, suggesting a potential reduction in gene flow between this region and the remaining South African sampling sites. This was further corroborated by the generated PCA, with EC found to be largely different from the remaining regions.

While neutral markers are highly valuable in elucidating information regarding genetic connectivity and demographic history, outlier loci have been shown to provide increased resolution and power to detect population structure as well as potential local adaptation (Nielsen et al., 2009a; White et al., 2010b; Limborg et al., 2012; Milano et al., 2014). For example, Atlantic herring (*C. harengus* – Limborg et al., 2012; Guo et al., 2016), European hake (*M. merluccius* - Milano et al., 2014) and Atlantic cod (*G. morhua* – Hemmer-Hansen et al., 2014), all exhibit markedly different patterns of structuring when based on neutral versus outlier loci, with the levels of differentiation being largely reduced in the former. Accordingly, structure analyses based on the reduced number of outlier loci revealed increased levels of differentiation and significant sub-structuring. Here, outlier-based analyses indicated a clear differentiation between EC and the remaining South African sampling sites (CB, TB and SC), suggesting the possible occurrence of two separate sub-populations along the South African coastline, in agreement with Henriques et al. (2017), and supporting

the observed differentiation with the full dataset. Moreover, the observed sub-populations were found to correspond largely with previously identified morphological stocks, namely the “Cape” and “South-East” stocks that showed variation in growth rates and otolith growth patterns (Payne, 1977, 1985). Specifically, the West coast sampling sites (CB and TB) were found to correspond to the “Cape” stock, described to occur from south of Lüderitz to Cape Point, whilst the highly differentiated EC sampling site aligns with the “South-East” stock, found within the vicinity of Algoa Bay (Payne, 1977, 1985).

In addition to morphological differentiation, two distinct spawning strategies and grounds have been reported between the West and East coast of South Africa (Olivar & Sabatés, 1989). The presence of multiple spawning sites and periods has been previously linked to patterns of differentiation and sub-structuring for several marine species (Lundy et al., 2000; Henriques et al., 2012, 2015). Therefore, the observed differentiation may similarly reflect the potential influence of divergent Kingklip spawning grounds and/or periods on observed patterns of sub-structuring (Lundy et al., 2000; Henriques et al., 2012, 2015). Finally, the genomic differentiation observed between the West and East coast aligns largely with the occurrence of two separate biogeographic provinces within the regions, with the East coast influenced by the warm Agulhas current, whilst the West coast is found to be under the influence of the cold, nutrient rich Benguela current (Hutchings et al., 2009; Teske et al., 2011). Whilst gene flow was found to occur between the two biogeographic regions, as suggested based on the neutral and full dataset’s results, observed putative adaptive divergence may be associated with environmental variation, and thereby differential selective pressures between the regions, with increasing evidence for spatially varying adaptive selection provided for several marine species, including the Atlantic herring (*C. harengus* – Limborg et al., 2012), the Purple sea urchin (*S. purpuratus* – Pespeni et al., 2010; Pespeni & Palumbi, 2013) and the European hake (*M. merluccius* – Milano et al., 2014). While conclusions regarding the potential selective forces shaping the observed patterns of adaptive divergence were not within the scope of this study, future seascape-genomic studies may allow for a more comprehensive understanding of the potential influence of environmental features and variation in shaping this divergence.

Broadly, the detected divergence of EC from the remaining South African sampling regions aligns with genetic, morphological, spawning and environmental variation and/or differentiation previously described between the two regions, supporting the hypothesis of two different sub-populations of Kingklip across South Africa.

South African versus Namibian genomic sub-structuring

With no previous assessments of the genetic or genomic relationship of South African versus Namibian Kingklip, this study provides novel insights into Kingklip population sub-structuring across the Benguela Current region. Population genetic analyses, based on the full and neutral datasets, revealed non-significant levels of differentiation between Namibian and South African sampling sites, with pairwise F_{ST} values as well as BAPs analyses suggesting the occurrence of high levels of gene flow across the entire system. These findings are in agreement with previous studies of the Deep-water Cape hake, *M. paradoxus*, co-occurring in Namibian and South African waters, with non-significant population structuring found across the Benguela system, yet contrast those of Shallow-water hake (*M. capensis*), Silver kob (*A. inodurus*) and Geelbek (*A. aequidens*), found to exhibit significant population sub-structure (Henriques et al., 2014, 2015, 2016; Mirimin et al., 2016).

Whereas neutral loci suggested the presence of a single population across the Benguela region, statistically significant population sub-structuring patterns were observed for the outlier loci dataset. Specifically, based on outlier-based F_{ST} values as well as PCAs, results showed the occurrence of two sub-populations along the Benguela Current system, with northern Namibia (Nam 1), to the north of the Lüderitz upwelling cell (26 °S), found to be significantly differentiated from the remaining southern Benguela regions occurring along the West coast (Nam 2, CB, TB and SC). Significant environmental variation, relating to salinity, upwelling patterns, bathymetric profiles and oxygen availability, is observed across this system, with seasonal upwelling and associated LOW events of the southern Benguela sub-system contrasting to the year-round low oxygen availability of the north (Shillington et al., 2006; Monteiro et al., 2008; Hutchings et al., 2009). Reported putative adaptive differentiation between the north and south Benguela sub-systems may thus result from environmental variation between these regions, with adaptive divergence having

been increasingly associated with differential selective pressures as a result of environmental variation between regions (Nielsen et al., 2009a; Limborg et al., 2012; Milano et al., 2012, 2014; Lal et al., 2017). More specifically, the potential occurrence of local adaptation and Isolation-by-Environment across the Benguela, has been reported for the Cape hake, *M. capensis*, with seascape analyses revealing associations with upwelling events and bathymetry (Henriques et al., 2016).

In addition to dividing the Benguela into a northern and southern sub-system, the Lüderitz upwelling cell has been identified as a potential barrier to gene flow for several marine species in this region, influencing the dispersal and transport of eggs, juveniles and adults (Henriques et al., 2012, 2014, 2015, 2016; Reid et al., 2016; Mirimin et al., 2016). Accordingly, with the establishment of adaptive divergence, the potential reduction of gene flow across this barrier may allow for local adaptive differences to accumulate, thereby resulting in the observed differentiation (Palumbi, 1994; Henriques et al., 2012; Tigano & Friesen., 2016). This however appears unlikely in the case of Kingklip, as high levels of gene flow across the Lüderitz upwelling region were observed (neutral SNPs). Instead, Kingklip divergence appears to be occurring in the face of gene flow, aligning with increasing evidence for the occurrence of adaptive divergence despite high levels of gene flow (Tigano & Friesen., 2016). With observed patterns of genetic divergence and population sub-structuring argued to be as a result of both historical (e.g. Pleistocene) and/or contemporary (e.g. postglacial) processes and barriers (Waters & Roy, 2004; Hemmer-Hansen et al., 2007; Henriques et al., 2014), two alternate hypotheses are thus proposed to explain the origin of initial Kingklip divergence, namely i) historical isolation followed by secondary contact or ii) a recent split of a single Kingklip population.

The first hypothesis involves the historical isolation of Kingklip, potentially across varying environments, followed by secondary contact. Associations between environmental heterogeneity and adaptive divergence are largely argued to be as a result of local adaptation to environmental variables, however, allopatric divergence followed by recent secondary contact is proposed as an additional/ alternative explanation, resulting in similar patterns of adaptive divergence (Bierne et al., 2011; Le Moan et al., 2016; Reid et al., 2016). Such an example is provided for the European anchovy (*Engraulis encrasicolus*), in which signals of past divergence are detected in genomic regions associated with local adaptation, yet are lost in the remainder of the

genome, with secondary contact and associated gene flow hypothesized to have led to the erosion of past signals of divergence (Le Moan et al., 2016). Overall, genomic regions and loci associated with selection and local adaptation (e.g. outlier loci) are proposed to retain signals of divergence for longer, as compared to the remainder of the genome (Gagnaire et al., 2015; Le Moan et al., 2016). Observed patterns of Kingklip adaptive divergence may thus be an artefact of past isolation followed by secondary contact, with the signals of past divergence maintained within genomic regions/loci associated with selection. While inferences regarding past isolation and secondary contact of Kingklip cannot be made, owing to the lack of historical (mtDNA) population sub-structuring studies between South Africa and Namibia, evidence for past isolation as well as secondary contact has been provided for several marine species found within the region (Sala-Bozano et al., 2009; Henriques et al., 2012; Reid et al., 2016). In particular, historical climatic fluctuations relating to changes in SST, sea level and oceanographic patterns of circulation are hypothesized to have influenced the population sub-structuring of marine species found within the region (von der Heyden et al., 2007; Henriques et al., 2012, 2014; Toms et al., 2014). Specifically, historical changes in environmental and oceanographic features of the Benguela are proposed to have resulted in the initial population divergence of the coastal Leervis (*L. amia*), with isolation maintained through contemporary features (Henriques et al., 2012). However, not all barriers to gene flow are permanent or impermeable, thereby allowing for secondary contact following divergence (Reid et al., 2016). Such an example is provided for the cosmopolitan Bluefish (*P. saltatrix*), with evidence for recent secondary contact across the Benguela region (Reid et al., 2016). As a result, similar isolation, resulting from historical climatic changes, and subsequent secondary contact may have occurred in Kingklip, providing a potential explanation for the observed signals of Kingklip adaptive divergence.

In contrast, patterns of Kingklip divergence may be a result of a recent split of a single population. With N_e , time since separation and migration rates found to influence genetic divergence, a recent isolation of populations with large N_e is argued to result in low levels of divergence (Saha et al., 2015; Le Moan et al., 2016). Accordingly, a contemporary (e.g. postglacial) split of a single Kingklip population may explain the lack of significant divergence observed for the full and neutral datasets, with not enough time having passed for sufficient divergence to have accumulated. In

particular, when divergence occurs under gene flow, as suggested, initial differentiation may be limited to a few genomic regions and/or loci, whilst the majority of the genome remains homogenized (Nosil et al., 2009; Yeaman & Otto, 2011; Hemmer-Hansen et al., 2013, 2014). One such example is the Atlantic cod (*G. morhua*), which is characterized by high levels of gene flow and large N_e but exhibits localized regions of strong population divergence between ecotypes and locations (Hemmer-Hansen et al., 2013). Accordingly, northern and southern Benguela Kingklip appear to be diverging at specific regions along the genome, as inferred by the putative outlier loci. These results not only align with general observations of divergence under gene flow but support the hypothesis that mutations and/or polymorphisms resulting in increased differentiation may accumulate in regions experiencing reduced recombination or divergent selection (Nosil et al., 2009; Bradbury et al., 2012). However as before, a lack of evidence and studies regarding the historical population sub-structuring of Kingklip across this region prevents conclusive inferences regarding to the origin of the observed genetic differentiation from being made.

Overall, regardless of how Kingklip divergence came about in the first place (historical versus contemporary), mechanisms and factors are acting to maintain signals of adaptive divergence between the northern and southern Benguela sub-systems, in the face of high levels of gene flow. With selection implicated as a potential key force in driving and maintaining the observed divergence, as discussed above, endogenous (intrinsic incompatibilities) and/or exogenous (local adaptation to environmental conditions) barriers may be significant in explaining observed differentiation, with the effects of each difficult to separate (Bierne et al., 2011; Hemmer-Hansen et al., 2013; Reid et al., 2016). In addition, homing behaviour, a phenomenon by which individuals return to specific, often natal, spawning grounds has been identified as a key mechanism underlying patterns of population divergence (Lundy et al., 2000). In particular, population sub-structuring has been linked to homing behaviour for several wide-ranging, high gene flow species such as the Mackerel (*S. scombrus* – Nesbø et al., 2000), Herring (*C. harengus* – McQuinn, 1997) and European hake (*M. merluccius* – Lundy et al., 2000). As such, given the evidence for high levels of gene flow across the Benguela region, such homing behaviour may similarly act to maintain observed patterns of Kingklip divergence. Moreover, taking into account the high degree of gene flow observed, it is argued that some level of homing behaviour is required to allow for

signals of divergence to be detected. Currently information regarding the occurrence of Kingklip homing behaviour, as well as the possible position of spawning sites/aggregations along the Namibian coastline, is unavailable. However, evidence for spawning aggregations either side of the Lüderitz upwelling cell has been provided for several marine fish found within the region, including European anchovy (*E. encrasicolus*), Horse mackerel (*Trachurus trachurus*), Round herring (*Etrumeus whiteheadi*) and Shallow-water hake (*M. capensis*) (Olivar & Fortuño, 1991; Olivar & Shelton, 1993; Sundby et al., 2001; Lett et al., 2007), with homing behaviour having been recorded for hake (Jansen et al., 2016). As such, taking into account the above arguments and evidence, potential Kingklip homing behaviour may play a significant role in maintaining divergence across the Lüderitz upwelling cell. Additionally, and often working in conjunction with homing behaviour, local larval retention may similarly act to maintain as well as promote divergence, through a reduction in gene flow and recruitment between spawning aggregations/sites (Waters & Roy, 2004; Bernadi, 2013; Milá et al., 2017). Upwelling cells have been shown to disrupt larvae and egg movement, transporting planktonic organisms (e.g. larvae and eggs) away from suitable habitats, and in turn promoting local larval retention on either side of the upwelling region (Waters & Roy, 2004; Lett et al., 2007; Henriques et al., 2014). In particular, the permanent Lüderitz upwelling cell has been hypothesized to constitute an impermeable barrier to the transport of Geelbek (*A. aequidens*) eggs and larvae (Henriques et al., 2014). Overall the maintenance of divergence observed between the northern and southern Benguela sub-systems is likely a result of contemporary and historical biological and oceanographic processes that point to complex interactions within this dynamic system.

Interestingly, no significant differentiation was observed between south Namibia (Nam 2) and the western South African regions (CB, TB), regardless of the analyses or dataset employed. These findings suggest the occurrence of two populations along the Benguela Current region, and thus point to a shared population between southern Namibia and West Coast South Africa.

With Kingklip occurring in both Namibian and South African waters, increased interest is placed on genetic/genomic relations across the political border off the Orange River. This area is characterized by several coastal and oceanographic features that can potentially influence population sub-structuring (Shannon et al., 1985, 1992; Hutchings

et al., 2009). Specifically, the freshwater outflow from the Orange River as well as a thermal barrier, represented by a cold-water body occurring at depths of 80 – 350 meters, are found within the vicinity (Shannon et al., 1992; Shillington et al., 2006; Hutchings et al., 2009). Freshwater outflow and/or thermal fronts have been identified as potential barriers to gene flow for several marine species (Muss et al., 2001; Rocha et al., 2002; White et al., 2010b, 2011), and thus genetic differentiation across this geopolitical border may be expected. However, these features do not appear to be effective in the case of Kingklip, with results indicating transboundary gene flow across the Orange River region. This lack of structure may be a result of the depth at which mature Kingklip occur (200 – 500 m; Badenhorst & Smale, 1991), as well as differences in life-history traits, as the majority of past studies focused on shallow water coastal species. Oceanographic features in the Benguela Current tend to be less prominent with depth. For example, Deep-water hake (*M. paradoxus*: 110 – 1000 m), does not display population sub-structuring across the Benguela system, while Shallow-water hake (*M. capensis*: 30 – 500m), was shown to have a significant reduction in gene flow across southern Namibia and the Orange River mouth (Henriques et al., 2016). However, despite Kingklip occurring mostly at the same depths as *M. capensis*, the observed genetic break appears to occur further north, off the coast of Lüderitz, at 26°S. While several potential barriers to gene flow are found across the Benguela region, the Lüderitz upwelling cell represents the most intense as well as persistent upwelling cell within the region (Parrish et al., 1983, Boyd, 1987; Lett et al., 2007). Characterised by high winds resulting in strong offshore drift as well as mixing within the water column, the Lüderitz upwelling cell has been identified as a year-round barrier to gene flow for several marine species (Henriques et al., 2012, 2014, 2015, 2016; Reid et al., 2016; Mirimin et al., 2016). While signals of adaptive divergence appear to be occurring in the face of gene flow, the potential influence of the Lüderitz upwelling cell on local larval retention may explain the location of the observed genetic break, with larvae found to occur on either side of the upwelling region, as previously discussed (Waters & Roy, 2004; Lett et al., 2007; Henriques et al., 2014). Furthermore, while information regarding the potential location of Kingklip spawning sites in relation to the observed genetic break is unavailable, the environmental conditions within the Lüderitz upwelling region are proposed to be unsuitable for the recruitment and survival of larvae, with the spawning aggregations of several fish species found to occur on either side of the upwelling region (Olivar &

Fortuño, 1991; Olivar & Shelton, 1993; Sundby et al., 2001; Lett et al., 2007). Accordingly, the potential spawning and associated larvae retention on either side of the Lüderitz upwelling region, in conjunction with the proposed homing behaviour and the different oceanographic conditions, may provide a possible explanation for the observed divergence between the northern and southern Benguela sub-system.

With water movement being mainly northward within the Benguela Current system, prevailing ocean currents may act to promote Kingklip dispersal and movement along the West coast, resulting in the observed homogeneity in this region, until they reach the breaking point designated by the Lüderitz upwelling cell. While Kingklip eggs have been described as pelagic, the Pelagic Larval Duration (PLD), is currently unknown (Brownell, 1979; Olivar & Sabatés, 1998). Consequently, due to the lack of information regarding Kingklip larval transport and genetic sub-structuring of early life stages, inferences regarding the potential influence of currents on Kingklip population structure must be made with caution. Nevertheless, it has been argued that prevailing Benguela currents may enhance the passive transport of Kingklip larvae, thereby resulting in genetic homogeneity across the southern Benguela region (Grant & Leslie, 2005).

Overall, results suggest the presence of two sub-populations within the Benguela region, corresponding to northern and southern Benguela sub-systems respectively. Moreover, these observed genomic sub-populations largely match previously identified morphological stocks (“Walvis” and “Cape” stocks – Payne, 1977), with differences in otolith morphology, meristic features and growth rate observed between the “Walvis stock”, north of Walvis Bay, and the “Cape stock”, south of Lüderitz.

Conclusion

The findings of this Chapter provide evidence for a three-stock hypothesis of Kingklip across southern Africa. Three separate Kingklip sub-populations were identified, namely “northern Benguela”, “southern Benguela” and “East Coast”, separated by two break points occurring within the vicinity of the Lüderitz upwelling cell and Cape Agulhas. Despite there being several potential barriers to gene flow across Kingklip’s distribution, results of the full and neutral dataset suggest that extensive gene flow occurs across the southern African coastline, thereby potentially creating a dynamic,

panmitic system composed of mixed stocks. As a result of this extensive gene flow, significant divergence was only observed in a few outlier loci. As with several marine based population sub-structuring studies, analyses based on a reduced number of outlier loci resulted in evidence for putative adaptive divergence between regions, providing insight into the possible influence of selection in shaping Kingklip population sub-structure.

By employing separate genome-wide datasets (neutral, outlier and full), valuable information with regards to the neutral (neutral dataset), adaptive (outlier dataset) and genome-wide (full; neutral and outlier dataset) levels of Kingklip population sub-structuring were provided. Furthermore, by including samples collected off the Namibian coastline this thesis provides novel insight into the genomic relationship between South African and Namibian Kingklip. Finally, by employing hundreds to thousands of SNPs, increased confidence in inferences regarding Kingklip population sub-structuring was obtained. Therefore, the outcomes of this chapter are aimed to assist the development of future fisheries policies as well as help with the accurate identification of appropriate management units.

CHAPTER 3: Molecular tools in action: Conservation recommendations and implications, as well as development towards a genomic tool for post-harvest control of Kingklip

Conservation recommendations and implications

Marine species represent one of the last wild protein resources globally, resulting in considerable anthropogenic pressure (WWF, 2011; Bernatchez et al., 2017). Within South Africa, approximately 54% of commercially valuable species are considered 'collapsed' (Bruce Mann, ORI, pers. comm), with an estimated 500 million tons of biomass harvested from the Benguela region over the last 200 years (Griffiths et al., 2005). Overexploitation may result in a loss of genetic diversity, reduced population and spawning biomass, alterations in growth rates and sizes, as well as significant changes to communities and ecosystems (Miethe et al., 2010; Briggs, 2011; Henriques et al., 2012, 2017; Pinsky & Palumbi, 2014). With evidence of such effects of past exploitation on Kingklip abundance, as well as genetic diversity, accurately assessing and managing Kingklip stocks should be a national priority, to ensure the long-term conservation of diversity and biomass (Henriques et al., 2017).

Fundamental to the management of wild fisheries is the identification and investigation of stock structure, where stocks represent the unit at which population assessments and management measures are applied (Begg, 1999; Cadrin & Secor, 2009; Duncan et al., 2015; Ovenden et al., 2015; Gilby et al., 2016; Izzo et al., 2017). The stock concept relies on the notion of individual responses to exploitation and fisheries pressures (Hauser & Carvalho, 2008; Benestan et al., 2015; Ovenden et al., 2015; Spies et al., 2015). As such, identifying stock structure is a pre-requisite for the accurate delimitation of harvest quotas and management units, with the aim of ensuring that each stock has a sustainable, and replenishable, harvest quota (Carvalho & Hauser, 1994; Duncan et al. 2015; Gilby et al., 2016). Disregarding patterns of sub-structuring may subsequently result in biologically unsound management, potentially leading to the over- or under-exploitation of certain stock components, as well as the loss of genetic diversity (Stephenson, 1999; Laikre et al. 2005; Fritsch et al., 2006; Pinsky & Palumbi, 2014; Henriques et al., 2017).

The NGS approach employed in this study allowed for a genome-wide investigation of patterns of Kingklip divergence and diversity, including assessing the influence of neutral and/or adaptive forces (see Chapter 1 and 2). The new findings of three putative populations of Kingklip across the southern African region, albeit identified based on outlier loci alone, can thus be used to aid future management decisions. Furthermore, the identification of adaptive divergence is argued to represent an important consideration with regards to fisheries management, as long-term sustainable fishery policies aim to conserve the biological complexity of stocks (Hawkins et al., 2016; Valenzuela-Quiñonez, 2016). The outlier loci identified in this study allowed to detect cryptic, and possibly adaptive divergence in Kingklip, which was overlooked when analysing only neutral loci (Bradbury et al., 2012; Hawkins et al., 2016; Valenzuela-Quiñonez, 2016; Bernatchez et al., 2017). Overall, the population structure analyses completed in Chapter 2 provide the most up-to-date and comprehensive information regarding Kingklip population sub-structuring along its southern African distribution. Based on the population sub-structuring results presented in this thesis consideration of, and changes to, current management strategies is advised. In particular and under the precautionary principle, it seems prudent to consider the independent management of each of the three identified stocks.

Evidence across all datasets and analyses provided strong support for transboundary gene flow between the West coast of South Africa and south of Namibia. Currently, Kingklip resources are managed separately between the two countries, with each employing its own management strategy (DAFF, 2016). Namibian Kingklip bycatch is managed through a bycatch-fee and gear restrictions, with fishing prohibited within the 200-meter isobath so as to protect juveniles (Boyer & Hampton, 2001). South African hake-directed fisheries employ an ecosystem-based management approach aimed at reducing Kingklip bycatch through the setting of a PUCL (DAFF, 2016). Managing this transboundary stock as two separate stocks may be cause for concern, as excessive fishing within one region is likely to influence the entire sub-population (Duncan et al., 2015). As such, potential misalignment in the management strategies employed by each country may result in a loss of biomass and/or genetic diversity. Based on this argument, as well as the comprehensive evidence for the existence of a transboundary stock provided within this study, joint management between the two countries is

advised. Such transboundary management between Namibia and South Africa was recently proposed for the co-occurring Deep-water hake (*M. paradoxus*) at the 2017 “Science and management forum of the Benguela Current Convention (BCC)” (SADSTIA, 2017). Accordingly, with Namibia and South Africa having vested interest in the effective management of Kingklip resources, a similar discussion with regards to Kingklip management between the BCC and regional partners is recommended. Such discussions, as well as a single stock assessment of the proposed “southern Benguela” sub-population, will represent a valuable step towards the effective and appropriate management of Kingklip resources.

With regards to South African Kingklip management, results presented within this study provided further validation for the existence of two South African sub-populations, occurring off the West and East coasts, respectively, and aligning with previously observed genetic, morphological and biological differences found between the two regions (Olivar & Sabatés, 1989; Punt & Japp, 1994; Henriques et al., 2017). Currently, South African Kingklip resources are managed following a one stock approach with a single PUCL set for the entire region (DAFF, 2016). In addition, a seasonal closure of the South-East coast, within the vicinity of Port Elizabeth, was enforced in 2008 so as to protect the spawning aggregation found within the region (DAFF, 2016). At this stage, South African Kingklip resources are argued to be optimally exploited and close to the Maximum Sustainable Yield (MSY), with the resource estimated to be at 40% of its pre-exploited level (DAFF, 2016). However, assessments of Kingklip abundance and biomass are found to be vulnerable to assumptions of underlying stock structure (Punt & Japp, 1994; Brandão & Butterworth, 2013).

With growing concern for the stock structure and status across the coastline, previous Replacement Yield (RY) assessments of the West and South coast found varying results (Brandão & Butterworth, 2008, 2013). When assessed separately, the West coast stock was deemed healthy, with a greater abundance (RY = 4 102 tons) and Precautionary Catch Limit (PCL = 4 302 tons) as compared to South coast stock (PCL = 1 614 tons, RY = 1 553 tons) (Brandão & Butterworth, 2013). In such circumstances, even if the PUCL limits the overall bycatch of Kingklip, the catches off the West and South-east coast may differ, potentially resulting in the over- or under-exploitation of certain regions, with the possibility for a collapse of the less productive stock/sub-

population (Hutchings, 2008; Casey et al., 2016). Accordingly, taking into account the observed genomic divergence between the two regions presented within this study, as well as the above-mentioned differences in stock status, a precautionary two stock management approach is suggested, thereby aligning with previous recommendations (Japp, 1989; Japp, 1990; Grant & Leslie, 2005; Henriques et al., 2017).

Furthermore, from a fisheries management perspective, it is often recommended that differences in phenotypes, biology and life-history traits be considered in addition to genetic components, so as to aid in the design of more comprehensive management plans (Griffiths, 1997; Winker, 2009; Mirimin et al., 2016). While seasonal closure is currently implemented off Port Elizabeth, so as to protect the recorded spawning aggregation, the potential occurrence of a second spawning aggregation off the West coast may be a valuable consideration (Japp, 1990), with similar seasonal closure advisable. Possible spawning aggregations have been observed along the West coast, within the vicinity of Dassen Island and Cape Point (Japp, pers. comms), as well as recorded differences in spatio-temporal spawning strategies (Olivar & Sabatés, 1989). Therefore, future studies investigating the potential existence of a second South African spawning aggregation are recommended.

Overall, the identification and proposed separate management of Kingklip sub-populations will not only assist in the conservation of Kingklip biomass, through the appropriate assessment and setting of harvest quotas, but will in turn aid in maintaining levels of genetic diversity (Laikre et al., 2005; Pinsky & Palumbi, 2014; Henriques et al., 2017). With genetic diversity found to largely influence a species' evolutionary and adaptive potential, a loss of diversity theoretically reduces its ability to both adapt and respond to new environmental pressures and/or conditions (Reiss et al., 2009; Briggs, 2011; Henriques et al., 2012; Pinsky & Palumbi, 2014). As such, in light of increasing evidence for species distributional shifts and recruitment failure in response to climate change, as well as the potential loss of diversity as a result of overexploitation (Henriques et al., 2017), maintaining Kingklip's genetic diversity into the future is a vital consideration with regards to fisheries management (O'Brien et al., 2000; Hauser et al., 2002; Rose, 2005; Reiss et al., 2009). Based on evidence for the importance of stock-/population-based management in conserving genetic diversity (Hutchinson, 2008; Dann et al., 2013), managing the adaptively divergent Kingklip sub-populations, as proposed above, may thus prove highly advantageous in maintaining

the levels of genome-wide diversity and thereby adaptive potential reported within this study (see Chapter 1).

While valuable information regarding the spatial patterns of Kingklip sub-structuring were provided within this study, stock boundaries may display temporal variability. It is thus recommended that future and continued sampling be performed so as to assess the temporal stability of the observed patterns. In addition, it is noted that evidence for adaptive divergence may reflect adaptation to local environmental conditions, as is widely debated (Hauser & Carvalho, 2008; Nielsen et al., 2011; Milano et al., 2011; Bradbury et al., 2012). As such, predicted climate change may result in Kingklip distributional shifts, as has been seen for several marine species (see O'Brein et al., 2000; Last et al., 2011; Poloczanska et al., 2013; Cure et al., 2017), thereby potentially influencing Kingklip fisheries management. Broadly, with potential temporal and spatial variability in stock structure, as well as evidence for considerable gene flow between the proposed stocks and possible local adaptation, adaptive management with annual assessments and regular input of scientific knowledge is recommended, so as to best respond to changes within this dynamic system.

Based on the results of this study, which represent the most up-to-date and the first genomic assessment of Kingklip sub-structuring along its southern African distribution, it is suggested that the current mismatch between Kingklip management units and underlying population sub-structure be addressed. Potential barriers to the uptake of this information, namely limited communication between scientists, government and industry is acknowledged, thus making it essential for regional and government organisations, such as the BCC, DAFF and NatMirc (National Marine Information and Research Centre), to facilitate discussions.

In addition to providing valuable information with regards to Kingklip sub-structuring (Chapter 2), the development of a genome wide dataset (Chapter 1) represents a valuable first step towards the development of a reduced and highly informative SNP marker panel (Kelley et al., 2016; Bernatchez et al., 2017). With a moderate number of markers having been shown to be effective in answering a range of fisheries related questions, the potential development of such a panel may act to provide a cost effective and valuable molecular tool for future stock assessments, thereby aiding in

the regular monitoring and assessment of Kingklip fisheries (Ovenden et al., 2015; Bernatchez et al., 2017).

Post-harvest control

With an estimated value of between \$10 to \$20 USD billion a year, Illegal, Unreported and Unregulated (IUU) fishing poses a global threat to the sustainable harvesting and management of marine resources both through the direct overexploitation of resources, as well as promotion of non-compliance within the industry (Ogden, 2008; Martinsohn & Ogden, 2009; von der Heyden et al., 2010; Helyar et al., 2012; Nielsen et al., 2012). In South Africa, von der Heyden et al. (2010) found that up to 50% of fish fillets marketed as “Kob” (*Argyrosomus* spp.), Yellowtail (*Seriola lalandi*), Dorado (*Coryphaena hippurus*) and Kingklip were mislabelled. In addition, an estimated 30% of fillets marketed as Kingklip, *G. capensis*, were found to be the closely related Pink ling (*G. blacodes*) originating from New Zealand (von der Heyden et al., 2010). The effective and successful management of fisheries thus requires not only direct fisheries management but also post-harvest management and control, with effective Monitoring, Control and Surveillance (MCS) in conjunction with reliable identification and tracking methods being paramount (Ogden, 2008; Martinsohn & Ogden, 2009; Helyar et al., 2012; Santaclara et al., 2014). Furthermore, in the case of bycatch management, information regarding how bycatch is divided between stocks/sub-populations provides a better understanding of the potential impacts of bycatch on said genetic stocks/sub-populations (Hasselman et al., 2016).

The reliable identification of harvested individuals and fish products is often difficult, due the removal of diagnostic morphological characteristics during harvest and processing (von der Heyden et al., 2014; Ovenden et al., 2015). Forensic genetic analyses provide a possible solution to this problem, with recent improvements in DNA techniques providing tools for the assignment of seafood products and harvested individuals to their genetic species and/or geographic origin (Martinsohn & Ogden, 2009; von der Heyden et al. 2010; Seeb et al., 2011). Genetic species identification methods have previously been developed for *Genypterus* species: by combining Polymerase Chain Reactions (PCR) and phylogenetic analysis (FINS; Forensically

Informative Nucleotide Sequencing), Santaclara et al. (2014) developed a methodology able to identify *Genypterus* species in a range of seafood products.

While such techniques are valuable for the correct labelling and marketing of seafood products, most fishery management plans and legislation identify the genetic stock and geographical region of origin as important management units (Martinson & Ogden, 2009; Ovenden et al., 2015). As such, tools allowing for the assignment of individual fish to their population/stock of origin are needed (von der Heyden et al. 2010, 2014). Genetically-based methods have proven highly valuable, with the application of multiple polymorphic markers applied across genetically differentiated stocks or populations providing a means to assign fish and fish products (Martinson & Ogden, 2009; Nielsen et al., 2012).

In particular, SNP-based methods have been shown to be highly informative and accurate with regards to individual assignment, with the ease of inter-laboratory calibration, standardization and the ability to identify outlier loci through genome-scans having promoted their use (Martinson & Ogden, 2009; Nielsen et al., 2012). Specifically, the identification and subsequent inclusion of outlier loci, potentially under selection, has been found by several studies to enhance individual assignment success regardless of their exact adaptive significance (Bradbury et al., 2012; Nielsen et al., 2012; Milano et al., 2014). For example, in a study by Freamo et al. (2011), the assignment of salmon to their population of origin using 14 outlier SNPs was found to be 10% more accurate than 67 neutral SNPs. Overall, the use of highly differentiated SNPs has thus proven to be an effective tool for assignment and identification (Nielsen et al., 2012; Milano et al., 2014). This is best demonstrated by the success of FishPoPTrace, an international project funded by the European Union and established with the purpose of employing SNP databases for the MCS of marine resources (Martinson & Ogden, 2009; Seeb et al., 2011). Specifically, this programme aimed to develop SNP marker panels to be used for the assignment of commercially important species (European hake; *M. merluccius*, Atlantic cod; *G. morhua*, Common sole; *Solea solea* and Atlantic herring; *C. harengus*) to their geographic origin (Martinson & Ogden, 2009; Seeb et al., 2011; Ovenden et al., 2015). The outputs of this project greatly contributed to fisheries control, product traceability, as well as the correct labelling of seafood products (Martinson & Ogden, 2009; Ovenden et al., 2015).

In order to develop molecular tools, in the form of SNP panels, to be used for the stock identification of individuals, several steps must be completed (Figure 14) (see also Martinsohn & Ogden, 2009). Firstly, genomic markers need to be identified through SNP discovery (Chapter 1). Using these genomic resources, population genetic analyses can then be performed, and the fine-scale population structure analysed (Chapter 2). The individual SNPs contributing most to observed genetic differentiation are subsequently identified to form the final SNP panels used for individual assignment and identification (post-harvest regulation). Therefore, by developing the first genome-wide SNP dataset for southern African Kingklip (Chapter 1) and investigating patterns of genomic population sub-structuring (Chapter 2) this study provides the necessary first steps towards the development of a reduced, highly informative marker panel to be used for the identification and assignment of harvested Kingklip individuals and products.

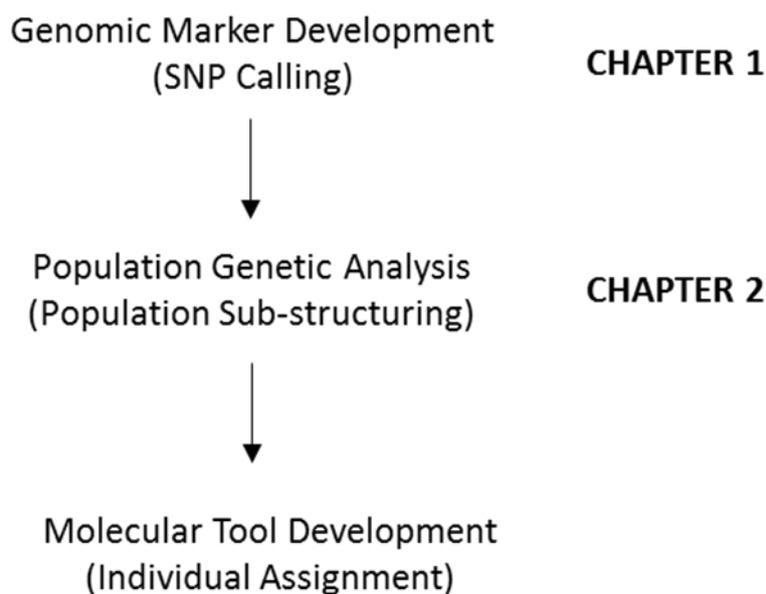


Figure 14: Development of molecular tools for stock identification and individual assignment of Kingklip (*Genypterus capensis*)

Steps towards the development and implementation of such a tool require the identification of loci contributing to the differentiation between designated management stocks/reporting groups. Overall, the development of such a tool will prove highly valuable, aiding in both the regulation and monitoring of Namibian and South African Kingklip harvesting and fisheries, by providing a means to determine the stock of origin of harvested Kingklip individuals and products.

While not within the scope of this study and requiring individual level sequencing as well as agreement regarding Kingklip management stocks, as discussed above, future studies aimed at identifying the most informative loci contributing to the differentiation of proposed management stocks, and subsequent individual level sequencing of the loci and/or the use of microchips, will allow for the assignment success of such a traceability tool to be investigated. Overall this study provided the much needed, and labour intensive, first steps, identifying hundreds of loci that could be used to distinguish proposed management stocks and populations, thus supporting an integrated management approach of Kingklip in southern Africa.

BIBLIOGRAPHY

- Ackerman, M.W., Habicht, C. & Seeb, L.W. (2011). Single-Nucleotide Polymorphisms (SNPs) under diversifying selection provide increased accuracy and precision in mixed-stock analyses of Sockeye Salmon from the Copper River, Alaska. *Transactions of the American Fisheries Society*, 140(3), 865-881. DOI: 10.1080/00028487.2011.588137.
- Akey, J.M., Zhang, K., Xiong, M. & Jin, L. (2003). The effect of Single Nucleotide Polymorphism identification strategies on estimates of linkage disequilibrium. *Molecular Biology and Evolution*, 20(2), 232-242.
- Allendorf, F.W., Hohenlohe, P.A. & Luikart, G. (2010). Genomics and the future of conservation genetics. *Nature Reviews Genetics*, 11(10), 697-709.
- Al-Nakeeb, K., Petersen, T.N. & Sicheritz-Pontén, T. (2017). Norgal: extraction and de novo assembly of mitochondrial DNA from whole-genome sequencing data. *BMC Bioinformatics*, 18, 510.
- Altmann, A., Weber, P., Bader, D., Preuß, M., Binder, E.B & Müller-Myhsok, B. (2012). A beginner's guide to SNP calling from high-throughput DNA-sequencing data. *Human Genetics*, 131(10), 1541-1554.
- Anand, S., Mangano, E., Barizzone, N., Bordoni, R., Sorosina, M., Clarelli, F., Corrado, L., Boneschi, F.M., D'Alfonso, S. & De Bellis, G. (2016). Next Generation Sequencing of pooled samples: Guideline for variants' filtering. *Scientific Reports*, 6, 33735.
- Anantharaman, V. & Aravind, L. (2006). The NYN domains: novel predicted RNAses with a PIN domain like fold. *RNA Biology*, 3(1), 18-27.
- Anderson, E.C., Ng, T.C., Crandall, E.C. & Garza, F.C. (2017). Genetic and individual assignment of tetraploid green sturgeon with SNP assay data. *Conservation Genetics*, 18, 1119-1130.
- Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>
- Babraham Bioinformatics. (2017). Trim-Galore!. Available online at: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore

- Badenhorst, A. (1988). Aspects of the South African longline fishery for kingklip *Genypterus capensis* and the Cape hakes *Merluccius capensis* and *M. paradoxus*. *South African Journal of Marine Science*, 6(1),33-42.
- Badenhorst, A. & Smale, M.J. (1991). The distribution and abundance of seven commercial trawlfish from the Cape south coast of South Africa, 1986–1990. *South African Journal of Marine Science*, 11(1), 377-393.
- Baerwald, M.R., Meek, M.H., Stephens, M.R., Nagarajan, R.P., Goodbla, A.M., Tomalty, K.M.H., Thorgaard, G.H. & Nichols, K.M. (2016). Migration- related phenotypic divergence is associated with epigenetic modifications in rainbow trout. *Molecular Ecology*, 25(8), 1785-1800.
- Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A. & Johnson, E.A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, 3(10), p.e3376.
- Bakun, A. (1996). Patterns in the ocean. Ocean processes and marine population dynamics. University of California Sea Grant, San Diego, CA, in cooperation with Centro de Investigaciones Biologicas de Noroeste, La Paz.
- Ballard, J.W.O. & Whitlock, M.C. (2004). The incomplete natural history of mitochondria. *Molecular Ecology*, 13, 729 - 744.
- Ballard, J.W.O., Chernoff, B. & James, A.C. (2002). Divergence of mitochondrial DNA is not corroborated by nuclear DNA, morphology, or behavior in *Drosophila simulans*. *Evolution*, 56, 527– 545.
- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Pribelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A. & Pevzner, P.A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 5, 455-477.
- Banks, S.C., Piggott, M.P., Williamson, J.E., Bové, U., Holbrook, N.J. & Beheregaray, L.B. (2007). Oceanic variability and coastal topography shape genetic structure in a long-dispersing sea urchin. *Ecology*, 88(12), 3055-3064.

- Barnett, D.W., Garrison, E.K., Quinlan, A.R., Strömberg, M.P. & Marth, G.T. (2011). BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, 27(12), 1691-1692.
- Barrio, A.M., Lamichhaney, S., Fan, G., Rafati, N., Pettersson, M., Zhang, H., Dainat, J., Ekman, D., Höpner, M., Jern, P., Martin, M., Nystedt, B., Liu, X., Chen, W., Liang, X., Shi, C., Fu, Y., Ma, K., Zhan, X., Feng, C., Gustafson, U., Rubin, C.J., Almén, M.S., Blass, M., Casini, M., Folkvord, A., Laikre, L., Ryman, N., Lee, S.M.L., Xu, X. & Andersson, L. (2016). The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. *eLIFE*, 5, e12081.
- Beckley, L.E. & van der Lingen, C.D. (1999). Biology, fishery and management of sardines (*Sardinops sagax*) in southern African waters. *Marine and Freshwater Research*, 50(8), 955-978.
- 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, 1–8.
- 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.
- Bensch, S., Irwin, D.E., Irwin, J.H., Kvist, L. & Åkesson, S. (2006). Conflicting patterns of mitochondrial and nuclear DNA diversity in *Phylloscopus* warblers. *Molecular Ecology*, 15, 161-171.
- Bernadi, G. (2013). Speciation in fishes. *Molecular Ecology*, 22, 5487–5502.
- 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., Suk, H.Y., Therkildsen, N.O. & Withler, R.E. (2017). *Trends in Ecology and Evolution*, 32(9), 665-680.
- Bierne, N., Welch, J., Loire, E., Bonhomme, F. & David, P. (2011). The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Molecular Ecology*, 20, 2044–2072.

Bisby F., Roskov Y., Culham A., Orrell T., Nicolson D., Paglinawan L., Bailly N., Appeltans W., Kirk P., Bourgoïn T., Baillargeon G. & Ovrard D. (2012). Species 2000 & ITIS Catalogue of Life, 2012 Annual Checklist. Available at: www.catalogueoflife.org/col/. Species 2000: Reading, UK.

Blankenberg, D., Von Kuster, G., Coraor, N., Anada, G., Lazarus, R., Mangan, M., Nekrutenko, A. & Taylor, J. (2010). Galaxy: A web-based genome analysis toll for experimentalists. *Current Protocols in Molecular Biology*, 89, 19.1021.

Boyd, A.J. (1987). The oceanography of the Namibian shelf. PhD dissertation, University of Cape Town, Rondebosch.

Boyer, D.C. & Hampton, I. (2001). An overview of the living marine resources of Namibia. *South African Journal of Marine Science*, 23, 5-33.

Bradbury, I.R., Hubert, S., Higgins, B., Bowman, S., Borza, T., Paterson, I.G., Snelgrove, P.V.R., Morris, C.J., Gregory, R.S., Hardie, D., Hutchings, J.A., Ruzzante, D.E., Taggart, C.T. & Bentzen, P. (2012). Genomic islands of divergence and their consequences for the resolution of spatial structure in an exploited marine fish. *Evolutionary Applications*, 6(3), 450-461.

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*, 275, 1803-1809.

Brandão, A. & Butterworth, D.S. (2008). An updated assessment of the South African kingklip resource including some sensitivity tests. In: Marine Resource Assessment and Management Group (Ed.). University of Cape Town, Cape Town.

Brandão, A. & Butterworth, D.S. (2013). A “replacement yield” model fit to catch and survey data for the south and west coasts kingklip resource of South Africa. In: Marine Resource Assessment and Management Group (Ed.). University of Cape Town, Cape Town.

Briggs, J.C. (2011). Marine extinctions and conservation. *Marine Biology*, 158, 484-488.

- Brownell, C. L. (1979). Stages in the early development of 40 marine fish species with pelagic eggs from the Cape of Good Hope. *Ichthyological bulletin of the J.L.B. Smith Institute of Ichthyology*, 40, 84.
- Cadrin, S.X. & Secor, D.H. (2009). Accounting for spatial population structure in stock assessment: past, present, and future. In: Beamish, R.J., Rothschild, B.J. (Eds.). *The Future of Fisheries Science in North America. Fish and Fisheries Series*, 31. Springer, Netherlands, 405–426.
- Cadrin, S.X. (2005). Morphometric Landmarks. In: Cadrin, S.X., Friedland, K.D. & Waldman, J.R. (Eds.). *Stock identification methods. Applications in Fishery Science*. Elsevier Academic Press, Amsterdam, 153 – 172.
- Campana, S.E. (2005). Otolith elemental composition as a natural marker of fish stocks. In: Cadrin, S.X., Friedland, K.D. & Waldman, J.R. (Eds.). *Stock Identification Methods Applications in Fishery Science*. Elsevier Academic Press, 227-245.
- Carreras, C., Ordóñez, V., Zane, L., Kruschel, C., Nasto, I., Macpherson, E. & Pascual, M. (2017). Population genomics of an endemic Mediterranean fish: differentiation by fine scale dispersal and adaptation. *Scientific Reports*, 7 (43417).
- Carrillo, M., Kim, S.K., Rajpurohit, S.K., Kulkarni, V. & Jagadeeswaran, P. (2010). Zebrafish von Willebrand Factor. *Blood Cells, Molecules and Diseases*, 45(4), 326-333.
- Carvalho, G. & Hauser, L. (1994). Molecular genetics and the stock concept in fisheries. *Reviews in Fish Biology and Fisheries*, 4(3), 326-350.
- Casey, J., Jardim, E. & Martinsohn, J.T.H. (2016). The role of genetics in fisheries management under the E.U. common fisheries policy. *Journal of Fish Biology*, 89, 2755-2767.
- Catchen, J.M., Hohenlohe, P.A., Bernatchez, L., Funk, W.C., Andrews, K.R. & Allendorf, F.W. (2017). Unbroken: RADseq remains a powerful tool for understanding the genetics of adaptation in natural populations. *Molecular Ecology Resources*, 17, 362-365.
- Chikhi, R. & Medvedev, P. (2014). Informed and automated k-mer size selection for genome assembly. *Bioinformatics*, 30(1), 31-37.

- Clemento, A.J., Crandall, E.D., Garza, J.C. & Anderson, E.C. (2014). Evaluation of a single nucleotide polymorphism baseline for genetic stock identification of Chinook Salmon (*Oncorhynchus tshawytscha*) in the California Current Large Marine Ecosystem. *Fishery Bulletin*, 112(2-3), 112-130.
- Corander, J. & Marttinen, P. (2006). Bayesian identification of admixture events using multilocus molecular markers. *Molecular Ecology*, 15, 2833-2843.
- Corander, J., Majander, K.K., Cheng, L. & Merilä, J. (2013). High degree of cryptic population differentiation in the Baltic Sea herring *Clupea harengus*. *Molecular Ecology*, 22, 2931-2940.
- Coyle, T. (1998). Stock identification and fisheries management: the importance of using several methods in a stock identification study. In: Hancock, D.A (Eds.) Taking stock: defining and managing shared resources. Australian Society for Fishery Biology, Sydney, 173–182
- Cure, K., Thomas, L., Hobbs, J.A., Fairclough, D.V. & Kennington, W.J. (2017). Genomic signatures of local adaptation reveal source-sink dynamics in a high gene flow fish species. *Scientific Reports*, 7(8618).
- DAFF. (2014). Status of the South African marine fishery resources. Department of Agriculture, Forestry and Fishery, Cape Town.
- DAFF. (2016). Status of the South African marine fishery resources. Department of Agriculture, Forestry and Fishery, Cape Town.
- 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. In: Fisheries Research and Development Corporation Final Report. Project No. 97/117. CSIRO Marine Research, Hobart.
- D'Aloia, C.C.D., Bogdanowicz, S.M., Harrison, R.G. & Buston, P.M. (2014). Seascape continuity plays an important role in determining patterns of spatial genetic structure in a coral reef fish. *Molecular Ecology*, 23(12), 2902-2913.

- Dambach, J., Raupach, M.J., Leese, F., Schwarzer, J. & Engler, J.O (2016). Ocean currents determine functional connectivity in an Antarctic deep-sea shrimp. *Marine Ecology*, 37(6), 1336-1344.
- Dann, T.H., Habicht, C., Baker, T.T. & Seeb, J.E. (2013). Exploiting genetic diversity to balance conservation and harvest of migratory salmon. *Canadian Journal of Fisheries and Aquatic Sciences*, 70, 785-793.
- Davey, J.L. & Blaxter, M.W. (2010). RADSeq: next-generation population genetics. *Briefings in Functional Genomics*, 9(5-6), 416-423.
- Davey, J.W., Cezard, T., Fuentes-Utrilla, P., Eland, C., Gharbi, K. & Blaxter, M.L. (2013). Special features of RAD Sequencing data: implications for genotyping. *Molecular Ecology*, 22, 3151-3164.
- de Moor, C.L., Johnston, S.J., Brandão, A., Rademeyer, R.A., Glazer, J.P., Furman, L.B. & Butterworth, D.S. (2015). A review of the assessments of the major fisheries resources in South Africa. *African Journal of Marine Science*, 37(3), 285-311.
- DeFaveri, J., Viitaniemi, H., Leder, E. & Merilä, J. (2013). Characterizing genic and nongenic molecular markers: comparison of microsatellites and SNPs. *Molecular Ecology Resources*, 13, 377-392.
- Dennenmoser, S., Vamosi, S.M., Nolte, A.W. & Rogers, S.M. (2017). Adaptive genomic divergence under high gene flow between freshwater and brackish-water ecotypes of prickly sculpin (*Cottus asper*) revealed by Pool-Seq. *Molecular Ecology*, 26, 25-42.
- DiBattista, J.D., Travers, M.J., Moore, G.I., Evans, R.D., Newman, S.J., Feng, M., Moyle, S.D., Rebecca, J.G., Saunders, T. & Berry, O. (2017). Seascape genomics reveals fine-scale patterns of dispersal for a reef fish along the ecologically divergent coast of Northwestern Australia. *Molecular Ecology*, 26, 6206-6223.
- Ding, J., Sidore, C., Butler, T.J., Wing, M.K., Qian, Y., Meirelles, O., Busonero, F., Tsoi, L.C., Maschio, A., Angius, A., Min Kang, H., Nagaraja, R., Cucca, F., Abecasis, G.R. & Schlessinger, D. (2015). Assessing mitochondrial DNA variation and copy number in lymphocytes of ~2,000 Sardinians using tailored sequencing analysis tools. *PLoS Genetics*, 11(7), e1005306.

- Duncan, M., James, N., Fennessy, S.T., Mutombene, R.J. & Mwale, M. (2015). Genetic structure and consequences of stock exploitation of *Chrysoblephus puniceus*, a commercially important sparid in the South West Indian Ocean. *Fisheries Research*, 164, 64-72.
- Durai, D.A & Schulz, M.H. (2016). Informed kmer selection for de novo transcriptome assembly. *Bioinformatics*, 32(11), 1670-1677.
- Eberl, R., Mateos, M., Grosberg, R.K., Santamaria, C.A & Hurtado, L.A. (2013). Phylogeography of the supralittoral isopod *Ligia occidentalis* around the Point Conception marine biogeographical boundary. *Journal of Biogeography*, 40(12), 2361-2372.
- Emanuel, B.P., Bustamante, R.H., Branch, G.M., Eekhout, S. & Odendaal, F.J. (1992). A zoogeographic and functional approach to the selection of marine reserves on the west coast of Africa. *South African Journal of Marine Science*, 12, 341-368.
- FAO. (2009). Deep-Sea Fisheries in the High Seas – Ensuring sustainable use of marine resources and the protection of vulnerable marine ecosystems. Food and Agriculture Organization, Rome.
- Fernández, R., Schubert, M., Vargas-Velázquez, A.M., Brownlow, A., Víkingsson, G.A., Siebert, U., Jensen, L.F., Øien, N., Wall, D., Rogan, E., Mikkelsen, B., Dabin, W., Alfarhan, A.H., Alquraishi, S.A., Al-Rasheid, K.A.S., Guillot, G. & Orlando, L. (2016). A genomewide catalogue of single nucleotide polymorphisms in white-beaked and Atlantic white-sided dolphins. *Molecular Ecology Resources*, 16, 266-276.
- Fischer, M.C., Rellstab, C., Leuzinger, M., Roumet, M., Gugerli, F., Shimizu, K.K., Holderegger, R. & Widmer, A. (2017). Estimating genomic diversity and population differentiation an empirical comparison of microsatellite and SNP variation in *Arabidopsis helleri*. *BMC Genomics*, 18(69).
- FishPopTrace. (2017). European Commission Joint Research Centre, FishPopTrace. Available online at: <https://fishpoptrace.jrc.ec.europa.eu/home>.
- Foll, M. & Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A bayesian perspective. *Genetics*, 180(2), 977-993.

- Freamo, H., O'Reilly, P., Berg, P.R., Lien, S. & Boulding, E.G. (2011). Outlier SNPs show more genetic structure between two Bay of Fundy metapopulations of Atlantic salmon than do neutral SNPs. *Molecular Ecology Resources*, 11, 254-267.
- Fritsch, M., Morizur, Y., Lambert, E., Bonhomme, F. & Guinand, B. (2006). Assessment of sea bass (*Dicentrarchus labrax*, L.) stock delimitation in the Bay of Biscay and the English Channel based on mark-recapture and genetic data. *Fisheries Research*, 83, 123-132.
- Fu, L., Cai, C., Cui, Y., Wu, J., Liang, J., Cheng, F. & Wang, X. (2016). Pooled mapping: an efficient method of calling variations for population samples with low-depth resequencing data. *Molecular Breeding*, 36(4).
- Funk, W.C., McKay, J.K., Hohenlohe, P.A. & Allendorf, F.W. (2012). Harnessing genomics for delineating conservation units. *Trends in Ecology and Evolution*, 27(9), 489-496.
- Futschik, A. & Schlötterer, C. (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*, 186(1), 207-218.
- Gaggiotti, O.E., Bekkevold, D., Jørgensen, H.B.H., Foll, M., Carvalho, G.R., Andre, C. & Ruzzante, D.E. (2009). Disentangling the effects of evolutionary, demographic, and environmental factors influencing genetic structure of natural populations: Atlantic herring as a case study. *Evolution*, 63(11), 2939-2951.
- Gagnaire, P.A., Broquet, T., Aurelle, D., Viard, F., Souissi, A., Bonhomme, F., Arnaud-Hoand, S. & Bierne, N. (2015). Using neutral, selected and hitchhiker loci to assess connectivity of marine populations in the genomic era. *Evolutionary Applications*, 8, 769-786.
- Gaither, M.R., Bowen, B.W., Rocha, L.A. & Griggs, J.C. (2016). Fishes that rule the world: circumtropical distributions revisited. *Fish and Fisheries*, 17, 664-679.
- Gaither, M.R., Gkafas, G.A., de Jong, M., Sarigol, F., Neat, F., Regnier, T., Moore, D., Gröcke, D.R., Hall, N., Liu, X., Kenny, J., Lucaci, A., Hughes, M., Haldenby, S. & Hoelzel, A.R. (2018). Genomics of habitat choice and adaptive evolution in a deep-sea fish. *Nature Ecology and Evolution*, 2, 680-687.

- Gale, A.J. (2011). Current understanding of hemostasis. *Toxicology Pathology*, 39(1), 273-280.
- Galindo, J., Grahame, J.W. & Butlin, R.K. (2010). An EST-based genome scan using 454 sequencing in the marine snail *Littorina saxatilis*. *Journal of Evolutionary Biology*, 22, 2004-2016.
- Garant, D., Forde, S.E. & Hendry, A.P. (2007). The multifarious effects of dispersal and gene flow on contemporary adaptation. *Functional Ecology*, 21, 434–443.
- Gilbey, J., Cauwelier, E., Coulson, M.W., Stradmeyer, L., Sampayo, J.N., Armstrong, A., Verspoor, E., Corrigan, L., Shelley, J. & Middlemas, S. (2016). Accuracy of assignment of Atlantic Salmon (*Salmo salar* L.) to rivers and regions in Scotland and Northeast England based on Single Nucleotide Polymorphism (SNP) markers. *PLoS ONE*, 11(10), e0164327.
- Glazier, A.E & Etter, R.J. (2014). Cryptic speciation along a bathymetric gradient. *Biological Journal of the Linnean Society*, 113(4), 897-913.
- Gleason, L.U. & Burton, R.S. (2016). Genomic evidence for ecological divergence against a background of population homogeneity in the marine snail *Chlorostoma funebris*. *Molecular Ecology*, 25, 3557-3573.
- Glover, K.A., Hansen, M.M., Lien, S., Als, T.D., Høyheim, B. & Skaala, Ø. (2010). A comparison of SNP and STR loci for delineating population structure and performing individual genetic assignment. *BMC Genetics*, 11(1), 2.
- Grant, W.S. & Leslie, R.W. (2005). Bayesian analysis of allozyme markers indicates a single genetic population of kingklip *Genypterus capensis* off South Africa. *African Journal of Marine Science*, 27(2), 479-485.
- Griffiths, M.H. (1997). The life history and stock separation of silver kob, *Argyrosomus inodorus*, in South African waters. *Fishery Bulletin*, 95, 47–67.
- Griffiths, C.L., Van Sittert, L., Best, P.B., Brown, A.C., Clark, B.M., Cook, P.A., Crawford, R.J.M., David, J.H.M., Davies, B., Griffiths, M.H., Hutchings, K., Jerardino, A., Kruger, N., Lamberth, S., Leslie, R.W., Melville-Smith, R., Tarr, R. & van der Lingen, C.D. (2005). Impacts of human activities on marine animal life in the

Benguela: a historical overview. In: *Oceanography and Marine Biology: an Annual Review*, 42. Press Taylor & Francis Group, Boca Raton, 303–392.

Gu, S., Hou, P., Liu, K., Niu, X., Wei, B., Mao, F. & Xu, Z. (2018). NOL8 the binding protein for beta-catenin, promoted the growth and migration of prostate cancer cells. *Chemico-Biological Interactions*, 294, 40-47.

Guillot, G., Leblois, R., Coulon, A. & Frantz, A.C. (2009). Statistical methods in spatial genetics. *Molecular Ecology*, 18, 4734-4756.

Guo, B., DeFaveri, J., Sotelo, G., Nair, A. & Merilä, J. (2015). Population genomic evidence for adaptive differentiation in Baltic Sea three-spined sticklebacks. *BMC Biology*, 13(1).

Guo, B., Li, Z. & Merilä, J. (2016). Population genomic evidence for adaptive differentiation in the Baltic Sea herring. *Molecular Ecology*, 25(12), 2833-2852.

Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. (2013). QUASt: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8), 1072-1075.

Hahn, C., Bachman, L. & Chevreur, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads - a baiting and iterative mapping approach. *Nucleic Acids Research*, 41(13), e129.

Hartl, D.L. & Clark, A.G. (2007) *Principles of Population Genetics*. Sinauer Associates, Sunderland, Massachusetts.

Hasselman, D.J., Anderson, E.C., Argo, E.E., Bethoney, N.D., Gephard, S.R., Post, D.M., Schondelmeier, B.P., Schultz, T.F., Willis, T.V. & Palkovacs, E.P. (2016). Genetic stock composition of marine bycatch reveals disproportional impacts on depleted river herring genetic stocks. *Canadian Journal of Fisheries and Aquatic Sciences*, 73, 951-963.

Hauser, L. & Carvalho, G. (2008). Paradigm shifts in marine fisheries genetics: ugly hypotheses slain by beautiful facts. *Fish and Fisheries*, 9(4), 333-362.

Hauser, L., Adcock, G.J., Smith, P.J., Bernal Ramirez, 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 of the United States of America*, 99(18), 11742-11747.

- Hawkins, S.J., Bohn, K., Sims, D.W., Ribeiro, P., Faria, J., Presa, P., Pita, A., Martins, G.M., Neto, A.I., Burrows, M.T. & Genner, M.J. (2016). Fisheries stocks from an ecological perspective: Disentangling ecological connectivity from genetic interchange. *Fisheries Research*, 179, 333-341.
- Helyar, S., Hemmer-Hansen, J., Bekkevold, D., Taylor, M.I., Ogden, R., Limborg, M.T., Cariani, A., Maes, G.E., Diopere, E., Carvalho, G.R. & Nielsen, E.E. (2011). Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources*, 11, 123-136.
- Helyar, S.J., Limborg, M.T., Bekkevold, D., Babbucci, M., van Houdt, J., Maes, G.E., Bargelloni, L., Nielsen, R.O., Taylor, M.I., Ogden, R., Cariani, A., Carvalho, G.R., Consortium, F. & Panitz, F. (2012). SNP discovery using next generation transcriptomic sequencing in Atlantic Herring (*Clupea harengus*). *PLoS ONE*, 7(8), e42089.
- Hemmer-Hansen, J., Nielsen, E.E., Gronkjaer, P. & Loeschcke, V. (2007). Evolutionary mechanisms shaping the genetic population structure of marine fishes; lessons from the European flounder (*Platichthys flesus* L.). *Molecular Ecology*, 16, 3104– 3118
- Hemmer-Hansen, J., Nielsen, E.E., Therkildsen, N.O., Taylor, M.I., Ogden, R., Geffen, A.J., Bekkevold, D., Helyar, S., Pampoulie, C., Johansen, T., FishPoPTrace Consortium & Carvalho, G.R. (2013). A genomic island linked to ecotype divergence in Atlantic cod. *Molecular Ecology*, 22, 2653-2667.
- Hemmer-Hansen, J., Therkildsen, N.O. & Pujolar, J.M. (2014). Population genomics of marine fishes: next-generation prospects and challenges. *Biological Bulletins*, 227, 117-132.
- Henriques, R., Nielsen, E., Durholtz, D., Japp, D. & von der Heyden, S. (2017). Genetic population sub-structuring of kingklip (*Genypterus capensis* – Ophidiidae), a commercially exploited demersal fish off South Africa. *Fisheries Research*, 187, 86-95.
- Henriques, R., Potts, W.M., Santos, C.V., Sauer, W.H.H. & Shaw, P.W. (2014). Population connectivity and phylogeography of a coastal fish, *Atractoscion*

aequidens (Sciaenidae), across the Benguela Current Region: evidence of an ancient vicariant event. *PLoS ONE*, 9(2), p.e87907.

Henriques, R., Potts, W.M, Sauer, W.H.H. & Shaw, P.W. (2012). Evidence of deep genetic divergence between populations of an important recreational fishery species, *Lichia amia* L. 1758, around southern Africa. *African Journal of Marine Science*, 34(4), 585-591.

Henriques, R., Potts, W.M, Sauer, W.H.H. & Shaw, P.W. (2015). Incipient genetic isolation of a temperate migratory coastal sciaenid fish (*Argyrosomus inodorus*) within the Benguela Cold Current system. *Marine Biology Research*, 11(4), 423-429.

Henriques, R., von der Heyden, S., Lipinski, M.R., du Toit, N., Kainge, P., Bloomer, P. & Matthee, C. (2016). Spatio-temporal genetic structure and the effects of long-term fishing in two partially sympatric offshore demersal fishes. *Molecular Ecology*, 25(23), 5843-5861.

Hess, J.O., Campbell, N.R., Docker, M.F., Baker, C., Jackson, A., Lampman, R., McIlraith, B., Moser, M.L., Statler, D.P., Young, W.P., Wildbill, A.J. & Narum, S.R. (2015). Use of genotyping by sequencing data to develop a high-throughput and multifunctional SNP panel for conservation applications in Pacific lamprey. *Molecular Ecology*, 15, 187-202.

Hess, J.E., Campbell, N.R., Population genomics of Pacific lamprey: adaptive variation in a highly dispersive species. *Molecular Ecology*, 22(11), 2898-2916.

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, 379–397.

Hohenlohe, P., Amish, S.J., Catchen, J.M., Allendorf, F. & Luikart, G. (2011). Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources*, 11, 117-122.

Huang, H.W., Mullikin, J.C. & Hansen, N.F. (2015). Evaluation of variant detection software for pooled next-generation sequence data. *BMC Bioinformatics*, 16(1), 235.

- Hubert, S., Higgins, B., Borza, T. & Bowman, S. (2010). Development of a SNP resource and a genetic linkage map for Atlantic cod (*Gadus morhua*). *BMC Genomics*, 11(1), 191.
- Hutchings, L., van der Lingen, C.D., Shannon, L.J., Crawford, R.J.M., Verheye, H.M.S., Bartholomae, C.H., van der Plas, A.K., Louw, D., Kreiner, A., Ostrowski, M., Fidel, Q., Barlow, R.G., Lamont, T., Coetzee, J., Shillington, F., Veitch, J., Currie, J. & Monteiro, P.M.S. (2009). The Benguela Current: An ecosystem of four components. *Progress in Oceanography*, 83(1-4), 15-32.
- Hutchinson, W.F. (2008). The dangers of ignoring stock complexity in fishery management: the case of the North Sea cod. *Biology Letters*, 4, 693-695.
- 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, 267-284.
- Jansen, T., Kristensen, K., Kainge, P., Durholtz, D., Strømme, T., Thygesen, U.H., Wilhelm, M.R., Kathena, J., Fairweather, T.P., Paulus, S., Degel, H., Lipinski, M.R. & Beyer, J.E. (2016). Migration, distribution and population (stock) structure of shallow-water hake (*Merluccius capensis*) in the Benguela Current Large Marine Ecosystem inferred using a geostatistical population model. *Fisheries Research*, 179, 156–167.
- Japp, D.W. (1989). An assessment of the South African longline fishery with emphasis on stock integrity of Kingklip *Genypterus capensis* (Pisces: Ophidiidae). MSc Thesis, Rhodes University.
- Japp, D.W. (1990). A new study on age and growth of kingklip *Genypterus capensis* off the south and west coasts of South Africa, with comments on its use for stock identification. *South African Journal of Marine Science*, 9(1), 223-237.
- Johansson, M.L., Alberto, F., Reed, C.D., Raimondi, P.T., Coelho, N.C., Young, M.A., Drake, P.T., Edwards, C.A., Cavanaugh, K., Assis, J., Ladah, L.B., Bell, T.W., Coyer, J.A., Sigel, D.A. & Serrao, E.A. (2015). Seascape drivers of *Macrocystis pyrifera* population genetic structure in northeast Pacific. *Molecular Ecology*, 24, 4866- 4885.

Jombart, T. (2008). Adagenet: a R package for multivariate analysis of genetic markers. *Bioinformatics*, 24(11), 1403-1405.

Jones, F.C., Grabherr, M.G., Chan, Y.F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M.C., White, S., Birney, E., Searle, S., Schmutz, J., Grimwood, J., Dickson, M.c., Myers, R.M., Miller, C.T., Summers, B.R., Knecht, A.K., Brady, S.D., Zhang, H., Pollen, A.A., Howes, T., Amemiya, C., Lander, E.S., Di Palma, F., Lindblad-Toh, K. & Kingsley, D.M. (2012a) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, 484,55–61.

Jones, F.C., Chan, Y.F., Schmutz, J., Grimwood, J., Brady, S.D., Southwick, A.M., Absher, D.M., Myers, R.M., Reimchen, T.E., Deagle, B.E., Schluter, D. & Kingsley, D.M. (2012b). A genome-wide SNP genotyping array reveals patterns of global and repeated species-pair divergence in Sticklebacks. *Current Biology*, 22, 83-90.

Karlsen, B.O., Klingan, K., Emblem, A., Jørgensen, T.E., Jueterbock, A., Furmanek, T., Hoarau, G., Johansen, S.D., Nordeide, J.T. & Moum, T. (2013). Genomic divergence between the migratory and stationary ecotypes of Atlantic cod. *Molecular Ecology*, 22(20).

Kassambara, A. (2017). Practical guide to principal component methods in R. In: Kassambara, A (Eds.). *Multivariate Analysis II. Statistical tools for high-throughput data analysis*.

Kassambara, A. & Mundt, F. (2017). factoextra: Extract and visualize the results of multivariate data analyses. Available online:
<https://rpkgs.datanovia.com/factoextra/index.html>

Keenan, K., McGinnity, P., Cross, T.F., Crozier, W.W. & Prodohl, P.A. (2013). diveRcity: An R package for the estimation and exploration of population genetics parameters and their associated errors. *Methods in Ecology and Evolution*, 4, 782-788.

Kelley, J.L., Brown, A.P., Therkildsen, N.O. & Foote, A.D. (2016) The life aquatic: advances in marine vertebrate genomics. *Nature Reviews Genetics*, 17(9), 523–534.

Knutsen, H., Jorde, P., Sannaes, H., Rus Hoelzel, A., Bergstad, O., Stefanni, S., Johansen, T. & Stenseth, N. (2009). Bathymetric barriers promoting genetic structure

in the deepwater demersal fish tusk (*Brosme brosme*). *Molecular Ecology*, 18(15), 3151-3162.

Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R.V., Nolte, V., Futschik, A., Kosiol, C. & Schlötterer, C. (2011b). PoPoolation: A toolbox for population genetic analysis of Next Generation Sequencing data from pooled individuals. *PLoS ONE*, 6(1), p.e15925.

Kofler, R., Pandey, R.W. & Schlotterer, C. (2011a). PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*, 27(24),3435-3436.

Kumar, S., Banks, T.W. & Cloutier, S. (2012). SNP discovery through next-generation sequencing and its applications. *International Journal of Plant Genomics*, 2012.

Lachance, J. & Tishkoff, S.A. (2013). SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. *Bioessays*, 35(9), 780-786.

Laikre, L., Palm, S. & Ryman, N. (2005). Genetic population structure of fishes: Implications for coastal zone management. *Ambio*, 34(2), 111-119.

Lal, M.M., Southgate, P.C., Jerry, D.R., Bosserelle, C. & Zenger, K.R. (2017). Swept away: ocean currents and seascape features influence genetic structure across the 18,000 Km Indo-Pacific distribution of a marine invertebrate, the black-lip pearl oyster *Pinctada margaritifera*. *BMC Genomics*, 18, 66.

Lamichhaney, S., Barrio, A.M., Rafati, N., Sundstrom, G., Rubin, C.R., Gilbert, E.R., Berglund, J., Wetterbom, A., Laikre, L., Webster, M.T., Grabherr, M., Ryman, N. & Andersson, L. (2012). Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. *Proceedings of the National Academy of Sciences*, 109(47), 19345-19350.

Langmead, B. & Salzberg, S. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9, 357-359.

Larson, W.A., Seeb, J.E., Pascal, C.E., Templin, W.D. & Seeb, L.W. (2014). Single-nucleotide polymorphisms (SNPs) identified through genotyping-by-sequencing improve genetic stock identification of Chinook salmon (*Oncorhynchus tshawytscha*)

from western Alaska. *Canadian Journal of Fisheries and Aquatic Sciences*, 71, 698-708.

Le Moan, A., Gagnaire, P.A. & Bonhomme, F. (2016). Parallel genetic divergence among coastal–marine ecotype pairs of European anchovy explained by differential introgression after secondary contact. *Molecular Ecology*, 25, 3187-3202.

Lett, C., Veitch, J., van der Lingen, C.D. & Hutchings, L. (2007). Assessment of an environmental barrier to transport of ichthyoplankton from the southern to the northern Benguela ecosystems. *Marine Ecology Progress Series*, 347, 247-259.

Li, H. & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754-1760.

Li, Y. & Wang, H. (2017). Advances of genotyping-by-sequencing in fisheries and aquaculture. *Review in Fisheries Biology and Fisheries*, 27, 535-559.

Li, Y., Cao, K. & Fu, C. (2018). Ten fish mitogenomes of the tribe Gobionini (Cypriniformes: Cyprinidae: Gobioninae). *Mitochondrial DNA Part B*, 3(2), 803-804.

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.

Liggins, L., Treml, E.A., Possingham, H.P. & Riginos, C. (2016). Seascape features, rather than dispersal traits, predict spatial genetic patterns in co-distributed reef fishes. *Journal of Biogeography*, 43, 256-267.

Limborg, M.T., Helyar, S.J., De Bruyn, M., Taylor, M.I., Nielsen, E., Ogden, R., Carvalho, G.R. & Bekkevold, D. (2012). Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish, the Atlantic herring (*Clupea harengus*). *Molecular Ecology*, 21(15), 3686-3703.

Limoge, F., Faivre, L., Gautier, T., Petit, J.M., Gautier, E., Masson, D., Jegou, G., El Chehadeh-Djebbar, S., Marle, N., Carmignac, V., Deckert, V., Brindisi, M.C., Edery, P., Ghoumid, J., Blair, E., Lagrost, L., Thauvin-Robinet, C. & Duplomb, L. (2015). Insulin response dysregulation explains abnormal fat storage and increased risk of diabetes mellitus type 2 in Cohen Syndrome. *Human Molecular Genetics*, 24(23), 6603-13.

- Lin-Jones, J., Parker, E., Wu, M., Dosé, A. & Burnside, B. (2004). Myosin 3A transgene expression produces abnormal actin filament bundles in transgenic *Xenopus laevis* rod photoreceptors. *Journal of Cell Science*, 117, 5825– 5834.
- Lin-Jones, J., Sohlberg, L., Dosé, A., Breckler, J., Hillman, D.W. & Burnside, B. (2009). Identification and Localization of Myosin Superfamily Members in Fish Retina and Retinal Pigmented Epithelium. *Journal of Comparative Neurology*, 513(2), 209-223.
- Lischer, H.E.L. & Excoffier, L. (2012). PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, 28(2), 298-299.
- Lowry, D.B., Hoban, S., Kelley, J.L., Lotterhos, K.E., Reed, L.K., Antolin, M.F. & Storfer, A. (2017). Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Molecular Ecology Resources*, 17, 142–152.
- Lundy, C., Rico, C. & Hewitt, G.M. (2000). Temporal and spatial genetic variation in spawning grounds of European hake (*Merluccius merluccius*) in the Bay of Biscay. *Molecular Ecology*, 9, 2067-2079.
- Luu, K., Bazini, E. & Blum, M.G.B. (2017). *pcadapt*: an R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources*, 17, 67-77.
- Martinson, J. & Ogden, R. (2009). FishPopTrace—Developing SNP-based population genetic assignment methods to investigate illegal fishing. *Forensic Science International: Genetics Supplement Series*, 2(1), 294-296.
- Mastretta-Yanes, A., Arrigo, N., Jorgensen, T.H., Piñeros, D. & Emerson, B.C. (2015). Restriction site-associated DNA sequencing, genotyping error estimation and *de novo* assembly optimization for population genetic inference. *Molecular Ecology*, 15, 28-41.
- McQuinn, I.H. (1997). Metapopulations of the Atlantic herring. *Reviews in Fish Biology and Fisheries*, 7, 297–329.

Miethe, T., Dytham, C., Dieckmann, U. & Pitchford, J.W. (2010). Marine reserves and the evolutionary effects of fishing on size at maturation. *ICES Journal of Marine Science*, 67, 412–425.

Milano, I., Babbucci, M., Cariani, A., Atanassova, M., Bekkevold, D., Carvalho, G.R., Espiñeira, M., Fiorentino, F., Garofalo, G., Geffen, A.J., Hansen, J.H., Helyar, S.J., Nielsen, E.E., Ogden, R., Patarnello, T., Stagioni, M., Tinti, F. & Bargelloni, L. (2014). Outlier SNP markers reveal fine-scale genetic structuring across European hake populations (*Merluccius merluccius*). *Molecular Ecology*, 23(1), 118-135.

Milano, I., Babbucci, M., Cariani, A., Atanassova, M., Bekkevold, D., Carvalho, G.R., Espiñeira, M., Fiorentino, F., Garofalo, G., Geffen, A.J., Hansen, J.H., Helyar, S.J., Nielsen, E.E., Ogden, R., Patarnello, T., Stagioni, M., FishPoPTrace Consortium, Tinti, F. & Bargelloni, L. (2014). Outlier SNP markers reveal fine-scale genetic structuring across European hake populations (*Merluccius merluccius*). *Molecular Ecology*, 23, 118-135.

Milano, I., Babbucci, M., Espineira, M., Atanassova, Geffen, A., Stagioni, M., Fiorentino, F., Panitz, F., Ogden, R., Nielsen, E.E., Taylor, M.I., Helyar, S.J., Carvalho, G.R., Maes, G.E., Cariani, A., Patarnello, T., FishPoPTrace Consortium, Tinti, F. & Bargelloni. (2012). Genomic tools for fishery and conservation of the European hake. *International Council for the Exploration of the Sea, Conference Meeting*.

Milano, I., Babbucci, M., Panitz, F., Ogden, R., Nielsen, R.O., Taylor, M.I., Helyar, S.J., Carvalho, G.R., Espiñeira, M., Atanassova, M., Tinti, F., Maes, G.E., Patarnello, T. & Bargelloni, L. (2011). Novel tools for conservation genomics: comparing two high-throughput approaches for SNP discovery in the transcriptome of the European Hake. *PLoS ONE*, 6(11), p.e28008.

Miller, D.C.M., Moloney, C.L., van der Lingen, C.D., Lett, C., Mullon, C. & Field, J.G. (2006). Modelling the effects of physical–biological interactions and spatial variability in spawning and nursery areas on transport and retention of sardine *Sardinops sagax* eggs and larvae in the southern Benguela ecosystem. *Journal of Marine Systems*, 62 (3-4), 212-229.

- Mirim, L., Kerwath, S., Macey, B., Lamberth, S.J., Cowley, P.D., van der Merwe, A.B., Bloomer, P. & Roodt-Wilding, R. (2016). Genetic analyses of overfished silver kob *Argyrosomus inodorus* (Scieanidae) stocks along the southern African coast. *Fisheries Research*, 176, 100-106.
- Monteiro, P.M.S., van der Plas, A.K., Melice, J.L. & Florenchie, P. (2008) Interannual hypoxia variability in a coastal upwelling system: ocean-shelf exchange, climate and ecosystem-state implications. *Deep-Sea Research Part I-Oceanographic Research Papers*, 55, 435–450.
- Montell, C. & Rubin, G.M. (1988). The *Drosophila ninaC* locus encodes two photoreceptor cell specific proteins with domains homologous to protein kinases and the myosin heavy chain head. *Cell*, 52, 757–772.
- Mullins, R.B. (2017). Population genomics analyses of yellowfin tuna *Thunnus Albacares* off South Africa reveals need for a shifted management boundary. MSc Thesis, Rhodes University.
- Muss, A., Robertson, D.R., Stepien, C.A., Wirtz, P. & Bowen, B.W. (2001). Phylogeography of *Ophioblennius*: the role of ocean currents and geography in reef fish evolution. *Evolution*, 55, 561–572.
- Nakamura, Y., Mori, K., Saitoh, K., Oshima, K., Mekuchi, M., Sugaya, T., Shigenobu, Y., Ojima, N., Muta, S., Fujiwara, A., Yasuike, M., Oohara, I., Hirakawa, H., Chowdhury, V.S., Kobayashi, T., Nakajima, K., Sano, M., Wada, T., Tashiro, K., Ikeo, K., Hattori, M., Kuhara, S., Gojobori, T. & Inouye, K. (2013). Evolutionary changes of multiple visual pigment genes in the complete genome of Pacific bluefin tuna. *PNAS*, 110(27), 11061-11066.
- Narum, S.R., Buerkle, C., Davey, J.W., Miller, M.R. & Hohenlohe, P.A. (2013). Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*, 22(11), 2841-2847.
- National Center for Biotechnology Information. (2018). Translate BLAST: Blastx. Available online at: <https://blast.ncbi.nlm.nih.gov/Blast>.
- Nesbo, C.L., Rueness, E.K., Iversen, S.A., Skagen, D.W. & Jakobsen, K.S. (2000). Phylogeography and population history of Atlantic Mackerel (*Scomber scombrus* L.):

a genealogical approach reveals genetic structuring among the eastern Atlantic stocks. *Proceedings of the Royal Society of London B*, 267, 281–292.

Nielsen, E., Cariani, A., Mac Aoidh, E., Maes, G.E., Milano, I., Ogden, R., Taylor, M., Hemmer-Hansen, J., Babbucci, M., Bargelloni, L., Bekkevold, D., Diopere, E., Grenfell, L., Helyar, S., Limborg, M.T., Martinsohn, J.T., McEwing, R., Panitz, F., Patarnello, T., Tinti, F., Van Houdt, J., Volckaert, F., Waples, R., Albin, J., Vieites Baptista, J., Barmintsev, V., Bautista, J., Bendixen, C., Bergé, J., Blohm, D., Cardazzo, B., Diez, A., Espiñeira, M., Geffen, A., Gonzalez, E., González-Lavín, N., Guarniero, I., Jérôme, M., Kochzius, M., Krey, G., Mouchel, O., Negrisolo, E., Piccinetti, C., Puyet, A., Rastorguev, S., Smith, J., Trentini, M., Verrez-Bagnis, V., Volkov, A., Zanzi, A. & Carvalho, G. (2012). Gene-associated markers provide tools for tackling illegal fishing and false eco-certification. *Nature Communications*, 3 (851).

Nielsen, E.E., Hemmer-Hansen, J., Larsen, P.F. & Bekkevold, D. (2009b). Population genomics of marine fishes: identifying adaptive variation in space and time. *Molecular Ecology*, 18, 4128-3150.

Nielsen, E., Hemmer-Hansen, J., Poulsen, N.A., Loeschcke, V., Moen, T., Johansen, T., Mittelholzer, C., Taranger, G., Ogden, R. & Carvalho, G. (2009a). Genomic signatures of local directional selection in a high gene flow marine organism; the Atlantic cod (*Gadus morhua*). *BMC Evolutionary Biology*, 9(1), 276.

Nielsen, R., Paul, J.S., Albrechtsen, A. & Song, Y.S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6), 443-451.

Nosil, P., Funk, D.J. & Oritz-Barrientos, D. (2009). Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, 18, 375-402.

O'Brien, C.M., Fox, C.J., Planque, B. & Casey, J. (2000). Climate variability North Sea cod. *Nature Brief Communications*, 404, 142.

Ogden, R. (2008). Fisheries forensics: the use of DNA tools for improving compliance, traceability and enforcement in the fishing industry. *Fish and Fisheries*, 9(4), 462-472.

- Olivar, M.P. & Fortuño, J.M. (1991). Guide to ichthyoplankton of the Southeast Atlantic (Benguela Current region). *Scientia Marina*, 55, 1–383.
- Olivar, M. & Sabatés, A. (1989). Early life history and spawning of *Genypterus capensis* (Smith, 1849) in the southern Benguela system. *South African Journal of Marine Science*, 8(1), 173-181.
- Olivar, M.P. & Shelton, P.A. (1993). Larval fish assemblages of the Benguela current. *Bulletin of Marine Science*, 53, 450–474.
- Operation Phakisa. (2014). Operation Phakisa. Available online at: <http://www.operationphakisa.gov.za/Pages/Home.aspx>
- 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.
- Palumbi, S.R. (1994). Genetic divergence, reproductive isolation, and marine speciation. *Annual Review of Ecology and Systematics*, 25, 547–572.
- Pante, E., Rohfritsch, A., Becquet, V., Belkhir, K., Bierne, N. & Garcia, P. (2013). Correction: SNP detection from *de novo* transcriptome sequencing in the bivalve *Macoma balthica*: marker development for evolutionary studies. *PLoS ONE*, 8(9).
- Parrish, R.H., Bakun, A., Husby, D.M. & Nelson, C.S. (1983). Comparative climatology of selected environmental processes. in relation to eastern boundary current pelagic fish reproduction. In: Sharp GD, Csirke J (Eds.). Proceedings of the expert consultation to examine changes in abundance and species composition of neritic fish resources. FAO Fish Rep, San Jose, Costa Rica, 731-77.
- Pascual, M., Rives, B., Schunter, C. & Macpherson, E. (2017). Impacts of life history traits on gene flow: A multispecies systematic review across oceanographic barriers in Mediterranean Sea. *PLoS ONE*, 12(5), e0176419.
- Pawson, M. & Jennings, S. (1996). A critique of methods for stock identification in marine capture fisheries. *Fisheries Research*, 25(3-4), 203-217.

- Payne, A.I.L. (1977). Stock differentiation and growth of the southern African kingklip *Genypterus capensis*. *Investl Rep. Sea Fish. Brch S. Afr.* 113 (32).
- Payne, A.I.L. (1985). Growth and stock differentiation of kingklip (*Genypterus capensis*) on the south-east coast of South Africa. *South African Journal of Zoology*, 20(2), 49-56.
- Payne, A.I.L. & Badenhorst, A. (1995). Other groundfish resources. In: Payne, A. I. L. and R. J. M. Crawford (Eds). *Oceans of Life off Southern Africa*, 2. Vlaeberg, Cape Town, 148–156.
- Pecquerie, L., Drapeau, L., Fréon, P., Coetzee, J.C., Leslie, R.W. & Griffiths, M.H. (2004). Distribution patterns of key fish species of the southern Benguela ecosystem: an approach combining fishery-dependent and fishery-independent data. *African Journal of Marine Science*, 26(1), 115-139.
- Pespeni, M.H., Oliver, T.A., Manier, M.K. & Palumbi, S.R. (2010). Restriction site tiling analysis: accurate discovery and quantitative genotyping of genome-wide polymorphisms using nucleotide arrays. *Genome Biology*, 11.
- Pespeni, M.H. & Palumbi, S.R. (2013). Signals of selection in outlier loci in a widely dispersing species across an environmental mosaic. *Molecular Ecology*, 22, 3580–3597.
- Phair, N., Toonen, R.J. & von der Heyden, S. (2018). Genomic signatures of adaptive divergence in the vulnerable African seagrass, *Zostera capensis*. IN PRESS
- Pinsky, M.L. & Palumbi, S.R. (2014). Meta-analysis reveals lower genetic diversity in overfished populations. *Molecular Ecology*, 23, 29-39.
- Pinsky, M.L., Reygondeau, G., Caddell, R., Palacios-Abrantes, J., Spijkers, J. & Cheng, W.W.L. (2018). Preparing ocean governance for species on the move. *SCIENCE*, 360 (6394), 1189.
- Poloczanska, E., Brown, C.J., Sydeman, W.J., Kiessling, W., Schoeman, D.S., Moore, P.J., Brander, K., Bruno, J.F., Buckley, L.B., Burrows, M.T., Duarte, C.M., Halpern, B.S., Holding, J., Kappel, C.V., O'Connor, M.I., Pandolfi, J.M., Parmesan,

C., Schwing, F., Thompson, S.A. & Richardson, A. (2013). Global imprint of climate change on marine life. *Nature Climate Change*. doi: 10.1038/NCLIMATE1958.

Porcelli, D., Butlin, R.K., Gaston, K.J., Joly, D. & Snook, R.R. (2015). The environmental genomics of metazoan thermal adaptation. *Heredity*, 114, 502–514.

Pritchard, J.K., Stephens, M. & Donnelly, P. (2000). Inference of population on structure using multilocus genotype data. *Genetics*, 155, 945-959.

Pujolar, J.M., Jacobsen, M.W., Als, T.D., Frydenberg, J., Munch, K., Jonsson, B., Jian, J.B., Cheng, L., Maes, G.E., Bernatchez, L. & Hansen, M.M. (2014). Genome-wide single-generation signatures of local selection in the panmictic European eel. *Molecular Ecology*, 23, 2514–2528.

Pujolar, J.M., Jacobsen, M.W., Frydenberg, J., Als, T.D., Larsen, P.F., Maes, G.E., Zane, L., Jian, J.b., Cheng, L. & Hansen, M.M. (2013). A resource of genome-wide single-nucleotide polymorphisms generated by RAD tag sequencing in the critically endangered European eel. *Molecular Ecology Resources*, 13, 706-714.

Punt, A.E. & Japp, D.W. (1994). Stock assessment of kingklip *Genypterus capensis* off South Africa. *South African Journal of Marine Science*, 14:1, 133-149.

R Core Development Team. (2008). R: A language and environment for statistical computing. Available online at: <http://www.R-project.org>

Reid, K., Hoareau, T.B., Graves, J.E., Potts, W.M., dos Santos, S.M.R., Klopper, A.W. & Bloomer, P. (2016). Secondary contact and asymmetrical gene flow in a cosmopolitan marine fish across the Benguela upwelling zone. *Heredity*, 117, 307-315.

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.

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, 361-395.

- Riginos, C., Crandall, E.D., Liggins, L., Bongaerts, P. & Trembl, E.A. (2016). Navigating the currents of seascape genomics: how spatial analyses can augment population genomic studies. *Current Zoology*, 62(6), 581-601.
- Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G. & Mesirov, J.P. (2011). Integrative Genomics Viewer. *Nature Biotechnology*, 29, 24-26.
- Rocha, L.A., Bass, A.L., Robertson, D.R. & Bowen, B.W. (2002). Adult habitat preferences, larval dispersal, and the comparative phylogeography of three Atlantic surgeon fishes (Teleostei: Acanthuridae). *Molecular Ecology*, 11, 243–252
- Rodríguez-Ezpeleta, N., Bradbury, I.R., Mendibil, I., Alvarez, P., Cotano, U. & Irigoien, X. (2016). Population structure of Atlantic mackerel inferred from RAD-seq-derived SNP markers: effects of sequence clustering parameters and hierarchical SNP selection. *Molecular Ecology Resources*, 16, 991-1001.
- Rose, G.A. (2005). On distributional responses of North Atlantic fish to climate change. *ICES Journal of Marine Science*, 62, 1360-1374.
- Rousset, F. (2008). GENEPOP' 007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources* 8, 103–106.
- Ruzzante, D.E., Walde, S.J., Gosse, J.C., Cussac, V.E., Habit, E., Zemplak, T.S. & Adams, E.D.M. (2008). Climate control on ancestral population dynamics: insight from Patagonian fish phylogeography. *Molecular Ecology Notes*, 6, 600-602.
- South African Deep-Sea Trawling Industry Association (SADSTIA). (2017). Annual Review.
- Saenz-Agudelo, P., Dibattista, J.D., Piatek, M.J., Gaither, M.R., Harrison, H.B., Nanninga, G.B. & Berumen, M.L. (2015). Seascape genetics along environmental gradients in the Arabian Peninsula: insights from ddRAD sequencing of anemonefishes. *Molecular Ecology*, 24(24), 6241-6255.
- Saha, A., Hauser, L., Kent, M., Planque, B., Neat, F., Kirubakaran, T.G., Huse, I., Homrum, E.I., Fevolden, S., Lien, S. & Johansen, T. (2015). Seascape genetics of

- saithe (*Pollachius virens*) across the North Atlantic using single nucleotide polymorphisms. *ICES Journal of Marine Science*, 72(9), 2732-2741.
- Sala-Bozano, M., Ketmaier, V. & Mariani, S. (2009). Contrasting signals from multiple markers illuminate population connectivity in a marine fish. *Molecular Ecology*, 18, 4811–4826.
- Salmerón, C. (2018). Adipogenesis in fish. *Journal of Experimental Biology*, 221, jeb161588.
- Santaclara, F.J., Pérez-Martín, R.I. & Sotelo, C.G. (2014). Developed of a method for the genetic identification of ling species (*Genypterus* spp.) in seafood products by FINS methodology. *Food Chemistry*, 143, 22-26.
- Schlötterer, C., Tobler, R., Kofler, R. & Nolte, V. (2014). Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics*, 15(11), 749-763.
- Seeb, J.E., Carvalho, G., Hauser, L., Naish, K., Roberts, S. & Seeb, L.W. (2011). Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Molecular Ecology Resources*, 11,1-8.
- Selkoe, K., D'Aloia, C.C., Crandall, E.D., Iacchei, M., Liggins, L., Puritz, J.B., von der Heyden, S. & Toonen, R.J. (2016). A decade of seascape genetics: contributions to basic and applied marine connectivity. *Marine Ecology Progress Series*, 554, 1-19.
- Selkoe, K., Henzler, C.M. & Gaines, S.D. (2008). Seascape genetics and the spatial ecology of marine populations. *Fish and Fisheries*, 9(4), 363-377.
- Selkoe, K., Watson, J.R., White, C., Horin, B., Iacchei, M., Mitarai, S., Siegel, D.A., Gaines, S.D. & Toonen, R.J. (2010). Taking the chaos out of genetic patchiness: seascape genetic reveals ecological and oceanographic drivers of genetic patterns in three temperate reef species. *Molecular Ecology*, 19, 3708-3726.
- Shafer, A.B.A., Peart, C.R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C.W. & Wolf, J.B.W. (2017). Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods in Ecology and Evolution*, 8, 907-917.

Shafer, A.B.B., Wolf, J.B.W., Alves, P.C., Bergström, L., Bruford, M.W., Brännström, I., Colling, G., Dalén, L., De Meester, L., Ekblom, R., Fawcett, K.d., Fior, S., Hajibabaei, M., Hill, J.A., Hoezel, A.R., Höglund, J., Jensen, E.L., Krause, J., Kristensen, T.N., Krützen, M., McKay, J.K., Norman, A.J., Ogden, R., Österling, E.M., Ouborg, N.J., Piccolo, J., Popovic, D., Primmer, C.R., Reed, F.A., Roumet, M., Salmona, J., Schenekar, T., Schwartz, M.K., Segelbacher, G., Senn, H., Thaulow, J., Valtonen, M., Veale, A., Vergeer, P., Vijay, N., Vila, C., Weissensteiner, M., Wennerstrom, L., Wheat, C.W. & Zelinski, P. (2014). Genomics and the challenging translation into conservation practice. *Trends in Ecology and Evolution*, 1-10.

Sham, P., Cherny, S. & Purcell, S. (2009). Application of genome-wide SNP data for uncovering pairwise relationships and quantitative trait loci. *Genetica*, 136, 237-243.

Shannon, L.V. (1985). The Benguela Ecosystem. 1. Evolution of the Benguela, physical features and processes. In: Barnes, M. (Ed.). *Oceanography and Marine Biology. An annual review*. University Press, Aberdeen, 105–182.

Shannon, L., Crawford, R.J.M., Pollock, D.E., Hutchings, L., Boyd, A.J., Taunton-Clark, J., Badenhorst, A., Melville-Smith, R., Augustyn, C.J., Cochrane, K.L., Hampton, I., Nelson, G., Japp, D.W. & Tarr, R. (1992). The 1980s – a decade of change in the Benguela ecosystem. *South African Journal of Marine Science*, 12(1), 271-296.

Shillington, F.A., Reason, C.J.C., Duncombe Rae, C.M., Florenchie, P. & Penven, P. (2006). Large scale physical variability of the Benguela Current Large Marine Ecosystem (BCLME). *Elsevier*, 14, 47-68.

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.

Smolka, M., Recheneder, P., Schatz, M.C., von Haeseler, A. & Sedlazeck, F.J. (2015). Teaser: Individualized benchmarking and optimization of read mapping results for NGS data. *Genome Biology*, 16 (235).

Spies, I., Spencer, P.D. & Punt, A.E. (2015). Where do we draw the line? A simulation approach for evaluating management of marine fish stocks with isolation-

by-distance stock structure. *Canadian Journal of Fisheries and Aquatic Sciences*, 72(7), 968-982.

Stephenson, R.L. (1999). Stock complexity in fisheries management: a perspective of emerging issues related to population sub-units. *Fisheries Research*, 43, 247–249.

Stephenson, T.A. & Stephenson, A. (1972). In: Life between tidemarks on rocky shores. Freeman, San Francisco, USA.

Stenevik, E., Verheye, H., Lipinski, M., Ostrowski, M. & Stromme, T. (2008). Drift routes of Cape hake eggs and larvae in the southern Benguela Current system. *Journal of Plankton Research*, 30(10), 1147-1156.

Strasburg, J.L., Sherman, N.A., Wright, K.M., Moyle, L.C., Willis, J.H. & Rieseberg, L.H. (2012). What can patterns of differentiation across plant genomes tell us about adaptation and speciation. *Philosophical Transactions of the Royal Society B*, 367, 364-373.

Sundby, S., Boyd, A.J., Hutchings, L., O'Toole, M.J., Thorisson, K. & Thorsen, A. (2001). Interaction between Cape hake spawning and the circulation in the northern Benguela upwelling ecosystem. *South African Journal of Marine Science*, 23, 317–336.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595.

Teske, P.R., von der Heyden, S., McQuaid, C.D. & Barker, N.P. (2011). A review of marine phylogeography in southern Africa. *South African Journal of Science*, 107, 45–53.

Tigano, A. & Friesen, V.I. (2016). Genomics of local adaptation with gene flow. *Molecular Ecology*, 25, 2144-2164.

Toms, J.A., Compton, J.C., Smale, M., von der Heyden, S. (2014). Variation in paleo-shorelines explains contemporary population genetic patterns of rocky shore species. *Biology Letters*, 10, 20140330

- Toonen, R.J., Puritz, J.B., Forsman, Z.H., Whitney, J.I., Fernandez-Silva, I., Andrews, K.R. & Bird, C.E. (2013). ezRAD: a simplified method for genomic genotyping in non-model organisms. *PeerJ*, 1, p.e203.
- Tucker, J.M., Schwartz, M.K., Truex, R.L., Wisely, S.M. & Allendorf, F.W. (2014). Sampling affects the detection of genetic subdivision and conservation implications for fisher in Sierra Nevada. *Conservation Genetics*.
- Turner, S.D. (2018). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *The Journal of Open Source Software*, 3(25), 731.
- Turpie, J.K., Beckley, L.E. & Katua, S.M. (2000). Biogeography and the selection of priority areas for conservation of South African coastal fishes. *Biological Conservation*, 92, 59-72.
- Valenzuela-Quiñonez, F. (2016). How fisheries management can benefit from genomics. *Briefings in Functional Genomics*, 15(5), 352-357.
- van der Lingen, C.D., Weston, L.F., Ssempe, N.N. & Reed, C.C. (2015). Incorporating parasite data in population structure studies of South African sardine *Sardinops sagax*. *Parasites in fisheries and mariculture*, 142(1), 156-167.
- 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, 1605458.
- von der Heyden, S., Barendse, J., Seebregts, A.J. & Matthee, C.A. (2010). Misleading the masses: detection of mislabelled and substituted frozen fish products in South Africa. *ICES Journal of Marine Science*, 67(1), 176-185.
- von der Heyden, S., Beger, M., Toonen, R.J., van Herwerden, L., Juinio-Meñez, M.A., Ravago-Gotanco, R., Fauvelot, C. & Bernardi, G. (2014). The application of genetics to marine management and conservation: examples from the Indo-Pacific. *Bulletin of Marine Science*, 90(1), 123-158.
- von der Heyden, S., Lipinski, M.R. & Matthee, C.A. (2007). Mitochondrial DNA analyses of the Cape hakes reveal an expanding, panmictic population for

- Merluccius capensis* and population structuring for mature fish in *Merluccius paradoxus*. *Molecular Phylogenetics and Evolution*, 42(2), 517-527.
- Wang, I.J. & Summers, K. (2010). Genetic structure is correlated with phenotypic divergence rather than geographic and ecological isolation. *Evolution*, 16, 175–182.
- Wang, J., Xue, D-X., Zhang, B-D., Li, Y-L., Liu, B.J. & Liu, J-X. (2016). Genome-wide SNP discovery, genotyping and their preliminary applications for population genetic inference in Spotted Sea Bass (*Lateolabrax maculatus*). *PLoS ONE*, 11(6), e0157809.
- 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, 1419–1439.
- Ward, R.D. & Reilly, A. (2001). Development of microsatellite loci for population studies of the pink ling, *Genypterus blacodes* (Teleostei: Ophidiidae). *Molecular Ecology Notes*, 1, 173-175.
- Ward R.D., Woodward, M. & Skibinski, D.O.F. (1994) A comparison of genetic diversity levels in marine, freshwater, and anadromous fishes. *Journal of Fish Biology*, 44, 213-232.
- Waters, J.M. & Roy, S. (2004). Phylogeography of a high-dispersal New Zealand sea-star: does upwelling block gene-flow. *Molecular Ecology*, 13, 2797-2806.
- White, T.A., Fotherby, H.A. & Hoelzel, A.R. (2011). Comparative assessment of population genetics and demographic history of two congeneric deep sea fish species living at different depths. *Marine Ecology Progress Series*, 434, 155–164.
- White, C., Selkoe, K.A., Watson, J., Siegel, D.A., Zacherl, D.C. & Toonen, R.J. (2010a). Ocean currents help explain population genetic structure. *Proceedings of the Royal Society B: Biological Sciences*, 277(1688), 1685-1694.
- White, T., Stamford, J. & Rus Hoelzel, A. (2010b). Local selection and population structure in a deep-sea fish, the roundnose grenadier (*Coryphaenoides rupestris*). *Molecular Ecology*, 19(2), 216-226.
- Whitlock, M.C. & McCauley, D.E. (1999). Indirect measures of gene flow and migration. *Heredity*, 82, 117–25.

Winker, K. (2009). Reuniting phenotype and genotype in biodiversity research. *Bioscience*, 59, 657–665.

Winnepenninckx, B., Backeljau, T. & Dewachter, R. (1993). Extraction of high molecular weight DNA from molluscs. *Trends in Genetics*, 9, 407-407.

World Wildlife Fund. (2011). Fisheries: facts and trends South Africa. In WWF Report.

Wright, S. (1931). A landmark paper in population genetics in which the effect of population size, mutation and migration on the abundance and distribution of genetic variation in populations is first quantitatively described. Evolution in Mendelian populations. *Genetics*, 16, 97–159.

Yeaman, S. & Otto, S.P. (2011). Establishment and maintenance of adaptive genetic divergence under migration, selection, and drift. *Evolution*, 65, 2123–2129.

Zhan, J., Pettway, R.E. & McDonald, B.A. (2012). The global genetic structure of the wheat pathogen *Mycosphaerella graminicola* is characterized by high nuclear diversity, low mitochondrial diversity, regular recombination, and gene flow. *Fungal Genetics and Biology*, 38, 286-297.

Zhang, W., Chen, J., Yang, Y., Tang, Y. Shang, J. & Shen, B. (2011). A practical comparison of de novo genome assembly software tools for Next-Generation Sequencing technologies. *PLoS ONE*, 6(3), e17915.

SUPPLEMENTARY MATERIAL

Supplementary Material 1: Scripts used for bioinformatic analyses and pipeline.

All scripts were run using MobaXterm portal through the HPC cluster provided by Stellenbosch University. The associated programmes used are provided for each script.

Quality control with TrimGalore! V0.4.4

```
trim_galore --paired -q 25 --length 50 -a
AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -a2
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT -stringency 10 -e 0.01
p1.read1.fastq p1.read2.fastq
```

Mapping onto reference genome/sequence with BWA 0.7.13

```
bwa index ref.fasta

bwa mem -R '@RG\tID:pop1\tSM:P1\tLB:library1' ref.fasta p1.read1_trim.fq
p1.read2_trim.fq -a -t16 -T20 > p1.sam
```

Mapping onto reference genome/sequence with Bowtie2

- a) bowtie2-build -f filter1.fasta filter1
- b) bowtie2 -f -N 1 -x filter1 -1 cmrGb42AF.fasta -2 cmrGb42AR.fasta -S 42A.N.sam

Convert SAM to BAM file format and sort BAM files in SAMtools 1.3

```
samtools view -bS p1_A.sam | samtools sort -o p1_A.sort.bam
```

Filtering of a) mapped, b) properly-paired and c) unique reads

- a) samtools view -b -F 0x04 p2_codmt.sort.bam > p2_mtdna5.sort.bam
- b) samtools view -b -f 0x01 -f 0x02 p2_mtdna5.sort.bam > p2_mtdna5.p3.sort.bam
- c) samtools view -b -q1 p2_mtdna5.p3.sort.bam > p2_mtdna5.p3.sort.u3.bam

Find optimal k-mer lengths for *de novo* assembly with KmerGenie

- a) Create "Reads.file.txt"

P1.read1.fq
P1.read2.fq
b) kmergenie reads_file.txt

de novo Assembly in SPAdes

```
INPUTS="p1.read1_trim.fq p1.read2_trim.fq p2.read1_trim.fq p2.read2_trim.fq  
p3.read1_trim.fq p3.read2_trim.fq p4.read1_trim.fq p4.read2_trim.fq p5.read1_trim.fq  
p5.read2_trim.fq p6.read1_trim.fq p6.read2_trim.fq p7.read1_trim.fq p7.read2_trim.fq  
p8.read1_trim.fq p8.read2_trim.fq"
```

```
spades.py \
```

```
--disable-gzip-output \
```

```
-t 32 \
```

```
-k 19,21,31 \
```

```
--pe1-1 p1.read1_trim.fq --pe1-2 p1.read2_trim.fq \
```

```
--pe2-1 p2.read1_trim.fq --pe2-2 p2.read2_trim.fq \
```

```
-o spades_3
```

Assessment of mapping statistics with SAMTools

```
samtools stats -c 1,1000,1 -q 20 -r ref3.fasta p2_A.sort.NA.bam
```

Create SAMTools mpileup file with SAMTools

```
samtools mpileup -d 10000 -Q 20 -B -f mrefD.fasta p1_mtdna.DA.final.bam  
p2_mtdna.DA.final.bam p3_mtdna.DA.final.bam p4_mtdna.DA.final.bam  
p5_mtdna.DA.final.bam p6_mtdna.DA.final.bam p7_mtdna.DA.final.bam  
p8_mtdna.DA.final.bam > ALL.mtdna.mpileup
```

Conversion of mpileup file format to sync file format with PoPoolation2

```
mpileup2sync.pl --fastq-type illumina --min-qual 20 --input ALL.mtdna.mpileup --  
output ALL.mtdna.orig.sync
```

Removal of mtDNA reads from original quality-controlled reads using BMAP

```
filterbyname.sh in=p8.read1_trim.fq in2=p8.read2_trim.fq out=p8.read1_trim.filter.fq
out2=p8.read2_trim.filter.fq names=total.list.remove.merged.sam include=f
minlen=50
```

Calculation of regional diversity measures with PoPoolation1 (values for contigs with SNPs only)

- a) perl /apps/PoPoolation/1.2.2/Variance-sliding.pl --fastq-type sanger --measure D --input p1.filter.pileup --min-count 2 --min-coverage 10 --max-coverage 500 --min-qual 10 --pool-size 40 --window-size 100 --step-size 100 --output p1.filter.210500.D
- b) more p1.filter.210500.D |awk '{if(\$5!="na"&&\$3!="0")print}' > p1.filter.210500.D.ls

Estimation of allele counts from sync files for SNP identification with PoPoolation2

```
snp-frequency-diff.pl --input ALL.filter.sync --output-prefix ALL.filter_diff --min-count 4
--min-coverage 25 --max-coverage 500
```

Creating list of biallelic SNPs only

```
more *rc|awk '{if ($4==2) print $1 '\t' $2}' > biallelic.SNP.list
```

Calculating number of biallelic SNPs per pool

- a) more ALL.filter.420500_diff_rc_pop |awk '{if(\$4==2)print\$1"\t"\$2"\t"\$10"\t"\$11"\t"\$12"\t"\$13"\t"\$14"\t"\$15"\t"\$16"\t"\$17}' > ALL.filter.420500.pop.maa.list
- b) more ALL.filter.420500.pop.maa.txt |awk '{if(\$1!=\$2)print\$1"\t"\$2}' > p1.filter.420500.pop.SNP.A.txt
- c) more p1.filter.420500.pop.SNP.A.txt |awk '{if(\$1!=0)print}' > p1.filter.420500.pop.SNP.txt

Calculating number of private SNPs per pool

```
more ALL.filter.420500.pop.maa.txt |awk '{if($1!=$2&&$3==&&$4&&$5==&&$6&&$7==&&$8&&$9==&&$10&&$11==&&$12&&$13==&&$14&&$15==&&$16)print}' > p1.filter.420500.pop.prv.txt
```

Creating simulated, Genepop files

a) Create GenePop files

```
perl /apps/PoPoolation/2.svn204/export/subsample_sync2GenePop.pl \  
--input ALL.mtdna.sync  
--output ${CHR}pos${POS}.GenePop  
--method fraction \  
--min-count 2 \  
--target-coverage 10 \  
--max-coverage 500 \  
--region ${CHR}:${POS}-${POS} \  
--diploid >>output 2>>errors  
rm ${CHR}pos${POS}.GenePop.params  
fi
```

b) Merge Genepop file

```
perl ${PBS_O_WORKDIR}/merge.gpop.pl
```

Supplementary Table S1: Major allele frequencies of shared outlier SNPs, identified for nuclear dataset, per pool. SNPs not found within pools indicated by -. Pool names as per Table 2.

SNP Position		Pool ID							
Node	Position	P1	P2	NAM 1	NAM 2	CB	TB	SC	EC
2	6000	0.80	0.97	0.83	0.94	0.86	0.95	-	0.79
2	6003	0.79	0.89	0.77	0.94	0.79	0.59	0.81	0.93
2	6014	0.73	0.89	0.84	0.94	0.79	0.62	0.84	0.93
6	5152	0.95	0.71	0.62	0.76	0.80	0.78	0.85	0.79
72	1475	0.59	0.60	0.59	0.52	0.81	0.79	0.63	0.59
72	1518	0.52	0.75	0.54	0.84	0.56	0.63	0.69	0.76
171	594	0.95	0.91	0.68	0.76	0.82	0.76	0.66	0.74
171	656	0.56	0.53	0.86	0.69	0.55	0.62	0.64	0.75
171	669	0.95	0.79	0.87	-	-	0.83	0.87	0.84
610	1886	0.88	-	0.92	0.74	0.88	0.87	0.80	0.82
929	2191	0.82	0.59	0.81	0.83	0.80	0.65	0.70	0.93
941	2617	0.88	0.60	0.79	0.74	0.63	0.90	0.55	0.53
1273	1696	0.60	0.53	0.54	0.56	0.54	0.52	0.85	0.55
2037	1439	-	0.94	0.93	0.76	0.93	-	0.93	0.92
2091	2369	0.82	0.87	0.65	0.87	0.78	0.92	0.92	0.83
2157	128	0.78	0.81	0.90	0.82	0.89	0.79	0.78	-
2475	1777	0.55	0.71	0.57	0.70	0.75	0.67	0.62	0.79
2507	81	0.70	0.60	0.90	0.85	0.73	0.51	0.65	0.70
3363	2164	0.51	0.50	0.78	0.59	0.57	0.53	0.59	0.56
3864	19	0.95	-	0.85	0.76	0.88	0.78	0.90	0.83
4112	1022	-	0.83	0.78	0.90	-	0.87	0.97	-
11982	1707	0.68	0.71	0.76	0.88	0.53	0.79	0.84	-
119258	675	0.53	0.82	0.90	-	0.87	0.63	0.63	0.80
159059	74	0.96	0.85	0.84	0.53	0.68	0.93	0.85	-
181211	594	0.51	0.51	0.90	0.54	0.55	0.54	0.88	0.83

Supplementary Table S2: Repeat motifs, primer sequences (Forward – F & Reverse – R) and GenBank accession number of 10 microsatellite primers employed (Ward & Reilly, 2010). Refer to Ward and Reilly (2001) for primer notes.

Locus	Motif	Primer Sequences	GenBank Accession No.
cmrGb4.11B	(GACA) ₁₁ (GACA) ₅ (GA) ₉	F- CCTGAGTGCTTAAAGAGGA R- GAGGAGGAGACGATGAAA	AF334145
cmrGb5.5	(GT) ₈ TT(GT) ₂₈	F- ACTCCTGGACTGGATCTAA R- TGCAAATTTTCATGTAAATG	AF334146
cmrGb5.9	(CA) ₁₁	F- AGGGTCACTTTTCAGTTTTA R- TGCAGAACACACTCCAC	AF334147
cmrGb4.2A	(TAAA) ₈	F- ATCGGGCAGTTCCTTGCTAT R- GGGAAAGCTTTTGTGAGCATC	AF334148
cmrGb5.2B	(CTTT) ₁₉	F- CGGTCTGAGCAATGATACGA R- TACAGAGGGGAGGTAAATCAAGTC	AF334149
cmrGb2.6.1	(GTT) ₉	F- AGAACTAAACCAGCAGAATC R- CACAACAAGAGGGAAGTC	AF334150
cmrGb5.8B	(GT) ₂₉	F- CACTTTGGGGCTTCTCCTC R- CCCGATTCATTCATCCATC	AF334151
cmrGb4.2B	(CT) ₁₆ T(CT) ₇ (GT) ₂₇	F- AGTTGGTGTTTGCCCTGA R- GTCTGGAGTGTTTTGGATCATT	AF334152
cmrGb5.8A	(GT) ₂₀ GA(GT) ₅	F- AACCTCTGGCATCCATTTTC R- CCCAAAGTGCTGCTACTG	AF334153
cmrGb5.2A	(GT) ₃₀ GC(GT) ₂ GC(GT) ₂ GC(GT) ₅	F- AAACAGTGTTTCGCGTACT R- CCTGACATGTGTCGTTGA	AF334157